CHAPTER

8 Cosmology

8.1 ■ MAXIMALLY SYMMETRIC UNIVERSES

Contemporary cosmological models are based on the idea that the universe is pretty much the same everywhere—a stance sometimes known as the **Copernican principle**. On the face of it, such a claim seems crazy; the center of the sun, for example, bears little resemblance to the desolate cold of interstellar space. But we take the Copernican principle to apply only on the very largest scales, where local variations in density are averaged over. Its validity on such scales is manifested in a number of different observations, such as number counts of galaxies and observations of diffuse X-ray and γ-ray backgrounds, but is most clear in the 3K cosmic microwave background (CMB). Although we now know that the microwave background radiation is not perfectly smooth (and nobody ever expected that it was), the deviations from regularity are on the order of 10^{-5} or less, certainly an adequate basis for an approximate description of spacetime on large scales.

The Copernican principle is related to two more mathematically precise properties that a manifold might have: isotropy and homogeneity. **Isotropy** applies at some specific point in the manifold, and states that the space looks the same no matter in what direction you look. More formally, a manifold M is isotropic around a point p if, for any two vectors V and W in T_pM, there is an isometry of M such that the pushforward of W under the isometry is parallel with V (not pushed forward). It is isotropy of space that is indicated by the observations of the microwave background.

Homogeneity is the statement that the metric is the same throughout the manifold. In other words, given any two points p and q in M, there is an isometry that takes p into q. Note that there is no necessary relationship between homogeneity and isotropy; a manifold can be homogeneous but nowhere isotropic (such as $\mathbf{R} \times S^2$ in the usual metric), or it can be isotropic around a point without being homogeneous (such as a cone, which is isotropic around its vertex but certainly not homogeneous). On the other hand, if a space is isotropic *everywhere*, then it is homogeneous. Likewise if it is isotropic around one point and also homogeneous, it will be isotropic around every point. Since there is ample observational evidence for isotropy, and the Copernican principle would have us believe that we are not the center of the universe and therefore observers elsewhere should also observe isotropy, we will henceforth assume both homogeneity and isotropy.

The usefulness of homogeneity and isotropy is that they imply that a space is maximally symmetric. Think of isotropy as invariance under rotations, and homogeneity as invariance under translations, suitably generalized. Then homogeneity and isotropy together imply that a space has its maximum possible number of Killing vectors. An extreme application of the Copernican principle would be to insist that spacetime itself is maximally symmetric. In fact this will turn out not to be true; observationally we know that the universe is homogeneous and isotropic in *space*, but not in all of *spacetime*. However, it is interesting to begin by considering spacetimes that are maximally symmetric (which are, after all, special cases of the more general situation in which only space is maximally symmetric). As we shall see, there is a sense in which such universes are "ground states" of general relativity. This discussion is less relevant to the observed universe than subsequent parts of this chapter, and empirically-minded readers are welcome to skip ahead to the next section.

We mentioned in Chapter 3 that the Riemann tensor for a maximally symmetric n-dimensional manifold with metric $g_{\mu\nu}$ can be written

$$R_{\rho\sigma\mu\nu} = \kappa(g_{\rho\mu}g_{\sigma\nu} - g_{\rho\nu}g_{\sigma\mu}), \tag{8.1}$$

where κ is a normalized measure of the Ricci curvature,

$$\kappa = \frac{R}{n(n-1)}, \tag{8.2}$$

and the Ricci scalar R will be a constant over the manifold. Since at any single point we can always put the metric into its canonical form ($g_{\mu\nu} = \eta_{\mu\nu}$), the kinds of maximally symmetric manifolds are characterized locally by the signature of the metric and the sign of the constant κ. The modifier "locally" is necessary to account for possible global differences, such as between the plane and the torus. We are interested in metrics of signature $(-+++)$. For vanishing curvature ($\kappa = 0$) the maximally symmetric spacetime is well known; it is simply Minkowski space, with metric

$$ds^2 = -\mathrm{d}t^2 + \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2. \tag{8.3}$$

The conformal diagram for Minkowski space is derived in Appendix H.

The maximally symmetric spacetime with positive curvature ($\kappa > 0$) is called **de Sitter space**. Consider a five-dimensional Minkowski space with metric $ds_5^2 = -\mathrm{d}u^2 + \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2 + \mathrm{d}w^2$, and embed a hyperboloid given by

$$-u^2 + x^2 + y^2 + z^2 + w^2 = \alpha^2. \tag{8.4}$$

Now induce coordinates $\{t, \chi, \theta, \phi\}$ on the hyperboloid via

$$u = \alpha \sinh(t/\alpha)$$

$$w = \alpha \cosh(t/\alpha) \cos\chi$$

$$x = \alpha \cosh(t/\alpha) \sin\chi \cos\theta$$
$$y = \alpha \cosh(t/\alpha) \sin\chi \sin\theta \cos\phi$$
$$z = \alpha \cosh(t/\alpha) \sin\chi \sin\theta \sin\phi. \tag{8.5}$$

The metric on the hyperboloid is then

$$ds^2 = -\mathrm{d}t^2 + \alpha^2 \cosh^2(t/\alpha)\left[\mathrm{d}\chi^2 + \sin^2\chi(\mathrm{d}\theta^2 + \sin^2\theta \mathrm{d}\phi^2)\right]. \tag{8.6}$$

We recognize the expression in round parentheses as the metric on a two-sphere, $d\Omega_2^2$, and the expression in square brackets as the metric on a three-sphere, $d\Omega_3^2$. Thus, de Sitter space describes a spatial three-sphere that initially shrinks, reaching a minimum size at $t = 0$, and then re-expands. Of course this particular description is inherited from a certain coordinate system; we will see that there are equally valid alternative descriptions.

These coordinates cover the entire manifold. You can generally check this by, for example, following the behavior of geodesics near the edges of the coordinate system; if the coordinates were incomplete, geodesics would appear to terminate in finite affine parameter. The topology of de Sitter is thus $\mathbf{R} \times S^3$. This makes it very simple to derive the conformal diagram, since the important step in constructing conformal diagrams is to write the metric in a form in which it is conformally related to the Einstein static universe (a spacetime with topology $\mathbf{R} \times S^3$, describing a spatial three-sphere of constant radius through time). Consider the coordinate transformation from t to t' via

$$\cosh(t/\alpha) = \frac{1}{\cos(t')}. \tag{8.7}$$

The metric (8.6) now becomes

$$ds^2 = \frac{\alpha^2}{\cos^2(t')} d\bar{s}^2, \tag{8.8}$$

where $d\bar{s}^2$ represents the metric on the Einstein static universe,

$$d\bar{s}^2 = -(\mathrm{d}t')^2 + \mathrm{d}\chi^2 + \sin^2\chi \, d\Omega_2^2. \tag{8.9}$$

The range of the new time coordinate is

$$-\pi/2 < t' < \pi/2. \tag{8.10}$$

The conformal diagram of de Sitter space will simply be a representation of the patch of the Einstein static universe to which de Sitter is conformally related. It looks like a square, as shown in Figure 8.1. A spacelike slice of constant t' represents a three-sphere; the dashed lines at the left and right edges are the north and south poles of this sphere. The diagonal lines represent null rays; a photon released at past infinity will get to precisely the antipodal point on the sphere at

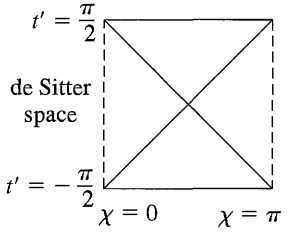

FIGURE 8.1 Conformal diagram for de Sitter spacetime. Spacelike slices are three-spheres, so that points on the diagram represent two-spheres except for those at left and right edges, which are points.

future infinity. Keep in mind that the spacetime "ends" to the past and the future only through the magic of conformal transformations; the actual de Sitter space extends indefinitely into the future and past. Note also that two points can have future (or past) light cones that are completely disconnected; this reflects the fact that the spherical spatial sections are expanding so rapidly that light from one point can never come into contact with light from the other.

A similar hyperboloid construction reveals the $\kappa < 0$ spacetime of maximal symmetry, known as **anti-de Sitter space**. Begin with a fictitious five-dimensional flat manifold with metric $ds_5^2 = -\mathrm{d}u^2 - \mathrm{d}v^2 + \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2$, and embed a hyperboloid given by

$$-u^2 - v^2 + x^2 + y^2 + z^2 = -\alpha^2. \tag{8.11}$$

Note all the minus signs. Then we can induce coordinates $\{t', \rho, \theta, \phi\}$ on the hyperboloid via

$$\begin{aligned} u &= \alpha \sin(t') \cosh(\rho) \\ v &= \alpha \cos(t') \cosh(\rho) \\ x &= \alpha \sinh(\rho) \cos\theta \\ y &= \alpha \sinh(\rho) \sin\theta \cos\phi \\ z &= \alpha \sinh(\rho) \sin\theta \sin\phi, \end{aligned} \tag{8.12}$$

yielding a metric on this hyperboloid of the form

$$ds^2 = \alpha^2\big(-\cosh^2(\rho)\,\mathrm{d}t'^2 + \mathrm{d}\rho^2 + \sinh^2(\rho)\,d\Omega_2^2\big). \tag{8.13}$$

These coordinates have a strange feature, namely that t' is periodic. From (8.12), t' and $t'+2\pi$ represent the same place on the hyperboloid. Since $\partial_{t'}$ is everywhere timelike, a curve with constant $\{\rho, \theta, \phi\}$ as t' increases will be a closed timelike curve. However, this is not an intrinsic property of the spacetime, merely an artifact of how we have derived the metric from a particular embedding. We are welcome to consider the "covering space" of this manifold, the spacetime with metric given by (8.13) in which we allow t' to range from $-\infty$ to ∞. There are no closed timelike curves in this space, which we will take to be the definition of anti-de Sitter space.

To derive the conformal diagram, perform a coordinate transformation analogous to that used for de Sitter, but now on the radial coordinate:

$$\cosh(\rho) = \frac{1}{\cos\chi}, \tag{8.14}$$

so that

$$ds^2 = \frac{\alpha^2}{\cos^2\chi} d\bar{s}^2, \tag{8.15}$$

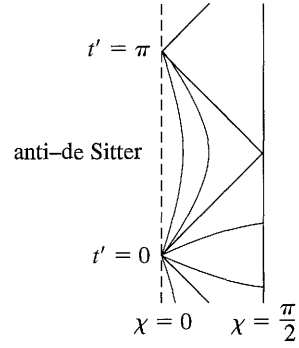

FIGURE 8.2 Conformal diagram for anti-de Sitter spacetime. Spacelike slices have the topology of $\mathbf{R}^3$, which we have represented in polar coordinates, so that points on the diagram stand for two-spheres except those at the left side, which stand for single points at the spatial origin. Infinity is a timelike surface at the right side.

where $d\bar{s}^2$ represents the metric on the Einstein static universe (8.9). Unlike in de Sitter, the radial coordinate now appears in the conformal factor. In addition, for anti-de Sitter, the t' coordinate goes from minus infinity to plus infinity, while the range of the radial coordinate is

$$0 \leq \chi < \frac{\pi}{2}. \tag{8.16}$$

Thus, anti-de Sitter space is conformally related to half of the Einstein static universe. The conformal diagram is shown in Figure 8.2, which illustrates a few representative timelike and spacelike geodesics passing through the point $t' = 0$, $\chi = 0$. Since χ only goes to $\pi/2$ rather than all the way to π, a spacelike slice of this spacetime has the topology of the interior of a hemisphere of S^3; that is, it is topologically $\mathbf{R}^3$ (and the entire spacetime therefore has the topology $\mathbf{R}^4$). Note that we have drawn the diagram in polar coordinates, such that a point on the left side represents a point at the spatial origin, while one on the right side represents a two-sphere at spatial infinity. Another popular representation is to draw the spacetime in cross-section, so that the spatial origin lies in the middle and the right and left sides together comprise spatial infinity.

An interesting feature of anti-de Sitter is that infinity takes the form of a timelike hypersurface, defined by $\chi = \pi/2$. Because infinity is timelike, the space is not globally hyperbolic, we do not have a well-posed initial value problem in terms of information specified on a spacelike slice, since information can always "flow in from infinity." Another interesting feature is that the exponential map is not onto the entire spacetime; geodesics, such as those drawn on the figure, which leave from a specified point do not cover the whole manifold. The future-pointing timelike geodesics, as indicated, can initially move radially outward from $t' = 0$, $\chi = 0$, but eventually refocus to the point $t' = \pi$, $\chi = 0$ and will then move radially outward once again.

As an aside, it is irresistible to point out that the timelike nature of infinity enables a remarkable feature of string theory, the "AdS/CFT correspondence." Here, AdS is of course the anti-de Sitter space we have been discussing, while CFT stands for a conformally-invariant field theory defined on the boundary [which is, for an n-dimensional AdS, an $(n-1)$-dimensional spacetime in its own right]. The AdS/CFT correspondence suggests that, in a certain limit, there is an equivalence between quantum gravity (or a supersymmetric version thereof) on an AdS background and a conformally-invariant nongravitational field theory defined on the boundary. Since we know a lot about nongravitational quantum field theory that we don't know about quantum gravity, this correspondence (if it is true, which seems likely but remains unproven) reveals a great deal about what can happen in quantum gravity.[1]

So we have three spacetimes of maximal symmetry: Minkowski ($\kappa = 0$), de Sitter ($\kappa > 0$), and anti-de Sitter ($\kappa < 0$). Are any one of these useful models for the real world? For that matter, are they solutions to Einstein's equation? Start by taking the trace of the Riemann tensor as given by (8.1), specifying to four dimensions:

$$R_{\mu\nu} = 3\kappa g_{\mu\nu}, \quad R = 12\kappa. \tag{8.17}$$

So the Ricci tensor is proportional to the metric in a maximally symmetric space. A spacetime with this property is sometimes called an Einstein space; the Einstein static universe is *not* an example of an Einstein space, which can sometimes be confusing. What is worse, we will later encounter the Einstein–de Sitter cosmology, which is not related to Einstein spaces, the Einstein static universe, or to de Sitter space. The Einstein tensor is

$$G_{\mu\nu} = R_{\mu\nu} - \tfrac{1}{2}Rg_{\mu\nu} = -3\kappa g_{\mu\nu}. \tag{8.18}$$

Therefore, Einstein's equation $G_{\mu\nu} = 8\pi G T_{\mu\nu}$ implies (in a maximally symmetric spacetime, not in general) that the energy-momentum tensor is proportional to the metric:

$$T_{\mu\nu} = -\frac{3\kappa}{8\pi G} g_{\mu\nu}. \tag{8.19}$$

Such an energy-momentum tensor corresponds to a vacuum energy or cosmological constant, as discussed in Chapter 4. The energy density and pressure are given by

$$\rho = -p = \frac{3\kappa}{8\pi G}. \tag{8.20}$$

If ρ is positive, we get a de Sitter solution; if ρ is negative, we get anti-de Sitter.

But in our universe, we have ordinary matter and radiation, as well as a possible vacuum energy. Our maximally symmetric spacetimes are not compatible

[1] For a comprehensive review article, see O. Aharony, S.S. Gubser, J.M. Maldacena, H. Ooguri, and Y. Oz, *Phys. Rept.* **323**, 183 (2000), `http://arxiv.org/hep-th/9905111`.

with a dynamically interesting amount of matter and/or radiation. Furthermore, since we observe the visible matter in the universe to be moving apart (the universe is expanding, as discussed below), the density of matter was higher in the past; so even if the matter contribution to the total energy were negligible today, it would have been appreciable in the earlier universe. The maximally symmetric spacetimes are therefore not reasonable models of the real world. They do, however, represent the (locally) unique solutions to Einstein's equation in the absence of any ordinary matter or gravitational radiation; it is in this sense that they may be thought of as ground states of general relativity.

8.2 ■ ROBERTSON–WALKER METRICS

To describe the real world, we are forced to give up the "perfect" Copernican principle, which implies symmetry throughout space and time, and postulate something more forgiving. It turns out to be straightforward, and consistent with observation, to posit that the universe is *spatially* homogeneous and isotropic, but evolving in time. In general relativity this translates into the statement that the universe can be foliated into spacelike slices such that each three-dimensional slice is maximally symmetric. We therefore consider our spacetime to be $\mathbf{R} \times \Sigma$, where $\mathbf{R}$ represents the time direction and Σ is a maximally symmetric three-manifold. The spacetime metric thus takes the form

$$ds^2 = -\mathrm{d}t^2 + R^2(t)d\sigma^2, \tag{8.21}$$

where t is the timelike coordinate, $R(t)$ is a function known as the **scale factor**, and $d\sigma^2$ is the metric on Σ, which can be expressed as

$$d\sigma^2 = \gamma_{ij}(u)\,\mathrm{d}u^i \mathrm{d}u^j, \tag{8.22}$$

where (u^1, u^2, u^3) are coordinates on Σ and γ_{ij} is a maximally symmetric three-dimensional metric. The scale factor tells us how big the spacelike slice Σ is at the moment t. (Don't confuse it with the curvature scalar.) The coordinates used here, in which the metric is free of cross terms $\mathrm{d}t\,\mathrm{d}u^i$ and coefficient of $\mathrm{d}t^2$ is independent of the u^i, are known as **comoving coordinates**, a special case of the Gaussian normal coordinates discussed in Appendix D. An observer who stays at constant u^i is also called "comoving." Only a comoving observer will think that the universe looks isotropic; in fact on Earth we are not quite comoving, and as a result we see a dipole anisotropy in the cosmic microwave background as a result of the conventional Doppler effect.

Our interest is therefore in maximally symmetric Euclidean three-metrics γ_{ij}. We know that maximally symmetric metrics obey

$${}^{(3)}R_{ijkl} = k(\gamma_{ik}\gamma_{jl} - \gamma_{il}\gamma_{jk}), \tag{8.23}$$

where for future convenience we have introduced

$$k = {}^{(3)}R/6, \tag{8.24}$$

and we put a superscript $^{(3)}$ on the Riemann tensor to remind us that it is associated with the three-metric γ_{ij}, not the metric of the entire spacetime. The Ricci tensor is then

$$^{(3)}R_{jl} = 2k\gamma_{jl}. \tag{8.25}$$

If the space is to be maximally symmetric, then it will certainly be spherically symmetric. We already know something about spherically symmetric spaces from our exploration of the Schwarzschild solution; the metric can be put in the form

$$d\sigma^2 = \gamma_{ij}\,\mathrm{d}u^i\,\mathrm{d}u^j = e^{2\beta(\bar{r})}\,\mathrm{d}\bar{r}^2 + \bar{r}^2 d\Omega^2, \tag{8.26}$$

where $\bar{r}$ is the radial coordinate and the metric on the two-sphere is $\mathrm{d}\Omega^2 = \mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2$ as usual. The components of the Ricci tensor for such a metric can be obtained from (5.14), the Ricci tensor for a static, spherically symmetric spacetime, by setting $\alpha = 0$ and $r = \bar{r}$, which gives

$$\begin{aligned} ^{(3)}R_{11} &= \frac{2}{\bar{r}}\partial_1\beta \\ ^{(3)}R_{22} &= e^{-2\beta}(\bar{r}\partial_1\beta - 1) + 1 \\ ^{(3)}R_{33} &= [e^{-2\beta}(\bar{r}\partial_1\beta - 1) + 1]\sin^2\theta\,. \end{aligned} \tag{8.27}$$

We set these proportional to the metric using (8.25), and can solve for $\beta(\bar{r})$:

$$\beta = -\tfrac{1}{2}\ln(1 - k\bar{r}^2), \tag{8.28}$$

which yields the metric on the three-surface Σ,

$$d\sigma^2 = \frac{\mathrm{d}\bar{r}^2}{1 - k\bar{r}^2} + \bar{r}^2 d\Omega^2. \tag{8.29}$$

Notice from (8.24) that the value of k sets the curvature, and therefore the size, of the spatial surfaces. It is common to normalize this so that

$$k \in \{+1, 0, -1\}, \tag{8.30}$$

and absorb the physical size of the manifold into the scale factor $R(t)$.

The $k = -1$ case corresponds to constant negative curvature on Σ, and is sometimes called **open**; the $k = 0$ case corresponds to no curvature on Σ, and is called **flat**; the $k = +1$ case corresponds to positive curvature on Σ, and is sometimes called **closed**. The physical interpretation of these cases is made more clear using an alternative form of the metric, obtained by introducing a new radial coordinate χ defined by

$$\mathrm{d}\chi = \frac{\mathrm{d}\bar{r}}{\sqrt{1 - k\bar{r}^2}}. \tag{8.31}$$

This can be integrated to obtain

$$\bar{r} = S_k(\chi), \tag{8.32}$$

where

$$S_k(\chi) \equiv \begin{cases} \sin(\chi), & k = +1 \\ \chi, & k = 0 \\ \sinh(\chi), & k = -1, \end{cases} \tag{8.33}$$

so that

$$d\sigma^2 = \mathrm{d}\chi^2 + S_k{}^2(\chi) d\Omega^2. \tag{8.34}$$

For the flat case $k = 0$, the metric on Σ becomes

$$\begin{aligned} d\sigma^2 &= \mathrm{d}\chi^2 + \chi^2 d\Omega^2 \\ &= \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2, \end{aligned} \tag{8.35}$$

which is simply flat Euclidean space. Globally, it could describe $\mathbf{R}^3$ or a more complicated manifold, such as the three-torus $S^1 \times S^1 \times S^1$. For the closed case $k = +1$ we have

$$d\sigma^2 = \mathrm{d}\chi^2 + \sin^2\chi \, d\Omega^2, \tag{8.36}$$

which is the metric of a three-sphere. In this case the only possible global structure is the complete three-sphere (except for the nonorientable manifold $\mathbf{RP}^3$, obtained by identifying antipodal points on S^3). Finally in the open $k = -1$ case we obtain

$$d\sigma^2 = \mathrm{d}\chi^2 + \sinh^2\chi \, d\Omega^2. \tag{8.37}$$

This is the metric for a three-dimensional space of constant negative curvature, a generalization of the hyperboloid discussed in Section 3.9. Globally such a space could extend forever (which is the origin of the word "open"), but it could also describe a nonsimply-connected compact space (so "open" is really not the most accurate description).

The metric on spacetime describes one of these maximally-symmetric hypersurfaces evolving in size, and can be written

$$ds^2 = -\mathrm{d}t^2 + R^2(t)\left[\frac{\mathrm{d}\bar{r}^2}{1 - k\bar{r}^2} + \bar{r}^2 d\Omega^2\right]. \tag{8.38}$$

This is the **Robertson–Walker (RW) metric**. We have not yet made use of Einstein's equation; that will determine the behavior of the scale factor $R(t)$. Note that the substitutions

$$R \to \lambda^{-1} R$$
$$\bar{r} \to \lambda \bar{r}$$
$$k \to \lambda^{-2} k \tag{8.39}$$

leave (8.38) invariant. Therefore we can choose a convenient normalization. In the variables where the curvature k is normalized to $\{+1, 0, -1\}$, the scale factor has units of distance and the radial coordinate $\bar{r}$ (or χ) is actually dimensionless; this is the most popular choice. We will flout the conventional wisdom and instead work with a dimensionless scale factor

$$a(t) = \frac{R(t)}{R_0}, \tag{8.40}$$

a coordinate with dimensions of distance

$$r = R_0 \bar{r}, \tag{8.41}$$

and a curvature parameter with dimensions of (length)$^{-2}$,

$$\kappa = \frac{k}{R_0^2}. \tag{8.42}$$

Note that κ can take on any value, not just $\{+1, 0, -1\}$. In these variables the Robertson–Walker metric is

$$\boxed{ds^2 = -\mathrm{d}t^2 + a^2(t)\left[\frac{\mathrm{d}r^2}{1-\kappa r^2} + r^2\, d\Omega^2\right].} \tag{8.43}$$

To convert to the more common notation, just plug in the relations (8.40), (8.41), and (8.42).

With the metric in hand, we can set about computing the connection coefficients and curvature tensor. Setting $\dot{a} \equiv da/dt$, the Christoffel symbols are given by

$$\Gamma^0_{11} = \frac{a\dot{a}}{1-\kappa r^2} \qquad \Gamma^1_{11} = \frac{\kappa r}{1-\kappa r^2}$$
$$\Gamma^0_{22} = a\dot{a}r^2 \qquad \Gamma^0_{33} = a\dot{a}r^2 \sin^2\theta$$
$$\Gamma^1_{01} = \Gamma^2_{02} \qquad \Gamma^3_{03} = \frac{\dot{a}}{a}$$
$$\Gamma^1_{22} = -r(1-\kappa r^2) \qquad \Gamma^1_{33} = -r(1-\kappa r^2)\sin^2\theta$$
$$\Gamma^2_{12} = \Gamma^3_{13} = \frac{1}{r}$$
$$\Gamma^2_{33} = -\sin\theta\,\cos\theta \qquad \Gamma^3_{23} = \cot\theta, \tag{8.44}$$

or related to these by symmetry. The nonzero components of the Ricci tensor are

$$\begin{aligned} R_{00} &= -3\frac{\ddot{a}}{a} \\ R_{11} &= \frac{a\ddot{a} + 2\dot{a}^2 + 2\kappa}{1 - \kappa r^2} \\ R_{22} &= r^2(a\ddot{a} + 2\dot{a}^2 + 2\kappa) \\ R_{33} &= r^2(a\ddot{a} + 2\dot{a}^2 + 2\kappa)\sin^2\theta, \end{aligned} \tag{8.45}$$

and the Ricci scalar is then

$$R = 6\left[\frac{\ddot{a}}{a} + \left(\frac{\dot{a}}{a}\right)^2 + \frac{\kappa}{a^2}\right]. \tag{8.46}$$

8.3 ■ THE FRIEDMANN EQUATION

The RW metric is defined for any behavior of the scale factor $a(t)$; our next step will be to plug it into Einstein's equation to derive the Friedmann equation(s) relating the scale factor to the energy-momentum of the universe. We will choose to model matter and energy by a perfect fluid. It is clear that, if a fluid that is isotropic in some frame leads to a metric that is isotropic in some frame, the two frames will coincide; that is, the fluid will be at rest in comoving coordinates. The four-velocity is then

$$U^\mu = (1, 0, 0, 0), \tag{8.47}$$

and the energy-momentum tensor

$$T_{\mu\nu} = (\rho + p)U_\mu U_\nu + p g_{\mu\nu} \tag{8.48}$$

becomes

$$T_{\mu\nu} = \begin{pmatrix} \rho & 0 & 0 & 0 \\ 0 & & & \\ 0 & & g_{ij}p & \\ 0 & & & \end{pmatrix}. \tag{8.49}$$

With one index raised this takes the convenient form

$$T^\mu{}_\nu = \mathrm{diag}(-\rho, p, p, p). \tag{8.50}$$

Note that the trace is given by

$$T = T^\mu{}_\mu = -\rho + 3p. \tag{8.51}$$

Before plugging in to Einstein's equation, it is educational to consider the zero component of the conservation of energy equation:

$$\begin{aligned} 0 &= \nabla_\mu T^\mu{}_0 \\ &= \partial_\mu T^\mu{}_0 + \Gamma^\mu_{\mu\lambda} T^\lambda{}_0 - \Gamma^\lambda_{\mu 0} T^\mu{}_\lambda \\ &= -\partial_0 \rho - 3\frac{\dot{a}}{a}(\rho + p). \end{aligned} \tag{8.52}$$

To make progress we can choose an **equation of state**, a relationship between ρ and p. Often the perfect fluids relevant to cosmology obey the simple equation of state

$$p = w\rho, \tag{8.53}$$

where w is a constant independent of time. Of course we are free to define the parameter $w = p/\rho$ whether or not it remains constant; if w varies, however, it is not really legitimate to call $p = w\rho$ the "equation of state." The conservation of energy equation becomes

$$\boxed{\frac{\dot{\rho}}{\rho} = -3(1+w)\frac{\dot{a}}{a}.} \tag{8.54}$$

If w is a constant, this can be integrated to obtain

$$\rho \propto a^{-3(1+w)}. \tag{8.55}$$

To get an idea about what values of w are allowed, refer to the discussion of energy conditions in Chapter 4. The Null Dominant Energy Condition, which allows for a vacuum energy of either sign but otherwise requires matter that cannot destabilize the vacuum, implies

$$|w| \leq 1. \tag{8.56}$$

While this requirement is by no means set in stone, it seems like a sensibly conservative starting point for investigations of what might happen in the real world.

The two most popular examples of cosmological fluids are known as **matter** and **radiation**. Matter is any set of collisionless, nonrelativistic particles, which will have essentially zero pressure:

$$p_{\mathrm{M}} = 0. \tag{8.57}$$

Examples include ordinary stars and galaxies, for which the pressure is negligible in comparison with the energy density. Matter is also known as *dust*, and universes whose energy density is mostly due to matter are known as **matter-dominated**. The energy density in matter falls off as

$$\rho_{\mathrm{M}} \propto a^{-3}. \tag{8.58}$$

This is simply interpreted as the decrease in the number density of particles as the universe expands. For matter the energy density is dominated by the rest energy,

which is proportional to the number density. Radiation may be used to describe either actual electromagnetic radiation, or massive particles moving at relative velocities sufficiently close to the speed of light that they become indistinguishable from photons (at least as far as their equation of state is concerned). Although an isotropic gas of relativistic particles is a perfect fluid and thus has an energy-momentum tensor given by (8.48), we also know that $T_{\mu\nu}$ for electromagnetism can be expressed in terms of the field strength as

$$T^{\mu\nu} = F^{\mu\lambda}F^{\nu}{}_{\lambda} - \tfrac{1}{4}g^{\mu\nu}F^{\lambda\sigma}F_{\lambda\sigma}. \tag{8.59}$$

The trace of this is given by

$$T^{\mu}{}_{\mu} = F^{\mu\lambda}F_{\mu\lambda} - \tfrac{1}{4}(4)F^{\lambda\sigma}F_{\lambda\sigma} = 0. \tag{8.60}$$

But this must also equal (8.51), so the equation of state is

$$p_{\mathrm{R}} = \tfrac{1}{3}\rho_{\mathrm{R}}. \tag{8.61}$$

A universe in which most of the energy density is in the form of radiation is known as **radiation-dominated**. The energy density in radiation falls off as

$$\rho_{\mathrm{R}} \propto a^{-4}. \tag{8.62}$$

Thus, the energy density in radiation falls off slightly faster than that in matter; this is because the number density of photons decreases in the same way as the number density of nonrelativistic particles, but individual photons also lose energy as a^{-1} as they redshift, which we will see later. Likewise, massive but relativistic particles will lose energy as they "slow down" in comoving coordinates. We believe that today the radiation energy density is much less than that of matter, with $\rho_{\mathrm{M}}/\rho_{\mathrm{R}} \sim 10^3$. However, in the past the universe was much smaller, and the energy density in radiation would have dominated at very early times.

As we have discussed, vacuum energy also takes the form of a perfect fluid, with an equation of state $p_\Lambda = -\rho_\Lambda$. The energy density is constant,

$$\rho_\Lambda \propto a^0. \tag{8.63}$$

Since the energy density in matter and radiation decreases as the universe expands, if there is a nonzero vacuum energy it tends to win out over the long term, as long as the universe doesn't start contracting. If this happens, we say that the universe becomes **vacuum-dominated**. de Sitter and anti-de Sitter are vacuum-dominated solutions.

We now turn to Einstein's equation. Recall that it can be written in the form (4.45):

$$R_{\mu\nu} = 8\pi G\left(T_{\mu\nu} - \tfrac{1}{2}g_{\mu\nu}T\right). \tag{8.64}$$

The $\mu\nu = 00$ equation is

$$-3\frac{\ddot{a}}{a} = 4\pi G(\rho + 3p), \tag{8.65}$$

and the $\mu\nu = ij$ equations give

$$\frac{\ddot{a}}{a} + 2\left(\frac{\dot{a}}{a}\right)^2 + 2\frac{\kappa}{a^2} = 4\pi G(\rho - p). \tag{8.66}$$

There is only one distinct equation from $\mu\nu = ij$, due to isotropy. We can use (8.65) to eliminate second derivatives in (8.66), and do a little cleaning up to obtain

$$\boxed{\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{\kappa}{a^2},} \tag{8.67}$$

and

$$\boxed{\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}(\rho + 3p).} \tag{8.68}$$

Together these are known as the **Friedmann equations**, and metrics of the form (8.43) obey these equations define Friedmann–Robertson–Walker (FRW) universes. In fact, if we know the dependence of ρ on a, the first of these (8.67) is enough to solve for $a(t)$; when you hear people refer to *the* Friedmann equation, this is the one to which they are referring, whereas (8.68) is sometimes called the *second* Friedmann equation.

A bunch of terminology is associated with the cosmological parameters, and we will just introduce the basics here. The rate of expansion is characterized by the **Hubble parameter**,

$$\boxed{H = \frac{\dot{a}}{a}.} \tag{8.69}$$

The value of the Hubble parameter at the present epoch is the Hubble constant, H_0. Current measurements lead us to believe that the Hubble constant is 70 ± 10 km/sec/Mpc. (Mpc stands for megaparsec, which is 3.09×10^{24} cm.) Since there is still some uncertainty in this value, we often parameterize the Hubble constant as

$$H_0 = 100h \text{ km/sec/Mpc}, \tag{8.70}$$

so that $h \approx 0.7$. Typical cosmological scales are set by the **Hubble length**

$$\begin{aligned} d_H &= H_0^{-1} c \\ &= 9.25 \times 10^{27} h^{-1} \text{ cm} \\ &= 3.00 \times 10^{3} h^{-1} \text{ Mpc}, \end{aligned} \tag{8.71}$$

and the **Hubble time**

$$\begin{aligned} t_H &= H_0^{-1} \\ &= 3.09 \times 10^{17} h^{-1} \text{ sec} \\ &= 9.78 \times 10^{9} h^{-1} \text{ yr}. \end{aligned} \tag{8.72}$$

Of course since we usually set $c = 1$, you will see H_0^{-1} referred to as both the Hubble length and the Hubble time. There is also the **deceleration parameter**,

$$q = -\frac{a\ddot{a}}{\dot{a}^2}, \tag{8.73}$$

which measures the rate of change of the rate of expansion.

Another useful quantity is the **density parameter**,

$$\boxed{\Omega = \frac{8\pi G}{3H^2}\rho = \frac{\rho}{\rho_{\text{crit}}},} \tag{8.74}$$

where the **critical density** is defined by

$$\rho_{\text{crit}} = \frac{3H^2}{8\pi G}. \tag{8.75}$$

This quantity, which will generally change with time, is called the *critical* density because the Friedmann equation (8.67) can be written

$$\Omega - 1 = \frac{\kappa}{H^2 a^2}. \tag{8.76}$$

The sign of κ is therefore determined by whether Ω is greater than, equal to, or less than, one. We have

$$\begin{array}{lllll} \rho < \rho_{\text{crit}} & \leftrightarrow & \Omega < 1 & \leftrightarrow & \kappa < 0 \quad \leftrightarrow \quad \text{open} \\ \rho = \rho_{\text{crit}} & \leftrightarrow & \Omega = 1 & \leftrightarrow & \kappa = 0 \quad \leftrightarrow \quad \text{flat} \\ \rho > \rho_{\text{crit}} & \leftrightarrow & \Omega > 1 & \leftrightarrow & \kappa > 0 \quad \leftrightarrow \quad \text{closed.} \end{array}$$

The density parameter, then, tells us which of the three Robertson–Walker geometries describes our universe. Determining it observationally is of crucial importance; recent measurements of the cosmic microwave background anisotropy lead us to believe that Ω is very close to unity.

8.4 ■ EVOLUTION OF THE SCALE FACTOR

Given a specification of the amounts of energy density ρ_i in different species i, along with their equations of state $p_i = p_i(\rho_i)$, and the amount of spatial curvature κ, one can solve the Friedmann equation (8.67) to obtain a complete history of the evolution of the scale factor, $a(t)$. In general we simply numerically integrate the Friedmann equation (which is just a first-order differential equation), but it is useful to get a feeling for the types of solutions appropriate to different cosmological parameters.

To simplify our task, let us imagine that all of the different components of energy density evolve as power laws,

$$\rho_i = \rho_{i0} a^{-n_i}. \tag{8.77}$$

Comparing to (8.55), this is equivalent to positing that each equation-of-state parameter $w_i = p_i/\rho_i$ is a constant equal to

$$w_i = \tfrac{1}{3} n_i - 1. \tag{8.78}$$

We can further streamline our expressions by treating the contribution of spatial curvature as a fictitious energy density

$$\rho_c \equiv -\frac{3\kappa}{8\pi G a^2}, \tag{8.79}$$

with a corresponding density parameter

$$\Omega_c = -\frac{\kappa}{H^2 a^2}. \tag{8.80}$$

It's *not* an energy density, of course, so don't forget that this is just notational sleight-of-hand. The behaviors of our favorite sources are summarized in the following table.

	w_i	n_i
matter	0	3
radiation	$\frac{1}{3}$	4
curvature	$-\frac{1}{3}$	2
vacuum	-1	0

(8.81)

In these variables, the Friedmann equation (8.67) can be written

$$H^2 = \frac{8\pi G}{3} \sum_{i(c)} \rho_i, \tag{8.82}$$

where the notation $\sum_{i(c)}$ indicates that we sum not only over all the actual components of energy density ρ_i, but also over the contribution of spatial curvature

ρ_c. Note that if we divide both sides by H^2, we obtain

$$1 = \sum_{i(c)} \Omega_i. \tag{8.83}$$

The right-hand side is *not* the total density parameter Ω, which only gets contributions from actual energy density (not curvature); we therefore have

$$\Omega_c = 1 - \Omega. \tag{8.84}$$

Let's begin by asking what can happen if all of the ρ_i's (including ρ_c) are nonnegative. Because H^2 is proportional to $\sum_{i(c)} \rho_i$, the universe will never undergo a transition from expanding to contracting so long as $\sum_{i(c)} \rho_i \neq 0$. We can also take the time derivative of the Hubble parameter,

$$\dot{H} = \frac{\ddot{a}}{a} - \left(\frac{\dot{a}}{a}\right)^2, \tag{8.85}$$

and plug in the two Friedmann equations (8.67) and (8.68) to obtain

$$\dot{H} = -4\pi G \sum_{i(c)} (1 + w_i)\rho_i. \tag{8.86}$$

Since we are imagining that $|w_i| \leq 1$, when all the ρ_i's are nonnegative we will always have $\dot{H} \leq 0$. In other words, the universe keeps expanding, but the expansion rate continually decreases (which suggests the excellent question, what made it so large in the first place?).

From (8.85) we see that $\ddot{a}$ can be positive and $\dot{H}$ be negative at the same time—the scale factor can be "accelerating" even though the expansion rate as measured by the Hubble parameter is decreasing (for example, if $a \propto t^2$). This is an unavoidable subtlety of non-Euclidean geometry. The Hubble parameter and the derivative of the scale factor are the answers to two different questions. If we set two test particles at a fixed initial distance, and ask by how much they have separated a short time thereafter, the answer is given by the Hubble parameter. If, on the other hand, we pick some fixed source, and ask how it appears to move away from us with time, the answer is given by the change in the scale factor. There are consequently two very different and equally legitimate senses of "accelerating" (or "decelerating"). In practice, "accelerating" usually refers to a situation in which $\ddot{a} > 0$, even if $\dot{H} < 0$. This discussion is not completely academic; as we will see below, our current real universe seems to be of this type.

It is by no means necessary that each ρ_i should be nonnegative. Matter and radiation arise from dynamical particles and fields, and we consequently expect that their energy densities will never be negative; if they could be, empty space could decay into a collection of positive- and negative-energy fields. But vacuum and curvature are different stories. Vacuum energy is nondynamical, so a negative value cannot induce any instabilities, while curvature is simply a property of the spatial geometry, and can have either sign. If we therefore have either a

negative vacuum energy or a positive spatial curvature (remember $\rho_c \propto -\kappa$), the Hubble parameter can vanish and even change sign. An example is provided by the de Sitter metric (8.6), which has a positive vacuum energy but also a positive spatial curvature; it describes a universe that initially collapses, reaches a turning point, and thereafter begins to expand.

The real world is an untidy place, consisting of numerous different kinds of energy density. Because different sources evolve at different rates, however, for long periods the energy density will be clearly dominated by one kind of source. It is therefore very useful to examine solutions to the Friedmann equation when there is only one kind of energy density $\rho \propto a^{-n}$. Because we are including spatial curvature as an effective energy source, this means we are considering either flat universes dominated by a single source, or completely empty universes with spatial curvature. The Friedmann equation then implies

$$\dot{a} \propto a^{1-n/2}. \tag{8.87}$$

This can be immediately integrated to obtain

$$\boxed{a \propto t^{2/n} \qquad (\text{for } \rho \propto a^{-n}).} \tag{8.88}$$

Consider for example a flat universe dominated by matter, $\Omega = \Omega_M = 1$; this is known as the Einstein–de Sitter model, and for a long time was the favorite (at least among theorists) to describe the real world. In an Einstein–de Sitter universe, the scale factor evolves as $a \propto t^{2/3}$. A flat radiation-dominated universe, meanwhile, evolves as $a \propto t^{1/2}$. The conformal diagram for any such universe with $n > 2$ is derived in Appendix H. Even though we believe there are nonzero amounts of matter, radiation, and vacuum energy in the real universe, these solutions are still very useful; as we discuss later, the universe was radiation-dominated at early times, and was matter dominated as the universe expanded from $a \sim 1/3000$ to $a \sim 1/2$.

These solutions all feature a singularity at $a = 0$, known as the **Big Bang**. It represents the creation of the universe from a singular state, not an explosion of matter into a pre-existing spacetime. It might be hoped that the perfect symmetry of our FRW universes is responsible for this singularity, but in fact that's not true; cosmological singularity theorems show that any universe with $\rho > 0$ and $p \geq 0$ must have begun at a singularity. Of course the energy density becomes arbitrarily high as $a \to 0$, and we don't expect classical general relativity to be an accurate description of nature in this regime; presumably quantum gravity becomes important, although it is unclear how at present.

Looking at (8.88), we see that a universe dominated by vacuum energy ($n = 0$) is clearly a special case. The scale factor then expands as an exponential rather than a power law; the entire metric is

$$ds^2 = -dt^2 + e^{Ht}[dx^2 + dy^2 + dz^2], \tag{8.89}$$

where the Hubble parameter H is a constant. Of course, in Section 8.1 we already described a cosmological spacetime with a positive cosmological constant: de Sitter space, which featured $\kappa > 0$ and $a \propto \cosh(t/\alpha)$. What is the relationship between that solution and the one here, with $\kappa = 0$ and $a \propto \exp(Ht)$? They are the same spacetime, represented in different coordinates. One way to verify this is to calculate the Riemann tensor for (8.89) and check that it has the characteristic form of a maximally symmetric spacetime, (8.1). Since maximally symmetric spacetimes with positive curvature are locally unique, the metrics (8.6) and (8.89) must describe the same manifold, or parts thereof. In fact, the coordinates of (8.89) only cover part of de Sitter; they are incomplete in the past. In the exercises you are asked to show that comoving geodesics in these coordinates reach $t = -\infty$ in finite affine parameter; they run into the edge of the coordinates. In the conformal diagram of Figure 8.1, these coordinates cover the upper-right triangular portion of the square. See Hawking and Ellis (1973) for a more complete description of different coordinate systems on de Sitter and anti-de Sitter.

Another interesting special case is the completely empty universe, with $\rho = 0$, but with spatial curvature. The Friedmann equation becomes

$$H^2 = -\frac{\kappa}{a^2}, \tag{8.90}$$

so the curvature κ must be negative. Thinking of curvature as a fictitious energy density $\rho_c \propto a^{-2}$, from (8.88) we know that such a universe will expand linearly, $a \propto t$. This spacetime is known as the **Milne universe**. However, just as with de Sitter, we know of another cosmological spacetime with $\rho = 0$—in this case, flat Minkowski space. Once again, the Milne spacetime is just a patch of Minkowski in a certain incomplete coordinate system. It can be thought of as the interior of the future light cone of some fixed point in Minkowski, foliated by negatively-curved hyperboloids. To check, it would suffice to calculate all of the components of the Riemann tensor, which turn out to vanish; any spacetime with vanishing Riemann curvature is locally Minkowski.

In contrast to these idealized solutions, a realistic cosmology will feature several forms of energy-momentum. In the current universe, we feel confident that the radiation density is significantly lower than the matter density, but that vacuum and matter are both dynamically important. It is therefore convenient to parameterize universes like ours by Ω_M and Ω_Λ, with the curvature fixed by $\Omega_c = 1 - \Omega_M - \Omega_\Lambda$. The expansion history of some particular examples of such universes is shown in Figure 8.3. As these universes expand, the relative influences of matter, curvature, and vacuum are altered, since the corresponding densities evolve at different rates:

$$\Omega_\Lambda \propto \Omega_c a^2 \propto \Omega_M a^3. \tag{8.91}$$

As $a \to 0$ in the past, curvature and vacuum will be negligible, and the universe will behave as Einstein–de Sitter. As $a \to \infty$ in the future, curvature and matter will be negligible, and the universe will asymptote to de Sitter; unless the scale

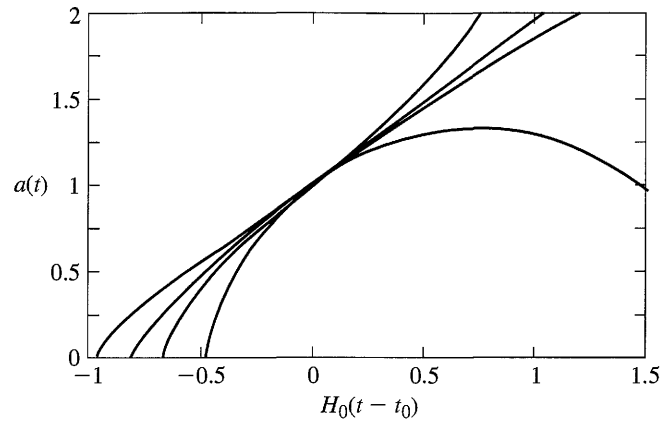

FIGURE 8.3 Expansion histories for different values of Ω_M and Ω_Λ. From top to bottom, the curves describe $(\Omega_M, \Omega_\Lambda) = (0.3, 0.7)$, $(0.3, 0.0)$, $(1.0, 0.0)$, and $(4.0, 0.0)$.

factor never reaches infinity, because the universe begins to recollapse at some finite time.

Recollapse will *always* occur if the vacuum energy is negative; as the universe expands, the vacuum energy eventually dominates, and the effect of $\Omega_\Lambda < 0$ is to cause deceleration and recollapse (just as the effect of $\Omega_\Lambda > 0$ is to push the universe apart). Recollapse is also possible with $\Omega_\Lambda \geq 0$, if Ω_M is sufficiently large that it halts the universal expansion before Ω_Λ has a chance to take over. The possibilities are expressed as different regions of the Ω_M/Ω_Λ parameter space in Figure 8.4. The diagonal line represents $\Omega_{total} = 1$, implying $\kappa = 0$.

To determine the dividing line between perpetual expansion and eventual recollapse, note that collapse requires the Hubble parameter to pass through zero as it changes from positive to negative. The scale factor a_* at which this turnaround occurs can be found by setting $H = 0$ in the Friedmann equation,

$$H^2 = 0 = \frac{8\pi G}{3}\left(\rho_{M0}a_*^{-3} + \rho_{\Lambda 0} + \rho_{c0}a_*^{-2}\right). \tag{8.92}$$

We can divide this by H_0^2, use $\Omega_{c0} = 1 - \Omega_{M0} - \Omega_{\Lambda 0}$, and rearrange a bit to obtain

$$\Omega_{\Lambda 0}a_*^3 + (1 - \Omega_{M0} - \Omega_\Lambda)a_* + \Omega_{M0} = 0. \tag{8.93}$$

This is a cubic equation for a_*, the scale factor at turnaround. Of course we don't actually care very much about a_*; what we care about are the values of $\Omega_{\Lambda 0}$, given Ω_{M0}, for which a real solution to (8.93) exists. Solving the cubic equation and doing some math, we find that the value of $\Omega_{\Lambda 0}$ for which the universe will expand forever is given by

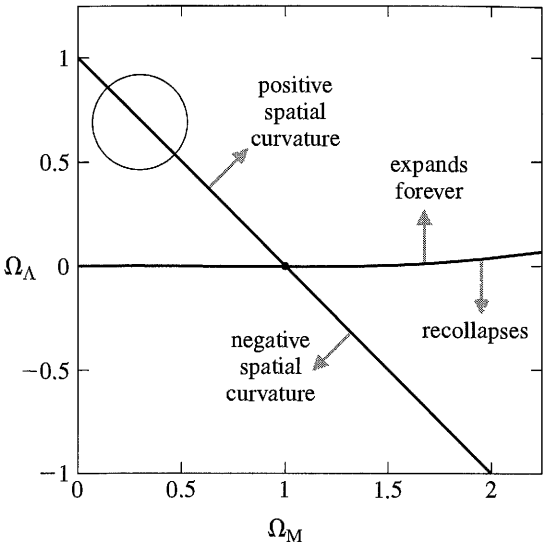

FIGURE 8.4 Properties of universes dominated by matter and vacuum energy, as a function of the density parameters Ω_M and Ω_Λ. The circular region in the upper-left corner represents roughly those values favored by experimental data (as of 2003).

$$\Omega_{\Lambda 0} \geq \begin{cases} 0 & 0 \leq \Omega_{M0} \leq 1 \\ 4\Omega_{M0}\cos^3\left[\frac{1}{3}\cos^{-1}\left(\frac{1-\Omega_{M0}}{\Omega_M 0}\right)+\frac{4\pi}{3}\right] & \Omega_{M0} > 1. \end{cases} \tag{8.94}$$

Note that, when $\Omega_{\Lambda 0} = 0$, open and flat universes ($\Omega_0 = \Omega_{M0} \leq 1$) will expand forever, while closed universes ($\Omega_0 = \Omega_{M0} > 1$) will recollapse. Traditional disdain for the the cosmological constant has led to a folk belief that this is a necessary correspondence; once the possibility of vacuum energy is admitted, however, any combination of spatial geometry and eventual fate is possible.

In the upper-left corner of Figure 8.4, we have indicated the currently favored values of the cosmological parameters: $\Omega_{M0} \sim 0.3$, $\Omega_{\Lambda 0} \sim 0.7$, as we will discuss in Section 8.7. This is well into the regime of perpetual expansion; if the vacuum energy remains truly constant (which it might not), our universe is fated to continue its expansion for all time.

We end this section by noting the difficulty of finding static solutions to the Friedmann equations. To be static, we must have not only $\dot{a} = 0$, but also $\ddot{a} = 0$. From (8.68), this can only happen if the pressure is

$$p = -\tfrac{1}{3}\rho, \tag{8.95}$$

and from (8.67), there must be a nonvanishing spatial curvature

$$\frac{\kappa}{a^2} = \frac{8\pi G}{3}\rho. \tag{8.96}$$

Because the energy density and pressure must be of opposite sign, these conditions can't be fulfilled if we only invoke matter or radiation. When Einstein first looked for cosmological solutions in GR, astronomers had not yet discovered that the universe was expanding, so the lack of static solutions was considered problematic. This provided the motivation for Einstein to introduce the cosmological constant; the static conditions can be satisfied by a combination of matter and vacuum energy, with

$$\rho_\Lambda = \tfrac{1}{2}\rho_M, \tag{8.97}$$

along with the appropriate positive spatial curvature. These parameters describe the **Einstein static universe**. Today we know that the universe is expanding, so this solution is of little empirical interest; it is, however, extremely useful to theorists, providing the basis for the construction of conformal diagrams.

8.5 ■ REDSHIFTS AND DISTANCES

It is clear that we would like to determine a number of quantities observationally to decide which of the FRW models corresponds to our universe. Obviously we would like to determine H_0, since that is related to the age of the universe. We would also like to know Ω, which determines κ through (8.76). To understand how these quantities might conceivably be measured, let's consider geodesic motion in an FRW universe. There are a number of spacelike Killing vectors, but no timelike Killing vector to give us a notion of conserved energy. There is, however, a Killing tensor. If $U^\mu = (1, 0, 0, 0)$ is the four-velocity of comoving observers, then the tensor

$$K_{\mu\nu} = a^2(g_{\mu\nu} + U_\mu U_\nu) \tag{8.98}$$

satisfies $\nabla_{(\sigma} K_{\mu\nu)} = 0$ (as you can check), and is therefore a Killing tensor. This means that if a particle has four-velocity $V^\mu = dx^\mu/d\lambda$, the quantity

$$K^2 = K_{\mu\nu}V^\mu V^\nu = a^2[V_\mu V^\mu + (U_\mu V^\mu)^2] \tag{8.99}$$

will be a constant along geodesics. Let's think about this, first for massive particles. Then we will have $V_\mu V^\mu = -1$, so

$$(V^0)^2 = 1 + |\vec{V}|^2, \tag{8.100}$$

where $|\vec{V}|^2 = g_{ij}V^iV^j$. We also have $U_\mu V^\mu = -V^0$, so (8.99) implies

$$|\vec{V}| = \frac{K}{a}. \tag{8.101}$$

The particle therefore "slows down" with respect to the comoving coordinates as the universe expands. In fact this is an actual slowing down, in the sense that a gas of particles with initially high relative velocities will cool down as the universe expands.

A similar thing happens to null geodesics. In this case $V_\mu V^\mu = 0$, and (8.99) implies

$$U_\mu V^\mu = \frac{K}{a}. \tag{8.102}$$

But the frequency of the photon as measured by a comoving observer is $\omega = -U_\mu V^\mu$. The frequency of the photon emitted with frequency $\omega_{\rm em}$ will therefore be observed with a lower frequency $\omega_{\rm obs}$ as the universe expands:

$$\frac{\omega_{\rm obs}}{\omega_{\rm em}} = \frac{a_{\rm em}}{a_{\rm obs}}. \tag{8.103}$$

Cosmologists like to speak of this in terms of the **redshift** z between the two events, defined by the fractional change in wavelength:

$$z_{\rm em} = \frac{\lambda_{\rm obs} - \lambda_{\rm em}}{\lambda_{\rm em}}. \tag{8.104}$$

If the observation takes place today ($a_{\rm obs} = a_0 = 1$), this implies

$$\boxed{a_{\rm em} = \frac{1}{1 + z_{\rm em}}.} \tag{8.105}$$

So the redshift of an object tells us the scale factor when the photon was emitted.

Notice that this redshift is not the same as the conventional Doppler effect; it is the expansion of space, not the relative velocities of the observer and emitter, which leads to the redshift. Nevertheless, if we observe galaxies over distances that are small compared to the Hubble radius H_0^{-1} and the radius of spatial curvature $\kappa^{-1/2}$, the expansion of the universe looks very much like a set of galaxies moving apart from each other and the redshift looks very much like the Doppler effect. Consequently, astronomers often think of the redshift in terms of a "velocity" $v = cz$, where c is the speed of light. Even though we know you can't really speak of the relative velocities between two objects at different points of a curved spacetime, the fiction works well over sufficiently short distances. Within this approximation, the "distance" d from us to a galaxy can be taken to be the **instantaneous physical distance** d_P (the distance, in physical units such as centimeters, between us and the location of the galaxy along our current spatial hypersurface). Let's write the RW metric in the form

$$ds^2 = -{\rm d}t^2 + a^2(t) R_0^2 \left[{\rm d}\chi^2 + S_k^2(\chi) d\Omega^2 \right], \tag{8.106}$$

where $S_k(\chi)$ is defined by (8.33), and $k \in \{+1, 0, -1\}$. In this form, the instantaneous physical distance as measured at time t between us ($\chi = 0$) and a galaxy at comoving radial coordinate χ is

$$d_P(t) = a(t) R_0 \chi, \tag{8.107}$$

where χ remains constant because we assume both we and the observed galaxy are perfectly comoving. (They might not be, in which case it is trivial to include the corrections due to so-called "peculiar velocities.") Of course "distance" is in quotes because there are several inequivalent useful notions of distance once we leave this approximation, but they all agree when d_P is small. Then the observed velocity (as inferred from the redshift) is simply

$$v = \dot{d}_P = \dot{a} R_0 \chi = \frac{\dot{a}}{a} d_P. \tag{8.108}$$

Evaluated today, this becomes

$$\boxed{v = H_0 d_P,} \tag{8.109}$$

the famous **Hubble law**: the observed recession velocity is directly proportional to the distance, for galaxies that are not too far away.

If the redshift is not very small, we have to think more carefully about what we mean by "distance" in cosmology. The instantaneous physical distance is a convenient construct, but not itself observable, since observations always refer to events on our past light cone, not our current spatial hypersurface. In Euclidean space there are a number of different ways to infer the distance of an object; we could for example compare its apparent brightness to its intrinsic luminosity, or its apparent angular velocity to its intrinsic transverse speed, or its apparent angular size to its physical extent. For each of these cases, we can define a kind of distance that is what we *would* infer if space were Euclidean and the universe were not expanding.

Let's start with the **luminosity distance** d_L, defined to satisfy

$$d_L^2 = \frac{L}{4\pi F}, \tag{8.110}$$

where L is the absolute luminosity of the source and F is the flux measured by the observer (the energy per unit time per unit area of some detector). This definition comes from the fact that in flat space, for a source at distance d the flux over the luminosity is just one over the area of a sphere centered around the source, $F/L = 1/A(d) = 1/4\pi d^2$. In an FRW universe, however, the flux will be diluted. Conservation of photons tells us that all of the photons emitted by the source will eventually pass through a sphere at comoving distance χ from the emitter. But the flux is diluted by two additional effects: the individual photons redshift by a factor $(1 + z)$, and the photons hit the sphere less frequently, since two photons emitted a time δt apart will be measured at a time $(1 + z)\delta t$ apart. Therefore we will have

$$\frac{F}{L} = \frac{1}{(1+z)^2 A}. \tag{8.111}$$

The area A of a sphere centered at comoving distance χ can be derived from the coefficient of $d\Omega^2$ in (8.106), yielding

$$A = 4\pi R_0^2 S_k{}^2(\chi), \tag{8.112}$$

where we have set $a(t) = 1$ because we are observing the photons today. Putting it all together yields

$$d_L = (1+z) R_0 S_k(\chi). \tag{8.113}$$

The luminosity distance d_L is something we might hope to measure, since there are some astrophysical sources whose absolute luminosities are known. But χ is not observable, so we have to remove that from our equation. On a null geodesic (chosen to be radial for convenience) we have

$$0 = ds^2 = -\mathrm{d}t^2 + a^2 R_0^2 \mathrm{d}\chi^2, \tag{8.114}$$

or

$$\chi = R_0^{-1} \int \frac{dt}{a} = R_0^{-1} \int \frac{da}{a^2 H(a)}, \tag{8.115}$$

where we have used $H = \dot{a}/a$. It is conventional to convert the scale factor to redshift using $a = 1/(1+z)$, so we have

$$\chi(z) = R_0^{-1} \int_0^z \frac{dz'}{H(z')}. \tag{8.116}$$

In order to evaluate the Hubble parameter in this integral we use the Friedmann equation (8.67), which we write as in the previous section as

$$H^2 = \frac{8\pi G}{3} \sum_{i(\mathrm{c})} \rho_i. \tag{8.117}$$

To simplify things, we may again assume that each density component evolves as a power law,

$$\rho_i(z) = \rho_{i0} a^{-n_i} = \rho_{i0} (1+z)^{n_i}, \tag{8.118}$$

Then we can write

$$H(z) = H_0 E(z), \tag{8.119}$$

where

$$E(z) = \left[\sum_{i(\mathrm{c})} \Omega_{i0} (1+z)^{n_i} \right]^{1/2}, \tag{8.120}$$

where the density parameters Ω_i are defined by (8.74). The equations below involving $E(z)$ will be true whether or not the energy sources evolve as power laws; if they do not, simply use $E(z) = H(z)/H_0$ [where $H(z)$ is determined by the Friedmann equation] rather than (8.120).

So the luminosity distance is

$$d_L(z) = (1+z)R_0 S_k\left[R_0^{-1}H_0^{-1}\int \frac{dz'}{E(z')}\right]. \tag{8.121}$$

Note that R_0 drops out when $k = 0$, which is good, because in that case it is a completely arbitrary parameter. Even when it is not arbitrary, it is still more common to speak in terms of $\Omega_{c0} = -k/R_0^2H_0^2$, which can be measured either directly through determinations of the spatial curvature, or by measuring the density parameter and using $\Omega_{c0} = 1 - \Omega_0$. In terms of this parameter we have

$$R_0 = H_0^{-1}\sqrt{-k\Omega_{c0}} = \frac{H_0^{-1}}{\sqrt{|\Omega_{c0}|}}. \tag{8.122}$$

We therefore write the luminosity distance in terms of measurable cosmological parameters as

$$\boxed{d_L(z) = (1+z)\frac{H_0^{-1}}{\sqrt{|\Omega_{c0}|}}S_k\left[\sqrt{|\Omega_{c0}|}\int \frac{dz'}{E(z')}\right].} \tag{8.123}$$

Although it appears unwieldy, this equation is of central importance in cosmology. Given the observables H_0 and Ω_{i0}, we can straightforwardly calculate the luminosity distance to an object at any redshift z; equally well, we can measure $d_L(z)$ for objects at a range of redshifts, and from that information extract H_0 and/or the Ω_{i0}'s.

Along with the luminosity distance are two other related distance measures. Just as the luminosity distance is the distance we infer from the intrinsic and observed luminosity of the source if we were in flat space, the **proper motion distance** d_M is the distance we infer from the intrinsic and observed motion of the source. It is defined to be

$$d_M = \frac{u}{\dot{\theta}}, \tag{8.124}$$

where u is the proper transverse velocity (something you would measure, for example, in k/s) and $\dot{\theta}$ is the observed angular velocity. The **angular diameter distance**, meanwhile, is the distance we infer from the intrinsic and observed size of the source; it is defined to be

$$d_A = \frac{R}{\theta}, \tag{8.125}$$

where R is the proper size of the object and θ is its observed angular diameter. In both cases we can derive formulas analogous to (8.123); fortunately, the unwieldy dependence on the cosmological parameters is common to all the distance measures, and we are left with a simple dependence on redshift:

$$d_L = (1+z)d_M = (1+z)^2 d_A, \tag{8.126}$$

as you are encouraged to check. So if we measure one such distance, we can easily convert to any other; or we can measure different distances independently and use (8.126) to test the consistency of the RW framework.

While we're contemplating distances, let's also consider the elapsed time between now and when the light from an object at redshift z was emitted. If the age of the universe today is t_0 and the age when the photon was emitted is t_*, the **lookback time** is

$$\begin{aligned} t_0 - t_* &= \int_{t_*}^{t_0} dt \\ &= \int_{a_*}^{1} \frac{da}{aH(a)} \\ &= H_0^{-1} \int_0^{z_*} \frac{dz'}{(1+z')E(z')}. \end{aligned} \tag{8.127}$$

For example, consider a flat ($k = 0$) matter-dominated ($\rho = \rho_{\rm M} = \rho_{\rm M0} a^{-3}$) universe. Then

$$E(z) = (1+z)^{3/2}, \tag{8.128}$$

so

$$\begin{aligned} t_0 - t_* &= H_0^{-1} \int_0^{z_*} \frac{dz'}{(1+z')^{5/2}} \\ &= \frac{2}{3} H_0^{-1} \left[1 - (1+z_*)^{-3/2} \right]. \end{aligned} \tag{8.129}$$

The total age of a matter-dominated universe is obtained by letting $t_* \to 0$ ($z_* \to \infty$),

$$t_0({\rm MD}) = \tfrac{2}{3} H_0^{-1}. \tag{8.130}$$

For universes that are not completely matter-dominated, the factor of $\frac{2}{3}$ will be not quite right, but for reasonable values of the cosmological parameters we usually get $t_0 \sim H_0^{-1}$.

8.6 ■ GRAVITATIONAL LENSING

In Chapter 7 we introduced the concept of gravitational lensing: the deflection and time delay of light by a Newtonian gravitational field. In addition to providing a

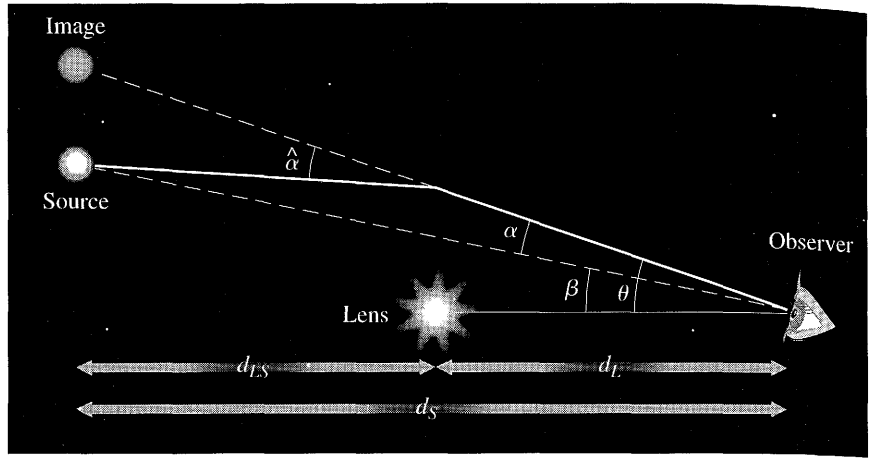

FIGURE 8.5 The geometry of gravitational lensing, encapsulated in the lens equation (8.132). The effect of the lens is to distort the angles β that would be observed in a flat Minkowski background into the angles θ.

test of GR in the Solar System, lensing occurs in numerous astrophysical contexts, and has become an indispensable part of modern cosmology.[2]

Two important features distinguish cosmological lensing from the case we discussed earlier: a Robertson–Walker metric replaces the Minkowski background, and the lenses themselves are often more complex than simple point masses. A typical lensing geometry is portrayed in Figure 8.5. Throughout this discussion we will assume that the lens is "thin"—much smaller in spatial extent than the distances between the source, lens, and observer. In this case we can sensibly speak of a unique distance to the lens, d_L, and between the lens and the source, d_{LS}.

We describe a (possibly complicated) image on the sky by a set of angles between different components of the image. These angles can be thought of as two-dimensional vectors on the sky. The effect of the lens is to distort the angles that would be observed in the absence of any deflection, such as the angle $\vec{\beta}$ between the source and the lens, into a new image characterized by a set of angles $\vec{\theta}$. We assume that the angles are small throughout. This map is described by the **reduced lensing angle** $\vec{\alpha} = \vec{\theta} - \vec{\beta}$. According to the geometry shown in Figure 8.5, it is related to the actual deflection angle $\widehat{\alpha}$ by

$$\vec{\alpha} = \frac{d_{LS}}{d_S}\widehat{\alpha}. \tag{8.131}$$

[2]An excellent overview of gravitational lensing, from which our discussion borrows, can be found in R. Narayan and M. Bartelmann, "Lectures on Gravitational Lensing," 13th Jerusalem Winter School in Theoretical Physics, `http://arxiv.org/astro-ph/9606001`.

We therefore get the **lens equation**

$$\boxed{\vec{\beta} = \vec{\theta} - \frac{d_{LS}}{d_S}\widehat{\alpha}.} \tag{8.132}$$

The lens equation simply describes ray-tracing in a perturbed spacetime.

Of course, we should think carefully about the "distances" d_i portrayed in the figure. Lensing occurs in an expanding universe, which might also have spatial curvature. The lens equation will nevertheless hold if we *define* the distances d_i to be such that the geometrical relations described by the lens equation hold. In other words, these are the distances that we would infer, given the angles and transverse physical sizes, in a static Euclidean spatial background. But this is precisely the definition of the angular diameter distance (8.125). We therefore take all distances in this section to be angular-diameter distances. Note that angular-diameter distances do not necessarily add, so that $d_S \neq d_L + d_{LS}$.

As a simple example, consider a point mass lens. In our investigation of the Newtonian limit in Chapter 7, we found that the deflection angle for a photon traveling through a gravitational potential Φ is given by

$$\widehat{\alpha} = 2\int \vec{\nabla}_\perp \Phi \, ds, \tag{8.133}$$

which for a point mass M at an impact parameter b becomes

$$\widehat{\alpha} = \frac{4GM}{b}. \tag{8.134}$$

The impact parameter can be expressed as $b = d_L\theta$. The lens equation (8.132) becomes

$$\beta = \theta - \frac{d_{LS}}{d_S d_L}\frac{4GM}{\theta}. \tag{8.135}$$

It is illuminating to consider the simplest situation, in which the source and lens are collinear ($\beta = 0$). In that case, the source will be lensed into an **Einstein ring** surrounding the lens, at an angular separation given by the **Einstein angle:**

$$\theta_{\rm E} = \sqrt{\frac{4GMd_{LS}}{d_L d_S}}. \tag{8.136}$$

The Einstein angle sets a characteristic scale for lensing, even in more complicated configurations. We can also define an associated distance scale, the **Einstein radius:**

$$R_{\rm E} = \sqrt{\frac{4GMd_L d_{LS}}{d_S}}. \tag{8.137}$$

When converting to centimeters or other physical units, don't forget that $c = 1$ in all of our equations. To get a feeling for the amount of lensing in typical astrophysical situations, we can consider two common occurrences: "microlensing" by approximately solar-mass objects within our galaxy, and cosmological lensing by galaxies or clusters. In the former case the Einstein angle will be of order milliarcseconds, while in the latter case it will be of order arcseconds:

$$\theta_{\rm E} = 0.9\sqrt{\left(\frac{M}{M_\odot}\right)\left(\frac{10\ {\rm kpc}}{D}\right)}\ {\rm milliarcsecs}$$
$$= 0.9\sqrt{\left(\frac{M}{10^{11}M_\odot}\right)\left(\frac{\rm Gpc}{D}\right)}\ {\rm arcsecs}. \tag{8.138}$$

Sticking for the moment with the point-mass lens, most often we will not be lucky enough to have source and lens perfectly aligned, although a number of spectacular examples of Einstein rings have been observed. Then we can solve (8.135) to obtain two values of the image angle,

$$\theta_\pm = \tfrac{1}{2}\left(\beta \pm \sqrt{\beta^2 + 4\theta_{\rm E}^2}\right). \tag{8.139}$$

The image at θ_+ will always be outside the Einstein angle, while θ_- will be inside. In fact this formula is somewhat misleading, as there will always be an odd number of images; for a point mass lens, the third image would be located at the same position as the lens itself.

Now let's consider more general lenses than point masses. We know that the deflection angle will be given in terms of the Newtonian gravitational potential by (8.133). We can define the **lensing potential** by integrating over past-directed geodesic paths emanating from the observer, as

$$\psi(\vec{\theta}) = 2\frac{d_{LS}}{d_L d_S}\int \Phi(d_L\vec{\theta}, s)\, ds. \tag{8.140}$$

In terms of the lensing potential, we can straightforwardly derive the reduced lensing angle by taking the gradient,

$$\vec{\alpha} = \vec{\nabla}_\theta \psi$$
$$= 2\frac{d_{LS}}{d_S}\int \vec{\nabla}_\perp \Phi\, ds. \tag{8.141}$$

Notice that the angular gradient $\vec{\nabla}_\theta$ is related to $\vec{\nabla}_\perp$, the gradient with respect to transverse distance at the location of the lens, by a factor of d_L. The thin-lens approximation allows us to collapse the integral to quantities evaluated at the location of the lens. We can also take the (two-dimensional) Laplacian of the lensing potential to obtain the **convergence** κ, via

$$\kappa(\vec{\theta}) \equiv \frac{1}{2}\nabla_\theta^2 \psi$$
$$= \frac{d_L d_{LS}}{d_S} \int \nabla^2 \Phi \, ds. \tag{8.142}$$

The convergence can be thought of as a measure of the integrated mass density. We can invert the above expressions to write both the lensing potential and the reduced deflection angle in terms of the convergence, as

$$\psi(\vec{\theta}) = \frac{1}{\pi} \int \kappa(\vec{\theta}') \ln|\vec{\theta} - \vec{\theta}'| \, d^2\theta' \tag{8.143}$$

and

$$\vec{\alpha}(\vec{\theta}) = \frac{1}{\pi} \int \kappa(\vec{\theta}') \frac{\vec{\theta} - \vec{\theta}'}{|\vec{\theta} - \vec{\theta}'|} d^2\theta'. \tag{8.144}$$

To check these equations, remember that the vectors are defined only in the two transverse dimensions.

The convergence describes the focusing of light rays by the gravitational lens. This focusing causes the source to appear larger (just as in a magnifying glass). According to Liouville's theorem of conservation of phase-space density for the photons emitted by the source, the surface brightness of the source will be conserved under lensing; the increase in size therefore leads to magnification of the brightness. At the same time, we can have distortion caused by twisting of the light rays through the lens, which leads to shear of the shape of the image. To describe both phenomena, we consider the 2×2 matrix of derivatives of the lens map,

$$A_{ij} \equiv \frac{\partial \beta^i}{\partial \theta^j}. \tag{8.145}$$

Note that there is no real distinction between upper and lower indices, as they are defined in a two-dimensional Euclidean plane. Since $\vec{\beta} = \vec{\theta} - \vec{\alpha}$, we have

$$A_{ij} = \delta_{ij} - \frac{\partial \alpha^i}{\partial \theta^j}$$
$$= \delta_{ij} - \psi_{ij}, \tag{8.146}$$

where we have introduced the notation

$$\psi_{ij} \equiv \frac{\partial^2 \psi}{\partial \theta^i \partial \theta^j}. \tag{8.147}$$

This matrix A encodes the local properties of the lensing map. Its inverse matrix is known as the *magnification tensor*,

$$M = \frac{\partial \vec{\theta}}{\partial \vec{\beta}} = A^{-1}. \tag{8.148}$$

Why does it get this name? The lens distorts an area element described by $\vec{\beta}$ into one described by $\vec{\theta}$, and the change in area is described by the Jacobian of this map, which is simply the determinant of M. This determinant is defined as the **magnification** μ,

$$\mu = |M| = \frac{1}{|A|}. \tag{8.149}$$

The absolute magnitude of μ tells us the actual change in brightness of the source; μ may be negative, which means that the parity of the image has been flipped. We speak of magnification because lensing is only noticeable if the lens and source are near to each other on the sky, in which case the focusing effect leads only to increases in the apparent brightness; a lens far away from the source (in position on the sky) would lead to a miniscule decrease in the luminosity that will never be noticed. (If there are multiple images, the sum of the brightnesses of all the images will exceed that of the undistorted source.)

The components of A can be decomposed into the effects of convergence and shear. For the convergence, from $\kappa = \frac{1}{2}\nabla_\theta^2 \psi$ we have

$$\kappa = \tfrac{1}{2}(\psi_{11} + \psi_{22}). \tag{8.150}$$

The **shear**, meanwhile, distorts the shape of the source; if an initially circular source is distorted into an ellipse of ellipticity γ and position angle ϕ, we define the two components of the shear to be

$$\begin{aligned} \gamma_1 &= \gamma \cos(2\phi) \\ \gamma_2 &= \gamma \sin(2\phi), \end{aligned} \tag{8.151}$$

so that the total shear is $\gamma = \sqrt{\gamma_1^2 + \gamma_2^2}$. In terms of the lensing potential the components are given by

$$\begin{aligned} \gamma_1 &= \tfrac{1}{2}(\psi_{11} - \psi_{22}) \\ \gamma_2 &= \psi_{12} = \psi_{21}. \end{aligned} \tag{8.152}$$

Inverting these relationships to find the components of A yields

$$A = \begin{pmatrix} 1 - \kappa - \gamma_1 & -\gamma_2 \\ -\gamma_2 & 1 - \kappa + \gamma_1 \end{pmatrix}. \tag{8.153}$$

We can therefore express the magnification in terms of the convergence and shear, as

$$\mu = \frac{1}{(1-\kappa)^2 - \gamma^2}. \tag{8.154}$$

These features of lensing are becoming increasingly important in observational cosmology. The obvious case of interest is so-called "strong lensing," when the source is within the Einstein radius of the lens, and multiple images are possible. By observing several images of a single source, we can infer properties of the lens mass distribution (for example, to search for dark matter); we can also use the time delay along different paths to measure the Hubble constant, and the statistical frequency of lensing to constrain other cosmological parameters. However, lensing need not be strong to have an important effect. "Weak lensing," when the source and lens are separated by more than an Einstein radius, will generally lead to small amounts of magnification and shear which are impossible to detect without a priori knowledge of the properties of the source. However, the shearing effect can be detected statistically, by looking at the shapes of thousands of galaxies that are assumed to be intrinsically random in their orientations. Shearing by weak lensing leads to correlated distortions in the shapes, which can reveal a great deal about the distribution of matter between the observer and the distant sources.

8.7 ■ OUR UNIVERSE

Throughout our discussion of the behavior of FRW cosmologies, we have alluded to the actual values of the cosmological parameters corresponding to the universe in which we live. Let us now be more systematic, and discuss both the universe we see today and a plausible extrapolation back to early times. Our discussion will necessarily be brief, both for reasons of space and because cosmology is an active area of research; look for recent review articles to get up-to-date descriptions of current views.

Many of our direct determinations of the expansion rate rely on the luminosity-distance formula (8.123) applied to some type of object whose intrinsic luminosity is assumed to be known, which we call **standard candles**. (Occasionally we measure the angular diameters of objects whose intrinsic size is assumed to be known: standard rulers.) The Hubble constant, for example, is measured with a variety of standard candles, and a consensus of different methods has converged on the value $H_0 = 70 \pm 10$ km/sec/Mpc, mentioned above. Deviations at high redshift from the linear Hubble law (8.109) can yield information about the density parameters Ω_{i0}, but only if we have very bright objects whose intrinsic luminosity is accurately known. These are provided by Type Ia supernovae, which are thought to be explosions of white dwarf stars that have accreted enough mass to surpass the Chandrasekhar limit. Since the Chandrasekhar limit is close to universal, the associated explosions are essentially of equal brightness (and some of the intrinsic variability can actually be accounted for by following the evolution of the brightness through time). It was measurements of SNe Ia at redshifts $z > 0.3$ that provided the first direct evidence for a nonzero cosmological constant; these observations imply that Ω_Λ is actually larger than Ω_M. Recall that matter is pressureless, $p_M = 0$, whereas vacuum energy is associated with a negative pressure, $p_\Lambda = -\rho_\Lambda$. Plugging into the second Friedmann equation (8.68) we find that a

universe with both matter and Λ obeys

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}(\rho_{\mathrm{M}} - 2\rho_{\Lambda}). \tag{8.155}$$

Thus, if ρ_{Λ} is sufficiently large compared to ρ_{M} (as the supernova observations indicate), we can have $\ddot{a} > 0$, an accelerating universe (in the sense described in Section 8.4).

The matter density itself is measured by a variety of methods, often involving measuring the density ρ_{M} by looking for the gravitational effects of clustered matter and then extrapolating to large scales. Because $\rho_{\mathrm{M}} = (3H^2/8\pi G)\Omega_{\mathrm{M}}$, limits obtained in this way are often quoted in terms of $\Omega_{\mathrm{M}}h^2$, where h is defined in (8.70). These days the uncertainty on H_0 appears to be small enough that it is fairly safe to take $h^2 \approx 0.5$, which we do henceforth. Most contemporary methods are consistent with the result

$$\Omega_{\mathrm{M0}} = 0.3 \pm 0.1. \tag{8.156}$$

Before there was good evidence for a cosmological constant, this low matter density was sometimes taken as an indication that space was negatively curved, $\kappa < 0$.

In addition to matter and cosmological constant, we also have radiation in the universe. Ordinary photons are the most obvious component of the radiation density, but any relativistic particle would contribute. For photons, most of the energy density resides in the cosmic microwave background, the leftover radiation from the Big Bang. Besides photons, the only obvious candidates for a radiation component are neutrinos. We expect that the number density of relic background neutrinos is comparable to that of photons; the photon density is likely to be somewhat larger, as photons can still be created after the number of neutrinos has become fixed. However, if the mass of the neutrinos is sufficiently large (greater than about 10^{-4} eV), they will have become nonrelativistic today, and contribute to matter rather than to radiation. Current ideas about neutrino masses suggest that this probably is the case, but it is not perfectly clear. Furthermore, it is conceivable that there are as-yet-undetected massless particles in addition to the ones we know about (although they can't be too abundant, or they would suppress the formation of large-scale structure.) Altogether, it seems likely that the total radiation density is of the same order of magnitude as the photon density; in this case we would have

$$\Omega_{\mathrm{R0}} \sim 10^{-4}. \tag{8.157}$$

As mentioned before, it is not surprising that the radiation density is lower than the matter density, as the former decays more rapidly as the universe expands. The radiation density goes as a^{-4}, while that in matter goes as a^{-3}; so the epoch of matter-radiation equality occurred at a redshift

$$z_{\mathrm{eq}} \approx \frac{\Omega_{\mathrm{M0}}}{\Omega_{\mathrm{R0}}} \sim 3 \times 10^3. \tag{8.158}$$

A further crucial constraint on the cosmological parameters comes from anisotropies in the temperature of the microwave background. The average temperature is $T_{\rm CMB} = 2.74$K, but in 1992 the COBE satellite discovered fluctuations from place to place at a level of $\Delta T/T \sim 10^{-5}$. These anisotropies arise from a number of sources, including gravitational redshift/blueshift from photons moving out of potential wells at recombination (the Sachs–Wolfe effect, dominant on large angular scales), intrinsic temperature fluctuations at the surface of last scattering (dominant on small angular scales), and the Doppler effect from motions of the plasma. The physics describing the evolution of CMB anisotropies is outside the scope of this book. A map of the CMB temperature over the entire sky clearly contains a great deal of information, but no theory predicts what the temperature at any given point is supposed to be. Instead, modern theories generally predict the expectation value of the amount of anisotropy on any given angular scale. We therefore decompose the anisotropy field into spherical harmonics,

$$\frac{\Delta T}{T}(\theta, \phi) = \sum_{lm} a_{lm} Y_{lm}(\theta, \phi). \tag{8.159}$$

The expectation value of $|a_{lm}|^2$ is likely to be independent of m; otherwise the statistical characteristics of the anisotropy will change from place to place on the sky (although we should keep an open mind). The relevant parameters to be measured are therefore

$$C_l = \langle |a_{lm}|^2 \rangle. \tag{8.160}$$

Since for any fixed l, there are $2l + 1$ possible values of m (from $-l$ to l), at all but the lowest l's there are enough independent measurements of the a_{lm}'s to accurately determine their expectation values. The irreducible uncertainty at very small l is known as cosmic variance.

Numerous experiments have measured the C_l's (the so-called CMB power spectrum), and improving these measurements is likely to be an important task for a number of years. (In addition to the temperature anisotropy, a great deal of information is contained in the polarization of the CMB, which is another target of considerable experimental effort.) To turn these observations into useful information, we need a specific theory to predict the CMB power spectrum as a function of the cosmological parameters. There are two leading possibilities (although one is much more leading than the other): either density perturbations are imprinted on all scales at extremely early times even modes for which the physical wavelength λ was much larger than the Hubble radius H^{-1}, or local dynamical mechanisms act as sources for anisotropies at all epochs. The latter possibility has essentially been ruled out by the CMB data; if anisotropies are produced continuously, we expect a relatively smooth, featureless spectrum of C_l's, whereas the observations indicate a significant amount of structure. It is therefore much more popular to imagine a primordial source of perturbations, such as inflation (discussed in the next section). Inflationary perturbations are adiabatic—perturbations in the mat-

ter density are correlated with those in the radiation density—and of nearly equal magnitude at all scales. With this input, we can make definite predictions for the C_l's as a function of all the cosmological parameters. Perhaps the most significant constraint obtained from experiments thus far is that universe is spatially flat, or nearly so; $|\Omega_{c0}| < 0.1$. Combined with the measurements of the matter density $\Omega_M \approx 0.3$, we conclude that the vacuum energy density parameter should be

$$\Omega_{\Lambda 0} = 0.7 \pm 0.1. \tag{8.161}$$

This is nicely consistent with the Type Ia supernova results described above; the concordance picture described here is that indicated in Figure 8.4. Converting from density parameter to physical energy density using $H_0 = 70$ km/sec/Mpc yields

$$\rho_{vac} \approx 10^{-8} \text{ erg/cm}^3, \tag{8.162}$$

as mentioned in our discussion of vacuum energy in Section 4.5.

One more remarkable feature completes our schematic picture of the present-day universe. We have mentioned that about 30% of the energy density in our universe consists of matter. But to a cosmologist, "matter" is any collection of nonrelativistic particles; the matter we infer from its gravitational influence need not be the same kind of ordinary matter we are familiar with from our experience on Earth. By **ordinary matter** we mean anything made from atoms and their constituents (protons, neutrons, and electrons); this would include all of the stars, planets, gas, and dust in the universe, immediately visible or otherwise. Occasionally such matter is referred to as *baryonic* matter, where baryons include protons, neutrons, and related particles (strongly interacting particles carrying a conserved quantum number known as baryon number). Of course electrons are conceptually an important part of ordinary matter, but by mass they are negligible compared to protons and neutrons:

$$\begin{aligned} m_p &= 0.938 \text{ GeV} \\ m_n &= 0.940 \text{ GeV} \\ m_e &= 0.511 \times 10^{-3} \text{ GeV}. \end{aligned} \tag{8.163}$$

In other words, the mass of ordinary matter comes overwhelmingly from baryons.

Ordinary baryonic matter, it turns out, is not nearly enough to account for the observed density $\Omega_M \approx 0.3$. Our current best estimates for the baryon density yield

$$\Omega_b = 0.04 \pm 0.02, \tag{8.164}$$

where these error bars are conservative by most standards. This determination comes from a variety of methods: direct counting of baryons (the least precise method), consistency with the CMB power spectrum (discussed above), and agreement with the predictions of the abundances of light elements for Big-Bang nucleosynthesis (discussed below). Most of the matter density must therefore be

in the form of **nonbaryonic dark matter**, which we will abbreviate to simply "dark matter." (Baryons can be dark, but it is increasingly common to reserve the terminology for the nonbaryonic component.) Essentially every known particle in the Standard Model of particle physics has been ruled out as a candidate for this dark matter. Fortunately, there are a number of plausible candidates beyond the Standard Model, including neutralinos (the lightest of the additional stable particles predicted by supersymmetry, with masses ≥ 100 GeV) and axions (light pseudoscalar particles arising from spontaneous breakdown of a hypothetical Peccei-Quinn symmetry invoked to explain conservation of CP in the strong interactions, with masses $\sim 10^{-4}$ eV). One of the few things we know about the dark matter is that it must be cold—not only is it nonrelativistic today, but it must have been that way for a very long time. If the dark matter were hot, it would have free-streamed out of overdense regions, suppressing the formation of galaxies. The other thing we know about cold dark matter (CDM) is that it should interact very weakly with ordinary matter, so as to have escaped detection thus far. Nevertheless, ambient dark matter particles may occasionally scatter off carefully shielded detectors in terrestrial laboratories; the attempt to directly detect dark matter by searching for the effects of such scatterings will be another significant experimental effort in the years to come.

The picture in which $\Omega_M = 0.3$ and $\Omega_\Lambda = 0.7$ seems to fit an impressive variety of observational data. The most surprising part of the picture is the cosmological constant. In Chapter 4 we mentioned that a naïve estimate of the vacuum energy yields a result many orders of magnitude larger than what has been measured. In fact there are three related puzzles: Why is the cosmological constant so much smaller than we expect? What is the origin of the small nonzero energy that comprises 70% of the current universe? And, why is the current value of the vacuum energy of the same order of magnitude as the matter density? The last problem is especially severe, as the vacuum energy and matter density evolve rapidly with respect to each other:

$$\frac{\Omega_\Lambda}{\Omega_M} \propto a^3. \tag{8.165}$$

If Ω_M and Ω_Λ are comparable today, in the past the vacuum energy would have been undetectably small, while in the future the matter density will be negligible. This "coincidence problem" has thus far proven to be a complete mystery. One suggested solution involves the "anthropic principle." If there are many distinct parts of the universe (in space, or even in branches of the wavefunction) in which the cosmological constant takes on very different values, intelligent life is most likely to arise in those places where the absolute magnitude is not too large—a large positive Λ would tear particles apart before galaxies could form, while a large negative Λ would cause the universe to recollapse before life could evolve. The anthropic explanation of the observed vacuum energy provides a good fit to the data, although the need to invoke such an elaborate scheme to explain this one quantity strikes some as slightly extravagant.

Another possibility that may (or may not) bear on the coincidence problem is the idea that we have not detected a nonzero cosmological constant, but rather a dynamical component that closely mimics the properties of vacuum energy. Consideration of this possibility has led cosmologists to coin the term **dark energy** to describe whatever it is that has been detected, whether it is dynamical or turns out to be a cosmological constant after all. What we know about the dark energy is that it is relatively smoothly distributed through space (or it would have been detected through its local gravitational field, just like dark matter) and is evolving slowly with time (or it would not make the universe accelerate, as indicated by the supernova data). A simple candidate for a dynamical source of dark energy is provided by a slowly-rolling scalar field. Consider a field ϕ with the usual action

$$S = \int d^4x\sqrt{-g}\left[-\frac{1}{2}g^{\mu\nu}\nabla_\mu\phi\nabla_\nu\phi - V(\phi)\right], \tag{8.166}$$

for which the energy-momentum tensor is

$$T_{\mu\nu} = \nabla_\mu\phi\nabla_\nu\phi + \left[\tfrac{1}{2}g^{\rho\sigma}\nabla_\rho\phi\nabla_\sigma\phi - V(\phi)\right]g_{\mu\nu} \tag{8.167}$$

and the equation of motion is

$$\Box\phi - \frac{dV}{d\phi} = 0. \tag{8.168}$$

Assume that the field is completely homogeneous through space ($\partial_i\phi = 0$). Then using the Christoffel symbols (8.44), we may express the d'Alembertian in terms of time derivatives and the Hubble constant to write (8.168) as

$$\ddot{\phi} + 3H\dot{\phi} + \frac{dV}{d\phi} = 0. \tag{8.169}$$

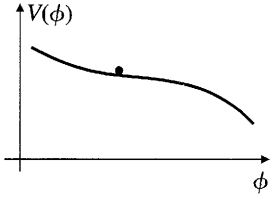

FIGURE 8.6 Potential energy for a slowly-rolling scalar field.

We see that the Hubble parameter acts as a friction term; the field will tend to roll down the potential, but when H is too large the motion will be damped. Therefore, a scalar field with a sufficiently shallow potential (as portrayed in Figure 8.6) will roll very slowly, leading to a kinetic energy much smaller than the potential energy $V(\phi)$. The energy-momentum tensor is then

$$T_{\mu\nu} \approx -V(\phi)g_{\mu\nu}, \tag{8.170}$$

where $\phi \approx$ constant. Comparing to (4.96), we see that the scalar field potential is mimicking a vacuum energy. As a simple example consider a quadratic potential, $V(\phi) = \frac{1}{2}m^2\phi^2$. Then (8.169) describes a damped harmonic oscillator, and overdamping will occur if $H > m$. But in particle-physics units, the Hubble constant today is $H_0 \approx 10^{-33}$ eV, so the mass of this scalar field would have to be incredibly tiny compared to the masses of the familiar elementary particles in equation (8.163). This seems to be an unnatural fine-tuning. Nevertheless, models

of dynamical dark energy are being actively explored, partially in the hope that they will lead somehow to a solution of the coincidence problem.

With this view of the contemporary situation, we can imagine what the early universe must have been like to have produced what we see today. For purposes of physical intuition it is often more helpful to keep track of the era under consideration by indicating the temperature rather than the redshift or time since the Big Bang. The temperature today is

$$T_0 = 2.74\ \mathrm{K} = 2.4 \times 10^{-4}\ \mathrm{eV}. \tag{8.171}$$

Of course, by "temperature" we mean the apparent blackbody temperature of the cosmic microwave background; in fact the CMB has not been in thermal equilibrium since recombination, so one should be careful in taking this concept too literally. Under adiabatic expansion, the temperature decreases as each relativistic particle redshifts, and we have $T \propto a^{-1}$. But there will be nonadiabatic phase transitions at specific moments in the early universe; in such circumstances the temperature doesn't actually increase, but decreases more gradually. To help relate the temperature, density, and scale factor, we introduce two different measures of the **effective number of relativistic degrees of freedom**: g_* and g_{*S} (where S stands for entropy). Consider a set of bosonic and fermionic species, each with their own effective temperature T_i, and number of spin states g_i. For example, a massless photon has two spin states, so $g_\gamma = 2$; a massive spin-$\frac{1}{2}$ fermion also has two spin states, so $g_{e^-} = g_{e^+} = 2$. The two different versions of the effective number of relativistic degrees of freedom obey

$$g_* = \sum_{\text{bosons}} g_i \left(\frac{T_i}{T}\right)^4 + \frac{7}{8} \sum_{\text{fermions}} g_i \left(\frac{T_i}{T}\right)^4 \tag{8.172}$$

and

$$g_{*S} = \sum_{\text{bosons}} g_i \left(\frac{T_i}{T}\right)^3 + \frac{7}{8} \sum_{\text{fermions}} g_i \left(\frac{T_i}{T}\right)^3. \tag{8.173}$$

The mysterious factors of $\frac{7}{8}$ arise from the difference between Bose and Fermi statistics when calculating the equilibrium distribution function. For any species in thermal equilibrium, the temperature T_i will be equal to the background temperature T; but we might have decoupled species at a lower temperature, which contribute less to the effective number of relativistic degrees of freedom. The reason why we need to define two different measures is that they play different roles; the first relates the temperature to the energy density (in relativistic species) via

$$\rho_{\mathrm{R}} = \frac{\pi^2}{30} g_* T^4, \tag{8.174}$$

while the second relates the temperature to the scale factor,

$$T \propto g_{*S}^{-1/3} a^{-1}. \tag{8.175}$$

In fact, g_* and g_{*S} are expected to be approximately equal so long as the relativistic degrees of freedom are those of the Standard Model of particle physics. A very rough guide is given by

$$g_* \approx g_{*S} \sim \begin{cases} 100 & T > 300 \text{ MeV} \\ 10 & 300 \text{ MeV} > T > 1 \text{ MeV} \\ 3 & T < 1 \text{ MeV}. \end{cases} \tag{8.176}$$

As we will discuss shortly, the events that change the effective number of relativistic degrees of freedom are the QCD phase transition at 300 MeV, and the annihilation of electron/positron pairs at 1 MeV.

With this background, let us consider the evolution of the universe from early times to today. To begin we imagine a Robertson–Walker metric with matter fields in thermal equilibrium at a temperature of 1 TeV = 1000 GeV. The high-temperature plasma is a complicated mixture of elementary particles (quarks, leptons, gauge and Higgs bosons). The dominant form of energy density will be relativistic particles, so the early universe is radiation-dominated. It is also very close to flat, since the curvature term in the Friedmann equation evolves more slowly than the matter and radiation densities. The Friedmann equation is therefore

$$\begin{aligned} H^2 &= \frac{8\pi G}{3}\rho_R \\ &\approx 0.1 g_* \frac{T^4}{\bar{m}_{\rm P}^2}, \end{aligned} \tag{8.177}$$

where the reduced Planck scale is $\bar{m}_{\rm P} = (8\pi G)^{-1/2} \approx 10^{18}$ GeV. If the radiation-dominated phase extends back to very early times, the age of the universe will be approximately $t \sim H^{-1}$, or

$$t \sim \frac{\bar{m}_{\rm P}}{T^2}. \tag{8.178}$$

In conventional units this becomes

$$t \sim 10^{-6} \left(\frac{\text{GeV}}{T}\right)^2 \text{ sec.} \tag{8.179}$$

Current experiments at particle accelerators have provided an accurate picture of what physics is like up to perhaps 100 GeV, so an additional order of magnitude is within the realm of reasonable extrapolation. At higher temperatures we are less sure what happens; there might be nothing very interesting between 1 TeV and the Planck scale, or this regime could be filled with all manner of surprises. Of course it is also conceivable that cosmology provides surprises at even lower temperatures, even though the Standard Model physics is well understood; in this section we are describing a conservative scenario, but as always it pays to keep an open mind.

A crucial feature of the Standard Model is the spontaneously broken symmetry of the electroweak sector. In cosmology, this symmetry breaking occurs at the electroweak phase transition, at $T \sim 200$ GeV. Above this temperature the symmetry is unbroken, so that elementary fermions (quarks and leptons) and the weak interaction gauge bosons are all massless, while below this temperature we have the pattern of masses familiar from low-energy experiments. The electroweak phase transition is not expected to leave any discernible impact on the late universe; one possible exception is baryogenesis, discussed below.

At these temperatures the strong interactions described by quantum chromodynamics (QCD) are not so strong. At low energies/temperatures, QCD exhibits "confinement"—quarks and gluons are bound into composite particles such as baryons and mesons. But above the QCD scale $\Lambda_{\rm QCD} \sim 300$ MeV, quarks and gluons are free particles. As the universe expands and cools, the confinement of strongly-interacting particles into bound states is responsible for the first drop in the effective number of relativistic degrees of freedom noted in (8.176). The QCD phase transition is not expected to leave a significant imprint on the observable universe.

Just as the strong interactions are not very strong at high temperatures, the weak interactions are not as weak as you might think; they are still weak in the sense of being accurately described by perturbation theory, but they occur rapidly enough to keep weakly-interacting particles, such as neutrinos, in thermal equilibrium. This ceases to be the case when $T \sim 1$ MeV. This is also approximately the temperature at which electrons and positrons become nonrelativistic and annihilate, decreasing the effective number of relativistic degrees of freedom, but the two events are unrelated. For temperatures below 1 MeV, we say the weak interactions are "frozen out"—the interaction rate drops below the expansion rate of the universe, so interactions happen too infrequently to keep particles in equilibrium. It may be the case that cold dark matter particles decouple from the plasma at this temperature. More confidently, we can infer that neutrons and protons cease to interconvert. The equilibrium abundance of neutrons at this temperature is about $\frac{1}{6}$ the abundance of protons (due to the slightly larger neutron mass). The neutrons have a finite lifetime ($\tau_n = 890$ sec) that is somewhat larger than the age of the universe at this epoch, $t(1\ \text{MeV}) \approx 1$ sec, but they begin to gradually decay into protons and leptons. Soon thereafter, however, we reach a temperature somewhat below 100 keV, and **Big-Bang Nucleosynthesis** (BBN) begins.

The nuclear binding energy per nucleon is typically of order 1 MeV, so you might expect that nucleosynthesis would occur earlier; however, the large number of photons per nucleon prevents nucleosynthesis from taking place until the temperature drops below 100 keV. At that point the neutron/proton ratio is approximately $\frac{1}{7}$. Of all the light nuclei, it is energetically favorable for the nucleons to reside in ^{4}He, and indeed that is what most of the free neutrons are converted into; for every two neutrons and fourteen protons, we end up with one helium nucleus and twelve protons. Thus, about 25% of the baryons by mass are converted to helium. In addition, there are trace amounts of deuterium (approximately 10^{-5} deuterons per proton), ^{3}He (also $\sim 10^{-5}$), and ^{7}Li ($\sim 10^{-10}$).

Of course these numbers are predictions, which are borne out by observations of the primordial abundances of light elements. (Heavier elements are not synthesized in the Big Bang, but require stellar processes in the later universe.) We have glossed over numerous crucial details, especially those that explain how the different abundances depend on the cosmological parameters. For example, imagine that we deviate from the Standard Model by introducing more than three light neutrino species. This would increase the radiation energy density at a fixed temperature through (8.174), which in turn decreases the timescales associated with a given temperature (since $t \sim H^{-1} \propto \rho_{\mathrm{R}}^{-1/2}$). Nucleosynthesis would therefore happen somewhat earlier, resulting in a higher abundance of neutrons, and hence in a larger abundance of ^{4}He. Observations of the primordial helium abundance, which are consistent with the Standard Model prediction, provided the first evidence that the number of light neutrinos is close to three. Similarly, all of the temperatures and timescales associated with nucleosynthesis depend on the baryon-to-photon ratio; agreement with the observed abundances requires that there be approximately 5×10^{-10} baryons per photon, which is the origin of the estimate (8.164) of the baryonic density parameter, and the associated need for nonbaryonic dark matter.

For our present purposes, perhaps the most profound feature of primordial nucleosynthesis is its sensitive dependence on the Friedmann relation between temperature and expansion rate, and hence on Einstein's equation. The success of BBN provides a stringent test of GR in a regime very far from our everyday experience. The fact that Einstein's theory, derived primarily from a need to reconcile gravitation with invariance under the Lorentz symmetries of electromagnetism, successfully describes the expansion of the universe when it was only one second old is a truly impressive accomplishment. To this day, BBN provides one of the most powerful constraints on alternative theories of gravity; in particular, it is the earliest epoch about which we have any direct observational signature.

Subsequent to nucleosynthesis, we have a plasma dominated by protons, electrons, and photons, with some helium and other nuclei. There is also dark matter, but it is assumed not to interact with the ordinary matter by this epoch. The next important event isn't until **recombination**, when electrons combine with protons (they combine with helium slightly earlier). Recombination happens at a temperature $T \approx 0.3$ eV; at this point the universe is matter-dominated. Again, since the binding energy of hydrogen is 13.6 eV, you might expect recombination to occur earlier, but the large photon/baryon ratio delays it. The crucial importance of recombination is that it marks the epoch at which the universe becomes transparent. The ambient photons interact strongly with free electrons, so that the photon mean free path is very short prior to recombination, but it becomes essentially infinite once the electrons and protons combine into neutral hydrogen. These ambient photons are visible today as the cosmic microwave background, which provides a snapshot of the universe at $T \approx 0.3$ eV, or a redshift $z \approx 1200$. Recombination is a somewhat gradual process, so any specification of when it happens is necessarily approximate.

Subsequent to recombination, the universe passes through a long period known as the "dark ages," as galaxies are gradually assembled through gravitational instability, but there are as yet no visible stars to light up the universe. The dark ages are a mysterious time; the processes by which stars and galaxies form are highly complicated and nonlinear, and new kinds of observations will undoubtedly be necessary before this era is well understood.

Our story has now brought us to the present day, but there are a couple of missing points we should go back and fill in. One is the asymmetry between matter and antimatter in the universe. Essentially all of the visible matter in the universe seems to be composed of protons, neutrons, and electrons, rather than their antiparticles; if distant galaxies were primarily antimatter, we would expect to observe high-energy photons from the occasional annihilation of protons with antiprotons at the boundaries of the matter/antimatter domains. While it is possible to build in an asymmetry as an initial condition, this seems somehow unsatisfying, and most physicists would prefer to find a dynamical mechanism of baryogenesis by which an initially matter/antimatter symmetric state could evolve into our present universe. Such broken symmetries are common in particle physics, and indeed numerous mechanisms for baryogenesis have been proposed (generally at temperatures at or above the electroweak scale). None of these specific schemes, however, has proven sufficiently compelling to be adopted as a standard scenario. It seems probable that we will need a better understanding of physics beyond the Standard Model to understand the origin of the baryon asymmetry.

The other missing feature we need to mention is that the universe is not, of course, perfectly homogeneous and isotropic; the current large-scale structure in the universe seems to have evolved from adiabatic and nearly scale-free perturbations present at very early times at the level of $\delta\rho/\rho \sim 10^{-5}$. Evidence for the adiabatic and scale-free nature of these perturbations comes from a combination of observations of the CMB and large-scale structure. Both the high degree of isotropy and homogeneity, and the small deviations therefrom, are simply imposed as mysterious initial conditions in the conventional cosmology. A possible dynamical origin for both is provided by the inflationary-universe scenario, to which we now turn.

8.8 ■ INFLATION

In the conventional understanding of the Big-Bang model, the universe is taken to be radiation-dominated at early times and matter-dominated at late times, with, as we now suspect, a very late transition to vacuum-domination. This picture has met with great success in describing a wide variety of observational data; nevertheless, we may still ask whether the initial conditions giving rise such a universe seem natural. This is the kind of question one might ask in cosmology but not in other sciences. Typically, as physicists we look for laws of nature, and imagine that we are free to specify initial conditions and ask how they evolve under such laws. But the universe seems to have only one set of initial conditions, so it

seems sensible to wonder if they are relatively generic or finely-tuned. Within the conventional picture, the early universe is indeed finely tuned to incredible precision. In particular, two features of our universe seem highly nongeneric: its spatial flatness, and its high degree of isotropy and homogeneity. It might be that this is just the universe we are stuck with, and it makes no sense to ask about the likelihood of different initial conditions. Alternatively, it might be that these conditions are more likely than they appear at first, if there is some dynamical mechanism that can take a wide spectrum of initial conditions and evolve them toward flatness and homogeneity/isotropy. The inflationary universe scenario provides such a mechanism (and more, besides), and has become a central organizing principle of modern cosmology, even if we are still far from demonstrating its truth.

Before describing inflation, let's describe the two problems of unnaturalness it claims to solve: the flatness problem and the horizon problem associated with homogeneity/isotropy. The **flatness problem** comes from considering the Friedmann equation in a universe with matter and radiation but no vacuum energy, which for later convenience we write in terms of the reduced Planck mass $\bar{m}_{\rm P} = (8\pi G)^{-1/2}$ as

$$H^2 = \frac{1}{3\bar{m}_{\rm P}^2}(\rho_{\rm M} + \rho_{\rm R}) - \frac{\kappa}{a^2}. \tag{8.180}$$

The curvature term $-\kappa/a^2$ is proportional to a^{-2} (obviously), while the energy density terms fall off faster with increasing scale factor, $\rho_{\rm M} \propto a^{-3}$ and $\rho_{\rm R} \propto a^{-4}$. This raises the question of why the ratio $(\kappa a^{-2})/(\rho/3\bar{m}_{\rm P}^2)$ isn't much larger than unity, given that a has increased by a factor of perhaps 10^{30} since the Planck epoch. Said another way, the point $\Omega = 1$ is a repulsive fixed point in a matter/radiation dominated universe—any deviation from this value will grow with time, so why do we observe $\Omega \sim 1$ today?

The **horizon problem** stems from the existence of particle horizons in FRW cosmologies, as illustrated in Figure 8.7. Horizons exist because there is only a finite amount of time since the Big Bang singularity, and thus only a finite distance that photons can travel within the age of the universe, as we briefly discussed in Chapter 2. Consider a photon moving along a radial trajectory in a flat universe (the generalization to nonflat universes is straightforward). A radial null path obeys

$$0 = ds^2 = -\mathrm{d}t^2 + a^2\,\mathrm{d}r^2, \tag{8.181}$$

so the comoving (coordinate) distance traveled by such a photon between times t_1 and t_2 is

$$\Delta r = \int_{t_1}^{t_2} \frac{dt}{a(t)}. \tag{8.182}$$

To get the physical distance as it would be measured by an observer at any time t, simply multiply by $a(t)$. For simplicity let's imagine we are in a matter-dominated

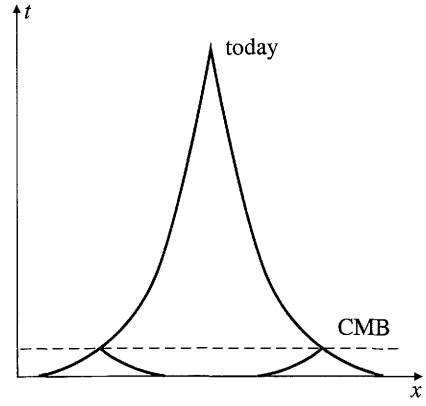

FIGURE 8.7 Past light cones in a universe expanding from a Big Bang singularity, illustrating particle horizons in cosmology. Points at recombination, observed today as parts of the cosmic microwave background on opposite sides of the sky, have nonoverlapping past light cones (in conventional cosmology); no causal signal could have influenced them to have the same temperature.

universe, for which

$$a = \left(\frac{t}{t_0}\right)^{2/3}. \tag{8.183}$$

Remember $a_0 = 1$. The Hubble parameter is therefore given by

$$\begin{aligned} H &= \tfrac{2}{3}t^{-1} \\ &= a^{-3/2}H_0. \end{aligned} \tag{8.184}$$

Then the photon travels a comoving distance

$$\Delta r = 2H_0^{-1}\left(\sqrt{a_2} - \sqrt{a_1}\right). \tag{8.185}$$

The comoving horizon size at any fixed value of the scale factor $a = a_*$ is the distance a photon travels since the Big Bang,

$$r_{\rm hor}(a_*) = 2H_0^{-1}\sqrt{a_*}. \tag{8.186}$$

The physical horizon size, as measured on the spatial hypersurface at a_*, is therefore simply

$$d_{\rm hor}(a_*) = a_* r_{\rm hor}(a_*) = 2H_*^{-1}. \tag{8.187}$$

Indeed, for any nearly-flat universe containing a mixture of matter and radiation, at any one epoch we will have

$$d_{\rm hor}(a_*) \sim H_*^{-1} = d_H(a_*), \tag{8.188}$$

where the Hubble distance d_H was introduced in (8.71). This approximate equality leads to a strong temptation to use the terms "horizon distance" and "Hubble distance" interchangeably; this temptation should be resisted, since inflation can render the former much larger than the latter, as we will soon demonstrate.

The horizon problem is simply the fact that the CMB is isotropic to a high degree of precision, even though widely separated points on the last scattering surface are completely outside each others' horizons. When we look at the CMB we are observing the universe at a scale factor $a_{\rm CMB} \approx 1/1200$; from (8.185), the comoving distance between a point on the CMB and an observer on Earth is

$$\begin{aligned} \Delta r &= 2H_0^{-1}\left(1 - \sqrt{a_{\rm CMB}}\right) \\ &\approx 2H_0^{-1}. \end{aligned} \tag{8.189}$$

However, the comoving horizon distance for such a point is

$$\begin{aligned} r_{\rm hor}(a_{\rm CMB}) &= 2H_0^{-1}\sqrt{a_{\rm CMB}} \\ &\approx 6 \times 10^{-2} H_0^{-1}. \end{aligned} \tag{8.190}$$

Hence, if we observe two widely-separated parts of the CMB, they will have nonoverlapping horizons; distinct patches of the CMB sky were causally disconnected at recombination. Nevertheless, they are observed to be at the same temperature to high precision. The question then is, how did they know ahead of time to coordinate their evolution in the right way, even though they were never in causal contact? We must somehow modify the causal structure of the conventional FRW cosmology.

Let's consider modifying the conventional picture by positing a period of **inflation**: an era of acceleration ($\ddot{a} > 0$) in the very early universe, driven by some component other than matter or radiation that redshifts away slowly as the universe expands. Then the flatness and horizon problems can be simultaneously solved. For simplicity consider the case where inflation is driven by a constant vacuum energy, leading to exponential expansion. Then, during the vacuum-dominated era, $\rho/3\bar{m}_{\rm P}^2 \propto a^0$ grows rapidly with respect to $-\kappa/a^2$, so the universe becomes flatter with time (Ω is driven to unity). If this process proceeds for a sufficiently long period, after which the vacuum energy is converted into matter and radiation, the density parameter will be sufficiently close to unity that it will not have had a chance to noticeably change into the present era. The horizon problem, meanwhile, can be traced to the fact that the physical distance between any two comoving objects grows as the scale factor, while the physical horizon size in a matter- or radiation-dominated universe grows more rapidly, as $d_{\rm hor} \sim a^{n/2} H_0^{-1}$. This can again be solved by an early period of exponential expansion, in which the true horizon size grows to a fantastic amount, so that our horizon today is actually much larger than the naïve estimate that it is equal to the Hubble radius H_0^{-1}.

In fact, a truly exponential expansion is not necessary; for any accelerated expansion, the spatial curvature will diminish with respect to the energy density,

and the horizon distance will grow rapidly. Typically we require that this accelerated period be sustained for 60 or more e-folds (where the number of e-folds is $N = \Delta \ln a$) which is what is needed to solve the horizon problem. It is easy to overshoot, and inflation generally makes the present-day universe spatially flat to incredible precision.

Now let's consider how we can get an inflationary phase in the early universe. The most straightforward way is to use the vacuum energy provided by the potential of a scalar field, the **inflaton**. Imagine a universe dominated by the energy of a spatially homogeneous scalar. The relevant equations of motion are precisely those of our discussion of dynamical dark energy in Section 8.7; the only difference is that the energy scale of inflation is much higher. We have the equation of motion for a scalar field in an RW metric,

$$\ddot{\phi} + 3H\dot{\phi} + V'(\phi) = 0, \tag{8.191}$$

as well as the Friedmann equation,

$$H^2 = \frac{1}{3\bar{m}_{\rm P}^2}\left(\frac{1}{2}\dot{\phi}^2 + V(\phi)\right). \tag{8.192}$$

We have ignored the curvature term, since inflation will flatten the universe anyway. Inflation can occur if the evolution of the field is sufficiently gradual that the potential energy dominates the kinetic energy, and the second derivative of ϕ is small enough to allow this state of affairs to be maintained for a sufficient period. Thus, we want

$$\begin{aligned} &\dot{\phi}^2 \ll V(\phi), \\ &|\ddot{\phi}| \ll |3H\dot{\phi}|,\ |V'|. \end{aligned} \tag{8.193}$$

Satisfying these conditions requires the smallness of two dimensionless quantities known as **slow-roll parameters**:

$$\begin{aligned} \epsilon &= \frac{1}{2}\bar{m}_{\rm P}^2\left(\frac{V'}{V}\right)^2, \\ \eta &= \bar{m}_{\rm P}^2\left(\frac{V''}{V}\right). \end{aligned} \tag{8.194}$$

Note that $\epsilon \geq 0$, while η can have either sign. Note also that these definitions are not universal; some people like to define them in terms of the Hubble parameter rather than the potential. Our choice describes whether a field has a chance to roll slowly for a while; the description in terms of the Hubble parameter describes whether the field actually is rolling slowly. When both of these quantities are small we can have a prolonged inflationary phase. They are not sufficient, however; no matter what the potential looks like, we can always choose initial conditions with $|\dot{\phi}|$ so large that slow-roll is never applicable. However, most initial conditions are attracted to an inflationary phase if the slow-roll parameters are small.

It isn't hard to invent potentials that satisfy the slow-roll conditions. Consider perhaps the simplest possible example,[3] $V(\phi) = \frac{1}{2}m^2\phi^2$. In this case

$$\epsilon = \eta = \frac{2\bar{m}_{\rm P}^2}{\phi^2}. \tag{8.195}$$

Clearly, for large enough ϕ, we can get the slow-roll parameters to be as small as we like. However, we have the constraint that the energy density should not be as high as the Planck scale, so that our classical analysis makes sense; this implies $\phi \ll \bar{m}_{\rm P}^2/m$. If we start the field at a value ϕ_i, the number of e-folds before inflation ends (that is, before the slow-roll parameters become of order unity) will be

$$\begin{aligned} N &= \int_{t_i}^{t_e} H\, dt \\ &\approx -\bar{m}_{\rm P}^{-2} \int_{\phi_i}^{\phi_e} \frac{V}{V'}\, d\phi \\ &\approx \frac{\phi_i^2}{4\bar{m}_{\rm P}^2} - \frac{1}{2}. \end{aligned} \tag{8.196}$$

The first equality is always true, the second uses the slow-roll approximation, and the third is the result for this particular model. To get 60 e-folds we therefore need $\phi_i > 16\bar{m}_{\rm P}$. Together with the upper limit on the energy density, we find that there is an upper limit on the mass parameter, $m \ll \bar{m}_{\rm P}/16$. In fact the size of the observed density fluctuations puts a more stringent upper limit on m, as we will discuss below. But there is no lower limit on m, so it is easy to obtain appropriate inflationary potentials only if we are willing to posit large hierarchies $m \ll \bar{m}_{\rm P}$, or equivalently a small dimensionless number $m/\bar{m}_{\rm P}$. Going through the same exercise with a $\lambda\phi^4$ potential would have yielded a similar conclusion, that λ would have had to be quite small; we often say that the inflaton must be weakly coupled. Of course, there is a sense in which we are cheating, since for field values $\phi > \bar{m}_{\rm P}$ we should expect additional terms in the effective potential, of the form $\bar{m}_{\rm P}^{4-n}\phi^n$ with $n > 4$, to become important. So in a *realistic* model it can be quite hard to get an appropriate potential.

At some point inflation ends, and the energy in the inflaton potential is converted into a thermalized gas of matter and radiation, a process known as "reheating." A proper understanding of the reheating process is of utmost importance, as it controls the production of various relics that we may or may not want in our universe. For example, one important beneficial aspect of inflation is that it can "inflate away" various relics that could be produced in the early universe, but are not observed today. A classic example occurs in the context of grand unified theories of particle physics, which generically predict the existence of super-

[3] We follow the exposition in A.R. Liddle, "An Introduction to Cosmological Inflation," `http://arxiv.org/astro-ph/9901124`.

heavy magnetic monopoles, with an abundance many orders of magnitude greater than allowed by observations. Historically, the monopole problem was the primary motivation for the invention of inflation by Guth; solutions to the flatness and horizon problems were considered a bonus. Inflation can dilute the monopole abundance appropriately, but they will be produced anew if the universe reheats to above the temperature of the grand-unification phase transition; fortunately, this is not a stringent constraint on most models. Similar considerations apply to other unwanted relics; in supersymmetric models, an especially worrisome problem is raised by the abundance of gravitinos (supersymmetric partners of the graviton). At the same time, it is necessary to reheat to a sufficiently high temperature to allow for some sort of baryogenesis scenario. For any specific implementation of inflation within a particle-physics model, it is crucial to check that unwanted relics are dispersed while wanted relics (such as baryons) are preserved.

A crucial element of inflationary scenarios is the production of density perturbations, which may be the origin of the CMB temperature anisotropies and the large-scale structure in galaxies that we observe today. The idea behind density perturbations generated by inflation is fairly straightforward. Inflation will attenuate any ambient particle density rapidly to zero, leaving behind only the vacuum. But the vacuum state in an accelerating universe has a nonzero temperature, the Gibbons–Hawking temperature, analogous to the Hawking temperature of a black hole. We won't be able to explore this subject in detail; here we simply outline the basic results.

For a universe dominated by a potential energy V the Gibbons–Hawking temperature is given by

$$T_{\rm GH} = \frac{H}{2\pi} \sim \frac{V^{1/2}}{\bar{m}_{\rm P}}. \tag{8.197}$$

Corresponding to this temperature are fluctuations in the inflaton field ϕ at each wavenumber k, with magnitude

$$|\Delta\phi|_k = T_{\rm GH}. \tag{8.198}$$

Since the potential is by hypothesis nearly flat, the fluctuations in ϕ lead to small fluctuations in the energy density,

$$\delta\rho = V'(\phi)\delta\phi. \tag{8.199}$$

Inflation therefore produces density perturbations on every scale. The amplitude of the perturbations is nearly equal at each wavenumber, but there will be slight deviations due to the gradual change in V as the inflaton rolls. Describing the perturbations is a messy subject, involving countless different notations. A sensible place to start is root-mean-square (RMS) density fluctuation,

$$\left.\frac{\delta\rho}{\rho}\right|_{\rm rms} = \sqrt{\left\langle \left(\frac{\delta\rho}{\rho}\right)^2 \right\rangle}, \tag{8.200}$$

where the angle brackets represent an average over spatial locations. For statistically isotropic perturbations (the expected amplitude is independent of direction), a bit of Fourier analysis allows us to write

$$\left(\left.\frac{\delta\rho}{\rho}\right|_{\rm rms}\right)^2 = \int \Delta^2(k)\, d(\ln k), \tag{8.201}$$

where we have introduced the dimensionless power spectrum,

$$\Delta^2(k) \equiv \frac{k^3|\delta_k|^2}{2\pi^2}, \tag{8.202}$$

and δ_k is the expectation value of the Fourier transform of the fractional density perturbation,

$$\delta_{\mathbf{k}} = \frac{1}{(2\pi)^{3/2}} \int e^{-i\mathbf{k}\cdot\mathbf{x}} \frac{\delta\rho}{\rho}\, d^3x, \tag{8.203}$$

which we've assumed to be isotropic. The dimensionless power spectrum is a function of time, as the amplitude for each mode evolves; it is most common to express the predictions of any specific model in terms of the amplitude of the perturbations at the moment when the physical wavelength of the mode, $\lambda = a/k$, is equal to the Hubble radius H^{-1},

$$A_{\rm S}^2(k) \equiv \left.\Delta^2(k)\right|_{k=aH}. \tag{8.204}$$

Thus, $A_{\rm S}(k)$ measures the amplitude for different modes at different times. For inflation driven by a slowly-rolling scalar field, $A_{\rm S}(k)$ is related to the potential via

$$A_{\rm S}^2(k) \sim \left.\frac{V^3}{\bar{m}_{\rm P}^6 (V')^2}\right|_{k=aH} \sim \left.\frac{V}{\bar{m}_{\rm P}^4 \epsilon}\right|_{k=aH}. \tag{8.205}$$

We have intentionally suppressed dimensionless numerical factors, which differ widely from reference to reference, in favor of highlighting the dependence on the potential.

The spectrum is given the subscript "S" because it describes scalar fluctuations in the metric. These are tied to the energy-momentum distribution, and the density fluctuations produced by inflation are adiabatic—fluctuations in the density of all species are correlated. The fluctuations are also Gaussian, in the sense that the phases of the Fourier modes describing fluctuations at different scales are uncorrelated. These aspects of inflationary perturbations—a nearly scale-free spectrum of adiabatic density fluctuations with a Gaussian distribution—are all consistent with current observations of the CMB and large-scale structure, and new data scheduled to be collected in years to come should greatly improve the precision of these tests.

It is not only the nearly-massless inflaton that is excited during inflation, but also any other nearly-massless particle. The other important example is the graviton, which corresponds to tensor perturbations in the metric (propagating excitations of the gravitational field). Tensor fluctuations have a spectrum

$$A_{\rm T}^2(k) \sim \left.\frac{V}{\bar{m}_{\rm P}^4}\right|_{k=aH}. \tag{8.206}$$

Importantly, the tensor amplitude depends only on the potential, not on its derivatives; observations of tensor perturbations would therefore give direct information about the energy scale of inflation.

For purposes of understanding observations, it is useful to parameterize the perturbation spectra in terms of observable quantities. We therefore write

$$A_{\rm S}^2(k) \propto k^{n_{\rm S}-1} \tag{8.207}$$

and

$$A_{\rm T}^2(k) \propto k^{n_{\rm T}}, \tag{8.208}$$

where $n_{\rm S}$ and $n_{\rm T}$ are the spectral indices. They are related to the slow-roll parameters of the potential by

$$n_{\rm S} = 1 - 6\epsilon + 2\eta \tag{8.209}$$

and

$$n_{\rm T} = -2\epsilon. \tag{8.210}$$

In models of the type we have considered (driven by single slowly-rolling scalar fields), there is a consistency relation relating the amplitudes and spectral indices of the scalar and tensor modes. It can be expressed in a convention-independent way as a relation between observable quantities, temperature fluctuations ΔT due to the different perturbations, as

$$\frac{(\Delta T/T)_{\rm T}^2}{(\Delta T/T)_{\rm S}^2} = -7n_{\rm T}. \tag{8.211}$$

The existence of tensor perturbations is a crucial prediction of inflation that may in principle be verifiable through observations of the polarization of the CMB. Polarization is also induced by ordinary density fluctuations, through the anisotropy of the Thompson scattering cross-section in an inhomogeneous plasma. Fortunately, we can imagine decomposing the polarization vector field on the sky into a curl-free part (E-modes) and a curl part (B-modes); the scalar perturbations lead to E-mode polarization, whereas tensor perturbations lead to B-modes (up to some inevitable processing in the post-recombination universe). CMB polarization has been detected; the challenge for the future will be to sepa-

rate out the scalar and tensor contributions, to test the prediction (8.211) of simple inflationary models. Of course this requires not only detecting the tensor-induced polarization, but measuring its spectral index with some precision.

Our current knowledge of the amplitude of the perturbations already gives us important information about the energy scale of inflation. The tensor perturbations depend on V alone, not its derivatives; if the CMB anisotropies seen by COBE are due to tensor fluctuations (possible, although unlikely), we can instantly derive $V_{\text{inflation}} \sim (10^{16}\ \text{GeV})^4$. Here, the value of V being constrained is that which was responsible for creating the observed fluctuations; namely, 60 e-folds before the end of inflation. This is remarkably reminiscent of the grand unification scale, which is very encouraging. Even in the more likely case that the perturbations observed in the CMB are scalar in nature, we can still write

$$V_{\text{inflation}}^{1/4} \sim \epsilon^{1/4} 10^{16}\ \text{GeV}, \tag{8.212}$$

where ϵ is the slow-roll parameter defined in (8.194). Although we expect ϵ to be small, the 1/4 in the exponent means that the dependence on ϵ is quite weak; unless this parameter is extraordinarily tiny, it is very likely that $V_{\text{inflation}}^{1/4} \sim 10^{15}$-$10^{16}$ GeV. The fact that we can have such information about such tremendous energy scales is a cause for great wonder.

8.9 ■ EXERCISES

1. Consider an $(N+n+1)$-dimensional spacetime with coordinates $\{t, x^I, y^i\}$, where I goes from 1 to N and i goes from 1 to n. Let the metric be

$$ds^2 = -\mathrm{d}t^2 + a^2(t)\delta_{IJ}\,\mathrm{d}x^I\mathrm{d}x^J + b^2(t)\gamma_{ij}(y)\,\mathrm{d}y^i\mathrm{d}y^j, \tag{8.213}$$

where δ_{IJ} is the usual Kronecker delta and $\gamma_{ij}(y)$ is the metric on an n-dimensional maximally symmetric spatial manifold. Imagine that we normalize the metric γ such that the curvature parameter

$$k = \frac{R(\gamma)}{n(n-1)} \tag{8.214}$$

is either $+1$, 0, or -1, where $R(\gamma)$ is the Ricci scalar corresponding to the metric γ_{ij}.

(a) Calculate the Ricci tensor for this metric.

(b) Define an energy-momentum tensor in terms of an energy density ρ and pressure in the x^I and y^i directions, $p^{(N)}$ and $p^{(n)}$:

$$T_{00} = \rho \tag{8.215}$$

$$T_{IJ} = a^2 p^{(N)} \delta_{IJ} \tag{8.216}$$

$$T_{ij} = b^2 p^{(n)} \gamma_{ij}. \tag{8.217}$$

Plug the metric and $T_{\mu\nu}$ into Einstein's equations to derive Friedmann-like equations for a and b (three independent equations in all).

(c) Derive equations for the energy density and the two pressures at a static solution where $\dot{a} = \dot{b} = \ddot{a} = \ddot{b} = 0$, in terms of k, n, and N. Use these to derive expressions for the equation-of-state parameters $w^{(N)} = p^{(N)}/\rho$ and $w^{(n)} = p^{(n)}/\rho$, valid at the static solution.

2. Consider de Sitter space in coordinates where the metric takes the form

$$ds^2 = -dt^2 + e^{Ht}[dx^2 + dy^2 + dz^2]. \tag{8.218}$$

Solve the geodesic equation for comoving observers ($x^i = \text{constant}$) to find the affine parameter as a function of t. Show that the geodesics reach $t = -\infty$ in a finite affine parameter, demonstrating that these coordinates fail to cover the entire manifold.

3. In Appendix F we discuss Raychaudhuri's equation. Show that, applied to a Robertson–Walker cosmology, the Raychaudhuri equation is equivalent to the second Friedmann equation, (8.68).

4. Consider the best-fit universe, with density parameters $\Omega_{R0} = 10^{-4}$, $\Omega_{M0} = 0.3$, $\Omega_{\Lambda 0} = 0.7$. Make a plot of the three Ω_i's as a function of the scale factor a, on a log scale, from $a = 10^{-35}$ to $a = 10^{35}$. Indicate the Planck time, nucleosynthesis, and today.

5. In a flat spacetime, objects of a fixed physical size subtend smaller and smaller angles as they are further and further away; in an expanding universe this is not necessarily so. Consider the angular size $\theta(z)$ of an object of physical size L at redshift z. In a matter-dominated flat universe, at what redshift is $\theta(z)/L$ a minimum? If all galaxies are at least 10 kpc across (and always have been), what is the minimum angular size of a galaxy in such a universe? Express your result both in terms of H_0, and plugging in $H_0 = 70$ km/s/Mpc.

6. In cosmology we tend to idealize nonrelativistic particles as having zero temperature T and pressure p. In reality, random motions will give them some temperature and pressure, satisfying $p \propto T\rho$.

   (a) How does the pressure of a gas of massive particles decay as a function of the scale factor?

   (b) Suppose neutrinos have a mass $m_\nu = 0.1$ eV, and a current temperature $T_{\nu 0} = 2$K. At about what redshift did the neutrinos go from being relativistic to nonrelativistic?

7. Suppose that the universe started out in a state of equipartition at the Planck time (so that the energy density in matter and radiation are of order the Planck density, and the temporal and spatial curvature radii are of order the Planck length). Neglecting any spatial inhomogeneity, calculate how long a positively curved universe will last, and how old a negatively curved universe would be when the temperature reaches 3K. How old would a flat universe be when the temperature reaches 3K? How old would a flat universe be by the time the expansion rate slows to $H_0 = 70$ km s^{-1} Mpc^{-1}?

CHAPTER

9 Quantum Field Theory in Curved Spacetime

9.1 ■ INTRODUCTION

Nobody believes that general relativity is the final word as far as gravity is concerned. The singularity theorems provide internal evidence that the theory is somehow incomplete; more convincing, however, is the fact that GR is a classical theory, while the world is fundamentally quantum-mechanical. The search for a working theory of quantum gravity drives a great deal of research in theoretical physics today, and much has been learned along the way, but convincing success remains elusive.

There are two parts to general relativity: the framework of spacetime curvature and its influence on matter, and the dynamics of the metric in response to energy-momentum (as described by Einstein's equation). Lacking a true theory of quantum gravity, we may still take the first part of GR—the idea that matter fields propagate on a curved spacetime background—and consider the case where those matter fields are quantum-mechanical. In other words, we take the metric to be fixed, rather than obeying some dynamical equations, and study quantum field theory (QFT) in that curved spacetime.

The epochal event in the study of QFT in curved spacetime was Hawking's realization in 1976 that black holes are not really black, but instead emit thermal radiation at a **Hawking temperature** proportional to the surface gravity κ,

$$\boxed{T = \frac{\kappa}{2\pi}.} \tag{9.1}$$

(Recall that our units set $\hbar = c = k = 1$; the Hawking temperature is actually proportional to $\hbar$ and inversely proportional to Boltzmann's constant k.) Since this remarkable discovery, QFT in curved spacetime has been put on a fairly rigorous theoretical footing, although its range of applicability is generally thought to be quite far away from any possible experimental probes. The Hawking temperature of a Schwarzschild black hole, for which $\kappa = 1/4GM$, can be written

$$T = \frac{1}{8\pi GM} = 1.2 \times 10^{26} K \left(\frac{1\text{ g}}{M}\right) = 6.0 \times 10^{-8} K \left(\frac{M_\odot}{M}\right), \tag{9.2}$$

where $M_\odot \sim 10^{33}$ g is the mass of the Sun. So the radiation from a realistic astrophysical black hole is at a much lower temperature even than the $3K$ cosmic microwave background, and thus would be hopelessly unobservable.

Recent observations in cosmology, however, have changed this situation somewhat. One example is the apparent discovery that the universe is accelerating, which is most readily interpreted as evidence for a nonzero vacuum energy (as discussed in Chapter 8). Although the magnitude of the vacuum energy remains a profound mystery, it seems clear that an understanding of how quantum-mechanical matter behaves in curved spacetime will play an important role in any eventual resolution to the puzzle. The other example comes from cosmological perturbations. Observations of the microwave background and large-scale structure provide strong evidence in favor of a nearly scale-free spectrum of primordial perturbations, including at wavelengths that would be much larger than the horizon size in a conventional cosmology. The leading theory for the origin of these perturbations comes from inflation. In the inflationary scenario, cosmological perturbations originate in the vacuum fluctuations of quantum fields in an inflating universe. If this picture is correct, what we are seeing in maps of the CMB is the imprint of primordial quantum fluctuations, greatly stretched by the expansion of the universe, and it is these fluctuations which eventually grew via gravitational instability into the galaxies and clusters we see today. At the very least, then, cosmological observations provide strong incentive for the study of QFT in curved spacetime.

Even without this empirical motivation, thought experiments based on QFT in curved spacetime have proven very fruitful in our tentative explorations of quantum gravity. In particular, the evaporation of black holes as predicted by Hawking radiation has led to the information-loss paradox, which we will discuss below. Since it is so difficult to do real experiments that bear directly on questions of quantum gravity, we must rely on thought experiments that focus on the tension between GR and quantum mechanics, much as Einstein used thought experiments in his attempts to reconcile classical dynamics with the Lorentz invariance of electromagnetism.

With these considerations in mind, the goal of the present chapter is to provide a brief introduction to some of the ideas and results of QFT in curved spacetime. Many introductory GR books do not cover this subject, usually because familiarity with ordinary QFT in flat spacetime should not be a prerequisite for studying GR. The happy fact is, however, that a familiarity with QFT in flat spacetime is by no means necessary for studying QFT in curved spacetime. This is because the features of QFT that are most interesting and useful in flat spacetime are almost completely distinct from those that are interesting and useful in curved spacetime. Deep down, a quantum field theory is simply an example of a quantum-mechanical system, just like a square well or a helium atom. Once a field theory is defined, applications in flat spacetime (to particle physics or condensed matter) will naturally focus on the issue of interactions between the various fields, often treated as perturbations around some natural vacuum state. In curved spacetime, however, we are generally interested in the effects of spacetime itself on the fields, for which the interactions are beside the point. We therefore can consider free (noninteracting) fields, but we will have to take great care in defining what an appropriate vacuum state should be. (Indeed, as we will see, almost all of

the states we deal with will be vacuum states!) Consequently, knowledge of QFT in flat spacetime is not only unnecessary for the present discussion, it probably won't even be of much help; the only prerequisite is a familiarity with the basics of ordinary quantum mechanics.

We will gradually work our way up to quantum field theory in curved spacetime, beginning with a review of the quantum mechanics of the system to which every physicist turns when the going gets rough: the simple harmonic oscillator. This is, of course, a paradigmatic example of the principles of the workings of quantum mechanics, but there is a bonus: When we next turn to field theory, we will find that the quantum mechanics of a free field in flat spacetime is precisely that of an infinite number of harmonic oscillators. (It is not that there is one oscillator at every point in space, but that each mode in the Fourier transform of the field acts like an harmonic oscillator.) The transition to field theory is then fairly straightforward. Once we grasp the basics of field theory, given our previous study of GR, it is not very difficult to generalize to curved spacetime, although a number of subtleties are encountered along the way. Our discussion will necessarily be somewhat superficial, focused on the goal of understanding the physical basis of Hawking radiation through an understanding of the Unruh effect in flat spacetime. In particular, we won't be discussing the important applications of QFT in curved spacetime to cosmology, nor will we be entering into detailed examination of renormalization and related issues. We will largely follow the discussion in Birrell and Davies (1982); look there or in Wald (1994) or in the review by Ford[1] for further discussion.

9.2 ■ QUANTUM MECHANICS

A quantum field theory is just a particular example of a quantum-mechanical system, so we can begin by reminding ourselves what that means. Of course, although the world is fundamentally quantum-mechanical, our intuition tends to align more readily with classical physics, so let's set the stage by thinking about classical mechanics. Any physical theory describing a certain system, classical or quantum, consists of the answers to three questions:

1. What are the possible states of the system? In classical mechanics, the space of states is typically given by a set of coordinates and momenta (what we might think of as "initial conditions" for the system). They can be specified exactly, and that is all there is to know about the state of the system.
2. What can we observe about the system? This question is often addressed only implicitly in classical mechanics, since the answer is trivial: any function of the coordinates and momenta qualifies as an observable.
3. How does the system evolve? This is usually expressed by a set of equations of motion. Given the state and the equations of motion, the subsequent

[1] L. H. Ford, "Quantum field theory in curved spacetime," (1997), `http://arxiv.org/gr-qc/9707062`.

evolution is uniquely defined; as a result, the space of initial conditions is equivalent to the space of classical solutions to the theory.

To make these ideas more concrete, and also because it will be directly relevant to our study of field theory, let's consider the simple harmonic oscillator. A simple harmonic oscillator may be thought of as a particle in one dimension subject to a quadratic potential. The state is specified by a single coordinate x, and a single momentum p. To get the equations of motion, we could start with the Lagrangian, which is written in terms of x and its time derivative $\dot{x}$ as

$$L = \tfrac{1}{2}\dot{x}^2 - \tfrac{1}{2}\omega^2 x^2, \tag{9.3}$$

where we have set the mass of the oscillator to unity for convenience. We can immediately derive the equation of motion

$$\ddot{x} + \omega^2 x = 0. \tag{9.4}$$

For the transition to quantum mechanics, however, it is more convenient to work in terms of the Hamiltonian, which is a function of x and p rather than x and $\dot{x}$. The Hamiltonian is related to the Lagrangian by a Legendre transformation,

$$H = p\dot{x} - L, \tag{9.5}$$

where the momentum satisfies

$$p = \frac{\partial L}{\partial \dot{x}} = \dot{x}. \tag{9.6}$$

We therefore have the Hamiltonian for the oscillator,

$$H = \tfrac{1}{2}p^2 + \tfrac{1}{2}\omega^2 x^2, \tag{9.7}$$

and Hamilton's equations

$$\frac{dx}{dt} = \partial_p H = p, \qquad \frac{dp}{dt} = -\partial_x H = -\omega^2 x, \tag{9.8}$$

serve as equations of motion. The solutions are, of course, straightforward; it is useful to express them as complex numbers

$$x(t) = x_0 e^{i\omega t + \alpha_0}, \tag{9.9}$$

where x_0 is the amplitude and α_0 is a phase. We can take the real part at the end of the day to get the physical answer.

Now we turn to quantum mechanics. Although quantum mechanics is profoundly different from classical mechanics, a given theory still consists of the answers to the same three questions listed above, with the answers taking somewhat different forms.

1. The state of the system is represented as an element of a **Hilbert space**. Mathematically, a Hilbert space is just a complex vector space equipped with a complex-valued inner product with the property that taking the inner product of two states in the opposite order is equivalent to complex conjugation. We denote elements of the Hilbert space as $|\psi\rangle$ and elements of the dual space as $\langle\psi|$, so that the inner product of $|\psi_1\rangle$ and $|\psi_2\rangle$ is $\langle\psi_2|\psi_1\rangle$, and obeys

$$\langle\psi_2|\psi_1\rangle^* = \langle\psi_1|\psi_2\rangle. \tag{9.10}$$

   (We are glossing over technical requirements concerning completeness of the space.) In quantum mechanics the Hilbert spaces of interest are very often infinite-dimensional. For example, if a classical system is represented by coordinate x and momentum p, the Hilbert space could be taken to consist of all square-integrable complex-valued functions of x, or equivalently all square-integrable complex-valued functions of p (but not both at once).
2. Observables are represented by **self-adjoint operators** on the Hilbert space. The definition of "self-adjoint" is actually very subtle, but in simple circumstances amounts to our usual understanding of an Hermitian operator,

$$A^\dagger = A, \tag{9.11}$$

   where $A^\dagger$ obeys

$$\langle\psi_2|A\psi_1\rangle = \langle A^\dagger\psi_2|\psi_1\rangle \tag{9.12}$$

   for all states $|\psi_1\rangle$, $|\psi_2\rangle$. Of course many operators will not be Hermitian, but observables should have this property. In general such operators do not commute, so we cannot simultaneously specify the precise values of everything we might want to measure about the system; there will be a complete set of commuting observables that represents all we can say about a system at once.
3. Evolution of the system may be represented in one of two ways: as unitary evolution of the state vector in Hilbert space (the **Schrödinger picture**), or by keeping the state fixed and allowing the observables to evolve according to equations of motion (the **Heisenberg picture**).

Strictly speaking, quantum mechanics is just different from classical mechanics; it is by no means necessary to start with a classical model and "quantize" it. Nevertheless, we usually do exactly that. Even for simple classical models, there is more than one way to construct a quantized version; these include canonical quantization and path-integral quantization, as well as more exotic procedures. What is worse, there is no simple map between classical and quantum theories; there are classical theories with no well-defined quantum counterpart, classical theories with multiple quantum versions, and quantum theories without any classical

analogue. For our present purposes, we may blithely ignore all of these subtleties, and proceed directly with canonical quantization.

Once again, the simple harmonic oscillator provides a useful example. Consider first the familiar Schrödinger picture, in which states are represented by complex-valued wave functions that evolve with time, such as $\psi(x,t)$. The wave function is really the set of components of the state vector $|\psi\rangle$, expressed in the "delta-function position basis" $|x\rangle$, so that $|\psi(t)\rangle = \int dx\, \psi(x,t)|x\rangle$. Canonical quantization consists of imposing the canonical commutation relation,

$$[\hat{x}, \hat{p}] = i, \tag{9.13}$$

on the coordinate operator $\hat{x}$ and its conjugate momentum $\hat{p}$. For states represented as wave functions depending on x and t, $\hat{x}$ is simply multiplication by x, so (9.13) can be implemented by setting

$$\hat{p} = -i\partial_x. \tag{9.14}$$

The Hamiltonian operator is

$$H = -\tfrac{1}{2}\partial_x^2 + \tfrac{1}{2}\omega^2 x^2, \tag{9.15}$$

and the equation of motion is the Schrödinger equation,

$$H\psi = i\partial_t \psi. \tag{9.16}$$

Since the Hamiltonian is time-independent, solutions to this equation separate into functions of space and functions of time, $\psi(x,t) = f(t)g(x)$. The solutions then come in a discrete set labeled by an integer $n \geq 0$, and we find (up to normalization)

$$\psi_n(x,t) = e^{-(1/2)\omega x^2} H_n(\sqrt{\omega}x) e^{-iE_n t}, \tag{9.17}$$

where H_n is a Hermite polynomial of degree n, and

$$E_n = \left(n + \tfrac{1}{2}\right)\omega. \tag{9.18}$$

These states are all eigenfunctions of H, and E_n is the energy eigenvalue. An arbitrary state of the oscillator will simply be a superposition of the energy eigenstates,

$$\psi(x,t) = \sum_n c_n \psi_n(x,t), \tag{9.19}$$

for some set of appropriately normalized coefficients c_n.

A number of important features of the quantum-mechanical oscillator are contained in this brief overview. There is a discrete spectrum of energy eigenstates; this is why it's called "quantum" mechanics (even though it is not hard to find

systems with continuous spectra). There is a ground state of lowest energy, plus a set of excited states uniquely labeled by their energy eigenvalue. The ground state has a nonvanishing energy,

$$E_0 = \tfrac{1}{2}\omega, \tag{9.20}$$

sometimes called the "zero-point" energy. It is interesting to note that the minimum energy of the classical system would have been zero, representing a particle with $x = 0$ and $p = 0$. The quantum zero-point energy can be traced to the Heisenberg uncertainty principle, which forbids us from localizing a state simultaneously in both position and momentum; there is consequently a minimum amount of "jiggle" in the oscillator, leading to a nonzero ground-state energy. On the other hand, we could certainly have chosen to examine an oscillator with a potential given by $V(x) = \frac{1}{2}\omega^2 x^2 - \frac{1}{2}\omega$; our analysis would have been identical, except that the factor of $\frac{1}{2}$ in (9.18) would have been missing, and the ground-state energy would have been zero. Quantum mechanics does not insist on a nonvanishing zero-point energy, it simply displaces the energy from the classical value.

An alternative way to solve the simple harmonic oscillator is to introduce creation and annihilation operators $\hat{a}^\dagger$ and $\hat{a}$ (often called raising and lowering operators), defined by

$$\hat{a} = \frac{1}{\sqrt{2\omega}}\left(\omega\hat{x} + i\hat{p}\right), \qquad \hat{a}^\dagger = \frac{1}{\sqrt{2\omega}}\left(\omega\hat{x} - i\hat{p}\right), \tag{9.21}$$

so that

$$\hat{x} = \frac{1}{\sqrt{2\omega}}(\hat{a} + \hat{a}^\dagger), \qquad \hat{p} = -i\sqrt{\frac{\omega}{2}}(\hat{a} - \hat{a}^\dagger). \tag{9.22}$$

Given our previous expressions for the commutation relations (9.13) and Hamiltonian (9.7), we can easily calculate the commutation relation for the creation and annihilation operators,

$$[\hat{a}, \hat{a}^\dagger] = 1, \tag{9.23}$$

and the new expression for the Hamiltonian,

$$H = \left(\hat{a}^\dagger\hat{a} + \tfrac{1}{2}\right)\omega. \tag{9.24}$$

The creation/annihilation operators commute with the Hamiltonian via

$$\begin{aligned}[H, \hat{a}] &= -\omega\hat{a} \\ [H, \hat{a}^\dagger] &= \omega\hat{a}^\dagger.\end{aligned} \tag{9.25}$$

Comparing this version of the Hamiltonian to the energy eigenvalues (9.18), we are inspired to define a number operator

$$\hat{n} = \hat{a}^\dagger\hat{a}. \tag{9.26}$$

Let's think about why the creation/annihilation operators and the number operator deserve their names. Consider an eigenstate $|n\rangle$ of the number operator,

$$\hat{n}|n\rangle = n|n\rangle, \tag{9.27}$$

where the $\hat{n}$ on the left stands for the number operator, while the first n on the right stands for the actual number n. (This formula is the most charming in all of quantum mechanics.) By playing with the commutation relations, it is easy to show that

$$\begin{aligned} \hat{n}\hat{a}^\dagger|n\rangle &= (n+1)\hat{a}^\dagger|n\rangle \\ \hat{n}\hat{a}|n\rangle &= (n-1)\hat{a}|n\rangle. \end{aligned} \tag{9.28}$$

Thus, when $\hat{a}^\dagger$ acts on $|n\rangle$, it gives another eigenstate of $\hat{n}$ with eigenvalue raised by 1, while $\hat{a}$ gives an eigenstate with eigenvalue lowered by 1. As before we can show that n takes integral values from 0 to ∞, so there must be a vacuum state $|0\rangle$ satisfying

$$\hat{a}|0\rangle = 0. \tag{9.29}$$

From this state we can construct all of the eigenstates by successive operation by creation operators,

$$|n\rangle = \frac{1}{\sqrt{n!}}\left(\hat{a}^\dagger\right)^n |0\rangle. \tag{9.30}$$

The number operator counts the number of excitations above the ground state. The set of eigenstates $|n\rangle$ acts as a basis; any state is an appropriate linear combination of these states. The creation and annihilation operators act on them according to

$$\begin{aligned} \hat{a}|n\rangle &= \sqrt{n}|n-1\rangle \\ \hat{a}^\dagger|n\rangle &= \sqrt{n+1}|n+1\rangle, \end{aligned} \tag{9.31}$$

and the energy of each state is of course given by (9.18). The basis states are taken to be time-independent, so a physical system obeying Schrödinger's equation will be described by a state

$$|\psi(t)\rangle = \sum_n c_n e^{-iE_n t}|n\rangle, \tag{9.32}$$

where again the c_n's are constant coefficients.

For purposes of smoothing the transition to field theory, it is useful to translate this Schrödinger-picture description into the Heisenberg picture, in which the states are fixed and the operators evolve with time. Given Schrödinger's equation (9.16), any state can be written formally as some fixed initial state acted on by a unitary time-evolution operator

$$|\psi(t)\rangle = U(t)|\psi(0)\rangle, \tag{9.33}$$

where

$$U(t) = e^{-i\int H\,dt}. \tag{9.34}$$

(By unitary we mean $U^\dagger U = 1$.) If the Hamiltonian is time-independent, of course, we simply have $U(t) = e^{-iHt}$. The Schrödinger-picture expression for the matrix element of a time-independent operator A between time-dependent states $|\psi_1(t)\rangle$ and $|\psi_2(t)\rangle$ can then be written as a Heisenberg-picture expression in terms of a time-dependent operator $A(t)$ and time-independent states as

$$\begin{aligned}\langle\psi_2(t)|A|\psi_1(t)\rangle &= \langle\psi_2(0)|U^\dagger(t)AU(t)|\psi_1(0)\rangle \\ &= \langle\psi_2|A(t)|\psi_1\rangle,\end{aligned} \tag{9.35}$$

where clearly the Heisenberg-picture operator is given by

$$A(t) = U^\dagger(t)AU(t). \tag{9.36}$$

Such an operator satisfies the **Heisenberg equation of motion**,

$$\frac{dA(t)}{dt} = i[H, A(t)], \tag{9.37}$$

which takes the place of Schrödinger's equation in this picture. For the harmonic oscillator, we would find

$$\frac{d\hat{a}}{dt} = -i\omega\hat{a}, \qquad \frac{d\hat{a}^\dagger}{dt} = i\omega\hat{a}^\dagger, \tag{9.38}$$

with solutions

$$\hat{a}(t) = e^{-i\omega t}\hat{a}(0), \qquad \hat{a}(t)^\dagger = e^{i\omega t}\hat{a}(0)^\dagger. \tag{9.39}$$

From this we immediately find

$$\hat{n}(t) = \hat{a}(t)^\dagger\hat{a}(t) = \hat{a}(0)^\dagger\hat{a}(0), \tag{9.40}$$

which reflects the fact that the number operator is conserved.

It is common to say that in the Heisenberg picture the states are time-independent; this is somewhat confusing, if nevertheless true. It might be better to say that the states extend throughout time, rather than only being defined at a fixed time. To make this more clear, consider a simple harmonic oscillator subject to an external influence, for example by simply adding a forcing term to the Hamiltonian,

$$H = \tfrac{1}{2}p^2 + \tfrac{1}{2}\omega^2x^2 + F(t), \tag{9.41}$$

where the function $F(t)$ vanishes outside an interval,

$$F(t) = \begin{cases} 0 & t < t_1 \\ F(t) & t_1 \le t \le t_2 \\ 0 & t_2 < t. \end{cases} \tag{9.42}$$

We can think of someone coming along and shaking our oscillator for a short while, and then leaving it alone after that. In the Schrödinger picture, we would say that an oscillator that started in its ground state would be excited by the external force, and the final state would not be the ground state. In the Heisenberg picture, however, we take the state to be a solution to the equation of motion for all times, and say that the number operator went from being zero to some other value.

For the oscillator subject to a transient external force, there are clearly a set of states that look like energy eigenstates at early times, although they don't look that way in the future; we might call such states the "in states" $|n_{\text{in}}\rangle$, with the property that

$$\hat{n}(t < t_1)|n_{\text{in}}\rangle = n|n_{\text{in}}\rangle. \tag{9.43}$$

There is also a separate set of states that look like energy eigenstates at late times, correspondingly called "out states" $|n_{\text{out}}\rangle$, and obeying

$$\hat{n}(t > t_2)|n_{\text{out}}\rangle = n|n_{\text{out}}\rangle. \tag{9.44}$$

Both sets of states exist at all times, but they look like energy eigenstates only in the appropriate asymptotic regime. Either set forms a basis for the entire Hilbert space, so in particular we could decompose one set in terms of the other. For example, by multiplying by a complete set of in states, we can write

$$|n_{\text{out}}\rangle = \sum_m \langle m_{\text{in}}|n_{\text{out}}\rangle |m_{\text{in}}\rangle. \tag{9.45}$$

The complex numbers $\langle m_{\text{in}}|n_{\text{out}}\rangle$ are matrix elements, which could, in principle, be calculated from the Hamiltonian (9.41); together they comprise the ***S*-matrix**. An observer equipped with a way to detect excitations of the oscillator would find that the number of excitations was changed by the applied force, and the S-matrix encodes the information necessary to characterize these changes between the asymptotic past and future. All of this discussion, needless to say, carries over essentially without modification to field theory. For particle physics, the role of the external force is played by the interactions between different particles, whereas for our purposes it will be played by the curvature of spacetime.

9.3 ■ QUANTUM FIELD THEORY IN FLAT SPACETIME

As we have already mentioned, quantum field theory is just a particular example of a quantum-mechanical system, in which we are quantizing a field (a function, or more generally some tensor field, defined on spacetime) rather than a single oscillator. We begin with the simplest possible example, of a free scalar field in flat spacetime; only a couple of generalizations are necessary to make the transition from a single oscillator to this field theory. Extending the theory to curved spacetime is straightforward as usual, involving writing the theory in a covariant form

and declaring it to be true. Once we lose the symmetries of Minkowski space, however, some of the ideas we think of as central in a quantum field theory will no longer seem so crucial; in particular, the notions of "vacuum" and "particles" will lose their privileged positions. (Expositions of quantum mechanics will occasionally make the point that waves and particles are complementary notions with different domains of validity, but don't be misled; in quantum field theory it is the fields that are truly fundamental, while the particles are approximate notions useful in certain restricted circumstances.) In this section we study QFT in flat spacetime, before generalizing to curved spacetime in the next section.

We start with the classical theory, in this case a real scalar field $\phi(x^\mu)$ in flat spacetime, just as we considered in Chapter 1, this time generalized to n dimensions. The action is the spacetime integral of the Lagrange density, $S = \int d^n x\, \mathcal{L}$; we will consider the Klein–Gordon Lagrangian

$$\mathcal{L} = -\tfrac{1}{2}\eta^{\mu\nu}\partial_\mu\phi\partial_\nu\phi - \tfrac{1}{2}m^2\phi^2. \tag{9.46}$$

It is not necessary to include the volume-element factor $\sqrt{|g|}$, since we are using inertial coordinates in Minkowski space, with metric

$$ds^2 = -\mathrm{d}t^2 + (\mathrm{d}\mathbf{x})^2. \tag{9.47}$$

The equation of motion is the Klein–Gordon equation,

$$\Box\phi - m^2\phi = 0. \tag{9.48}$$

Translation into a Hamiltonian description for the field theory is straightforward. The conjugate momentum for a field is simply the derivative of the Lagrange density with respect to the time derivative of that field,

$$\pi = \frac{\partial\mathcal{L}}{\partial(\partial_0\phi)}. \tag{9.49}$$

For the Klein–Gordon Lagrangian (9.46), this is

$$\pi = \dot{\phi}. \tag{9.50}$$

Of course, referring to the time derivative assumes that we have chosen a particular inertial frame; consequently, the Hamiltonian procedure necessarily violates manifest Lorentz invariance. If we are careful, however, observable quantities in the resulting theory will still be Lorentz-invariant. The Hamiltonian itself can be expressed as an integral over space of a Hamiltonian density,

$$H = \int d^{n-1}x\, \mathcal{H}, \tag{9.51}$$

which is related to the Lagrangian by a Legendre transformation,

$$\begin{aligned}\mathcal{H}(\phi,\pi) &= \pi\dot{\phi} - \mathcal{L}(\phi,\partial_\mu\phi)\\ &= \tfrac{1}{2}\pi^2 + \tfrac{1}{2}(\nabla\phi)^2 + \tfrac{1}{2}m^2\phi^2,\end{aligned} \tag{9.52}$$

where $(\nabla\phi)^2 = \delta^{ij}(\partial_i\phi)(\partial_j\phi)$. The correspondence between this field theory and the harmonic oscillator should be clear: the field value $\phi(x)$ plays the role of the coordinate x, with momentum field $\pi(x)$ instead of a single momentum p. Instead of the state being specified by two numbers (x and p) at some fixed time, we would have to give field values [$\phi(x^i)$ and $\pi(x^i)$] all over space at some fixed time as initial data, and there is an additional gradient term that was missing in the oscillator case; but otherwise the formalism is very similar.

We should emphasize that $\phi(x^\mu)$ is *not* a wave function; it is a dynamical variable, generalizing the single degree of freedom x in the case of the harmonic oscillator. In a Schrödinger-picture quantization of the field theory, we would define a complex wave functional $\Psi[\phi(x^\mu)]$, which would represent the probability amplitude for finding the field in each configuration. Instead, however, we will use the Heisenberg picture, so that our primary concern will be to promote ϕ to a quantum operator.

First, we should complete the classical analysis by actually solving this theory. It is not hard to write down solutions to the Klein–Gordon equation. One good example is a plane wave,

$$\phi(x^\mu) = \phi_0 e^{ik_\mu x^\mu} = \phi_0 e^{-i\omega t + i\mathbf{k}\cdot\mathbf{x}}, \tag{9.53}$$

where the wave vector has components

$$k^\mu = (\omega, \mathbf{k}), \tag{9.54}$$

and the frequency must satisfy the dispersion relation

$$\omega^2 = \mathbf{k}^2 + m^2. \tag{9.55}$$

There is a clear similarity between such a solution and that for the simple harmonic oscillator, given by (9.9). But there is also an important difference: For the oscillator, there is only one independent solution. Because the oscillator has a unique frequency, when we add two solutions with specified amplitude x_0 and phase α_0, they combine to give a third solution with the same frequency but different amplitude and phase. This is no longer true in field theory. Given (9.55), the frequency is determined by the spatial wave vector $\mathbf{k}$, at least up to sign. Therefore, instead of a single kind of solution, we have a set parameterized by $\mathbf{k}$ and the sign of ω.

However, we can still write down the most general solution by constructing a complete, orthonormal set of modes in terms of which any solution may be expressed. To make sense of "orthonormal," we need to define an inner product on the space of solutions to the Klein–Gordon equation. Although the modes themselves are functions of spacetime, the appropriate inner product can be expressed as an integral over a constant-time hypersurface Σ_t,

$$(\phi_1, \phi_2) = -i \int_{\Sigma_t} \left(\phi_1 \partial_t \phi_2^* - \phi_2^* \partial_t \phi_1\right) d^{n-1}x. \tag{9.56}$$

As we would hope, the inner product is actually *independent* of the hypersurface Σ_t over which the integral is taken, as you can easily check by using Stokes's theorem and the Klein–Gordon equation. Applying this inner product to two plane waves of different wave vectors gives

$$\begin{aligned}
&(e^{ik_1^\mu x_\mu}, e^{ik_2^\mu x_\mu}) \\
&\quad = -i \int_{\Sigma_t} \left(e^{-i\omega_1 t + i\mathbf{k}_1\cdot\mathbf{x}} \partial_t e^{i\omega_2 t - i\mathbf{k}_2\cdot\mathbf{x}} - e^{i\omega_2 t - i\mathbf{k}_2\cdot\mathbf{x}} \partial_t e^{-i\omega_1 t + i\mathbf{k}_1\cdot\mathbf{x}} \right) d^{n-1}x \\
&\quad = (\omega_2 + \omega_1) e^{-i(\omega_1 - \omega_2)t} \int_{\Sigma_t} e^{i(\mathbf{k}_1 - \mathbf{k}_2)\cdot\mathbf{x}} d^{n-1}x \\
&\quad = (\omega_2 + \omega_1) e^{-i(\omega_1 - \omega_2)t} (2\pi)^{n-1} \delta^{(n-1)}(\mathbf{k}_1 - \mathbf{k}_2),
\end{aligned} \tag{9.57}$$

where we have used

$$\int e^{i\mathbf{k}\cdot\mathbf{x}} d^{n-1}x = (2\pi)^{n-1} \delta^{(n-1)}(\mathbf{k}). \tag{9.58}$$

The inner product thus vanishes unless the spatial wave vectors $\mathbf{k}$, and hence the frequencies ω, are equal for both modes. An orthonormal set of mode solutions is thus given by

$$f_{\mathbf{k}}(x^\mu) = \frac{e^{ik_\mu x^\mu}}{[(2\pi)^{n-1} 2\omega]^{1/2}}, \tag{9.59}$$

with k^μ obeying (9.55), so that

$$(f_{\mathbf{k}_1}, f_{\mathbf{k}_2}) = \delta^{(n-1)}(\mathbf{k}_1 - \mathbf{k}_2). \tag{9.60}$$

Given the dispersion relation (9.55), $\mathbf{k}$ only determines the frequency up to an overall sign. Our strategy will be to insist that ω always be a positive number, and complete the set of modes by including the complex conjugates $f_{\mathbf{k}}^*(x^\mu)$. (Complex conjugation changes the sign of the $\mathbf{k}$ term in the exponent as well as the ω term, but the components of $\mathbf{k}$ are defined from $-\infty$ to ∞ already.) The $f_{\mathbf{k}}$ modes are said to be positive-frequency, meaning they satisfy

$$\partial_t f_{\mathbf{k}} = -i\omega f_{\mathbf{k}}, \qquad \omega > 0, \tag{9.61}$$

while the $f_{\mathbf{k}}^*$ modes are negative-frequency, satisfying

$$\partial_t f_{\mathbf{k}}^* = i\omega f_{\mathbf{k}}^*, \qquad \omega > 0. \tag{9.62}$$

(Be careful; these modes are called negative-frequency even though $\omega > 0$, because the time derivative pulls down a factor $+i\omega$ rather than $-i\omega$.) The complex conjugate modes are orthogonal to the original modes,

$$(f_{\mathbf{k}_1}, f_{\mathbf{k}_2}^*) = 0, \tag{9.63}$$

and orthonormal with each other but with a negative norm,

$$(f^*_{\mathbf{k}_1}, f^*_{\mathbf{k}_2}) = -\delta^{(n-1)}(\mathbf{k}_1 - \mathbf{k}_2). \tag{9.64}$$

Together, the modes $f_{\mathbf{k}}$ and $f^*_{\mathbf{k}}$ form a complete set, in terms of which we can expand any solution to the Klein–Gordon equation.

To canonically quantize this theory, we promote our classical variables (the fields and their conjugate momenta) to operators acting on a Hilbert space, and impose the canonical commutation relations on equal-time hypersurfaces:

$$\begin{aligned}
&[\phi(t, \mathbf{x}), \phi(t, \mathbf{x}')] = 0 \\
&[\pi(t, \mathbf{x}), \pi(t, \mathbf{x}')] = 0 \\
&[\phi(t, \mathbf{x}), \pi(t, \mathbf{x}')] = i\delta^{(n-1)}(\mathbf{x} - \mathbf{x}').
\end{aligned} \tag{9.65}$$

In field theory we need to state explicitly that the field and its momentum commute with themselves throughout space; for a single oscillator this is implicit, since there is only a single coordinate and momentum, each of which will necessarily commute with itself. The delta function implies that operators at equal times commute everywhere except at coincident spatial points; this feature arises from the demands of causality (operators at spacelike separation cannot influence each other).

Just as classical solutions to the Klein–Gordon equation can be expanded in terms of the modes (9.59), so can the quantum operator field $\phi(t, \mathbf{x})$. Denoting the coefficients of the mode expansion of the field operator by $\hat{a}^\dagger_{\mathbf{k}}$ and $\hat{a}_{\mathbf{k}}$, we have

$$\phi(t, \mathbf{x}) = \int d^{n-1}k \left[\hat{a}_{\mathbf{k}} f_{\mathbf{k}}(t, \mathbf{x}) + \hat{a}^\dagger_{\mathbf{k}} f^*_{\mathbf{k}}(t, \mathbf{x})\right]. \tag{9.66}$$

Plugging this expansion into (9.65), we find that the operators $\hat{a}^\dagger_{\mathbf{k}}$ and $\hat{a}_{\mathbf{k}}$ obey commutation relations

$$\begin{aligned}
&[\hat{a}_{\mathbf{k}}, \hat{a}_{\mathbf{k}'}] = 0 \\
&[\hat{a}^\dagger_{\mathbf{k}}, \hat{a}^\dagger_{\mathbf{k}'}] = 0 \\
&[\hat{a}_{\mathbf{k}}, \hat{a}^\dagger_{\mathbf{k}'}] = \delta^{(n-1)}(\mathbf{k} - \mathbf{k}').
\end{aligned} \tag{9.67}$$

These operators thus obey the commutation relations characteristic of creation and annihilation operators, familiar from (9.23) for the simple harmonic oscillator. The difference, of course, is that there are an infinite number of such operators, indexed by $\mathbf{k}$. We can see the relevance of dividing the modes into positive- and negative-frequency; the positive-frequency modes are coefficients of annihilation operators, while negative-frequency modes are coefficients of creation operators. The idea of positive- and negative-frequency modes will turn out to generalize to static spacetimes, although not to arbitrary spacetimes.

In the case of the harmonic oscillator, we used the creation and annihilation operators to define a basis for the Hilbert space in which the basis states were

eigenstates of the number operator. The same procedure works for the free scalar field, although now we have to keep track of separate numbers of excitations for each spatial wave vector $\mathbf{k}$. There will be a single vacuum state $|0\rangle$, characterized by the fact that it is annihilated by each $\hat{a}_{\mathbf{k}}$,

$$\hat{a}_{\mathbf{k}}|0\rangle = 0 \qquad \text{for all } \mathbf{k}. \tag{9.68}$$

A state with $n_{\mathbf{k}}$ particles with identical momenta $\mathbf{k}$ is created by repeated action by $\hat{a}^{\dagger}_{\mathbf{k}}$,

$$|n_{\mathbf{k}}\rangle = \frac{1}{\sqrt{n_{\mathbf{k}}!}}\left(\hat{a}^{\dagger}_{\mathbf{k}}\right)^{n_{\mathbf{k}}}|0\rangle, \tag{9.69}$$

while a state with n_i excitations of various momenta $\mathbf{k}_i$ would be

$$|n_1, n_2, \ldots, n_j\rangle = \frac{1}{\sqrt{n_1! n_2! \cdots, n_j!}}\left(\hat{a}^{\dagger}_{\mathbf{k}_1}\right)^{n_1}\left(\hat{a}^{\dagger}_{\mathbf{k}_2}\right)^{n_2}\cdots\left(\hat{a}^{\dagger}_{\mathbf{k}_j}\right)^{n_j}|0\rangle. \tag{9.70}$$

Acting on such a state, the creation and annihilation operators change the number of excitations, as expected:

$$\begin{aligned}\hat{a}_{\mathbf{k}_i}|n_1, n_2, \ldots, n_i, \ldots, n_j\rangle &= \sqrt{n_i}|n_1, n_2, \ldots, n_i - 1, \ldots, n_j\rangle \\ \hat{a}^{\dagger}_{\mathbf{k}_i}|n_1, n_2, \ldots, n_i, \ldots, n_j\rangle &= \sqrt{n_i + 1}|n_1, n_2, \ldots, n_i + 1, \ldots, n_j\rangle.\end{aligned} \tag{9.71}$$

We can define a number operator for each wave vector,

$$\hat{n}_{\mathbf{k}} = \hat{a}^{\dagger}_{\mathbf{k}}\hat{a}_{\mathbf{k}}, \tag{9.72}$$

which obeys

$$\hat{n}_{\mathbf{k}_i}|n_1, n_2, \ldots, n_i, \ldots, n_j\rangle = n_i|n_1, n_2, \ldots, n_i, \ldots, n_j\rangle. \tag{9.73}$$

The states that are eigenstates of the number operators form a basis for the entire Hilbert space, known as the **Fock basis**; the space constructed from this basis is often called "Fock space," but of course it is just the original Hilbert space.

One thing we might want to investigate is how our Fock basis behaves under Lorentz transformations. We have clearly been taking advantage of the symmetries of Minkowski space, for example in using plane waves as a basis for solutions to the Klein–Gordon equation. The crucial aspect of these modes is our ability to distinguish between positive and negative frequencies, allowing for an interpretation of their coefficients in the mode expansion of ϕ as annihilation and creation operators. Now consider a boost by velocity $\mathbf{v} = d\mathbf{x}/dt$, leading to new coordinates $x^{\mu'}$ given by

$$t' = \gamma t - \gamma \mathbf{v}\cdot\mathbf{x}, \qquad \mathbf{x}' = \gamma\mathbf{x} - \gamma\mathbf{v}t, \tag{9.74}$$

where $\gamma = 1/\sqrt{1 - v^2}$, and the inverse transformation is given by

$$t = \gamma t' + \gamma\mathbf{v}\cdot\mathbf{x}', \qquad \mathbf{x} = \gamma\mathbf{x}' + \gamma\mathbf{v}t'. \tag{9.75}$$

The time derivative of our mode functions in the boosted frame is

$$\begin{aligned}\partial_{t'} f_{\mathbf{k}} &= \frac{\partial x^\mu}{\partial t'}\partial_\mu f_{\mathbf{k}} \\ &= \gamma(-i\omega) f_{\mathbf{k}} + \gamma \mathbf{v}\cdot(i\mathbf{k}) f_{\mathbf{k}} \\ &= -i\omega' f_{\mathbf{k}}\end{aligned} \tag{9.76}$$

where

$$\omega' = \gamma\omega - \gamma \mathbf{v}\cdot\mathbf{k} \tag{9.77}$$

is simply the frequency in the boosted frame. Clearly, then, a state describing a collection of particles with certain momenta is boosted into a state describing the same particles, but with boosted momenta. Thus, the total number operator in the two frames will coincide, and in particular the vacuum state will coincide. In this sense, our original choice of inertial frame was irrelevant. In the next section we will see that our ability to find positive- and negative-frequency solutions can be traced to the existence of a timelike Killing vector ∂_t in Minkowski spacetime, while the invariance of the Fock space under changes of basis can be traced to the fact that all such timelike Killing vectors are related by Lorentz transformations. Therefore, even if the frequency of a mode depends on the choice of inertial frame, the decomposition into positive and negative frequencies is invariant.

We would like to express the Hamiltonian

$$H = \int d^{n-1}x \left[\frac{1}{2}\dot{\phi}^2 + \frac{1}{2}(\nabla\phi)^2 + \frac{1}{2}m^2\phi^2\right] \tag{9.78}$$

in terms of the creation and annihilation operators, just as we did for the harmonic oscillator. We can analyze this expression term-by-term, starting with the ϕ^2 term for simplicity:

$$\begin{aligned}&\frac{1}{2}m^2\int d^{n-1}x\,\phi^2 \\ &\quad= \frac{1}{2}m^2\int d^{n-1}x\,d^{n-1}k\,d^{n-1}k'\left(\hat{a}_{\mathbf{k}} f_{\mathbf{k}} + \hat{a}^\dagger_{\mathbf{k}} f^*_{\mathbf{k}}\right)\left(\hat{a}_{\mathbf{k}'} f_{\mathbf{k}'} + \hat{a}^\dagger_{\mathbf{k}'} f^*_{\mathbf{k}'}\right) \\ &\quad= \frac{1}{2}m^2\int d^{n-1}x\,d^{n-1}k\,d^{n-1}k'\left(\hat{a}_{\mathbf{k}}\hat{a}_{\mathbf{k}'} f_{\mathbf{k}} f_{\mathbf{k}'} + \hat{a}^\dagger_{\mathbf{k}}\hat{a}_{\mathbf{k}'} f^*_{\mathbf{k}} f_{\mathbf{k}'}\right. \\ &\qquad\left. + \hat{a}_{\mathbf{k}}\hat{a}^\dagger_{\mathbf{k}'} f_{\mathbf{k}} f^*_{\mathbf{k}'} + \hat{a}^\dagger_{\mathbf{k}}\hat{a}^\dagger_{\mathbf{k}'} f^*_{\mathbf{k}} f^*_{\mathbf{k}'}\right).\end{aligned} \tag{9.79}$$

Zooming in on the first term in parentheses, and ignoring for the moment the integral over $\mathbf{k}$, we can plug in the explicit form of the mode functions (9.59) to obtain

$$\int d^{n-1}x\,d^{n-1}k'\,\hat{a}_{\mathbf{k}}\hat{a}_{\mathbf{k}'} f_{\mathbf{k}} f_{\mathbf{k}'} = \int d^{n-1}x\,d^{n-1}k'\,\hat{a}_{\mathbf{k}}\hat{a}_{\mathbf{k}'}\frac{e^{-i(\omega+\omega')t}e^{i(\mathbf{k}+\mathbf{k}')\cdot\mathbf{x}}}{2(2\pi)^{n-1}\sqrt{\omega\omega'}}$$

$$= \int d^{n-1}k' \, \hat{a}_{\mathbf{k}} \hat{a}_{\mathbf{k}'} \frac{e^{-i(\omega+\omega')t}}{2\sqrt{\omega\omega'}} \delta^{(n-1)}(\mathbf{k}+\mathbf{k}')$$

$$= \hat{a}_{\mathbf{k}} \hat{a}_{-\mathbf{k}} \frac{e^{-2i\omega t}}{2\omega}, \tag{9.80}$$

where we have used (9.58) again. Evaluating the other terms in (9.79) similarly, we find that the potential-energy contribution to the Hamiltonian therefore becomes

$$\frac{1}{2}m^2 \int d^{n-1}x \, \phi^2 = \frac{1}{2}m^2 \int d^{n-1}k \left(\frac{1}{2\omega}\right) \left[\hat{a}_{\mathbf{k}} \hat{a}_{-\mathbf{k}} e^{-2i\omega t} + \hat{a}^\dagger_{\mathbf{k}} \hat{a}_{\mathbf{k}} + \hat{a}_{\mathbf{k}} \hat{a}^\dagger_{\mathbf{k}} + \hat{a}^\dagger_{\mathbf{k}} \hat{a}^\dagger_{-\mathbf{k}} e^{2i\omega t}\right]. \tag{9.81}$$

For the kinetic-energy and gradient-energy pieces, the derivatives pull down factors of ω and $\mathbf{k}$ respectively; we obtain

$$\frac{1}{2}\int d^{n-1}x \, \dot{\phi}^2 = \frac{1}{2}\int d^{n-1}k \left(\frac{\omega}{2}\right) \left[-\hat{a}_{\mathbf{k}} \hat{a}_{-\mathbf{k}} e^{-2i\omega t} + \hat{a}^\dagger_{\mathbf{k}} \hat{a}_{\mathbf{k}} + \hat{a}_{\mathbf{k}} \hat{a}^\dagger_{\mathbf{k}} - \hat{a}^\dagger_{\mathbf{k}} \hat{a}^\dagger_{-\mathbf{k}} e^{2i\omega t}\right] \tag{9.82}$$

and

$$\frac{1}{2}\int d^{n-1}x \, (\nabla\phi)^2 = \frac{1}{2}\int d^{n-1}k \left(\frac{\mathbf{k}^2}{2\omega}\right) \left[\hat{a}_{\mathbf{k}} \hat{a}_{-\mathbf{k}} e^{-2i\omega t} + \hat{a}^\dagger_{\mathbf{k}} \hat{a}_{\mathbf{k}} + \hat{a}_{\mathbf{k}} \hat{a}^\dagger_{\mathbf{k}} + \hat{a}^\dagger_{\mathbf{k}} \hat{a}^\dagger_{-\mathbf{k}} e^{2i\omega t}\right]. \tag{9.83}$$

Using $\omega^2 = \mathbf{k}^2 + m^2$, we can put it all together to write the Hamiltonian for the scalar field theory as

$$H = \frac{1}{2}\int d^{n-1}k \left[\hat{a}^\dagger_{\mathbf{k}} \hat{a}_{\mathbf{k}} + \hat{a}_{\mathbf{k}} \hat{a}^\dagger_{\mathbf{k}}\right] \omega$$

$$= \int d^{n-1}k \left[\hat{n}_{\mathbf{k}} + \frac{1}{2}\delta^{(n-1)}(0)\right] \omega, \tag{9.84}$$

where the last step invokes the commutation relation (9.67) and the number operator $\hat{n}_{\mathbf{k}} = \hat{a}^\dagger_{\mathbf{k}} \hat{a}_{\mathbf{k}}$. By similar logic, we can construct an operator corresponding to the spatial components of the total momentum, which works out to be

$$P^i = \int d^{n-1}k \, \hat{n}_{\mathbf{k}} k^i. \tag{9.85}$$

As we might expect, energy eigenstates will be those with fixed numbers of excitations, each of which carries an energy ω. The excitations in the Fock basis

are interpreted as particles. This is how particles arise in a quantum field theory: energy eigenstates are collections of particles with definite momenta. Of course, our modes are plane waves that extend throughout space, not the localized tracks in bubble chambers that come to mind when we think of particles. What is worse, in a curved spacetime the wave equation will not have plane-wave solutions of definite frequency that we can interpret as particles. The solution to both issues is to think operationally, in terms of what would be observed by an experimental apparatus. The best strategy is to define a sensible notion of a particle detector that reduces to our intuitive picture in flat spacetime, and then define "particles" as "what a particle detector detects." For a properly defined particle detector, our plane wave modes can be shown to "leave tracks" in the way we would hope; in an array of such detectors, if a plane wave sets off one detector, there is a high probability that it will set off other detectors along a path from the first one in a direction given by the wave vector. (We should point out that, if you visit an actual particle accelerator at a place like Fermilab or CERN, the detectors bear little resemblance to those invented by theorists studying quantum field theory in curved spacetime; deep down, however, there is a fundamental similarity.) For a discussion of particle detectors see Birrell and Davies (1982).

You might worry about the factor $\delta^{(n-1)}(0)$ in the Hamiltonian (9.84), and well you should. It means that the Hamiltonian is infinite even when measured in the vacuum state $|0\rangle$. This term is the field-theory analogue of the harmonic-oscillator zero-point energy (9.20). In our discussion of the cosmological constant in Chapter 4, we mentioned that quantum fluctuations induced a formally infinite displacement of the classical vacuum energy; this infinite contribution to the scalar-field Hamiltonian will, when gravity is included, show up as a divergent cosmological constant. The fact that it is an integral over an infinite range of $\mathbf{k}$ of the infinite quantity $\delta^{(n-1)}(0)$ can be translated into the statement that the total energy is an integral over an infinitely big space of an infinite energy density. But the energy density contributed by high-frequency modes is the real problem, not the infinite volume; if we regularized the calculation by performing it in a box of volume L^{n-1}, we would find

$$\frac{1}{2}\int d^{n-1}k\,\delta^{(n-1)}(0)\omega \to \frac{1}{2}\left(\frac{L}{2\pi}\right)^{n-1}\sum_{\mathbf{k}}\omega, \tag{9.86}$$

which diverges even for finite L, since $\mathbf{k}$ (and thus ω) can be arbitrarily large. Putting a cutoff at some high momentum $k_{\max}$ would recover (4.104).

In the case of the simple harmonic oscillator, we pointed out that the zero-point energy could have been avoided had we chosen a classical potential with a negative minimum; the quantum-mechanical contribution does not necessarily represent the true answer, only the displacement of the energy from its classical value. The same holds in field theory; we are free to define our original classical scalar field theory so that the quantum-mechanical vacuum energy vanishes. However, we cannot simply subtract off a finite energy mode by mode, since our freedom is only to add a single constant to the potential, and thus to the Hamil-

tonian density (9.52). To obtain a finite Hamiltonian for the vacuum state, this constant would have to be infinite. There is nothing wrong with subtracting off an infinite constant; it is a venerable technique in quantum field theory, known as "renormalization." At times renormalization can seem scary or somehow illegitimate, but in truth it is perfectly sensible; infinities only arise in the relationship between quantum theories and their classical counterparts, not in any observable quantities. Since Nature presumably doesn't know or care about our fondness for classical mechanics, there should be nothing deeply disturbing about renormalization.

Of course, once we renormalize to obtain a finite vacuum energy, this energy could be anything we like; it is completely arbitrary. This continues to hold for quantum field theory in curved spacetime; we might not be able to decompose the field into modes of definite frequency, and it is therefore impossible to assign a vacuum energy contribution to each mode, but a careful analysis allows one to renormalize the vacuum energy to whatever number you like. Again, nothing profound has happened; the vacuum energy was completely arbitrary in our classical model in the first place, we simply chose it to be zero for convenience. The cosmological constant problem does not arise because quantum mechanics contributes a huge amount of vacuum energy, since this contribution can be straightforwardly renormalized away; the problem arises because there is no reason for the resulting arbitrary number to be close to zero. As discussed before, from the point of view of effective field theory the problem is somewhat sharper, since there is a logical expectation for the scale of the vacuum energy, namely the Planck scale at which unknown quantum-gravity effects should be contributing. Throughout this chapter, however, we will only be concerned with the propagation of quantum fields in fixed spacetime backgrounds, not in using the quantum energy-momentum tensor as a source for Einstein's equation; we can therefore choose to ignore the cosmological constant problem.

9.4 ■ QUANTUM FIELD THEORY IN CURVED SPACETIME

In Chapter 4 we discussed how easy it is to generalize physical theories from flat to curved spacetime—we simply express the theories in a coordinate-invariant form, and assert that they remain true when spacetime is curved. This procedure remains valid for quantum field theory, although we will need to give up on some of the concepts that seemed indispensable in flat spacetime.

We start with the Lagrange density of a scalar field in curved spacetime,

$$\mathcal{L} = \sqrt{-g}\left(-\tfrac{1}{2}g^{\mu\nu}\nabla_\mu\phi\nabla_\nu\phi - \tfrac{1}{2}m^2\phi^2 - \xi R\phi^2\right). \tag{9.87}$$

Aside from the predictable appearance of the metric $g_{\mu\nu}$ and its determinant, we have also included a direct coupling to the curvature scalar R, parameterized by a constant ξ. Since ξ is dimensionless, there is no reason to expect that it is small;

indeed, it should naturally be of order unity. In the literature there are two favorite choices for the value of ξ: **minimal coupling** simply turns off the direct interaction with R,

$$\xi = 0, \tag{9.88}$$

while **conformal coupling** sets

$$\xi = \frac{(n-2)}{4(n-1)}, \tag{9.89}$$

which is $\xi = \frac{1}{6}$ in four dimensions. Using the formulas in Appendix G, it is easy to check that when ξ takes on this value and $m = 0$, the scalar field theory is invariant under conformal transformations $g_{\mu\nu} \to \omega^2(x) g_{\mu\nu}$. In fact, there is no good reason to choose either minimal or conformal coupling in the real world; no symmetry is enhanced by minimal coupling, and conformal invariance is certainly not a symmetry of most physical theories. (Since conformal transformations are local changes of scale, theories characterized by dimensionful parameters such as masses will generally not be conformally invariant.) Even if a classical theory is conformally invariant, quantization can break this symmetry, which happens for example in the theory of quantum chromodynamics (QCD) coupled to massless quarks. Generally, in four dimensions it is difficult to find exactly conformally invariant interacting theories, although some models with high degrees of supersymmetry are known to be conformally invariant.

We may proceed to quantize the theory as before. The conjugate momentum is

$$\pi = \frac{\partial \mathcal{L}}{\partial(\nabla_0 \phi)}, \tag{9.90}$$

which for the Lagrangian (9.87) is

$$\pi = \sqrt{-g} \nabla_0 \phi. \tag{9.91}$$

We can impose canonical commutation relations

$$\begin{aligned} [\phi(t, \mathbf{x}), \phi(t, \mathbf{x}')] &= 0 \\ [\pi(t, \mathbf{x}), \pi(t, \mathbf{x}')] &= 0 \\ [\phi(t, \mathbf{x}), \pi(t, \mathbf{x}')] &= \frac{i}{\sqrt{-g}} \delta^{(n-1)}(\mathbf{x} - \mathbf{x}'). \end{aligned} \tag{9.92}$$

The equation of motion for the scalar field is

$$\Box \phi - m^2 \phi - \xi R \phi = 0. \tag{9.93}$$

For a spacelike hypersurface Σ with induced metric γ_{ij} and unit normal vector n^μ, the inner product on solutions to this equation is

$$(\phi_1, \phi_2) = -i \int_\Sigma \left(\phi_1 \nabla_\mu \phi_2^* - \phi_2^* \nabla_\mu \phi_1\right) n^\mu \sqrt{\gamma}\, d^{n-1}x, \tag{9.94}$$

which is independent of the choice of Σ.

So far, so good. To continue the steps we took in flat space, we would now introduce a set of positive- and negative-frequency modes forming a complete basis for solutions to (9.93), expand the field operator ϕ in terms of these modes, and interpret the operator coefficients as creation and annihilation operators. It is at this point where our procedure breaks down. Since there will generically not be any timelike Killing vector, we will not in general be able to find solutions to the wave equation that separate into time-dependent and space-dependent factors, and correspondingly cannot classify modes as positive- or negative-frequency. We can find a set of basis modes, but the problem is that there will generally be many such sets, with no way to prefer one over any others, and the notion of a vacuum or number operator will depend sensitively on which set we choose.

Let's see what we can do. We will always be able to find a set of solutions $f_i(x^\mu)$ to (9.93) that are orthonormal,

$$(f_i, f_j) = \delta_{ij}, \tag{9.95}$$

and corresponding conjugate modes with negative norm,

$$(f_i^*, f_j^*) = -\delta_{ij}. \tag{9.96}$$

The index i may be continuous or discrete; for the moment we will adopt notation appropriate to the discrete case. These modes can be chosen to be a complete set, so that we may expand our field as

$$\phi = \sum_i \left(\hat{a}_i f_i + \hat{a}_i^\dagger f_i^*\right). \tag{9.97}$$

The coefficients $\hat{a}_i$ and $\hat{a}_i^\dagger$ have commutation relations

$$\begin{aligned} [\hat{a}_i, \hat{a}_j] &= 0 \\ [\hat{a}_i^\dagger, \hat{a}_j^\dagger] &= 0 \\ [\hat{a}_i, \hat{a}_j^\dagger] &= \delta_{ij}. \end{aligned} \tag{9.98}$$

There will be a vacuum state $|0_f\rangle$ that is annihilated by all the annihilation operators,

$$\hat{a}_i |0_f\rangle = 0 \qquad \text{for all } i. \tag{9.99}$$

From this vacuum state we can define an entire Fock basis for the Hilbert space. As before, a state with n_i excitations is created by repeated action by $\hat{a}_i^\dagger$,

$$|n_i\rangle = \frac{1}{\sqrt{n_i!}}\left(\hat{a}_i^\dagger\right)^{n_i}|0_f\rangle, \tag{9.100}$$

and likewise for states with different kinds of excitations. We can even define a number operator for each mode,

$$\hat{n}_{fi} = \hat{a}_i^\dagger \hat{a}_i. \tag{9.101}$$

The subscript f on the vacuum state and the number operator reminds us that they are defined with respect to the set of modes f_i.

This apparatus seems quite similar to what we had in flat space; why can't we declare the excitations created by $\hat{a}_i^\dagger$ to be particles and be done with it? We could, but we must face the fact that there are other choices we could have made; the basis modes $f_i(x^\mu)$ are highly nonunique. Consider an alternative set of modes $g_i(x^\mu)$ with all of the properties that our original modes possessed, including forming (along with conjugate modes g_i^*) a complete basis with respect to which we can expand our field operator,

$$\phi = \sum_i \left(\hat{b}_i g_i + \hat{b}_i^\dagger g_i^*\right). \tag{9.102}$$

The annihilation and creation operators $\hat{b}_i$ and $\hat{b}_i^\dagger$ have commutation relations

$$\begin{aligned}[\hat{b}_i, \hat{b}_j] &= 0 \\ [\hat{b}_i^\dagger, \hat{b}_j^\dagger] &= 0 \\ [\hat{b}_i, \hat{b}_j^\dagger] &= \delta_{ij},\end{aligned} \tag{9.103}$$

and there will be a vacuum state $|0_g\rangle$ that is annihilated by all the annihilation operators,

$$\hat{b}_i|0_g\rangle = 0 \qquad \text{for all } i. \tag{9.104}$$

We can construct a Fock basis by repeated application of creation operators on this vacuum, and define a number operator

$$\hat{n}_{gi} = \hat{b}_i^\dagger \hat{b}_i. \tag{9.105}$$

What we have lost in the transition from flat to curved spacetime is any reason to prefer one set of modes over any other. In flat spacetime, we were able to pick out a natural set of modes by demanding that they be positive-frequency with respect to the time coordinate, as defined by (9.61). The time coordinate is not unique, since we are free to perform Lorentz transformations; but we saw that the vacuum state and total number operators are invariant under such transformations. Thus, every inertial observer will agree on what is the vacuum state, and how many particles are around.

In the more general context we are considering now, if one observer defines particles with respect to a set of modes f_i and another observer uses a set of modes g_i, they will typically disagree on how many particles are observed (or even if particles are observed at all). To see this, it is convenient to expand each set of modes in terms of the other,

$$\begin{aligned} g_i &= \sum_j \left(\alpha_{ij} f_j + \beta_{ij} f_j^*\right) \\ f_i &= \sum_j \left(\alpha_{ji}^* g_j - \beta_{ji} g_j^*\right). \end{aligned} \tag{9.106}$$

The transformation from one set of basis modes into another is known as a **Bogolubov transformation**, and the matrices α_{ij} and β_{ij} implementing the transformation are Bogolubov coefficients. Using the orthonormality of the mode functions, they can be expressed as

$$\begin{aligned} \alpha_{ij} &= (g_i, f_j) \\ \beta_{ij} &= -(g_i, f_j^*). \end{aligned} \tag{9.107}$$

They satisfy their own normalization conditions,

$$\begin{aligned} \sum_j \left(\alpha_{ik}\alpha_{jk}^* - \beta_{ik}\beta_{jk}^*\right) &= \delta_{ij} \\ \sum_j \left(\alpha_{ik}\beta_{jk} - \beta_{ik}\alpha_{jk}\right) &= 0. \end{aligned} \tag{9.108}$$

As well as describing a transformation between modes, the Bogolubov coefficients can be used to transform between the operators,

$$\begin{aligned} \hat{a}_i &= \sum_j \left(\alpha_{ji}\hat{b}_j + \beta_{ji}^*\hat{b}_j^\dagger\right) \\ \hat{b}_i &= \sum_j \left(\alpha_{ij}^*\hat{a}_j - \beta_{ij}^*\hat{a}_j^\dagger\right). \end{aligned} \tag{9.109}$$

Now imagine that the system is in the f-vacuum $|0_f\rangle$, in which no f-particles would be observed; we would like to know how many particles are observed by an observer using the g-modes. We therefore calculate the expectation value of the g number operator in the f-vacuum:

$$\begin{aligned} \langle 0_f|\hat{n}_{gi}|0_f\rangle &= \langle 0_f|b_i^\dagger b_i|0_f\rangle \\ &= \left\langle 0_f\middle|\sum_{jk}\left(\alpha_{ij}\hat{a}_j^\dagger - \beta_{ij}\hat{a}_j\right)\left(\alpha_{ik}^*\hat{a}_k - \beta_{ik}^*\hat{a}_k^\dagger\right)\middle|0_f\right\rangle \\ &= \sum_{jk}\left(-\beta_{ij}\right)\left(-\beta_{ik}^*\right)\langle 0_f|\hat{a}_j\hat{a}_k^\dagger|0_f\rangle \end{aligned}$$

$$= \sum_{jk} \beta_{ij}\beta^*_{ik} \langle 0_f | \left(\hat{a}^\dagger_k \hat{a}_j + \delta_{jk} \right) | 0_f \rangle$$

$$= \sum_{jk} \beta_{ij}\beta^*_{ik}\delta_{jk} \langle 0_f | 0_f \rangle$$

$$= \sum_{j} \beta_{ij}\beta^*_{ij}. \tag{9.110}$$

The number of g-particles in the f-vacuum can thus be expressed in terms of the Bogolubov coefficients as

$$\langle 0_f | \hat{n}_{gi} | 0_f \rangle = \sum_j |\beta_{ij}|^2. \tag{9.111}$$

There is no reason for this to vanish; what looks like an empty vacuum from one perspective will be bubbling with particles according to another. If any of the β_{ij} are nonvanishing, the vacuum states will not coincide. We can understand why this is by looking at (9.109), where we see that β_{ij} describes the admixture of creation operators from one basis into the annihilation operators in the other basis.

This talk about modes and number operators may seem unnecessarily abstract; certainly, if an actual particle detector is traveling along some trajectory in a possibly-curved spacetime, it will either detect particles or not, without knowing what set of basis modes we are using for field theory. How do we know what definition of "particles" is actually being used by such a detector? The answer is that a detector measures the proper time τ along its trajectory, and will define positive- and negative-frequency with respect to that proper time. Thus, if a set of modes f_i can be found that obey

$$\frac{D}{d\tau} f_i = -i\omega f_i, \tag{9.112}$$

we can use these modes to calculate how many particles the detector will see. Of course, it will generally not be possible to find such modes all over the spacetime. The one time that it might be possible is in a *static* spacetime, when we have a hypersurface-orthogonal timelike Killing vector K^μ. In that case we can choose coordinates in which the metric components are independent of the time coordinate t, and there are no time-space cross terms:

$$\partial_0 g_{\mu\nu} = 0, \qquad g_{0i} = 0. \tag{9.113}$$

(Indices i, j are now spatial components, not mode labels.) For such a metric, the d'Alembertian acting on some mode function $f(t, \mathbf{x})$ works out to be

$$\Box f = \left[g^{00}\partial_0^2 + \tfrac{1}{2} g^{00} g^{ij} (\partial_i g_{00}) \partial_j + g^{ij}\partial_i\partial_j - g^{ij}\Gamma^k_{ij}\partial_k \right] f. \tag{9.114}$$

The equation of motion (9.93) can thus be written in the form

$$\partial_0^2 f = -\left(g^{00}\right)^{-1}\left[g^{ij}\partial_i\partial_j + \tfrac{1}{2}g^{00}g^{ij}(\partial_i g_{00})\partial_j - g^{ij}\Gamma^k_{ij}\partial_k - (m^2 + \xi R)\right] f. \tag{9.115}$$

The operator on the left is a pure time derivative, while the operator on the right involves only spatial derivatives and functions of space alone. We can therefore find separable solutions

$$f_\omega(t, \mathbf{x}) = e^{-i\omega t}\bar{f}_\omega(\mathbf{x}), \tag{9.116}$$

which can be described as positive-frequency,

$$\partial_t f_\omega(t, \mathbf{x}) = -i\omega f_\omega(t, \mathbf{x}), \qquad \omega > 0. \tag{9.117}$$

This relation can be recast in a coordinate-invariant form as

$$\mathcal{L}_K f_\omega = K^\mu \partial_\mu f_\omega = -i\omega f_\omega, \qquad \omega > 0, \tag{9.118}$$

where $\mathcal{L}_K f_\omega$ denotes the Lie derivative of f_ω along K. There will also be negative-frequency conjugate modes,

$$\mathcal{L}_K f^*_\omega = K^\mu \partial_\mu f^*_\omega = i\omega f^*_\omega, \qquad \omega > 0. \tag{9.119}$$

Together, the modes (f_ω, f^*_ω) will form a basis for solutions to the wave equation in a static background. The existence of such modes won't help us unless they are relevant for our detector; if the detector's trajectory follows along orbits of the Killing field (the four-velocity $U^\mu = dx^\mu/d\tau$ is proportional to K^μ), the proper time will be proportional to the Killing time t, and modes that are positive-frequency with respect to this Killing vector will serve as a natural basis for describing Fock space. We will see this phenomenon at work in our discussion of the Unruh effect in the next section.

In the last section we mentioned the need to renormalize the vacuum energy in quantum field theory. This requirement still exists in curved spacetime, but an appropriate renormalization procedure is harder to construct, since there is no preferred mode basis. Nevertheless, algebraic methods have been developed to define a renormalized energy-momentum tensor rigorously, at least in certain cases; we won't delve into this subject in detail, but should at least present some of the underlying philosophy. The basic idea is that, even in the presence of curvature, spacetime should look Minkowskian on small enough scales. Because the vacuum-energy divergence we found in flat spacetime was due to short-wavelength modes, we should be able to match the behavior of fields in curved spacetime on very small scales to those in flat spacetime, and subtract off any divergences that appear. In particular, we consider the two-point function of a quantum field ϕ in some state $|\psi\rangle$,

$$G(x_1, x_2) = \langle\psi|\phi(x_1)\phi(x_2)|\psi\rangle, \tag{9.120}$$

where x_1 and x_2 are two spacetime points. The two-point function in the Minkowski vacuum becomes singular as x_1 and x_2 are brought close to each other. We would like to characterize this singularity, and insist that it hold for any regular state in curved spacetime. By "brought close to each other" we mean that $\sigma(x_1, x_2)$, the squared distance along the shortest geodesic connecting the two points, goes to zero. In the limit as x_1 and x_2 are very close, the squared geodesic distance is simply

$$\sigma(x_1, x_2) = g_{\mu\nu}(x_1^\mu - x_2^\mu)(x_1^\nu - x_2^\nu), \qquad x_1 \to x_2. \tag{9.121}$$

Of course, in a Lorentzian manifold, the geodesic distance will vanish when points are null separated, not only when they are coincident. We therefore include a small imaginary part and take the limit as it goes to zero, by defining

$$\sigma_\epsilon(x_1, x_2) = \sigma(x_1, x_2) + 2i\epsilon(t_1 - t_2) + \epsilon^2. \tag{9.122}$$

Here, t is the timelike coordinate, and the limit as $\epsilon \to 0^+$ is assumed. (The manifest coordinate-dependence of this formula will be irrelevant in this limit.) Then it turns out that there is a unique singularity structure for the natural vacuum in Minkowski spacetime, such that the two-point function (in four dimensions) contains a leading singularity of the form $1/(4\pi^2\sigma_\epsilon)$ and a subleading one proportional to $\ln\sigma_\epsilon$, with all other terms being regular. We therefore require that any physically reasonable quantum state in curved spacetime obey

$$G(x_1, x_2) = \frac{U(x_1, x_2)}{4\pi^2\sigma_\epsilon} + V(x_1, x_2)\ln\sigma_\epsilon + W(x_1, x_2), \tag{9.123}$$

where the functions $U(x_1, x_2)$, $V(x_1, x_2)$, and $W(x_1, x_2)$ are all regular at $x_1 = x_2$, and $U(x, x) = 1$. A state with this property is said to be a **Hadamard state**. It can be shown that the renormalized energy-momentum tensor is well-defined and nonsingular in all Hadamard states, and furthermore that it will be singular in any non-Hadamard state. If the Hadamard condition is obeyed on some partial Cauchy surface, it will also be obeyed everywhere in the domain of dependence; in other words, the energy-momentum tensor may become singular on a horizon, but not within the Cauchy development of some well-posed initial data. States of this form, therefore, seem appropriate for consideration in QFT on curved spacetime. For details see Wald (1994).

We see that QFT in curved spacetime shares most of the basic features of QFT in flat spacetime; the crucial difference involves what we cannot do, namely decide on a natural set of basis modes that all inertial observers would identify as particles. At the end of Section 9.2 we briefly discussed an oscillator subject to a transient force, and how to define an S-matrix relating number eigenstates at

early times to number eigenstates at late times. The same set of ideas translates directly to quantum field theory. If we have a situation in which spacetime is static in the asymptotic past and future, but with some disturbance in between, we can define in- and out-states that are energy eigenstates at early and late times, and a set of Bogolubov coefficients describing how the in-vacuum (for example) will be described as a multiparticle configuration in terms of the out-states. This phenomenon goes by the name of particle production by gravitational fields; relevant physical examples include the early universe and black holes.[2]

9.5 ■ THE UNRUH EFFECT

We must admit that, having put so much effort into understanding the basics of quantum field theory in curved spacetime, we won't actually do any detailed calculations in a curved background. Instead, we will investigate a phenomenon that relies on the ideas we have introduced, but is manifested even in flat spacetime: the Unruh effect, which states that an accelerating observer in the traditional Minkowski vacuum state will observe a thermal spectrum of particles. Historically, the Unruh effect was discovered in an attempt to understand the physics underlying the Hawking effect (thermal radiation in the presence of a black hole event horizon). Our strategy will be to carefully derive the Unruh effect, and in the next section argue under reasonable assumptions that this implies the Hawking effect, which is more difficult to derive directly just because it's harder to solve wave equations in curved spacetime than in flat spacetime.

The basic idea of the Unruh effect is simple: it is a manifestation of the idea that observers with different notions of positive- and negative-frequency modes will disagree on the particle content of a given state. For a uniformly accelerated observer in Minkowski space, the trajectory will move along orbits of a timelike Killing vector, but not that of the usual time-translation symmetry. We can therefore expand the field in modes appropriate to the accelerated observer, and calculate the number operator in the ordinary Minkowski vacuum, where we will find a thermal spectrum of particles. Different sets of explanatory words can be attached to this result; the basic lesson to learn is that what we think of as an inert vacuum actually has the character of a thermal state.

In the interest of discarding all possible complications to get at the underlying phenomenon, we consider a quantum field theory that is as simple as it can be without becoming completely trivial: a massless ($m = 0$) scalar field in two spacetime dimensions ($n = 2$). In two dimensions, conformal coupling and minimal coupling coincide, so we do not include any direct interaction with the curvature scalar. (We're in flat spacetime, so such a coupling wouldn't have any effect anyway.) The relevant wave equation is thus

$$\Box\phi = 0. \tag{9.124}$$

[2]Interestingly, the first discussion of particle production in curved spacetime was given by Schrödinger himself; see E. Schrödinger (1939), *Physica* (Utrecht) **6**, 899.

Before diving into the quantization of this field theory, let's think about two-dimensional Minkowski space as seen by a uniformly accelerating observer. We know that the metric can be written in inertial coordinates as

$$ds^2 = -\mathrm{d}t^2 + \mathrm{d}x^2. \tag{9.125}$$

Consider an observer moving at a uniform acceleration of magnitude α in the x-direction. We claim that the resulting trajectory $x^\mu(\tau)$ will be given by

$$\begin{aligned} t(\tau) &= \frac{1}{\alpha}\sinh(\alpha\tau) \\ x(\tau) &= \frac{1}{\alpha}\cosh(\alpha\tau). \end{aligned} \tag{9.126}$$

Let's verify that this path corresponds to constant acceleration. The acceleration two-vector is given in the globally inertial coordinate system by

$$a^\mu = \frac{D^2x^\mu}{d\tau^2} = \frac{d^2x^\mu}{d\tau^2}, \tag{9.127}$$

where the covariant derivative along the path is equal to the ordinary derivative because the Christoffel symbols vanish in these coordinates. The components of a^μ are thus

$$\begin{aligned} a^t &= \alpha\sinh(\alpha\tau) \\ a^x &= \alpha\cosh(\alpha\tau), \end{aligned} \tag{9.128}$$

and the magnitude is

$$\sqrt{a_\mu a^\mu} = \sqrt{-\alpha^2\sinh^2(\alpha\tau) + \alpha^2\cosh^2(\alpha\tau)} = \alpha. \tag{9.129}$$

The path therefore corresponds to a constant acceleration of magnitude α, as desired. The trajectory of our accelerated observer obeys the relation

$$x^2(\tau) = t^2(\tau) + \alpha^2, \tag{9.130}$$

and thus describes an hyperboloid asymptoting to null paths $x = -t$ in the past and $x = t$ in the future. The accelerated observer travels from past null infinity to future null infinity, rather than timelike infinity as would be reached by geodesic observers.

We can choose new coordinates (η, ξ) on two-dimensional Minkowski space that are adapted to uniformly accelerated motion. Let

$$t = \frac{1}{a}e^{a\xi}\sinh(a\eta), \qquad x = \frac{1}{a}e^{a\xi}\cosh(a\eta) \qquad (x > |t|). \tag{9.131}$$

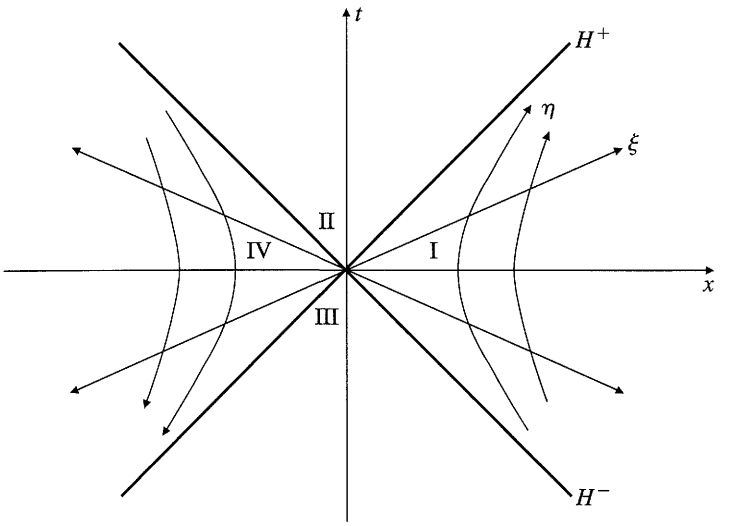

FIGURE 9.1 Minkowski spacetime in Rindler coordinates. Region I is the region accessible to an observer undergoing constant acceleration in the $+x$-direction. The coordinates (η, ξ) can be used in region I, or separately in region IV, where they point in the opposite sense. The vector field ∂_η corresponds to the generator of Lorentz boost symmetry. The horizons $H^\pm$ are Killing horizons for this vector field, and also represent boundaries of the past and future as witnessed by the Rindler observer.

The new coordinates have ranges

$$-\infty < \eta, \xi < +\infty, \tag{9.132}$$

and cover the wedge $x > |t|$, labeled as region I in Figure 9.1. In these coordinates, the constant-acceleration path (9.126) is given by

$$\eta(\tau) = \frac{\alpha}{a}\tau$$
$$\xi(\tau) = \frac{1}{a}\ln\left(\frac{a}{\alpha}\right), \tag{9.133}$$

so that the proper time is proportional to η and the spatial coordinate ξ is constant. In particular, an observer with $\alpha = a$ moves along the path

$$\eta = \tau, \qquad \xi = 0. \tag{9.134}$$

The metric in these coordinates takes the form

$$ds^2 = e^{2a\xi}(-d\eta^2 + d\xi^2). \tag{9.135}$$

Region I, with this metric, is known as **Rindler space**, even though it is obviously just a part of Minkowski space. A **Rindler observer** is one moving along

a constant-acceleration path, as in (9.133). The causal structure of Rindler space resembles the region $r > 2GM$ of the maximally extended Schwarzschild solution of Figure 5.12. In particular, the null line $x = t$, labeled H^+ in Figure 9.1, is a future Cauchy horizon for any $\eta =$ constant spacelike hypersurface in region I; similarly, H^- is a past Cauchy horizon. These horizons are reminiscent of the event horizons in the Kruskal diagram, with static observers ($r =$ constant) in Schwarzschild being related to constant-acceleration paths in Rindler space.

The metric components in (9.135) are independent of η, so we immediately know that ∂_η is a Killing vector. But of course this is just Minkowski spacetime, so we think we know what all of the Killing vectors are. Indeed, if we express ∂_η in the (t, x) coordinates, we find

$$\begin{aligned}\partial_\eta &= \frac{\partial t}{\partial \eta}\partial_t + \frac{\partial x}{\partial \eta}\partial_x \\ &= e^{a\xi}\left[\cosh(a\eta)\partial_t + \sinh(a\eta)\partial_x\right] \\ &= a(x\partial_t + t\partial_x).\end{aligned} \tag{9.136}$$

This is nothing more or less than the Killing field associated with a boost in the x-direction. It is clear from this expression that this Killing field naturally extends throughout the spacetime; in regions II and III it is spacelike, while in region IV it is timelike but past-directed. The horizons we have identified are actually Killing horizons for ∂_η. The redshift factor, defined in (6.12) as the magnitude of the norm of the Killing vector, is

$$V = e^{a\xi}. \tag{9.137}$$

The surface gravity $\kappa = \sqrt{\nabla_\mu V \nabla^\mu V}$ of this Killing horizon is thus

$$\kappa = a. \tag{9.138}$$

There is no real gravitational force, since we're in flat space; but this surface gravity characterizes the acceleration of Rindler observers.

We can also define coordinates (η, ξ) in region IV by flipping the signs in (9.131),

$$t = -\frac{1}{a}e^{a\xi}\sinh(a\eta), \qquad x = -\frac{1}{a}e^{a\xi}\cosh(a\eta) \qquad (x < |t|). \tag{9.139}$$

The sign guarantees that ∂_η and ∂_t point in opposite directions in region IV. Strictly speaking, we cannot use (η, ξ) simultaneously in regions I and IV, since the ranges of these coordinates are the same in each region, but we will be okay so long as we explicitly indicate to which region we are referring. The reason why it's better to use the same set of coordinate labels twice, rather than simply introducing new coordinates, is that the metric (9.135) will apply to both region I and region IV.

Along the surface $t = 0$, ∂_η is a hypersurface-orthogonal timelike Killing vector, except for the single point $x = 0$ where it vanishes. This vector can therefore be used to define a set of positive- and negative-frequency modes, on which we can build a Fock basis for the scalar-field Hilbert space. The massless Klein–Gordon equation in Rindler coordinates takes the form

$$\Box\phi = e^{-2a\xi}(-\partial_\eta^2 + \partial_\xi^2)\phi = 0. \tag{9.140}$$

A normalized plane wave $g_k = (4\pi\omega)^{-1/2}e^{-i\omega\eta+ik\xi}$, with $\omega = |k|$, solves this equation and apparently has positive frequency, in the sense that $\partial_\eta g_k = -i\omega g_k$. But we need our modes to be positive-frequency with respect to a future-directed Killing vector, and in region IV that role is played by $\partial_{(-\eta)} = -\partial_\eta$ rather than ∂_η. To deal with this annoyance, we introduce two sets of modes, one with support in region I and the other in region IV:

$$g_k^{(1)} = \begin{cases} \dfrac{1}{\sqrt{4\pi\omega}}e^{-i\omega\eta+ik\xi} & \text{I} \\ 0 & \text{IV} \end{cases}$$
$$g_k^{(2)} = \begin{cases} 0 & \text{I} \\ \dfrac{1}{\sqrt{4\pi\omega}}e^{+i\omega\eta+ik\xi} & \text{IV} \end{cases} \tag{9.141}$$

We take $\omega = |k|$ in each case; in two dimensions, the spatial wave vector is just the single number k. Each set of modes is positive-frequency with respect to the appropriate future-directed timelike Killing vector,

$$\begin{aligned} \partial_\eta g_k^{(1)} &= -i\omega g_k^{(1)} \\ \partial_{(-\eta)} g_k^{(2)} &= -i\omega g_k^{(2)}, \qquad \omega > 0. \end{aligned} \tag{9.142}$$

These two sets, along with their conjugates, form a complete set of basis modes for any solutions to the wave equation throughout the spacetime. (The single point $x = t = 0$ is a set of measure zero, so we shouldn't have to worry about it.) Both sets are nonvanishing in regions II and III of the Rindler diagram; this is obscured by writing them in terms of the coordinates η and ξ, but these functions can be analytically extended into the future and past regions. Denoting the associated annihilation operators as $\hat{b}_k^{(1,2)}$, we can write

$$\phi = \int dk \left(\hat{b}_k^{(1)} g_k^{(1)} + \hat{b}_k^{(1)\dagger} g_k^{(1)*} + \hat{b}_k^{(2)} g_k^{(2)} + \hat{b}_k^{(2)\dagger} g_k^{(2)*}\right). \tag{9.143}$$

This expansion is an alternative to our expression (9.66) in terms of the original Minkowski modes, which in two dimensions takes the form

$$\phi = \int dk \left(\hat{a}_k f_k + \hat{a}_k^\dagger f_k^*\right). \tag{9.144}$$

It is straightforward to check that the modes (9.141) are properly normalized with respect to the inner product (9.94). In the metric (9.135), the future-directed unit normal to the surface $\eta = 0$ is normalized to

$$-1 = g_{\mu\nu} n^\mu n^\nu = -e^{2a\xi} (n^0)^2, \tag{9.145}$$

or

$$n^0 = e^{-a\xi}. \tag{9.146}$$

Meanwhile, the spatial metric determinant satisfies

$$\sqrt{\gamma} = e^{a\xi}. \tag{9.147}$$

We therefore have $n^0 \sqrt{\gamma} = 1$, and the calculation of the inner product of the Rindler modes follows precisely that of ordinary Minkowski modes. We end up with

$$\begin{aligned} (g^{(1)}_{k_1}, g^{(1)}_{k_2}) &= \delta(k_1 - k_2) \\ (g^{(2)}_{k_1}, g^{(2)}_{k_2}) &= \delta(k_1 - k_2) \\ (g^{(1)}_{k_1}, g^{(2)}_{k_2}) &= 0, \end{aligned} \tag{9.148}$$

and similarly for the conjugate modes.

There are thus two sets of modes, Minkowski and Rindler, with which we can expand solutions to the Klein–Gordon equation in a flat two-dimensional spacetime. Although the Hilbert space for the theory is the same in either representation, its interpretation as a Fock space will be different; in particular, the vacuum states will be different. The Minkowski vacuum $|0_{\mathrm{M}}\rangle$, satisfying

$$\hat{a}_k |0_{\mathrm{M}}\rangle = 0, \tag{9.149}$$

will be described as a multi-particle state in the Rindler representation; likewise, the Rindler vacuum $|0_{\mathrm{R}}\rangle$, satisfying

$$\hat{b}^{(1)}_k |0_{\mathrm{R}}\rangle = \hat{b}^{(2)}_k |0_{\mathrm{R}}\rangle = 0, \tag{9.150}$$

will be described as a multi-particle state in the Minkowski representation. At a practical level, the difference arises because an individual Rindler mode can never be written as a sum of positive-frequency Minkowski modes; at $t = 0$ the Rindler modes only have support on the half-line, and such a function cannot be expanded in purely positive-frequency plane waves. Thus, the Rindler annihilation operators used to define $|0_{\mathrm{R}}\rangle$ are necessarily superpositions of Minkowski creation and annihilation operators, so the two vacua cannot coincide.

A Rindler observer will be static with respect to orbits of the boost Killing vector ∂_η. Such an observer in region I will therefore describe particles in terms

of the Rindler modes $g_k^{(1)}$, and in particular will observe a state in the Rindler vacuum to be devoid of particles, a state $\hat{b}_k^{(1)\dagger}|0_R\rangle$ to contain a single particle of frequency $\omega = |k|$, and so on. Conversely, a Rindler observer traveling through the Minkowski vacuum state will detect a background of particles, even though an inertial observer would describe the state as being completely empty. What kind of particles would the Rindler observer detect? We know how to answer this question: Calculate the Bogolubov coefficients relating the Minkowski and Rindler modes, and use them to determine the expectation value of the Rindler number operator in the Minkowski vacuum. This is straightforward but tedious, so we will take a shortcut due to Unruh. We will find a set of modes that share the same vacuum state as the Minkowski modes (although the description of excited states may be different), but for which the overlap with the Rindler modes is more direct. The way to do this is to start with the Rindler modes, analytically extend them to the entire spacetime, and express this extension in terms of the original Rindler modes.

To see how this works, notice from (9.131) and (9.139) that we have the following relationships between the Minkowski coordinates (t, x) and Rindler coordinates (η, ξ) in regions I and IV:

$$e^{-a(\eta-\xi)} = \begin{cases} a(-t+x) & \text{I} \\ a(t-x) & \text{IV} \end{cases}$$
$$e^{a(\eta+\xi)} = \begin{cases} a(t+x) & \text{I} \\ a(-t-x) & \text{IV} \end{cases} \tag{9.151}$$

We can therefore express the spacetime dependence of a mode $g_k^{(1)}$ with $k > 0$ (so $\omega = k$) in terms of Minkowski coordinates in region I as

$$\begin{aligned} \sqrt{4\pi\omega}\, g_k^{(1)} &= e^{-i\omega\eta + ik\xi} \\ &= e^{-i\omega(\eta-\xi)} \\ &= a^{i\omega/a}(-t+x)^{i\omega/a}. \end{aligned} \tag{9.152}$$

The analytic extension of this function throughout spacetime is straightforward; we simply use this final expression for any values of (t, x). But we would like to express the result in terms of the original Rindler modes everywhere; since the $g_k^{(1)}$ modes vanish in region IV, we need to bring the modes $g_k^{(2)}$ into play. When we express them in terms of the Minkowski coordinates in region IV, for $k > 0$ we obtain

$$\begin{aligned} \sqrt{4\pi\omega}\, g_k^{(2)} &= e^{+i\omega\eta + ik\xi} \\ &= e^{+i\omega(\eta+\xi)} \\ &= a^{-i\omega/a}(-t-x)^{-i\omega/a}. \end{aligned} \tag{9.153}$$

This doesn't match the behavior of (9.152) that we want. But if we take the complex conjugate and reverse the wave number, we obtain

$$\begin{aligned}\sqrt{4\pi\omega}\, g_{-k}^{(2)*} &= e^{-i\omega\eta+ik\xi} \\ &= e^{-i\omega(\eta-\xi)} \\ &= a^{i\omega/a}(t-x)^{i\omega/a} \\ &= a^{i\omega/a}[e^{-i\pi}(-t+x)]^{i\omega/a} \\ &= a^{i\omega/a}e^{\pi\omega/a}(-t+x)^{i\omega/a}. \end{aligned} \tag{9.154}$$

The combination

$$\sqrt{4\pi\omega}\left(g_k^{(1)} + e^{-\pi\omega/a} g_{-k}^{(2)*}\right) = a^{i\omega/a}(-t+x)^{i\omega/a} \tag{9.155}$$

is therefore well-defined along the whole surface $t = 0$. We have explicitly examined the case $k > 0$, but an identical result obtains for $k < 0$.

A properly normalized version of this mode is given by

$$h_k^{(1)} = \frac{1}{\sqrt{2\sinh\left(\frac{\pi\omega}{a}\right)}}\left(e^{\pi\omega/2a} g_k^{(1)} + e^{-\pi\omega/2a} g_{-k}^{(2)*}\right). \tag{9.156}$$

This is an appropriate analytic extension of the $g_k^{(1)}$ modes; to get a complete set, we need to include the extensions of the $g_k^{(2)}$ modes, which by an analogous argument are given by

$$h_k^{(2)} = \frac{1}{\sqrt{2\sinh\left(\frac{\pi\omega}{a}\right)}}\left(e^{\pi\omega/2a} g_k^{(2)} + e^{-\pi\omega/2a} g_{-k}^{(1)*}\right). \tag{9.157}$$

To verify the normalization, for example for $h_k^{(1)}$, we use (9.148):

$$\begin{aligned}\left(h_{k_1}^{(1)}, h_{k_2}^{(1)}\right) &= \frac{1}{2\sqrt{\sinh\left(\frac{\pi\omega_1}{a}\right)\sinh\left(\frac{\pi\omega_2}{a}\right)}}\Big[e^{\pi(\omega_1+\omega_2)/2a}\left(g_{k_1}^{(1)}, g_{k_2}^{(1)}\right) \\ &\quad + e^{-\pi(\omega_1+\omega_2)/2a}\left(g_{-k_1}^{(2)*}, g_{-k_2}^{(2)*}\right)\Big] \\ &= \frac{1}{2\sqrt{\sinh\left(\frac{\pi\omega_1}{a}\right)\sinh\left(\frac{\pi\omega_2}{a}\right)}}\Big[e^{\pi(\omega_1+\omega_2)/2a}\delta(k_1-k_2) \\ &\quad + e^{-\pi(\omega_1+\omega_2)/2a}\delta(-k_1+k_2)\Big] \\ &= \frac{e^{\pi\omega_1/a} - e^{-\pi\omega_1/a}}{2\sinh\left(\frac{\pi\omega_1}{a}\right)}\delta(k_1-k_2)\end{aligned}$$

$$= \delta(k_1 - k_2), \tag{9.158}$$

just as we would like.

We can now expand our field in these modes,

$$\phi = \int dk \left(\hat{c}_k^{(1)} h_k^{(1)} + \hat{c}_k^{(1)\dagger} h_k^{(1)*} + \hat{c}_k^{(2)} h_k^{(2)} + \hat{c}_k^{(2)\dagger} h_k^{(2)*} \right). \tag{9.159}$$

From our discussion of Bogolubov transformations in Section 9.4, we know that the expressions (9.156) and (9.157) for the $h_k^{(1,2)}$ modes in terms of the $g_k^{(1,2)}$ modes implies corresponding expressions for the Rindler operators $\hat{b}_k^{(1,2)}$ in terms of the operators $\hat{c}_k^{(1,2)}$, as

$$\begin{aligned} \hat{b}_k^{(1)} &= \frac{1}{\sqrt{2\sinh\left(\frac{\pi\omega}{a}\right)}} \left(e^{\pi\omega/2a} \hat{c}_k^{(1)} + e^{-\pi\omega/2a} \hat{c}_{-k}^{(2)\dagger} \right) \\ \hat{b}_k^{(2)} &= \frac{1}{\sqrt{2\sinh\left(\frac{\pi\omega}{a}\right)}} \left(e^{\pi\omega/2a} \hat{c}_k^{(2)} + e^{-\pi\omega/2a} \hat{c}_{-k}^{(1)\dagger} \right). \end{aligned} \tag{9.160}$$

We can therefore express the Rindler number operator in region I,

$$\hat{n}_{\mathrm{R}}^{(1)}(k) = \hat{b}_k^{(1)\dagger} \hat{b}_k^{(1)}, \tag{9.161}$$

in terms of the new operators $\hat{c}_k^{(1,2)}$.

The original positive-frequency Minkowski plane-wave modes with $k > 0$, $f_k \propto e^{-i\omega(t-x)}$, are analytic and bounded for complex (t, x) so long as $\mathrm{Im}(t - x) \leq 0$. (Such modes are called "right-moving," as they describe waves propagating to the right.) The same holds for our new modes $h_k^{(1)}$ so long as we take the branch cut for the imaginary power to lie in the upper-half complex $(t - x)$ plane, as we can see from examination of (9.152) and (9.154); this is consistent with our setting $-1 = e^{-i\pi}$ in (9.154). Similar considerations apply to the $h_k^{(2)}$ modes, which are analytic and bounded in the lower-half complex $(t + x)$ plane, as are the positive-frequency Minkowski plane-wave modes with $k < 0$ (left-moving). Consequently, unlike the original Rindler modes $g_k^{(1,2)}$, we know that the modes $h_k^{(1,2)}$ can be expressed purely in terms of positive-frequency Minkowski modes f_k. They therefore share the same vacuum state $|0_{\mathrm{M}}\rangle$, so that

$$\hat{c}_k^{(1)} |0_{\mathrm{M}}\rangle = \hat{c}_k^{(2)} |0_{\mathrm{M}}\rangle = 0. \tag{9.162}$$

The excited states will not coincide, but that won't bother us, since we are interested in what a Rindler observer sees when the state is precisely in the Minkowski vacuum. An observer in region I, for example, will observe particles defined by the operators $\hat{b}_k^{(1)}$; the expected number of such particles of frequency ω will be given by

$$\langle 0_{\mathrm{M}} | \hat{n}_{\mathrm{R}}^{(1)}(k) | 0_{\mathrm{M}} \rangle = \langle 0_{\mathrm{M}} | \hat{b}_k^{(1)\dagger} \hat{b}_k^{(1)} | 0_{\mathrm{M}} \rangle$$

$$= \frac{1}{2\sinh\left(\frac{\pi\omega}{a}\right)} \langle 0_M | e^{-\pi\omega/a} \hat{c}^{(1)}_{-k} \hat{c}^{(1)\dagger}_{-k} | 0_M \rangle$$

$$= \frac{e^{-\pi\omega/a}}{2\sinh\left(\frac{\pi\omega}{a}\right)} \delta(0)$$

$$= \frac{1}{e^{2\pi\omega/a} - 1} \delta(0), \tag{9.163}$$

where we have used the fact that a $\hat{c}^{(1)\dagger}_k |0_M\rangle$ is a normalized one-particle state,

$$\langle 0_M | \hat{c}^{(1)}_k \hat{c}^{(1)\dagger}_k | 0_M \rangle = \delta(0). \tag{9.164}$$

The delta function in (9.163) is merely an artifact of our use of (nonsquare-integrable) plane wave basis modes; had we constructed normalized wave packets, we would have obtained a finite result with an identical spectrum.

The result (9.163) is a Planck spectrum with temperature

$$T = \frac{a}{2\pi}. \tag{9.165}$$

Thus, *an observer moving with uniform acceleration through the Minkowski vacuum observes a thermal spectrum of particles.* This is the **Unruh effect**. Of course, there is more to thermal radiation than just the spectrum (9.163); to be truly thermal, we should check that there are no hidden correlations in the observed particles. This has been verified; the radiation detected by a Rindler observer is truly thermal. At the most basic level, the Unruh effect shows how two different sets of observers (inertial and Rindler) will describe the same state in very different terms; at a slightly deeper level, it reveals the essentially thermal nature of the vacuum in quantum field theory.

The temperature $T = a/2\pi$ is what would be measured by an observer moving along the path $\xi = 0$, which feels an acceleration $\alpha = a$. Using (9.133), we know that any other path with $\xi =$ constant feels an acceleration

$$\alpha = ae^{-a\xi} \tag{9.166}$$

and thus should measure thermal radiation with a temperature $\alpha/2\pi$. This is consistent with our discussion in Chapter 6 of the redshift witnessed by static observers moving along orbits of some Killing vector K^μ; we found that radiation emitted with frequency ω_1 at a point x_1 would be observed at a point x_2 with a frequency

$$\omega_2 = \frac{V_1}{V_2}\omega_1, \tag{9.167}$$

where the redshift factor V is the norm of the Killing vector. In (9.137) we found that the redshift factor associated with ∂_η is $V = e^{a\xi}$, so that

$$\omega_2 = e^{a(\xi_1 - \xi_2)}\omega_1. \tag{9.168}$$

Thus, if an observer at $\xi_1 = 0$ detects a temperature $T = a/2\pi$, the observer at $\xi_2 = \xi$ will see it to be redshifted to a temperature $T = ae^{-a\xi}/2\pi$, just as in (9.166). In particular, the temperature redshifts all the way to zero as $\xi \to +\infty$. This makes sense, since a Rindler observer at infinity will be nearly inertial, and will define the same notion of vacuum and particles as an ordinary Minkowski observer.

The Unruh effect tells us that an accelerated observer will detect particles in the Minkowski vacuum state. An inertial observer, of course, would describe the same state as being completely empty; indeed, the expectation value of the energy-momentum tensor would be $\langle T_{\mu\nu} \rangle = 0$. But if there is no energy-momentum, how can the Rindler observers detect particles? This is a subtle issue, but by no means a contradiction. If the Rindler observer is to detect background particles, she must carry a detector—some sort of apparatus coupled to the particles being detected. But if a detector is being maintained at constant acceleration, energy is not conserved; we need to do work constantly on the detector to keep it accelerating. From the point of view of the Minkowski observer, the Rindler detector *emits* as well as absorbs particles; once the coupling is introduced, the possibility of emission is unavoidable. When the detector registers a particle, the inertial observer would say that it had emitted a particle and felt a radiation-reaction force in response. Ultimately, then, the energy needed to excite the Rindler detector does not come from the background energy-momentum tensor, but from the energy we put into the detector to keep it accelerating.

9.6 ■ THE HAWKING EFFECT AND BLACK HOLE EVAPORATION

Even though it occurs in flat spacetime, the Unruh effect teaches us the most important lesson of QFT in curved spacetime, the idea that "vacuum" and "particles" are observer-dependent notions rather than fundamental concepts. In fact, given our understanding of the Unruh effect, we can see almost immediately how the Hawking effect arises. This should not be too surprising, as we have already noted the similarity between the causal structure of Rindler space and that of the maximally-extended Schwarzschild spacetime describing an eternal black hole. We will therefore be able to argue in favor of Hawking radiation without ever doing an explicit calculation in curved spacetime; of course, there are many features that you might like to investigate in more detail, for which the full power of the curved metric is necessary. In addition to Birrell and Davies (1982) and Wald (1994), there are good review articles where you can find a more full discussion of the issues discussed here.[3] Our derivation of Hawking radiation follows that of Jacobson.

[3]T.A. Jacobson, "Introductory Lectures on Black Hole Thermodynamics," Lectures at University of Utrecht (1996), `http://www.fys.ruu.nl/~wwwthe/lectures/itfuu-0196.ps`; R.M. Wald, "The thermodynamics of black holes," *Living Rev. Rel.* **4**, 6 (2001), `http://arxiv.org/gr-qc/9912119`; J. Traschen, "An introduction to black hole evaporation" (2000), `http://arxiv.org/gr-qc/0010055`.

Consider a static observer at radius $r_1 > 2GM$ outside a Schwarzschild black hole. Such an observer moves along orbits of the timelike Killing vector $K = \partial_t$. In Chapter 6 we showed that the redshift factor $V = \sqrt{-K_\mu K^\mu}$ for static observers in Schwarzschild is given by

$$V = \sqrt{1 - \frac{2GM}{r}}, \tag{9.169}$$

with a corresponding magnitude of the acceleration given by

$$a = \frac{GM}{r\sqrt{r - 2GM}}. \tag{9.170}$$

For observers very close to the event horizon, $r_1 - 2GM \ll 2GM$, this acceleration becomes very large compared to the scale set by the Schwarzschild radius,

$$a_1 \gg \frac{1}{2GM}. \tag{9.171}$$

The Schwarzschild radius in turn sets the radius of curvature of spacetime near the horizon. Therefore, as observed over length- and timescales set by $a_1^{-1} \ll 2GM$, spacetime looks essentially flat. Let us make the crucial assumption that the quantum state of some scalar field ϕ looks like the Minkowski vacuum (free of any particles) as seen by *freely-falling* observers near the black hole. This assumption is reasonable, since the event horizon is not a local barrier; a freely-falling observer sees nothing special happen when crossing the horizon. Then the static observer looks just like a constant-acceleration observer in flat spacetime, and will detect Unruh radiation at a temperature $T_1 = a_1/2\pi$.

Now consider a static observer at infinity, or at least a distance r_2 large compared to $2GM$. In that case there is no sense in which the spacetime curvature can be neglected over timescales $a_2^{-1} \gg 2GM$, so there is no reason to expect that they will see radiation with a temperature $a_2/2\pi$, where a_2 is evaluated at r_2. But the radiation observed near the horizon will propagate to infinity with an appropriate redshift. We can apply the argument used at the end of the last section to determine what such an observer should see; they should detect thermal radiation redshifted to a temperature

$$T_2 = \frac{V_1}{V_2} T_1 = \frac{V_1}{V_2} \frac{a}{2\pi}. \tag{9.172}$$

At infinity we have $V_2 \to 1$, so the observed temperature is

$$T = \lim_{r_1 \to 2GM} \frac{V_1 a_1}{2\pi} = \frac{\kappa}{2\pi}, \tag{9.173}$$

where $\kappa = \lim(Va)$ is the surface gravity; for Schwarzschild, $\kappa = 1/4GM$. Unlike for accelerating observers in flat spacetime, in Schwarzschild the static Killing vector has finite norm at infinity, and the radiation near the horizon redshifts to a finite value rather than all the way to zero. Observers far from the black

hole thus see a flux of thermal radiation emitted from the black hole at a temperature proportional to its surface gravity. This is the celebrated **Hawking effect**, and the radiation itself is known as Hawking radiation.

Despite its slickness, there is nothing dishonest about this derivation of the Hawking effect. In particular, the relation to acceleration makes it clear why the temperature is proportional to the black hole surface gravity (which continues to hold for more general black holes, not only Schwarzschild). However, we need to be clear about the assumption we made that the vacuum state near the horizon looks nonsingular to freely-falling observers. In technical terms, the renormalized energy-momentum tensor is taken to be finite at the horizon, or equivalently, the two-point function obeys the Hadamard condition (9.123).

The meaning of this assumption becomes more clear by considering possible vacuum states in the maximally extended Schwarzschild geometry. Such states are not necessarily physically relevant to a realistic black hole formed by gravitational collapse, but the possibilities that arise in the idealized case carry instructive lessons for the real world. We will only describe the states, not specify them quantitatively or derive any of their properties; for more details see the references above.

In searching for a vacuum state, we might begin by looking for a state that is regular [in the Hadamard sense, (9.123)] throughout spacetime. For maximally extended Schwarzschild, such a state was found by Hartle and Hawking, so we call it the **Hartle–Hawking vacuum**; indeed, this is the unique vacuum state that is regular everywhere and invariant under the Schwarzschild Killing vector ∂_t representing time translations at infinity. In particular, recalling the conformal diagram of Schwarzschild shown in Figure 5.16, the Hartle–Hawking vacuum is regular on the past and future event horizons $H^{\pm}$ at $r = 2GM$, and also on past and future null infinity $\mathscr{I}^{\pm}$. From the consideration of static observers as outlined above, we should then expect that the Hartle–Hawking vacuum features thermal radiation being emitted from the black hole, and indeed this turns out to be true. However, a close examination of this state reveals that there is an equal flux of thermal radiation coming in from past null infinity ($\mathscr{I}^{-}$) toward the black hole; in other words, it represents a black hole in thermal equilibrium with its environment. This is not what we would use to model a realistic black hole in our universe. Another vacuum, more closely analogous to that of a black hole formed via gravitational collapse, is the **Unruh vacuum**, which is nonsingular on H^{+} (and therefore predicts outgoing Hawking radiation), but exhibits no incoming radiation from $\mathscr{I}^{-}$. The Unruh vacuum turns out to be singular on the past horizon H^{-} of Schwarzschild; this doesn't bother us if we are only using it as a model for realistic black holes, since a spacetime featuring gravitational collapse as in Figure 5.17 would not have a white hole or any past horizons. Finally, we might look for a vacuum state in which no particles come into the black hole, nor escape to infinity; in other words, vanishing flux at $\mathscr{I}^{\pm}$. There is such a state, called the **Boulware vacuum**. The existence of such a state seems to be in conflict with our argument for the Hawking effect from the Unruh effect, except that a careful analysis reveals that the Boulware vacuum is singular both on H^{-} and H^{+}. Thus,

the assumption that the vacuum is regular as seen by freely-falling observers near the horizon is violated in this state.

So a careful examination of vacuum states in an eternal Schwarzschild metric is consistent with our reasoning from the Unruh effect; states that are regular on H^+ predict Hawking radiation of the expected form. Note that the existence of an event horizon is crucial to the argument; without such an horizon, the requirement that the state be regular on the horizon has no force. Consider for example a neutron star, whose radius may be close to the Schwarzschild radius but for which the spacetime is free of any horizons. Neutron stars do not emit any Hawking radiation. One way to understand this is to recognize that a static neutron-star metric features a Killing vector that is timelike everywhere, and can be used to define positive-frequency modes that extend throughout the spacetime and match the Minkowski modes at infinity. The resulting vacuum state would actually resemble the Boulware vacuum, free of flux at $\mathscr{I}^{\pm}$; the fact that the full Boulware vacuum is singular on the horizon doesn't bother us in the neutron-star case, since there aren't any horizons.

To be absolutely sure that we have correctly chosen a vacuum state appropriate to realistic black holes, we should consider gravitational collapse in a spacetime that is nearly Minkowskian in the past and Schwarzschild in the future, as in Figure 5.17. If the vacuum takes the standard Minkowski form on $\mathscr{I}^-$, we can ask how the modes propagate through the collapse geometry to $\mathscr{I}^+$, defining an S-matrix as in (9.45) to determine what would be seen by asymptotic observers. This is in fact what Hawking did when he first discovered black hole radiation; the calculations involve some messy algebra but are basically straightforward, with the same answer for the temperature as we derived above.

Of course, from a complete calculation we can learn more than just the blackbody temperature; we might ask, for example, what happens when the wavelength of the emitted radiation is comparable to the Schwarzschild radius, in which case our approximations clearly break down. If we were to carefully investigate the emission of arbitrary species of particles from any kind of black hole (that is, allowing for both charge and spin), we would find that the spectrum of emitted radiation takes the form

$$\langle \hat{n}_\omega \rangle = \frac{\Gamma(\omega)}{e^{2\pi(\omega-\mu)/\kappa} \pm 1}. \tag{9.174}$$

Here, κ is of course the surface gravity. The parameter μ is a chemical potential, characterizing the tendency of the black hole to shed its conserved quantum numbers; a charged black hole preferentially emits particles with the same-sign charge as the hole, while a rotating black hole preferentially emits particles with the same-sign angular momentum as the hole. Hawking radiation therefore tends to bring black holes to a Schwarzschild state. $\Gamma(\omega)$ is a greybody factor, which can be thought of as arising from backscattering of wavepackets off of the gravitational field and into the black hole. In the high-frequency limit the wavelength is very small and backscattering can be neglected; at very low frequencies the wavelength becomes greater than the Schwarzschild radius and backscattering

becomes important. Although an analytic expression for the greybody factor is hard to derive, in the limiting cases of large and small frequencies the greybody factor for a scalar field obeys

$$\Gamma(\omega) \to 1, \qquad \omega \gg \frac{1}{GM}$$
$$\Gamma(\omega) \to \frac{A}{4\pi}\omega^2, \quad \omega \ll \frac{1}{GM}, \tag{9.175}$$

where A is the area of the black hole.

The discovery that black holes emit thermal radiation is certainly surprising from the point of view of classical general relativity, where we emphasized the impossibility of escape to infinity from points inside the event horizon. One picturesque way to understand what is going on is to think of vacuum fluctuations in terms of Feynman diagrams, with the fluctuations being represented by virtual particle/antiparticle pairs popping in and out of existence. This picture is also helpful, for example, in understanding observed phenomena such as the Lamb shift, in which atomic spectra are affected by the interaction of photons with virtual electron/positron pairs. Normally, the pairs will always annihilate, and their effect is only indirect, through a renormalization of processes coupled to the virtual particles. In the presence of an event horizon, however, occasionally one member of a virtual pair will fall into the black hole while its partner escapes to infinity, as depicted in Figure 9.2. In this picture, it is these escaping virtual particles that we observe as Hawking radiation. The total energy of the virtual pair must add to zero, but the infalling particle can have a negative energy as viewed from infinity, because the asymptotically-timelike Killing vector is spacelike inside the horizon. The picture is somewhat informal, but provides a useful heuristic for what is going on.

Once we know the formula for the temperature of a black hole we can fix the proportionality constants in the relationships between black hole parameters and

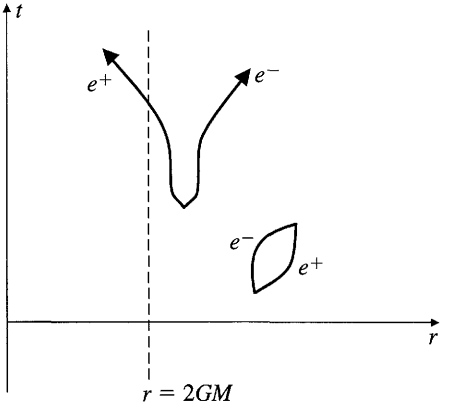

FIGURE 9.2 Vacuum fluctuations occasionally result in one of a particle/antiparticle pair falling into the event horizon, and the other escaping to infinity as Hawking radiation.

thermodynamic variables, as listed in (6.118). Hawking radiation essentially consummates the marriage of black hole mechanics and thermodynamics; stationary black holes act just like bodies of energy $E = M$ in thermal equilibrium with temperature $T = \kappa/2\pi$ and entropy $S = A/4G$. This is a very large entropy indeed. For matter fields in the universe, the entropy is approximately equal to the number of relativistic particles; within one Hubble radius, this number works out to be

$$S_{\mathrm{M}} \sim 10^{88}. \tag{9.176}$$

Meanwhile, the entropy of a black hole is the area of its horizon measured in Planck units (remember that we have been setting $\hbar = 1$ all along). We can convert to astrophysical units to obtain

$$S_{\mathrm{BH}} \sim 10^{90} \left(\frac{M}{10^6 M_\odot} \right)^2. \tag{9.177}$$

Thus, a single million-solar-mass black hole (such as can be found at the center of our galaxy, and many other galaxies) has more entropy than all of the matter in the visible universe. The total entropy of the universe is much smaller than we could make it, just by putting more mass into black holes. (When cosmologists say that the entropy S_{M} is large, they mean it is surprising that so much entropy is found within one curvature radius.) Presumably the reason why we are in such a low-entropy state has to do with initial conditions, and perhaps with inflation.

Coming back to black hole mechanics, we see a puzzle: The entropy of a macroscopic black hole will be huge, but from a statistical-mechanical point of view the entropy is supposed to measure the logarithm of the number of accessible states. A classical black hole is specified by a small number of parameters (mass, charge, and spin), so it is hard to know what those states could be. Nevertheless, we could take the attitude that this discrepancy doesn't really matter, since any information about the state of a black hole would presumably be hidden behind the event horizon.

The inclusion of quantum mechanics makes the puzzle worse rather than better, because black holes will not only radiate but also evaporate. When we started our investigation of QFT in curved spacetime, one of the rules we set was that we would assume a fixed background metric, and not worry about the effect of the energy-momentum tensor of the quantum fields themselves. Nevertheless, even in quantum mechanics we have conservation of energy (in the sense, for example, of a conserved ADM mass in an asymptotically flat spacetime). Hence, when Hawking radiation escapes to infinity, we may safely conclude that it will carry energy away from the black hole, which must therefore shrink in mass. (This phenomenon does not violate the area theorem, since the quantum field energy-momentum tensor will not obey the weak energy condition near the horizon.) As the mass shrinks, the surface gravity increases, and with it the temperature; there is a runaway process in which the entire mass evaporates away in a finite time.

Plugging in the numbers gives a lifetime of order

$$\tau_{\rm BH} \sim \left(\frac{M}{m_{\rm P}}\right)^3 t_{\rm P} \sim \left(\frac{M}{M_\odot}\right)^3 \times 10^{71}\ {\rm sec}, \tag{9.178}$$

where $m_{\rm P} \sim 10^{-5}$ g is the Planck mass and $t_{\rm P} \sim 10^{-43}$ sec is the Planck time. Since the Hubble time is $H_0^{-1} \sim 10^{18}$ sec, a solar-mass black hole has a lifetime of order 10^{53} times the age of the universe. This seems like a long time, but we are speaking of questions of principle here.

You can see why the question of the black hole entropy has become so severe: Once the black hole has evaporated, we can no longer appeal to the event horizon as a way to hide purported states of the black hole. There is no black hole any more, just the Hawking radiation it produced. The fact that this radiation is supposed to be precisely thermal (no hidden correlations in the outgoing particles) means that it has no way of conveying the vast amount of information needed to specify the states implied by our entropy calculation. Thus, if we assemble two very different original states and collapse them into two black holes of the same mass, charge, and spin, they will radiate away into two indistinguishable clouds of Hawking particles. The information that went into the specification of the system before it became a black hole seems to have been erased; this is the **information loss paradox**. Both quantum field theory and general relativity feature unitary evolution—the information required to specify a state at early times is precisely equal to that needed to specify a state at later times, since they are connected by the equations of motion. But in the process of combining QFT with GR this unitarity has apparently been violated. It seems likely that we have made an inappropriate assumption somewhere in our argument, but it is hard to see where.

One way of conveying the essence of the information loss paradox is to consider a hypothetical conformal diagram for an evaporating black hole, shown in Figure 9.3. We don't really know what the full spacetime should look like, but here we have made the plausible assumptions that a singularity forms, along with an associated event horizon, both of which disappear when the black hole has fully evaporated, leaving behind a spacetime with a Minkowskian causal structure. The problem is then obvious if we think in terms of Cauchy surfaces. The future domain of dependence of an achronal surface stretching from spacelike infinity i^0 to a point with $r = 0$ to the past of the singularity would be the entire spacetime, so such a surface would be a Cauchy surface. But a similar surface stretching to a point with $r = 0$ to the future of the singularity would not be a Cauchy surface, since the region behind the event horizon would not be in its domain of dependence. Thus, the past cannot be retrodicted from the future, due to the disappearance of information into the singularity. In other words, this process seems to be time-irreversible (in a microscopic sense, not merely a statistical sense), even though the dynamical laws that were used to predict it were fully invariant under time reversal.

In addressing the information loss paradox, keep in mind that our analysis of black-hole evaporation has only been in the context of a hybrid theory of quantum

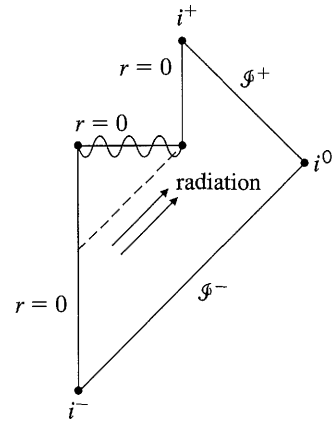

FIGURE 9.3 Hypothetical conformal diagram for an evaporating black hole. Energy is carried away by the Hawking radiation, so that the black hole eventually evaporates away entirely, leaving a future with the causal structure of Minkowski space. Information that falls past the event horizon into the singularity appears to be lost.

field theory coupled to general relativity, not in a realistic theory of quantum gravity. What might be going on in the real world? One possibility is that information really is lost, unitarity is violated, and we just have to learn to live with it. Many physicists find the introduction of such a fundamental breakdown of predictability to be unpalatable, and arguments have also been made that unitarity violations would necessarily lead to violations of energy conservation. Another possibility is that unitarity appears to be violated in our world, but only because the information that entered the black hole has somehow escaped to a disconnected region of space (a baby universe). General relativity predics a singularity at the center of the black hole, not creation of a disconnected region, but clearly we are in a regime where quantum effects will dramatically alter our classical expectations, so we should keep an open mind.

Some evidence against information loss comes from string theory. String theory is naturally defined in 10 or 11 spacetime dimensions, and features not only one-dimensional extended objects (strings), but also various types of higher-dimensional extended objects known collectively as "branes." A crucial aspect of string theory is a high degree of supersymmetry relating bosons to fermions. In the real world supersymmetry must be spontaneously broken if it exists at all, since we don't observe a bosonic version of the electron with the same mass and charge. But as a tool for thought experiments, supersymmetry is invaluable. Supersymmetric configurations of strings and branes can be assembled that describe black hole geometries in various dimensions. In string theory there is a free parameter (really a scalar field), the string coupling, that controls the strength of gravity as well as the strength of other forces. If we consider a configuration describing a black hole at a certain value of the string coupling, as we decrease the coupling the Schwarzschild radius will eventually shrink below the size of

the configuration, which thus turns into a collection of weakly-coupled strings and branes. Due to the high degree of supersymmetry, we can be confident that various characteristics of the state remain unchanged as we vary the string coupling; in particular, we expect that the number of degrees of freedom (and thus the entropy) is unaltered. But in the weakly-coupled regime there is no black hole, we simply have a "gas" of conventional degrees of freedom (admittedly, of extended objects in higher dimensions), whose entropy we should be able to reliably calculate.

Strominger and Vafa considered this process for a particular type of five-dimensional supersymmetric black hole with different kinds of charges.[4] They found a remarkable result: the number of degrees of freedom of the system at weak coupling matches precisely that which would be predicted based on the entropy of the black hole at strong coupling. Since the black hole entropy depends nontrivially on the charges of the configuration, it seems unlikely that this agreement is simply an accident. Subsequent investigations have extended this analysis to other kinds of black holes, for which agreement continues to be found. Furthermore, we can even calculate the greybody factors expected for the black hole by considering scattering off of the weakly-coupled system; again, the result matches the strong-coupling expectation. Thus, in string theory at least, there is excellent reason to believe that the degrees of freedom implied by black hole radiation are really there.

Unfortunately, the string theory counting of states provides little direct understanding of how information about the black hole state could somehow be conveyed to the outgoing Hawking radiation. Nevertheless, we should certainly take seriously the possibility that this is what happens, even if there are severe difficulties in imagining how such a process might actually work. The difficulties arise when considering some information, perhaps in the form of a volume of an encyclopedia, being tossed into a large black hole, long before it has evaporated away. At this stage the black hole temperature is low, there is very little surface gravity, and the spacetime curvature near the event horizon is quite small. From the point of view of the encyclopedia, nothing special happens at the horizon, and we should expect it to fall through essentially unmolested. In particular, it is hard to imagine how the information in the encyclopedia can be transferred to the Hawking radiation being emitted at early times. In unitary evolution, the information cannot be duplicated; either it falls past the horizon with the encyclopedia, or it needs to be effectively extracted just before the horizon is crossed, which seems implausible. We might hope that the information accompanies the encyclopedia into a region near the singularity, and is somehow preserved there until late times when the hole is very small. But by then most of the radiating particles have already been emitted, and the number of states accessible to the final burst of radiation will generally be smaller than required to describe the different states that could have fallen into the hole.

[4] A. Strominger and C. Vafa, "Microscopic origin of the Bekenstein-Hawking entropy," *Phys. Lett.* B **379**, 99 (1996), `http://arxiv.org/hep-th/9601029`. For reviews see Johnson (2003) or A.W. Peet, "TASI lectures on black holes in string theory," (2001), `http://arxiv.org/hep-th/0008241`.

To imagine that the information is somehow encoded in the outgoing radiation, it therefore seems necessary to encode correlations in the Hawking particles even at early times. We just argued that this is hard to do, given that the horizon is an unremarkable place when the black hole is large. One conceivable way out of this dilemma is to take the dramatic step of giving up on local quantum field theory. In other words, we have been making the implicit assumption that information can be sensibly described as being located in some region of space; this is an indisputable feature of ordinary quantum field theories. But perhaps quantum gravity is different, and the information contained in the black hole is somehow spread out nonlocally across the horizon. By itself this suggestion doesn't lead directly to a mechanism for getting the information into the outgoing Hawking radiation, but it does call into question some of the arguments we have given for why it would be difficult to do so.

A particular realization of nonlocality goes under the name of the **holographic principle**. This is the idea, suggested originally by 't Hooft and Susskind, that the number of degrees of freedom in a region of space is not proportional to the volume of the region (as would be expected in a local field theory), but rather to the area of the boundary of the region.[5] The inspiration comes of course from black hole entropy, which scales as the area of the event horizon; if the entropy counts the number of accessible states, holography would account for why it is the area rather than the enclosed volume that matters. You might worry about how to deal with closed universes, in which a region might consist of almost all of space but have a very small boundary, but a more covariant version of the holographic principle may be formulated by replacing the region of space by a set of "light-sheets" extending inward from the boundary. The great triumph of holography has been in the AdS/CFT correspondence, mentioned in Chapter 8. There, the physics of quantum gravity in an anti-de Sitter background is equivalent to a conformal field theory without gravity defined on the boundary of AdS, which has one lower dimension. One can imagine that all of the physical phenomena we observe in the universe could be described by the nonlocal holographic projection of some ordinary nongravitational theory defined in lower dimensions; it is by no means clear how we should go about constructing such a correspondence or connecting it with observations, but considerations of cosmology and the large-scale structure of the universe might be a promising place to start.

These remarks about black hole entropy, string theory, and holography are obviously not intended as a careful introduction to what is a very active area of research. Rather, they are meant to indicate some of the possibilities being explored at the forefront of gravitational physics. Classical general relativity is the most beautiful physical theory invented to date, but we have every right to expect that a synthesis of GR with other areas of physics will reveal layers of beauty we can only now imagine.

[5]For a review see R. Bousso, "The Holographic Principle" (2002), `http://arxiv.org/hep-th/0203101`.

APPENDIX

A Maps between Manifolds

When we discussed manifolds in Chapter 2, we introduced maps between two different manifolds and how maps could be composed. Here we will investigate such maps in much greater detail, focusing on the use of such maps in carrying along tensor fields from one manifold to another. The manifolds in question might end up being a submanifold and the bigger space in which it is embedded, or we might just have two different copies of the same abstract manifold being mapped to each other.

Consider two manifolds M and N, possibly of different dimension, with coordinate systems x^μ and y^α, respectively. We imagine that we have a map $\phi : M \to N$ and a function $f : N \to \mathbf{R}$. Obviously we can compose ϕ with f to construct a map $(f \circ \phi) : M \to \mathbf{R}$, which is simply a function on M. Such a construction is sufficiently useful that it gets its own name; we define the **pullback** of f by ϕ, denoted $\phi^* f$, by

$$\phi^* f = (f \circ \phi). \tag{A.1}$$

The name makes sense, since we think of ϕ^* as "pulling back" the function f from N to M (see Figure A.1).

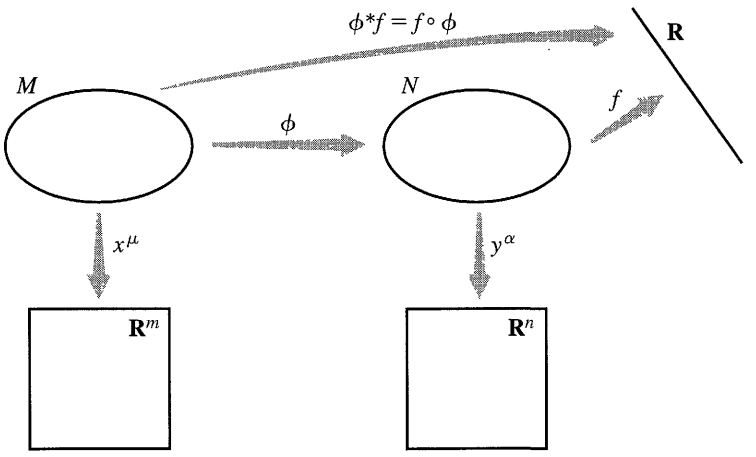

FIGURE A.1 The pullback of a function f from N to M by a map $\phi : M \to N$ is simply the composition of ϕ with f.

We can pull functions back, but we cannot push them forward. If we have a function $g : M \to \mathbf{R}$, there is no way we can compose g with ϕ to create a function on N; the arrows don't fit together correctly. But recall that a vector can be thought of as a derivative operator that maps smooth functions to real numbers. This allows us to define the **pushforward** of a vector; if $V(p)$ is a vector at a point p on M, we define the pushforward vector $\phi_* V$ at the point $\phi(p)$ on N by giving its action on functions on N:

$$(\phi_* V)(f) = V(\phi^* f). \tag{A.2}$$

So to push forward a vector field we say "the action of $\phi_* V$ on any function is simply the action of V on the pullback of that function."[1]

This discussion is a little abstract, and it would be nice to have a more concrete description. We know that a basis for vectors on M is given by the set of partial derivatives $\partial_\mu = \partial/\partial x^\mu$, and a basis on N is given by the set of partial derivatives $\partial_\alpha = \partial/\partial y^\alpha$. Therefore we would like to relate the components of $V = V^\mu \partial_\mu$ to those of $(\phi_* V) = (\phi_* V)^\alpha \partial_\alpha$. We can find the sought-after relation by applying the pushed-forward vector to a test function and using the chain rule (2.12):

$$\begin{aligned}(\phi_* V)^\alpha \partial_\alpha f &= V^\mu \partial_\mu (\phi^* f) \\ &= V^\mu \partial_\mu (f \circ \phi) \\ &= V^\mu \frac{\partial y^\alpha}{\partial x^\mu} \partial_\alpha f.\end{aligned} \tag{A.3}$$

This simple formula makes it irresistible to think of the pushforward operation ϕ_* as a matrix operator, $(\phi_* V)^\alpha = (\phi_*)^\alpha{}_\mu V^\mu$, with the matrix being given by

$$(\phi_*)^\alpha{}_\mu = \frac{\partial y^\alpha}{\partial x^\mu}. \tag{A.4}$$

The behavior of a vector under a pushforward thus bears an unmistakable resemblance to the vector transformation law under change of coordinates. In fact it is a generalization, since when M and N are the same manifold the constructions are (as we shall discuss) identical; but don't be fooled, since in general μ and α have different allowed values, and there is no reason for the matrix $\partial y^\alpha/\partial x^\mu$ to be invertible.

It is a rewarding exercise to convince yourself that, although you can push vectors forward from M to N (given a map $\phi : M \to N$), you cannot in general pull them back—just keep trying to invent an appropriate construction until the futility of the attempt becomes clear. Since one-forms are dual to vectors, you should not be surprised to hear that one-forms can be pulled back (but not in general pushed forward). To do this, remember that one-forms are linear maps from vectors to the real numbers. The pullback $\phi^* \omega$ of a one-form ω on N can

[1] Unfortunately the location of the asterisks is not completely standard; some references use a superscript * for pushforward and a subscript * for pullback, so be careful.

therefore be defined by its action on a vector V on M, by equating it with the action of ω itself on the pushforward of V:

$$(\phi^*\omega)(V) = \omega(\phi_* V). \tag{A.5}$$

Once again, there is a simple matrix description of the pullback operator on forms, $(\phi^*\omega)_\mu = (\phi^*)_\mu{}^\alpha \omega_\alpha$, which we can derive using the chain rule. It is given by

$$(\phi^*)_\mu{}^\alpha = \frac{\partial y^\alpha}{\partial x^\mu}. \tag{A.6}$$

That is, it is the same matrix as the pushforward (A.4), but of course a different index is contracted when the matrix acts to pull back one-forms.

There is a way of thinking about why pullbacks and pushforwards work on some objects but not others, which may be helpful. If we denote the set of smooth functions on M by $\mathcal{F}(M)$, then a vector $V(p)$ at a point p on M (that is, an element of the tangent space T_pM) can be thought of as an operator from $\mathcal{F}(M)$ to $\mathbf{R}$. But we already know that the pullback operator on functions maps $\mathcal{F}(N)$ to $\mathcal{F}(M)$, just as ϕ itself maps M to N, but in the opposite direction. Therefore we can define the pushforward ϕ_* acting on vectors simply by composing maps, as we first defined the pullback of functions; this is shown in Figure A.2. Similarly, if T_qN is the tangent space at a point q on N, then a one-form ω at q (that is, an element of the cotangent space T_q^*N) can be thought of as an operator from T_qN to $\mathbf{R}$. Since the pushforward ϕ_* maps T_pM to $T_{\phi(p)}N$, the pullback ϕ^* of a one-form can also be thought of as mere composition of maps, as indicated in Figure A.3. If this is not helpful, don't worry about it. But do keep straight what exists and what doesn't; the actual concepts are simple, it's just forgetting which map goes what way that leads to confusion.

You will recall further that a $(0, l)$ tensor—one with l lower indices and no upper ones—is a linear map from the direct product of l vectors to $\mathbf{R}$. We can therefore pull back not only one-forms, but tensors with an arbitrary number of lower indices. The definition is simply the action of the original tensor on the pushed-forward vectors:

$$(\phi^*T)(V^{(1)}, V^{(2)}, \ldots, V^{(l)}) = T(\phi_* V^{(1)}, \phi_* V^{(2)}, \ldots, \phi_* V^{(l)}), \tag{A.7}$$

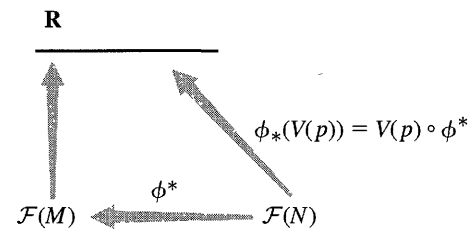

FIGURE A.2 Pushing forward a vector, thought of as composition of a map between the spaces of functions on N and M, and a map from functions on M to $\mathbf{R}$.

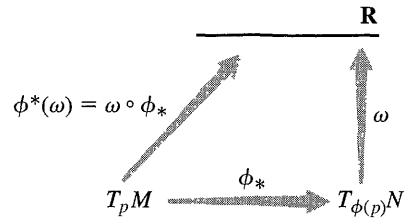

FIGURE A.3 Pulling back a one-form, thought of as composition of a map between tangent spaces T_pM and $T_{\phi(p)}N$ and a map from $T_{\phi p}N$ to **R**.

where $T_{\alpha_1\cdots\alpha_l}$ is a $(0, l)$ tensor on N. We can similarly push forward any $(k, 0)$ tensor $S^{\mu_1\cdots\mu_k}$ by acting it on pulled-back one-forms:

$$(\phi_* S)(\omega^{(1)}, \omega^{(2)}, \ldots, \omega^{(k)}) = S(\phi^*\omega^{(1)}, \phi^*\omega^{(2)}, \ldots, \phi^*\omega^{(k)}). \tag{A.8}$$

Fortunately, the matrix representations of the pushforward (A.4) and pullback (A.6) extend to the higher-rank tensors simply by assigning one matrix to each index; thus, for the pullback of a $(0, l)$ tensor, we have

$$\boxed{(\phi^* T)_{\mu_1\cdots\mu_l} = \frac{\partial y^{\alpha_1}}{\partial x^{\mu_1}} \cdots \frac{\partial y^{\alpha_l}}{\partial x^{\mu_l}} T_{\alpha_1\cdots\alpha_l},} \tag{A.9}$$

while for the pushforward of a $(k, 0)$ tensor we have

$$\boxed{(\phi_* S)^{\alpha_1\cdots\alpha_k} = \frac{\partial y^{\alpha_1}}{\partial x^{\mu_1}} \cdots \frac{\partial y^{\alpha_k}}{\partial x^{\mu_k}} S^{\mu_1\cdots\mu_k}.} \tag{A.10}$$

Our complete picture is therefore as portrayed in Figure A.4. Note that tensors with both upper and lower indices can generally be neither pushed forward nor pulled back.

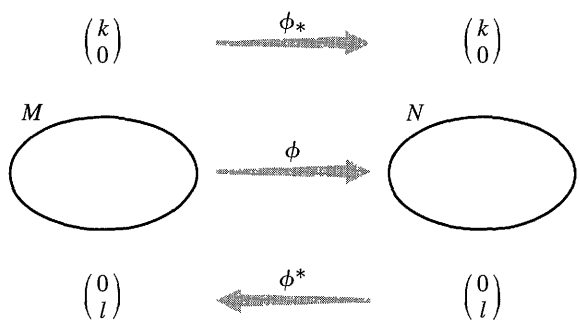

FIGURE A.4 A map $\phi : M \to N$ allows us to pull back $(0, l)$ tensors and push forward $(k, 0)$ tensors.

This machinery becomes somewhat less imposing once we see it at work in a simple example. One common occurrence of a map between two manifolds is when M is actually a submanifold of N, which we will discuss more carefully in Appendix C. The basic idea is that there is a map from M to N that just takes an element of M to the "same" element of N. Consider the two-sphere embedded in $\mathbf{R}^3$, thought of as the locus of points a unit distance from the origin. If we put coordinates $x^\mu = (\theta, \phi)$ on $M = S^2$ and $y^\alpha = (x, y, z)$ on $N = \mathbf{R}^3$, the map $\phi : M \to N$ is given by

$$\phi(\theta, \phi) = (\sin\theta \cos\phi, \sin\theta \sin\phi, \cos\theta). \tag{A.11}$$

Sticking the sphere into $\mathbf{R}^3$ in this way induces a metric on S^2, which is just the pullback of the flat-space metric. The simple-minded way to find this is to start with the metric $ds^2 = dx^2 + dy^2 + dz^2$ on $\mathbf{R}^3$ and substitute (A.11) into this expression, yielding a metric $d\theta^2 + \sin^2\theta\, d\phi^2$ on S^2. Let's see how this answer comes about using the more respectable formalism. (Of course it would be easier if we worked in spherical coordinates on $\mathbf{R}^3$, but doing it the hard way is more illustrative.) The matrix of partial derivatives is given by

$$\frac{\partial y^\alpha}{\partial x^\mu} = \begin{pmatrix} \cos\theta\cos\phi & \cos\theta\sin\phi & -\sin\theta \\ -\sin\theta\sin\phi & \sin\theta\cos\phi & 0 \end{pmatrix}. \tag{A.12}$$

The metric on S^2 is obtained by simply pulling back the metric from $\mathbf{R}^3$,

$$\begin{aligned} (\phi^* g)_{\mu\nu} &= \frac{\partial y^\alpha}{\partial x^\mu}\frac{\partial y^\beta}{\partial x^\nu} g_{\alpha\beta} \\ &= \begin{pmatrix} 1 & 0 \\ 0 & \sin^2\theta \end{pmatrix}, \end{aligned} \tag{A.13}$$

as you can easily check. So the answer really is the same as you would get by naive substitution, but now we know why.

APPENDIX

B Diffeomorphisms and Lie Derivatives

In this Appendix we continue the explorations of the previous one, now focusing on the special case when the two manifolds are actually the same. Thus far, we have been careful to emphasize that a map $\phi : M \to N$ can be used to pull certain things back (A.9) and push other things forward (A.10). The reason why it generally doesn't work both ways can be traced to the fact that ϕ might not be invertible. If ϕ is invertible (and both ϕ and ϕ^{-1} are smooth, which we always implicitly assume), then it defines a diffeomorphism between M and N. This can only be the case if M and N are actually the same abstract manifold; indeed, the existence of a diffeomorphism is the definition of two manifolds being the same. The beauty of diffeomorphisms is that we can use both ϕ and ϕ^{-1} to move tensors from M to N; this will allow us to define the pushforward and pullback of arbitrary tensors. Specifically, for a (k, l) tensor field $T^{\mu_1 \cdots \mu_k}{}_{\nu_1 \cdots \nu_l}$ on M, we define the pushforward by

$$\begin{aligned}(\phi_* T)(\omega^{(1)}, \ldots, \omega^{(k)}, V^{(1)}, \ldots, V^{(l)}) \\ = T(\phi^* \omega^{(1)}, \ldots, \phi^* \omega^{(k)}, [\phi^{-1}]_* V^{(1)}, \ldots, [\phi^{-1}]_* V^{(l)}),\end{aligned} \tag{B.1}$$

where the $\omega^{(i)}$'s are one-forms on N and the $V^{(i)}$'s are vectors on N. In components this becomes

$$(\phi_* T)^{\alpha_1 \cdots \alpha_k}{}_{\beta_1 \cdots \beta_l} = \frac{\partial y^{\alpha_1}}{\partial x^{\mu_1}} \cdots \frac{\partial y^{\alpha_k}}{\partial x^{\mu_k}} \frac{\partial x^{\nu_1}}{\partial y^{\beta_1}} \cdots \frac{\partial x^{\nu_l}}{\partial y^{\beta_l}} T^{\mu_1 \cdots \mu_k}{}_{\nu_1 \cdots \nu_l}. \tag{B.2}$$

The appearance of the inverse matrix $\partial x^\nu / \partial y^\beta$ is legitimate because ϕ is invertible. Note that we could also define the pullback in the obvious way, but there is no need to write separate equations because the pullback ϕ^* is the same as the pushforward via the inverse map, $[\phi^{-1}]_*$.

We are now in a position to explain the relationship between diffeomorphisms and coordinate transformations: they are two different ways of doing precisely the same thing. If you like, diffeomorphisms are "active coordinate transformations," while traditional coordinate transformations are "passive." Consider an n-dimensional manifold M with coordinate functions $x^\mu : M \to \mathbf{R}^n$. To change coordinates we can either simply introduce new functions $y^\mu : M \to \mathbf{R}^n$ ("keep the manifold fixed, change the coordinate maps"), or we could just as well introduce a diffeomorphism $\phi : M \to M$, after which the coordinates would just be the pullbacks $(\phi^* x)^\mu : M \to \mathbf{R}^n$ ("move the points on the manifold, and then

evaluate the coordinates of the new points"), as shown in Figure B.1. In this sense, (B.2) really is the tensor transformation law, just thought of from a different point of view.

Since a diffeomorphism allows us to pull back and push forward arbitrary tensors, it provides another way of comparing tensors at different points on a manifold. Given a diffeomorphism $\phi : M \to M$ and a tensor field $T^{\mu_1 \cdots \mu_k}{}_{\nu_1 \cdots \nu_l}(x)$, we can form the difference between the value of the tensor at some point p and $\phi^*[T^{\mu_1 \cdots \mu_k}{}_{\nu_1 \cdots \nu_l}(\phi(p))]$, its value at $\phi(p)$ pulled back to p. This suggests that we could define another kind of derivative operator on tensor fields, one that categorizes the rate of change of the tensor under the flow of the diffeomorphism. For that, however, a single discrete diffeomorphism is insufficient; we require a one-parameter family of diffeomorphisms, ϕ_t. This family can be thought of as a smooth map $\mathbf{R} \times M \to M$, such that for each $t \in \mathbf{R}$ we have a diffeomorphism ϕ_t, satisfying $\phi_s \circ \phi_t = \phi_{s+t}$. This last condition implies that ϕ_0 is the identity map.

One-parameter families of diffeomorphisms can be thought of as arising from vector fields (and vice-versa). If we consider what happens to the point p under the entire family ϕ_t, it is clear that it describes a curve in M; since the same thing will be true of every point on M, these curves fill the manifold (although there can be degeneracies where the diffeomorphisms have fixed points). We can define a vector field $V^\mu(x)$ to be the set of tangent vectors to each of these curves at every point, evaluated at $t = 0$. An example on S^2 is provided by the diffeomorphism $\phi_t(\theta, \phi) = (\theta, \phi + t)$, shown in Figure B.2. We can reverse the construction to define a one-parameter family of diffeomorphisms from any vector field. Given a vector field $V^\mu(x)$, we define the **integral curves** of the vector field to be those curves $x^\mu(t)$ that solve

$$\frac{dx^\mu}{dt} = V^\mu. \tag{B.3}$$

Note that this familiar-looking equation is now to be interpreted in the opposite sense from our usual way; we are given the vectors, from which we define the curves. Solutions to (B.3) are guaranteed to exist as long as we don't do anything silly like run into the edge of our manifold; the proof amounts to finding a coordinate system in which the problem reduces to the fundamental theorem

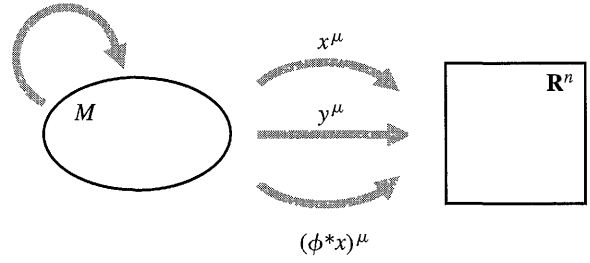

FIGURE B.1 A coordinate change induced by the diffeomorphism $\phi : M \to M$.

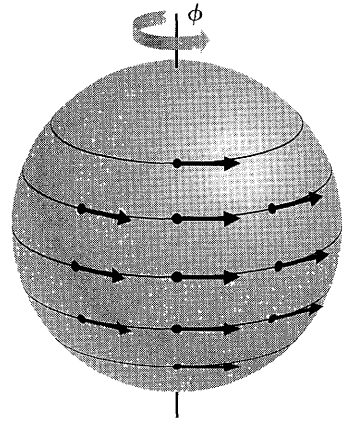

FIGURE B.2 A diffeomorphism on the two-sphere, given by a rotation about its axis.

of ordinary differential equations. Our diffeomorphisms ϕ_t represent "flow down the integral curves," and the associated vector field is referred to as the **generator** of the diffeomorphism. (Confusingly, vector fields and their integral curves also appear in the context of null hypersurfaces, where it is the curves rather than the vector fields that are called "generators.") Integral curves are used all the time in elementary physics, just not given the name. The "lines of magnetic flux" traced out by iron filings in the presence of a magnet are simply the integral curves of the magnetic field vector **B**.

Given a vector field $V^\mu(x)$, then, we have a family of diffeomorphisms parameterized by t, and we can ask how fast a tensor changes as we travel down the integral curves. For each t we can define this change as the difference between the pullback of the tensor to p and its original value at p,

$$\Delta_t T^{\mu_1\cdots\mu_k}{}_{\nu_1\cdots\nu_l}(p) = \phi_t^*[T^{\mu_1\cdots\mu_k}{}_{\nu_1\cdots\nu_l}(\phi_t(p))] - T^{\mu_1\cdots\mu_k}{}_{\nu_1\cdots\nu_l}(p). \qquad \text{(B.4)}$$

Note that both terms on the right-hand side are tensors at p, as shown in Figure B.3. We then define the **Lie derivative** of the tensor along the vector field as

$$\mathcal{L}_V T^{\mu_1\cdots\mu_k}{}_{\nu_1\cdots\nu_l} = \lim_{t\to 0}\left(\frac{\Delta_t T^{\mu_1\cdots\mu_k}{}_{\nu_1\cdots\nu_l}}{t}\right). \qquad \text{(B.5)}$$

The Lie derivative is a map from (k, l) tensor fields to (k, l) tensor fields, which is manifestly independent of coordinates. Since the definition essentially amounts to the conventional definition of an ordinary derivative applied to the component functions of the tensor, it should be clear that it is linear,

$$\mathcal{L}_V(aT + bS) = a\mathcal{L}_V T + b\mathcal{L}_V S, \qquad \text{(B.6)}$$

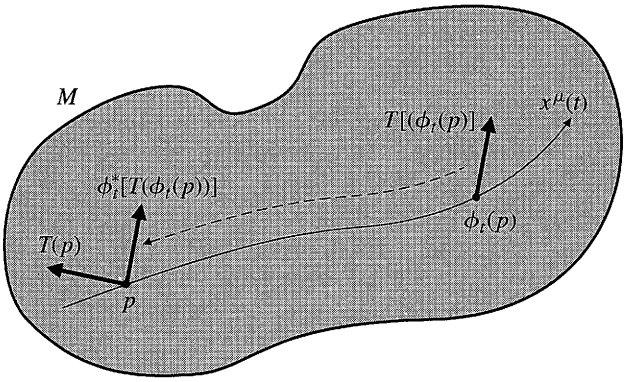

FIGURE B.3 The rate of change of a tensor along the integral curves of a vector field is computed by comparing the original tensor $T(p)$ at a point p to the value of T a point $\phi_t(p)$ by pulling $T(\phi_t(p))$ back to p.

and obeys the Leibniz rule,

$$\mathcal{L}_V(T \otimes S) = (\mathcal{L}_V T) \otimes S + T \otimes (\mathcal{L}_V S), \tag{B.7}$$

where S and T are tensors and a and b are constants. The Lie derivative is in fact a more primitive notion than the covariant derivative, since it does not require specification of a connection (although it does require a vector field, of course). A moment's reflection will convince you that it reduces to the ordinary directional derivative on functions,

$$\mathcal{L}_V f = V(f) = V^\mu \partial_\mu f. \tag{B.8}$$

To discuss the action of the Lie derivative on tensors in terms of other operations we know, it is convenient to choose a coordinate system adapted to our problem. Specifically, we will work in coordinates $x^\mu = (x^1, \ldots x^n)$, such that x^1 is the parameter along the integral curves and the other coordinates are chosen any way we like. Then the vector field takes the form $V = \partial/\partial x^1$; that is, it has components $V^\mu = (1, 0, 0, \ldots, 0)$. The magic of this coordinate system is that a diffeomorphism by t amounts to a coordinate transformation from x^μ to $y^\mu = (x^1 + t, x^2, \ldots, x^n)$. Thus, from (A.6) the pullback matrix is simply

$$(\phi_{t*})_\mu{}^\nu = \delta_\mu^\nu, \tag{B.9}$$

and the components of the tensor pulled back from $\phi_t(p)$ to p are simply

$$\phi_{t*}[T^{\mu_1 \cdots \mu_k}{}_{\nu_1 \cdots \nu_l}(\phi_t(p))] = T^{\mu_1 \cdots \mu_k}{}_{\nu_1 \cdots \nu_l}(x^1 + t, x^2, \ldots, x^n). \tag{B.10}$$

In this coordinate system, then, the Lie derivative becomes

$$\mathcal{L}_V T^{\mu_1 \cdots \mu_k}{}_{\nu_1 \cdots \nu_l} = \frac{\partial}{\partial x^1} T^{\mu_1 \cdots \mu_k}{}_{\nu_1 \cdots \nu_l}, \tag{B.11}$$

and in particular the derivative of a vector field $U^\mu(x)$ is

$$\mathcal{L}_V U^\mu = \frac{\partial U^\mu}{\partial x^1}. \tag{B.12}$$

Although this expression is clearly not covariant, we know that the commutator $[V, U]$ is a well-defined tensor, and in this coordinate system

$$\begin{aligned} [V, U]^\mu &= V^\nu \partial_\nu U^\mu - U^\nu \partial_\nu V^\mu \\ &= \frac{\partial U^\mu}{\partial x^1}. \end{aligned} \tag{B.13}$$

Therefore the Lie derivative of U with respect to V has the same components in this coordinate system as the commutator of V and U; but since both are vectors, they must be equal in any coordinate system:

$$\mathcal{L}_V U^\mu = [V, U]^\mu. \tag{B.14}$$

As an immediate consequence, we have $\mathcal{L}_V U = -\mathcal{L}_U V$. It is because of (B.14) that the commutator is sometimes called the **Lie bracket.**

To derive the action of $\mathcal{L}_V$ on a one-form ω_μ, begin by considering the action on the scalar $\omega_\mu U^\mu$ for an arbitrary vector field U^μ. First use the fact that the Lie derivative with respect to a vector field reduces to the action of the vector itself when applied to a scalar:

$$\begin{aligned} \mathcal{L}_V(\omega_\mu U^\mu) &= V(\omega_\mu U^\mu) \\ &= V^\nu \partial_\nu (\omega_\mu U^\mu) \\ &= V^\nu (\partial_\nu \omega_\mu) U^\mu + V^\nu \omega_\mu (\partial_\nu U^\mu). \end{aligned} \tag{B.15}$$

Then use the Leibniz rule on the original scalar:

$$\begin{aligned} \mathcal{L}_V(\omega_\mu U^\mu) &= (\mathcal{L}_V \omega)_\mu U^\mu + \omega_\mu (\mathcal{L}_V U)^\mu \\ &= (\mathcal{L}_V \omega)_\mu U^\mu + \omega_\mu V^\nu \partial_\nu U^\mu - \omega_\mu U^\nu \partial_\nu V^\mu. \end{aligned} \tag{B.16}$$

Setting these expressions equal to each other and requiring that equality hold for arbitrary U^μ, we see that

$$\mathcal{L}_V \omega_\mu = V^\nu \partial_\nu \omega_\mu + (\partial_\mu V^\nu) \omega_\nu, \tag{B.17}$$

which (like the definition of the commutator) is completely covariant, although not manifestly so.

By a similar procedure we can define the Lie derivative of an arbitrary tensor field. The answer can be written

$$\begin{aligned}\mathcal{L}_V T^{\mu_1\mu_2\cdots\mu_k}{}_{\nu_1\nu_2\cdots\nu_l} &= V^\sigma \partial_\sigma T^{\mu_1\mu_2\cdots\mu_k}{}_{\nu_1\nu_2\cdots\nu_l} \\ &\quad - (\partial_\lambda V^{\mu_1}) T^{\lambda\mu_2\cdots\mu_k}{}_{\nu_1\nu_2\cdots\nu_l} \\ &\quad - (\partial_\lambda V^{\mu_2}) T^{\mu_1\lambda\cdots\mu_k}{}_{\nu_1\nu_2\cdots\nu_l} - \cdots \\ &\quad + (\partial_{\nu_1} V^\lambda) T^{\mu_1\mu_2\cdots\mu_k}{}_{\lambda\nu_2\cdots\nu_l} \\ &\quad + (\partial_{\nu_2} V^\lambda) T^{\mu_1\mu_2\cdots\mu_k}{}_{\nu_1\lambda\cdots\nu_l} + \cdots .\end{aligned} \tag{B.18}$$

Once again, this expression is covariant, despite appearances. It would undoubtedly be comforting, however, to have an equivalent expression that looked manifestly tensorial. In fact it turns out that we can write

$$\begin{aligned}\mathcal{L}_V T^{\mu_1\mu_2\cdots\mu_k}{}_{\nu_1\nu_2\cdots\nu_l} &= V^\sigma \nabla_\sigma T^{\mu_1\mu_2\cdots\mu_k}{}_{\nu_1\nu_2\cdots\nu_l} \\ &\quad - (\nabla_\lambda V^{\mu_1}) T^{\lambda\mu_2\cdots\mu_k}{}_{\nu_1\nu_2\cdots\nu_l} \\ &\quad - (\nabla_\lambda V^{\mu_2}) T^{\mu_1\lambda\cdots\mu_k}{}_{\nu_1\nu_2\cdots\nu_l} - \cdots \\ &\quad + (\nabla_{\nu_1} V^\lambda) T^{\mu_1\mu_2\cdots\mu_k}{}_{\lambda\nu_2\cdots\nu_l} \\ &\quad + (\nabla_{\nu_2} V^\lambda) T^{\mu_1\mu_2\cdots\mu_k}{}_{\nu_1\lambda\cdots\nu_l} + \cdots ,\end{aligned} \tag{B.19}$$

where ∇_μ represents *any* symmetric (torsion-free) covariant derivative (including, of course, one derived from a metric). You can check that all of the terms that would involve connection coefficients if we were to expand (B.19) would cancel, leaving only (B.18). Both versions of the formula for a Lie derivative are useful at different times. A particularly useful formula is for the Lie derivative of the metric:

$$\begin{aligned}\mathcal{L}_V g_{\mu\nu} &= V^\sigma \nabla_\sigma g_{\mu\nu} + (\nabla_\mu V^\lambda) g_{\lambda\nu} + (\nabla_\nu V^\lambda) g_{\mu\lambda} \\ &= \nabla_\mu V_\nu + \nabla_\nu V_\mu ,\end{aligned} \tag{B.20}$$

or

$$\boxed{\mathcal{L}_V g_{\mu\nu} = 2\nabla_{(\mu} V_{\nu)},} \tag{B.21}$$

where ∇_μ is the covariant derivative derived from $g_{\mu\nu}$.

Let's put some of these ideas into the context of general relativity. You will often hear it proclaimed that GR is a "diffeomorphism invariant" theory. What this means is that, if the universe is represented by a manifold M with metric $g_{\mu\nu}$ and matter fields ψ, and $\phi : M \to M$ is a diffeomorphism, then the sets $(M, g_{\mu\nu}, \psi)$ and $(M, \phi^* g_{\mu\nu}, \phi^* \psi)$ represent the same physical situation. Since diffeomorphisms are just active coordinate transformations, this is a highbrow

way of saying that the theory is coordinate invariant. Although such a statement is true, it is a source of great misunderstanding, for the simple fact that it conveys very little information. Any semi-respectable theory of physics is coordinate invariant, including those based on special relativity or Newtonian mechanics; GR is not unique in this regard. When people say that GR is diffeomorphism invariant, more likely than not they have one of two (closely related) concepts in mind: the theory is free of "prior geometry," and there is no *preferred* coordinate system for spacetime. The first of these stems from the fact that the metric is a dynamical variable, and along with it the connection and volume element and so forth. Nothing is given to us ahead of time, unlike in classical mechanics or SR. As a consequence, there is no way to simplify life by sticking to a specific coordinate system adapted to some absolute elements of the geometry. This state of affairs forces us to be very careful; it is possible that two purportedly distinct configurations (of matter and metric) in GR are actually "the same," related by a diffeomorphism. In a path integral approach to quantum gravity, where we would like to sum over all possible configurations, special care must be taken not to overcount by allowing physically indistinguishable configurations to contribute more than once. In SR or Newtonian mechanics, meanwhile, the existence of a preferred set of coordinates saves us from such ambiguities. The fact that GR has no preferred coordinate system is often garbled into the statement that it is coordinate invariant (or "generally covariant," or "diffeomorphism invariant"); both things are true, but one has more content than the other.

On the other hand, the fact of diffeomorphism invariance can be put to good use. Recall that the complete action for gravity coupled to a set of matter fields ψ^i is given by a sum of the Hilbert action for GR plus the matter action,

$$S = \frac{1}{16\pi G} S_H[g_{\mu\nu}] + S_M[g_{\mu\nu}, \psi^i]. \tag{B.22}$$

The Hilbert action S_H is diffeomorphism invariant when considered in isolation, so the matter action S_M must also be if the action as a whole is to be invariant. We can write the variation in S_M under a diffeomorphism as

$$\delta S_M = \int d^n x \frac{\delta S_M}{\delta g_{\mu\nu}} \delta g_{\mu\nu} + \int d^n x \frac{\delta S_M}{\delta \psi^i} \delta \psi^i. \tag{B.23}$$

We are not considering arbitrary variations of the fields, only those that result from a diffeomorphism. Nevertheless, the matter equations of motion tell us that the variation of S_M with respect to ψ^i will vanish for any variation, since the gravitational part of the action doesn't involve the matter fields. Hence, for a diffeomorphism invariant theory the first term on the right-hand side of (B.23) must also vanish. If the diffeomorphism is generated by a vector field $V^\mu(x)$, the infinitesimal change in the metric is simply given by its Lie derivative along V^μ; by (B.20) we have

$$\begin{aligned}\delta g_{\mu\nu} &= \mathcal{L}_V g_{\mu\nu} \\ &= 2\nabla_{(\mu} V_{\nu)}.\end{aligned} \tag{B.24}$$

Setting $\delta S_M = 0$ then implies

$$\begin{aligned}0 &= \int d^n x \frac{\delta S_M}{\delta g_{\mu\nu}} \nabla_\mu V_\nu \\ &= -\int d^n x \sqrt{-g} V_\nu \nabla_\mu \left(\frac{1}{\sqrt{-g}} \frac{\delta S_M}{\delta g_{\mu\nu}} \right),\end{aligned} \tag{B.25}$$

where we are able to drop the symmetrization of $\nabla_{(\mu} V_{\nu)}$ since $\delta S_M / \delta g_{\mu\nu}$ is already symmetric. Demanding that (B.25) hold for diffeomorphisms generated by arbitrary vector fields V^μ, and using the definition (4.73) of the energy-momentum tensor, we obtain precisely the law of energy-momentum conservation,

$$\nabla_\mu T^{\mu\nu} = 0. \tag{B.26}$$

Conservation of $T_{\mu\nu}$ is a powerful statement, and it might seem surprising that we derived it from as weak a requirement as diffeomorphism invariance. Actually we sneaked in a much stronger assumption, namely that there is a clean separation between the "matter" and "gravitational" actions (in the sense that no matter fields appeared in the gravitational action). If there were, for example, a scalar field multiplying the curvature scalar and also appearing in the matter action (as in the scalar-tensor theories discussed in Chapter 4), this assumption would have been violated, and $T_{\mu\nu}$ would not be conserved by itself.

Recall that in Chapter 3 we spoke of symmetries and Killing vectors, with repeated appeals to look in the Appendices. Now that we understand more about diffeomorphisms, it is perfectly straightforward to understand symmetries. We say that a diffeomorphism ϕ is a **symmetry** of some tensor T if the tensor is invariant after being pulled back under ϕ:

$$\phi^* T = T. \tag{B.27}$$

Although symmetries may be discrete, it is also common to have a one-parameter family of symmetries ϕ_t. If the family is generated by a vector field $V^\mu(x)$, then (B.27) amounts to

$$\mathcal{L}_V T = 0. \tag{B.28}$$

By (B.12), one implication of a symmetry is that, if T is symmetric under some one-parameter family of diffeomorphisms, we can always find a coordinate system in which the components of T are all independent of one of the coordinates (the integral curve coordinate of the vector field). The converse is also true; if all of the components are independent of one of the coordinates, then the partial

derivative vector field associated with that coordinate generates a symmetry of the tensor.

The most important symmetries are those of the metric, for which $\phi^* g_{\mu\nu} = g_{\mu\nu}$. A diffeomorphism of this type is called an isometry. If a one-parameter family of isometries is generated by a vector field $K^\mu(x)$, then K^μ turns out to be a Killing vector field. The condition that K^μ be a Killing vector is thus

$$\mathcal{L}_K g_{\mu\nu} = 0, \tag{B.29}$$

or from (B.20),

$$\nabla_{(\mu} K_{\nu)} = 0. \tag{B.30}$$

We recognize this last version as Killing's equation, (3.174). From our discussion in Chapter 3 we know that, if a spacetime has a Killing vector, we can find a coordinate system in which the metric is independent of one of the coordinates, and the quantity $p_\mu K^\mu$ will be constant along geodesics with tangent vector p^μ. Once we have set up the machinery of diffeomorphisms and Lie derivatives, the derivation of Killing vectors proceeds much more elegantly.

B.1 ■ EXERCISES

8. In Euclidean three-space, find and draw the integral curves of the vector fields

$$A = \frac{y-x}{r}\frac{\partial}{\partial x} - \frac{x+y}{r}\frac{\partial}{\partial y}$$

and

$$B = xy\frac{\partial}{\partial x} - y^2\frac{\partial}{\partial y}.$$

Calculate $C = \mathcal{L}_A B$ and draw the integral curves of C.

APPENDIX

C Submanifolds

The notion of a submanifold, some subset of another manifold which might be (and usually is) of lower dimension, is intuitively straightforward; it should come as no surprise, however, to learn that a certain amount of formalism comes along for the ride. Submanifolds arise all the time in general relativity—as boundaries of spacetimes, hypersurfaces at fixed time, spaces into which larger spaces are foliated by the action of symmetries—so it is worth our effort to understand how they work.

Consider an n-dimensional manifold M and an m-dimensional manifold S, with $m \leq n$, and a map $\phi : S \to M$. If the map ϕ is both C^∞ and one-to-one, and the inverse $\phi^{-1} : \phi[M] \to S$ is also one-to-one, then we say that the the image $\phi[M]$ is an **embedded submanifold** of M. If ϕ is one-to-one locally but not necessarily globally (that is, there may be self-intersections of $\phi[M]$ in M), then we say that $\phi[M]$ is an **immersed submanifold** of M. When we speak of "submanifolds" without any particular modifier, we are imagining that they are embedded. An m-dimensional submanifold of an n-dimensional manifold is said to be of **codimension** $n - m$.

As discussed in Appendix A, the map $\phi : S \to M$ can be used to push forward $(k, 0)$ tensors from S to M, and to pull back $(0, l)$ tensors from M to S. In particular, given a point $q \in S$ and its image $\phi(q) \in M$, the tangent space $T_{\phi(q)}\phi[S]$ is naturally identified as an m-dimensional subspace of the n-dimensional vector space $T_{\phi(q)}M$. If you think about the definition of a vector as the directional derivative along a curve, this makes perfect sense; any curve $\gamma : \mathbf{R} \to S$ clearly defines a curve in M via composition ($\phi \circ \gamma : \mathbf{R} \to M$), which in turn defines a directional derivative. Similarly, differential forms in M can be pulled back to S by restricting their action to vectors in the subspace $T_{\phi(q)}\phi[S]$.

Another way to define submanifolds is as places where a collection of functions takes on some specified fixed set of values. An m-dimensional submanifold of M can be specified in terms of $n - m$ functions $f^a(x)$, where a runs from 1 to $n - m$, as the set of points x, where the f^a's are equal to some constants f^a_*:

$$\begin{aligned} f^1(x) &= f^1_* \\ f^2(x) &= f^2_* \\ &\vdots \\ f^{n-m}(x) &= f^{n-m}_*. \end{aligned} \tag{C.1}$$

The functions should be nondegenerate, so that the submanifold really is of dimension m. Notice that the submanifold defined in this way is an actual subset of M; it is equivalent to what we called $\phi[S]$ in our previous definition. For convenience, we will henceforth tend to blur the distinction between the original space and its embedding as a submanifold, and simply refer to "the submanifold S."

To see the relationship between the two definitions of a submanifold, imagine constructing a set of coordinates $x^\mu = \{f^a, y^\alpha\}$ in a neighborhood of $\phi[S] \subset M$, consisting of the $n-m$ functions f^a and an additional m function y^α. Then we can pull back the functions y^α to serve as coordinates on S, and the map $\phi : S \to M$ is simply given by

$$\phi : (y^\alpha) \to (f^a_*, y^\alpha). \tag{C.2}$$

A simple example is the two-sphere S^2, which in fact we defined as the set of points a unit distance from the origin in $\mathbf{R}^3$. In polar coordinates (r, θ, ϕ), this is equivalent to the requirement $r = 1$, so the coordinate r plays the role of the function $f(x)$, while θ and ϕ are induced coordinates on S^2.

We have already mentioned in (B.3) that specifying a single vector field leads to a family of integral curves, which are simply one-dimensional submanifolds. We might imagine generalizing this construction by using a set of several vector fields to define higher-dimensional submanifolds. Imagine we have an n-dimensional manifold M, an m-dimensional submanifold S, and a set of p linearly independent vector fields $V^\mu_{(a)}$, with $p \geq m$. Then the notion that these vector fields "fit together to define S" means that each vector is tangent to S everywhere, so that the $V^\mu_{(a)}$'s span each tangent space T_pS; we say that S is an **integral submanifold** of the vector fields. However, any given set of vector fields may or may not actually fit together to define such submanifolds. Whether they do or not is revealed by **Frobenius's theorem**: a set of vector fields $V^\mu_{(a)}$ fit together to define integral submanifolds if and only if all of their commutators are in the space spanned by the $V^\mu_{(a)}$'s; that is, if

$$[V_{(a)}, V_{(b)}]^\mu = \alpha^c V^\mu_{(c)} \tag{C.3}$$

for some set of coefficients $\alpha^c(x)$. (In the language of group theory, this means that the vector fields form a Lie algebra.) We won't provide a proof, but hopefully the result makes some mathematical sense. If the vector fields are going to fit together to form a submanifold S, they must remain tangent to S everywhere. But the commutator $[V, W]$ is equivalent to the Lie derivative $\mathcal{L}_V W$, which measures how W changes as we travel along V. If this Lie derivative doesn't remain in the space defined by the vectors, it means that W starts sticking out of the submanifold S. Examples of vector fields fitting together to form submanifolds are easy to come by; in Section 5.2 we discussed how the three Killing vectors associated with spherical symmetry define a foliation of a three-dimensional space into two-spheres. (Notice that the dimensionality of the integral submanifold can be less

than the number of vector fields.) For a discussion of Frobenius's theorem, see Schutz (1980).

An interesting alternative formulation of Frobenius's theorem uses differential forms. First notice that any set of p linearly independent one-forms $\omega_\mu^{(a)}$ defines an $(n-p)$-dimensional vector subspace of T_pM, called the **annihilator** of the set of forms, consisting of those vectors $V^\mu \in T_pM$ satisfying

$$\omega_\mu^{(a)} V^\mu = 0 \tag{C.4}$$

for all $\omega_\mu^{(a)}$. So instead of asking whether a collection of vector fields fit together to define a submanifold, we could ask whether a collection of one-forms $\omega_\mu^{(a)}$ define a set of vector subspaces that fit together as tangent spaces to a set of submanifolds. To understand when this happens, recall the definition (C.1) of an m-dimensional submanifold as a place where a set of $p = n-m$ functions $f^a(x)$ are set equal to constants. A constant function is one for which the exterior derivative $(\mathrm{d}f^a)_\mu = \nabla_\mu f^a$ vanishes; but if a function is constant only along some submanifold, that means that

$$\mathrm{d}f^a(V) = V^\mu \nabla_\mu f^a = 0 \tag{C.5}$$

for all vectors V^μ tangent to the submanifold, $V^\mu \in T_pS$. It also goes the other way; if a vector V^μ is annihilated by all of the gradients $\nabla_\mu f^a$, it is necessarily tangent to the corresponding submanifold S. Therefore, if a set of one-forms are each exact, $\omega_\mu^{(a)} = \nabla_\mu f^a$, the vector spaces they annihilate will certainly define submanifolds, namely those along which the f^a's are constant. But if a set of p one-forms annihilates a certain subspace, so will any other set of p one-forms that are linear combinations of the originals. We therefore say that a set of one-forms $\omega_\mu^{(a)}$ is **surface-forming** if every member can be expressed as a linear combination of a set of exact forms; that is, if there exist functions $g^a{}_b(x)$ and $f^a(x)$ such that

$$\omega_\mu^{(a)} = \sum_b g^a{}_b \nabla_\mu f^b. \tag{C.6}$$

Of course, when handed a set of forms, it might be hard to tell whether there exist functions such that this condition is satisfied; this is where the dual formulation of Frobenius's theorem comes in. This version of the theorem states that a set of one-forms $\omega_\mu^{(a)}$ is surface-forming if and only if every pair of vectors in the annihilator of the set is also annihilated by the exterior derivatives $\mathrm{d}\omega^{(a)}$. In other words, the set $\omega_\mu^{(a)}$ will satisfy (C.6) if and only if, for every pair of vectors V^μ and W^μ satisfying $\omega_\mu^{(a)} V^\mu = 0$ and $\omega_\mu^{(a)} W^\mu = 0$ for all a, we also have

$$\nabla_{[\mu} \omega_{\nu]}^{(a)} V^\mu W^\nu = 0. \tag{C.7}$$

A set of forms $\omega_\mu^{(a)}$ satisfying this condition is sometimes called "closed," which is obviously a generalization of the notion of a single form being closed (namely, that its exterior derivative vanishes). We won't prove the equivalence of the dual formulation of Frobenius' theorem with the vector-field version, but it clearly involves acting our set of forms on the vector-field commutator (C.3).

APPENDIX

D Hypersurfaces

A **hypersurface** is an $(n-1)$-dimensional (codimension one) submanifold Σ of an n-dimensional manifold M. (Of course if $n = 3$, Σ might as well just be called a "surface," but we'll continue to use "hyper-" for consistency.) Hypersurfaces are of great utility in general relativity, and a lot of formalism goes along with them. In this Appendix we collect a set of results in the study of hypersurfaces: normal vectors, generators of null hypersurfaces, Frobenius's theorem for hypersurfaces, Gaussian normal coordinates, induced metrics, projection tensors, extrinsic curvature, and manifolds with boundary. It's something of a smorgasbord, with all the messiness that implies, but hopefully appetizing and nutritious as well.

One way to specify a hypersurface Σ is by setting single function to a constant,

$$f(x) = f_* . \tag{D.1}$$

The vector field

$$\zeta^\mu = g^{\mu\nu}\nabla_\mu f \tag{D.2}$$

will be normal to the surface, in the sense that it is orthogonal to all vectors in $T_p\Sigma \subset T_pM$. If ζ^μ is timelike, the hypersurface is said to be spacelike; if ζ^μ is spacelike the hypersurface is timelike, and if ζ^μ is null the hypersurface is also null. Any vector field proportional to a normal vector field,

$$\xi^\mu = h(x)\nabla^\mu f \tag{D.3}$$

for some function $h(x)$, will itself be a normal vector field; since the normal vector is unique up to scaling, any normal vector can be written in this form. For timelike and spacelike hypersurfaces, we can therefore define a normalized version of the normal vector,

$$n^\mu = \pm\frac{\zeta^\mu}{|\zeta_\mu\zeta^\mu|^{1/2}} . \tag{D.4}$$

Then $n^\mu n_\mu = -1$ for spacelike surfaces and $n^\mu n_\mu = +1$ for timelike surfaces; up to an overall orientation, such a normal vector field is unique. For spacelike surfaces the sign is typically chosen so as to make n^μ be future-directed.

Null hypersurfaces have a special feature: they can be divided into a set of null geodesics, called **generators** of the hypersurface. Let's see how this works.

Notice that the normal vector ζ^μ is tangent to Σ as well as normal to it, since null vectors are orthogonal to themselves. Therefore the integral curves $x^\mu(\alpha)$, satisfying

$$\zeta^\mu = \frac{dx^\mu}{d\alpha}, \tag{D.5}$$

will be null curves contained in the hypersurface. These curves $x^\mu(\alpha)$ necessarily turn out to be geodesics, although α might not be an affine parameter. To verify this claim, recall that the general form of the geodesic equation can be expressed as

$$\zeta^\mu \nabla_\mu \zeta_\nu = \eta(\alpha)\zeta_\nu, \tag{D.6}$$

where $\eta(\alpha)$ is a function that will vanish if α is an affine parameter. We simply plug in (D.2) and calculate:

$$\begin{aligned} \zeta^\mu \nabla_\mu \zeta_\nu &= \zeta^\mu \nabla_\mu \nabla_\nu f \\ &= \zeta^\mu \nabla_\nu \nabla_\mu f \\ &= \zeta^\mu \nabla_\nu \zeta_\mu \\ &= \tfrac{1}{2} \nabla_\nu (\zeta^\mu \zeta_\mu). \end{aligned} \tag{D.7}$$

In the second line we used the torsion-free condition, that covariant derivatives acting on scalars commute. Note that, even though $\zeta^\mu \zeta_\mu = 0$ on Σ itself, we can't be sure that $\nabla_\nu(\zeta^\mu \zeta_\mu)$ vanishes, since $\zeta^\mu \zeta_\mu$ might be nonzero off the hypersurface. If the gradient vanishes, (D.7) is the geodesic equation, and we're done. But if it doesn't vanish, we can use $\zeta^\mu \zeta_\mu = 0$ as an alternative way to define the submanifold Σ, and its derivative defines a normal vector. Therefore, we must have

$$\nabla_\mu (\zeta^\nu \zeta_\nu) = g \nabla_\mu f = g \zeta_\mu, \tag{D.8}$$

where $g(x)$ is some scalar function. We then plug into (D.7) to get

$$\zeta^\mu \nabla_\mu \zeta_\nu = \tfrac{1}{2} g \zeta_\nu, \tag{D.9}$$

which is equivalent to the geodesic equation (D.6). Of course, once we know that a path $x^\mu(\alpha)$ is a geodesic, we are free to re-parameterize it with an affine parameter $\lambda(\alpha)$. Equivalently, we scale the normal vector field by a scalar function $h(x)$,

$$\xi^\mu = h \zeta^\mu, \tag{D.10}$$

such that $\zeta^\mu \nabla_\mu \zeta^\nu = 0$. It is conventional to do exactly this, and use the corresponding tangent vectors

$$\xi^\mu = \frac{dx^\mu}{d\lambda} \tag{D.11}$$

as normal vectors to Σ. The null geodesics $x^\mu(\lambda)$, whose union is the null hypersurface Σ, are the generators of Σ.

From (D.3) we know that a vector field normal to a hypersurface can be written in the form $\xi^\mu = h\nabla^\mu f$. In the exercises for Chapter 4 you were asked to show that this implies

$$\xi_{[\mu}\nabla_\nu\xi_{\sigma]} = 0, \tag{D.12}$$

or in differential forms notation,

$$\xi \wedge d\xi = 0. \tag{D.13}$$

The converse, that any vector field satisfying this equation is orthogonal to a hypersurface, is harder to show from first principles, but is a direct consequence of the dual formulation of Frobenius's theorem. Imagine we have two vectors V^μ and W^μ, both of which are annihilated by a one-form ξ_μ obeying (D.12). From Frobenius's theorem (C.7), ξ_μ will define a hypersurface if and only if

$$\nabla_{[\mu}\xi_{\nu]}V^\mu W^\nu = 0. \tag{D.14}$$

Applying the expression in (D.12) to $V^\mu W^\nu$ and expanding the antisymmetrization brackets, we get

$$\begin{aligned}\xi_{[\mu}\nabla_\nu\xi_{\sigma]}V^\mu W^\nu &= \tfrac{1}{3}(\xi_\mu\nabla_{[\nu}\xi_{\sigma]} + \xi_\nu\nabla_{[\sigma}\xi_{\mu]})V^\mu W^\nu + \tfrac{1}{3}\xi_\sigma\nabla_{[\mu}\xi_{\nu]}V^\mu W^\nu \\ &= \tfrac{1}{3}\xi_\sigma\nabla_{[\mu}\xi_{\nu]}V^\mu W^\nu,\end{aligned} \tag{D.15}$$

where in the last line we used the fact that V^μ and W^μ are annihilated by ξ_μ. But since $\nabla_{[\mu}\xi_{\nu]}V^\mu W^\nu$ is a scalar and ξ_σ is a nonvanishing one-form, the only way (D.15) can vanish is if (D.14) holds. Therefore, (D.12) will be true if and only if ξ_μ is hypersurface-orthogonal.

It is often convenient to put a coordinate system on a manifold (or part of it) that is naturally adapted to some hypersurface Σ; Gaussian normal coordinates provide a convenient way to do just that. First choose coordinates $y^i = \{y^1, \ldots, y^{n-1}\}$ on Σ. At each point $p \in \Sigma$, construct the (unique) geodesic for which n^μ is the tangent vector at p. Let z be the affine parameter on each geodesic. [This parameter is unique if n^μ is normalized and $z(p) = 0$.] Any point q in a neighborhood of Σ lives on one such geodesic. To each such point we assign coordinates $\{z, y^1, \ldots, y^{n-1}\}$, where the y^i's are the coordinates of the point p connected to q by the geodesic we have constructed. These coordinates $\{z, y^1, \ldots, y^{n-1}\}$ are **Gaussian normal coordinates** (not to be confused with "Riemann normal coordinates," constructed by following geodesics in all directions from a single point p). These coordinates will eventually fail to be well-defined if we reach a point where geodesics focus and intersect, but they will always exist in some region including Σ. All of our statements about Gaussian normal coordinates should be taken as applying in the region where they are well-defined.

Associated with the coordinate functions $\{z, y^1, \ldots, y^{n-1}\}$ are coordinate-basis vector fields $\{\partial_z, \partial_1, \ldots, \partial_{n-1}\}$. For notational convenience let's label these vector fields

$$\begin{aligned}(\partial_z)^\mu &= n^\mu, \\ (\partial_i)^\mu &= Y^\mu_{(i)},\end{aligned} \tag{D.16}$$

where the first line makes sense because ∂_z is simply the extension along the geodesics of the original normal vector n^μ. With respect to these basis vectors, the metric takes on a simple form. To start, we know that

$$g_{zz} = ds^2(\partial_z, \partial_z) = n_\mu n^\mu = \pm 1, \tag{D.17}$$

since n^μ is just the normalized tangent vector to the geodesics emanating from Σ. To encapsulate the sign ambiguity, let's label this σ:

$$\sigma = n_\mu n^\mu = \pm 1. \tag{D.18}$$

But it is also the case that $g_{zi} = n_\mu Y^\mu_{(i)} = 0$, as we can straightforwardly check. Start at the original surface Σ, where $n_\mu Y^\mu_{(i)} = 0$ by hypothesis (since n^μ is normal to Σ). Then we calculate

$$\begin{aligned}\frac{D}{dz}(n_\mu Y^\mu_{(i)}) &= n^\nu \nabla_\nu (n_\mu Y^\mu_{(i)}) \\ &= n^\nu n_\mu \nabla_\nu Y^\mu_{(i)} \\ &= Y^\nu_{(i)} n_\mu \nabla_\nu n^\mu \\ &= \frac{1}{2} Y^\nu_{(i)} \nabla_\nu (n_\mu n^\mu) \\ &= 0.\end{aligned} \tag{D.19}$$

Let's explain this derivation line-by-line. The first line is simply the definition of the directional covariant derivative D/dz. The second uses the Leibniz rule, plus the fact that n_μ is parallel-transported along the geodesic ($n^\nu \nabla_\nu n_\mu = 0$). The third line uses the fact that n^μ and $Y^\nu_{(i)}$ are both coordinate basis vectors, so their Lie bracket vanishes: $[n, Y_{(i)}]^\mu = n^\nu \nabla_\nu Y^\mu_{(i)} - Y^\nu_{(i)} \nabla_\nu n^\mu = 0$. The fourth line again uses Leibniz and the fact that n_μ is parallel-transported, while the fifth simply reflects the fact that $n_\mu n^\mu = \sigma$ is a constant.

We can therefore write the metric in Gaussian normal coordinates as

$$\boxed{ds^2 = \sigma \mathrm{d}z^2 + \gamma_{ij} \mathrm{d}y^i \mathrm{d}y^j,} \tag{D.20}$$

where $\gamma_{ij} = g(\partial_i, \partial_j)$ will in general be a function of all the coordinates $\{z, y^1, \ldots, y^{n-1}\}$. We haven't made any assumptions whatsoever about the geometry; we have simply chosen a coordinate system in which the metric takes a

certain form. Notice that setting z = constant defines a family of hypersurfaces diffeomorphic to the original surface Σ; the lack of off-diagonal terms g_{zi} in (D.20) reflects the fact that the vector field n^μ is orthogonal to all of these surfaces, not just the original one. Gaussian normal coordinates are by no means exotic; we use them all the time. Simple examples include inertial coordinates on Minkowski space,

$$ds^2 = -\mathrm{d}t^2 + \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2, \tag{D.21}$$

or polar coordinates in Euclidean 3-space,

$$ds^2 = \mathrm{d}r^2 + r^2\mathrm{d}\theta^2 + r^2 \sin^2\theta \, \mathrm{d}\phi^2. \tag{D.22}$$

Ordinary Robertson–Walker coordinates in cosmology provide a slightly less trivial example,

$$ds^2 = -\mathrm{d}t^2 + a^2(t)\left[\frac{\mathrm{d}r^2}{1-\kappa r^2} + r^2 d\Omega^2\right]. \tag{D.23}$$

Of course, the RW geometries are highly symmetric (homogeneous and isotropic). But, since we have just seen that Gaussian normal coordinates can always be defined, we know that we can describe a perfectly general geometry by altering the spatial components of the metric. This provides one popular way of describing cosmological perturbations; we define "synchronous gauge" for flat spatial sections as

$$ds^2 = -\mathrm{d}t^2 + a^2(t)(\delta_{ij} + h_{ij})\mathrm{d}x^i \mathrm{d}x^j, \tag{D.24}$$

where $h_{ij}(t, \mathbf{x})$ is the metric perturbation. (The generalization to curved spatial sections is immediate.) Again, we have not made any assumptions about the geometry, only chosen a potentially convenient coordinate system.

Recall that the map $\phi : \Sigma \to M$ that embeds any submanifold allows us to pull back the metric from M to Σ. Given coordinates y^i on Σ and x^μ on M, we define the **induced metric** on the submanifold as

$$(\phi^* g)_{ij} = \frac{\partial x^\mu}{\partial y^i}\frac{\partial x^\nu}{\partial y^j} g_{\mu\nu}. \tag{D.25}$$

In the case where the submanifold is a hypersurface, this induced metric is precisely the same as the γ_{ij} appearing in (D.20). To see this, notice that Gaussian normal coordinates are a special case of the natural embedding coordinates described by (C.2). We have a hypersurface Σ defined by $z = z_*$ on M, with coordinates y^i defined on Σ, and a map $\phi : \Sigma \to M$ given by

$$\phi : y^i \to x^\mu = (z_*, y^i). \tag{D.26}$$

Given the form of the metric (D.20) on M, it is immediate that under this map the pullback (D.25) is simply

$$(\phi^* g)_{ij} = \gamma_{ij}. \tag{D.27}$$

Keep in mind that this equation should only be evaluated in Gaussian normal coordinates; otherwise the right-hand side doesn't even make sense.

Along with an induced metric, submanifolds inherit an induced volume element from the manifold in which they are embedded. Recall that a volume element on an n-dimensional manifold with metric $g_{\mu\nu}$ is given by the Levi–Civita tensor, which can be expressed as

$$\epsilon = \sqrt{|g|}\, \mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^n. \tag{D.28}$$

To get a volume element on a submanifold Σ, it is convenient to introduce Gaussian normal coordinates $(z, y^1, \ldots y^{n-1})$, in which the metric takes the form (D.20). The volume element $\widehat{\epsilon}$ on Σ will then take the form

$$\widehat{\epsilon} = \sqrt{|\gamma|}\, \mathrm{d}y^1 \wedge \cdots \wedge \mathrm{d}y^{n-1}. \tag{D.29}$$

(By choosing the first coordinate to be the one normal to the hypersurface, we have implicitly chosen a convention for how the orientation on M defines an orientation on Σ.) In these coordinates we have

$$\sqrt{|g|} = \sqrt{|\gamma|}, \tag{D.30}$$

and the volume element on M therefore becomes

$$\epsilon = \sqrt{|\gamma|}\, \mathrm{d}z \wedge \mathrm{d}y^1 \wedge \cdots \wedge \mathrm{d}y^{n-1}. \tag{D.31}$$

We can relate the two volume elements by using the normal vector to Σ, which has components

$$n^\mu = (1, 0, \ldots, 0). \tag{D.32}$$

The contraction of ϵ with n^μ can be denoted

$$[\epsilon(n)]_{\mu_1 \cdots \mu_{n-1}} = n^\lambda \epsilon_{\lambda \mu_1 \cdots \mu_{n-1}}. \tag{D.33}$$

It is then clear that, in these coordinates, we have

$$\begin{aligned} \epsilon(n) &= \sqrt{|\gamma|}\, \mathrm{d}y^1 \wedge \cdots \wedge \mathrm{d}y^{n-1} \\ &= \widehat{\epsilon}. \end{aligned} \tag{D.34}$$

Thus, the induced volume element has components

$$\widehat{\epsilon}_{\mu_1 \cdots \mu_{n-1}} = n^\lambda \epsilon_{\lambda \mu_1 \cdots \mu_{n-1}}. \tag{D.35}$$

But this is a relation between tensors, so will be true in any coordinate system. We can also reconstruct ϵ from $\widehat{\epsilon}$ and n^μ, via

$$\frac{1}{n} \epsilon_{\nu \mu_1 \cdots \mu_{n-1}} = n_{[\nu} \widehat{\epsilon}_{\mu_1 \cdots \mu_{n-1}]}, \tag{D.36}$$

as can easily be checked by contracting with n^ν. The notion of a submanifold volume element will be crucial in our discussion of Stokes's theorem below.

Another concept closely related to the induced metric on a hypersurface is that of the **projection tensor** for a hypersurface Σ with unit normal vector n^μ, given by

$$P_{\mu\nu} = g_{\mu\nu} - \sigma n_\mu n_\nu, \tag{D.37}$$

where $\sigma = n_\mu n^\mu$. Let's collect some useful properties of this object. Given any vector V^μ in T_pM, $P_{\mu\nu}$ will project it tangent to the hypersurface (that is, orthogonal to n^μ):

$$\begin{aligned}(P_{\mu\nu}V^\mu)n^\nu &= g_{\mu\nu}V^\mu n^\nu - \sigma n_\mu n_\nu V^\mu n^\nu \\ &= V^\mu n_\mu - \sigma^2 V^\mu n_\mu \\ &= 0.\end{aligned} \tag{D.38}$$

Acting on any two vectors V^μ and W^ν that are already tangent to Σ, the projection tensor acts like the metric:

$$\begin{aligned}P_{\mu\nu}V^\mu W^\nu &= g_{\mu\nu}V^\mu W^\nu - \sigma n_\mu n_\nu V^\mu W^\nu \\ &= g_{\mu\nu}V^\mu W^\nu.\end{aligned} \tag{D.39}$$

Finally, the projection tensor is idempotent; acting two (or more) times produces the same result as only acting once:

$$\begin{aligned}P^\mu{}_\lambda P^\lambda{}_\nu &= (\delta^\mu_\lambda - \sigma n^\mu n_\lambda)(\delta^\lambda_\nu - \sigma n^\lambda n_\nu) \\ &= \delta^\mu_\nu - \sigma n^\mu n_\nu - \sigma n^\mu n_\nu + \sigma^3 n^\mu n_\nu \\ &= P^\mu{}_\nu.\end{aligned} \tag{D.40}$$

$P_{\mu\nu}$ is sometimes called the **first fundamental form** of the hypersurface. Because it really does act like the metric for vectors tangent to Σ, and hypersurfaces are often spacelike, you will sometimes see it referred to as the "spatial metric."

Long ago when we first spoke of manifolds and curvature, we were careful to distinguish between the "intrinsic" curvature of a space, as measured by the Riemann tensor, and the "extrinsic" curvature, which depends on how the space is embedded in some larger space. For example, a two-torus can have a flat metric, but any embedding in $\mathbf{R}^3$ makes it look curved. We are now in a position to give a formal definition of this notion, which makes sense for hypersurfaces. Let's assume we have a family of hypersurfaces Σ with unit vector field n^μ, and we extend n^μ through a region (any way we like). Then the **extrinsic curvature** of Σ is simply given by the Lie derivative of the projection tensor along the normal vector field,

$$K_{\mu\nu} = \tfrac{1}{2}\mathcal{L}_n P_{\mu\nu}. \tag{D.41}$$

The extrinsic curvature, sometimes called the **second fundamental form** of the submanifold, is thus interpreted as the rate of change of the projection tensor (the spatial metric, if Σ is spacelike) as we travel along the normal vector field; it is independent of the extension of n^μ away from Σ. It is the work of a few lines to show that this definition is equivalent to the projected Lie derivative of the metric itself,

$$K_{\mu\nu} = \tfrac{1}{2} P^\alpha{}_\mu P^\beta{}_\nu \mathcal{L}_n g_{\alpha\beta}. \tag{D.42}$$

We know from (B.20) that the Lie derivative of $g_{\mu\nu}$ is given by the symmetrized covariant derivative of the normal vector, so we have

$$K_{\mu\nu} = P^\alpha{}_\mu P^\beta{}_\nu \nabla_{(\alpha} n_{\beta)}. \tag{D.43}$$

Since we are not assuming that the integral curves of n^μ are geodesics, we can define the acceleration as

$$a^\mu = n^\nu \nabla_\nu n^\mu. \tag{D.44}$$

Then it is the work of a few more lines to show that (D.43) is equivalent to

$$\boxed{K_{\mu\nu} = \nabla_\mu n_\nu - \sigma n_\mu a_\nu.} \tag{D.45}$$

The extrinsic curvature has a number of nice properties. It is symmetric,

$$K_{\mu\nu} = K_{\nu\mu}, \tag{D.46}$$

which looks obvious from (D.41), although not from (D.45). You can check that (D.45) really is symmetric, taking advantage of the fact that n^μ is hypersurface-orthogonal. The extrinsic curvature is also orthogonal to the normal direction ("purely spatial"),

$$\begin{aligned} n^\mu K_{\mu\nu} &= n^\mu \nabla_\mu n_\nu - \sigma n^\mu n_\mu a_\nu \\ &= a_\nu - \sigma^2 a_\nu \\ &= 0. \end{aligned} \tag{D.47}$$

We can define a covariant derivative acting along the hypersurface, $\widehat{\nabla}_\mu$, by taking an ordinary covariant derivative and projecting it. For example, on a (1, 1) tensor $X^\mu{}_\nu$ we would have

$$\widehat{\nabla}_\sigma X^\mu{}_\nu = P^\alpha{}_\sigma P^\mu{}_\beta P^\gamma{}_\nu \nabla_\alpha X^\beta{}_\gamma. \tag{D.48}$$

From this we can construct the curvature tensor on the hypersurface $\widehat{R}^\rho{}_{\sigma\mu\nu}$, for example by considering the commutator of covariant derivatives acting on a vector field V^μ, which is tangent to the hypersurface ($P^\mu{}_\nu V^\nu = V^\mu$),

$$[\widehat{\nabla}_\mu, \widehat{\nabla}_\nu] V^\rho = \widehat{R}^\rho{}_{\sigma\mu\nu} V^\sigma. \tag{D.49}$$

Two important equations relate the n-dimensional Riemann curvature to the hypersurface Riemann curvature and the extrinsic curvature. **Gauss's equation** is

$$\widehat{R}^{\rho}{}_{\sigma\mu\nu} = P^{\rho}{}_{\alpha}P^{\beta}{}_{\rho}P^{\gamma}{}_{\mu}P^{\delta}{}_{\nu}R^{\alpha}{}_{\beta\gamma\delta} + \sigma(K^{\rho}{}_{\mu}K_{\sigma\nu} - K^{\rho}{}_{\nu}K_{\sigma\mu}). \tag{D.50}$$

We can take the appropriate traces to get the hypersurface curvature scalar,

$$\widehat{R} = P^{\sigma\nu}\widehat{R}^{\lambda}{}_{\sigma\lambda\nu} = R - \sigma(2R_{\mu\nu}n^{\mu}n^{\nu} + K^2 - K^{\mu\nu}K_{\mu\nu}), \tag{D.51}$$

where $K = g^{\mu\nu}K_{\mu\nu}$. We also have **Codacci's equation**,

$$\widehat{\nabla}_{[\mu}K_{\nu]}{}^{\mu} = \tfrac{1}{2}P^{\sigma}{}_{\nu}R_{\rho\sigma}n^{\rho}. \tag{D.52}$$

Together, (D.50) and (D.52) are, imaginatively enough, called the Gauss–Codacci equations.

To stave off confusion, we should note that the definition of extrinsic curvature tends to vary from reference to reference. In some sources the normal vector field is taken to be geodesic everywhere ($a^{\mu} = 0$); things then simplify considerably, and it's straightforward to show that in this case we have

$$\begin{aligned} K_{\mu\nu} &= \tfrac{1}{2}\mathcal{L}_n P_{\mu\nu} \\ &= \tfrac{1}{2}\mathcal{L}_n g_{\mu\nu} \\ &= \nabla_{\mu}n_{\nu}. \end{aligned} \tag{D.53}$$

(If we are given an entire set of hypersurfaces ahead of time, we cannot simply assume that integral curves of the unit normal vector field are geodesics. However, we are often given just a single surface, in which case we are allowed to extend the normal vector field off the surface by solving the geodesic equation.) Other references prefer to think of the extrinsic curvature as a tensor $\widehat{K}_{ij}$ living on Σ rather than in M. If we have an embedding $\phi : y^i \to x^{\mu}$, this version of the extrinsic curvature is given by the pullback,

$$\begin{aligned} \widehat{K}_{ij} &= (\phi^* K)_{ij} \\ &= \frac{\partial x^{\mu}}{\partial y^i}\frac{\partial x^{\nu}}{\partial y^j}K_{\mu\nu}. \end{aligned} \tag{D.54}$$

Finally, some sources like to define the extrinsic curvature to be minus our definition. It should be straightforward to convert back and forth between the different conventions.

To conclude our discussion, we mention that a very common appearance of hypersurfaces is as the **boundary** of a closed region N of a manifold M, conventionally denoted ∂N. If for example N consists of all the elements of $\mathbf{R}^n$ that lie at a distance from the origin $r \leq 1$, the boundary ∂N is clearly the $(n-1)$-sphere defined by $r = 1$. We may extend this notion to cases where we are not considering a closed region, but an entire manifold with a boundary attached. A **manifold with**

boundary is a set equipped with an atlas of coordinate charts, exactly as in our definition of a manifold in Chapter 2, except that the charts are taken to be maps to the upper half of $\mathbf{R}^n$: the set of n-tuples $\{x^1, \ldots, x^n\}$ with $x^1 \geq 0$. The boundary ∂M is the set of points that are mapped to $x^1 = 0$ by the charts. Then ∂M is naturally an $(n-1)$-dimensional submanifold (without boundary). An example of a boundary of a manifold will appear in our later discussion of conformal diagrams, in which conformal infinity can be thought of as a boundary to spacetime. By continuity, we can treat the boundary as a hypersurface, including inducing metrics and so on; occasionally we need to be careful in taking derivatives on the boundary, but for the most part we can trust our intuition.

APPENDIX

E Stokes's Theorem

In Section 2.10 we introduced the idea that integration on a manifold maps n-form fields to the real numbers. This point of view leads to an elegant statement of one of the most powerful theorems of differential geometry: Stokes's theorem. This theorem is the generalization of the fundamental theorem of calculus, $\int_b^a dx = a - b$. Imagine that we have an n-dimensional region M (which might be an entire manifold) with boundary ∂M, and an $(n-1)$-form ω on M. We will soon explain what is meant by the boundary of a manifold. Then $\mathrm{d}\omega$ is an n-form, which can be integrated over M, while ω itself can be integrated over ∂M. Stokes's theorem is simply

$$\boxed{\int_M \mathrm{d}\omega = \int_{\partial M} \omega.} \tag{E.1}$$

Different special cases of this theorem include not only the fundamental theorem of calculus, but also the theorems of Green, Gauss, and Stokes, familiar from vector calculus in three dimensions.

The presentation (E.1) of Stokes's theorem is extremely elegant, almost too elegant to be useful. We can, fortunately, recast it in pedestrian coordinate-and-index notation. It is convenient to first write the $(n-1)$-form ω as the Hodge dual of a one-form V,

$$\omega = *V, \tag{E.2}$$

with components

$$\begin{aligned} \omega_{\mu_1 \cdots \mu_{n-1}} &= (*V)_{\mu_1 \cdots \mu_{n-1}} \\ &= \epsilon^{\nu}{}_{\mu_1 \cdots \mu_{n-1}} V_\nu \\ &= \epsilon_{\nu \mu_1 \cdots \mu_{n-1}} V^\nu, \end{aligned} \tag{E.3}$$

where ϵ is the Levi–Civita n-form on M and we have raised the index on V in the last line. If we wanted to construct V from ω, we apply the Hodge operator again to obtain

$$V = (-1)^{s+n-1} **V = (-1)^{s+n-1} * \omega, \tag{E.4}$$

where s equals -1 for Lorentzian signatures and $+1$ for Euclidean signatures. The exterior derivative of $\omega = *V$ is an n-form, given by

$$\begin{aligned}(\mathrm{d}\omega)_{\lambda\mu_1\cdots\mu_{n-1}} &= (\mathrm{d} * V)_{\lambda\mu_1\cdots\mu_{n-1}} \\ &= n\nabla_{[\lambda}(\epsilon_{|\nu|\mu_1\cdots\mu_{n-1}]}V^{\nu}) \\ &= n\epsilon_{\nu[\mu_1\cdots\mu_{n-1}}\nabla_{\lambda]}V^{\nu},\end{aligned} \tag{E.5}$$

where n is the dimensionality of the region, not to be confused with the normal vector n^{μ} to the boundary. But any n-form can be written as a function $f(x)$ times ϵ, or equivalently as the Hodge dual of $f(x)$,

$$\mathrm{d}\omega = f\epsilon = *f. \tag{E.6}$$

Taking the dual of both sides gives

$$f = (-1)^s * *f = (-1)^s * \mathrm{d}\omega. \tag{E.7}$$

In our case,

$$\begin{aligned}*\mathrm{d}\omega &= *\mathrm{d} * V \\ &= \frac{1}{n!}\epsilon^{\lambda\mu_1\cdots\mu_{n-1}}\left(n\epsilon_{\nu[\mu_1\cdots\mu_{n-1}}\nabla_{\lambda]}V^{\nu}\right) \\ &= \frac{1}{(n-1)!}(-1)^s(n-1)!\delta^{\lambda}_{\nu}\nabla_{\lambda}V^{\nu} \\ &= (-1)^s\nabla_{\nu}V^{\nu}.\end{aligned} \tag{E.8}$$

Finally we recall that the Levi–Civita tensor is simply the volume element,

$$\begin{aligned}\epsilon &= \sqrt{|g|}\,\mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^n \\ &= \sqrt{|g|}\,d^n x.\end{aligned} \tag{E.9}$$

Putting it all together, we find

$$\mathrm{d}\omega = \nabla_{\nu}V^{\nu}\sqrt{|g|}\,d^n x. \tag{E.10}$$

So the exterior derivative of an $(n-1)$-form on an n manifold is just a slick way of representing the divergence of a vector (times the metric volume element).

To make sense of the right-hand side of (E.1), we recall from the previous Appendix that the induced volume element on a hypersurface (such as the boundary) is given by

$$\widehat{\epsilon} = \sqrt{|\gamma|}\,d^{n-1}y, \tag{E.11}$$

where γ_{ij} is the induced metric on the boundary in coordinates y^i. The components of $\widehat{\epsilon}$ in the x^{μ} coordinates on M are

$$\widehat{\epsilon}_{\mu_1\cdots\mu_{n-1}} = n^{\lambda}\epsilon_{\lambda\mu_1\cdots\mu_{n-1}}, \tag{E.12}$$

where n^μ is the unit normal to the boundary. For a general hypersurface, the sign of n^μ is arbitrary; when the hypersurface is the boundary of a region, however, we have a notion of inward-pointing and outward-pointing. A crucial point is that, to correctly recover Stokes's theorem, n^μ should be chosen to be inward-pointing if the boundary is timelike, and outward-pointing if it's spacelike. Since ω is an $(n-1)$-form, it must be proportional to $\widehat{\epsilon}$ when restricted to the $(n-1)$-dimensional boundary. Following in the path of the previous paragraph, we derive

$$\omega = n_\mu V^\mu \sqrt{|\gamma|}\, d^{n-1}y. \tag{E.13}$$

Stokes's theorem therefore relates the divergence of the vector field to its value on the boundary:

$$\boxed{\int_M d^n x \sqrt{|g|}\, \nabla_\mu V^\mu = \int_{\partial M} d^{n-1} y \sqrt{|\gamma|}\, n_\mu V^\mu.} \tag{E.14}$$

This is the most common version of Stokes's theorem in general relativity.

You shouldn't get the impression that we need to descend to index notation to put Stokes's theorem to use. As a simple counterexample, let's show that the charge associated with a conserved current is "conserved" in a very general sense: Not only is it independent of time in some specific coordinate system, but also the charge passing through a spacelike hypersurface Σ is (under reasonable assumptions) completely independent of the choice of hypersurface. Start by imagining that we have a current J^μ that is conserved, by which we mean

$$\nabla_\mu J^\mu = 0. \tag{E.15}$$

In terms of the one-form $J_\mu = g_{\mu\nu} J^\nu$, we can translate the conservation condition into

$$\mathrm{d}(*J) = 0. \tag{E.16}$$

We then define the charge passing through a hypersurface Σ via

$$Q_\Sigma = -\int_\Sigma *J. \tag{E.17}$$

Typically we will choose Σ to be a hypersurface of constant time, so that Q_Σ is the total charge throughout space at that moment in time; but the formula is applicable more generally. The minus sign is a convention, which can be understood by converting (temporarily) to components. Comparing to (E.2) and (E.13), we can turn (E.17) into

$$Q_\Sigma = -\int_\Sigma d^{n-1} y \sqrt{|\gamma|}\, n_\mu J^\mu. \tag{E.18}$$

We see that the minus sign serves to compensate for the minus sign that the time component of n^μ picks up when we lower the index, so that a positive charge

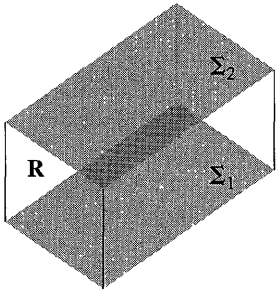

FIGURE E.1 A region R of spacetime with spatial boundaries at infinity; the future and past boundaries include the two spatial hypersurfaces Σ_1 and Σ_2.

density $\rho = J^0$ yields a positive integrated total charge. Next imagine a four-dimensional spacetime region R, defined as the region between two spatial hypersurfaces Σ_1 and Σ_2, as shown in Figure E.1; the part of the boundary connecting these two hypersurfaces is assumed to be off at infinity where all of the fields vanish, and can be ignored. The conservation law (E.16) and Stokes's theorem (E.1) then give us

$$\begin{aligned} 0 &= \int_R \mathrm{d}(*J) \\ &= \int_{\partial R} *J \\ &= \int_{\Sigma_1} *J - \int_{\Sigma_2} *J \\ &= Q_1 - Q_2. \end{aligned} \tag{E.19}$$

The minus sign in the third line is due to the orientation on Σ_2 inherited from R; the normal vector is pointing inward, which is opposite from what would be the conventional choice in an integral over Σ_2. We see that Q_Σ will be the same over any spacelike hypersurface Σ chosen such that the current vanishes at infinity. Thus, Stokes's theorem shows how the existence of a divergenceless current implies the existence of a conserved charge.

Another use of Stokes's theorem (corresponding to the conventional use of Gauss's theorem in three-dimensional Euclidean space) is to actually calculate this charge Q by integrating over the hypersurface. Thinking momentarily about the real world, let's consider Maxwell's equations in a four-dimensional spacetime. These equations describe how the electromagnetic field strength tensor $F_{\mu\nu}$ responds to the conserved current four-vector,

$$\nabla_\mu F^{\nu\mu} = J^\nu. \tag{E.20}$$

We can therefore plug $\nabla_\mu F^{\nu\mu}$ into (E.18) to calculate the charge:

$$Q = -\int_\Sigma d^3y \sqrt{|\gamma|}\, n_\mu \nabla_\nu F^{\mu\nu}. \tag{E.21}$$

Whenever we are faced with the divergence of an antisymmetric tensor field $F^{\mu\nu} = -F^{\nu\mu}$ integrated over a hypersurface Σ, we can follow similar steps to those used to arrive at (E.14), to relate the divergence to the value of $F^{\mu\nu}$ on the boundary, this time at spatial infinity (if the hypersurface is timelike):

$$\boxed{\int_\Sigma d^{n-1}y \sqrt{|\gamma|} n_\mu \nabla_\nu F^{\mu\nu} = \int_{\partial\Sigma} d^{n-2}z \sqrt{|\gamma^{(\partial\Sigma)}|} n_\mu \sigma_\nu F^{\mu\nu},} \tag{E.22}$$

where the z^a's are coordinates on $\partial\Sigma$, $\gamma_{ab}^{(\partial\Sigma)}$ is the induced metric on $\partial\Sigma$, and σ^μ is the unit normal to $\partial\Sigma$. You might worry about the integral over $\partial\Sigma$, since the

boundary of a boundary is zero; but Σ is not the entire boundary of any region, just a piece of one, so it can certainly have a boundary of its own.

Just to make sure we know what we're doing, let's verify that we can actually recover the charge of a point particle in Minkowski space. We write the metric in polar coordinates,

$$ds^2 = -\mathrm{d}t^2 + \mathrm{d}r^2 + r^2\mathrm{d}\theta^2 + r^2 \sin^2\theta \, \mathrm{d}\phi^2. \tag{E.23}$$

The electric field of a charge q in our units (Lorentz–Heaviside conventions, where there are no 4π's in Maxwell's equations) is

$$E^r = \frac{q}{4\pi r^2}, \tag{E.24}$$

with other components vanishing; this is related to the field strength tensor by

$$F^{tr} = -F^{rt} = E^r. \tag{E.25}$$

The unit normal vectors are

$$n^\mu = (1, 0, 0, 0), \quad \sigma^\mu = (0, 1, 0, 0), \tag{E.26}$$

so that

$$n_\mu \sigma_\nu F^{\mu\nu} = -E^r = -\frac{q}{4\pi r^2}. \tag{E.27}$$

The metric on the two-sphere at spatial infinity is

$$\gamma^{(S^2)}_{ab} \mathrm{d}z^a \mathrm{d}z^b = r^2\mathrm{d}\theta^2 + r^2 \sin^2\theta \, \mathrm{d}\phi^2, \tag{E.28}$$

so the volume element is

$$d^2z\sqrt{\gamma^{(S^2)}} = r^2 \sin\theta \, d\theta \, d\phi. \tag{E.29}$$

Plugging (E.27), (E.29), and (E.21) into (E.22) gives

$$\begin{aligned} Q &= -\lim_{r\to\infty} \int_{S^2} d\theta \, d\phi \, r^2 \sin\theta \left(-\frac{q}{4\pi r^2}\right) \\ &= q, \end{aligned} \tag{E.30}$$

which is just the answer we're looking for.

APPENDIX

F Geodesic Congruences

In Section 3.10 we derived the geodesic deviation equation, governing the evolution of a separation vector connecting a one-parameter family of neighboring geodesics. A more comprehensive picture of the behavior of neighboring geodesics comes from considering not just a one-parameter family, but an entire **congruence** of geodesics. A congruence is a set of curves in an open region of spacetime such that every point in the region lies on precisely one curve. We can think of a geodesic congruence as tracing the paths of a set of noninteracting particles moving through spacetime with nonintersecting paths. If the geodesics cross, the congruence necessarily comes to an end at that point. Clearly, in a multidimensional congruence there is a lot of information to keep track of; we will be interested in the local behavior in the neighborhood of a single geodesic, for which things become quite tractable.

Let $U^\mu = dx^\mu/d\tau$ be the tangent vector field to a four-dimensional timelike geodesic congruence; equivalently, the four-velocity field of some pressureless fluid. (If the fluid were not pressureless, integral curves of U^μ would not in general describe geodesics.) Null geodesics present special problems, which we will return to later; for now stick with the timelike case. For reference we recall that the tangent field is normalized and obeys the geodesic equation:

$$U_\mu U^\mu = -1, \quad U^\lambda \nabla_\lambda U^\mu = 0. \tag{F.1}$$

When we discussed the geodesic deviation equation in Section 3.10, we considered a separation vector V^μ pointing from one geodesic to a neighboring one, and found that it obeyed

$$\frac{DV^\mu}{d\tau} \equiv U^\nu \nabla_\nu V^\mu = B^\mu{}_\nu V^\nu, \tag{F.2}$$

where

$$B^\mu{}_\nu = \nabla_\nu U^\mu. \tag{F.3}$$

(In Chapter 3 we used T instead of U, and S instead of V.) The tensor $B_{\mu\nu}$ therefore can be thought of as measuring the failure of V^μ to be parallel-transported along the congruence; in other words, it describes the extent to which neighboring geodesics deviate from remaining perfectly parallel.

To deal with an entire congruence, rather than just a one-parameter family of curves, we can imagine setting up a set of three normal vectors orthogonal to our

timelike geodesics, and following their evolution. The failure of this set of vectors to be parallel-transported will tell us how nearby geodesics in the congruence are evolving. Equivalently, we can imagine a small sphere of test particles centered at some point, and we want to describe quantitatively the evolution of these particles with respect to their central geodesic. Fortunately, all we have to keep track of is the behavior of $B_{\mu\nu}$.

Given the vector field U^μ, at each point p we consider the subspace of T_pM corresponding to vectors normal to U^μ. Any vector in T_pM can be projected into this subspace via the projection tensor

$$P^\mu{}_\nu = \delta^\mu_\nu + U^\mu U_\nu, \tag{F.4}$$

familiar from our discussion of submanifolds in Appendix D. In this case we are not projecting onto a submanifold, only onto a vector subspace of the tangent space, but the idea is the same. We notice that $B_{\mu\nu}$ is already in the normal subspace, since

$$\begin{aligned} U^\mu B_{\mu\nu} &= U^\mu \nabla_\nu U_\mu = 0 \\ U^\nu B_{\mu\nu} &= U^\nu \nabla_\nu U_\mu = 0. \end{aligned} \tag{F.5}$$

The first of these follows from $\nabla_\nu(U^\mu U_\mu) = \nabla_\nu(-1) = 0$, while the second follows from the geodesic equation. We should not confuse $B_{\mu\nu}$ with the extrinsic curvature $K_{\mu\nu}$ from (D.53); the difference is that our tangent vector field U^μ will generally not be orthogonal to any hypersurface.

As a $(0, 2)$ tensor, $B_{\mu\nu}$ can be decomposed into symmetric and antisymmetric parts, and the symmetric part can further be decomposed into a trace and a trace-free part. Since $B_{\mu\nu}$ is in the normal subspace, we can use $P_{\mu\nu}$ to take the trace in this decomposition. The result can be written

$$B_{\mu\nu} = \tfrac{1}{3}\theta P_{\mu\nu} + \sigma_{\mu\nu} + \omega_{\mu\nu}. \tag{F.6}$$

Here we have introduced three quantities describing the decomposition, starting with the **expansion** θ of the congruence,

$$\theta = P^{\mu\nu} B_{\mu\nu} = \nabla_\mu U^\mu, \tag{F.7}$$

which is simply the trace of $B_{\mu\nu}$. The expansion describes the change in volume of the sphere of test particles centered on our geodesic. It is clearly a scalar, which makes sense, since the overall expansion/contraction of the volume is described by a single number. The **shear** $\sigma_{\mu\nu}$ is given by

$$\sigma_{\mu\nu} = B_{(\mu\nu)} - \tfrac{1}{3}\theta P_{\mu\nu}. \tag{F.8}$$

It is symmetric and traceless. The shear represents a distortion in the shape of our collection of test particles, from an initial sphere into an ellipsoid; symmetry represents the fact that elongation along (say) the x-direction is the same as

elongation along the $-x$ direction. Finally we have the **rotation** $\omega_{\mu\nu}$, given by

$$\omega_{\mu\nu} = B_{[\mu\nu]}. \tag{F.9}$$

It is an antisymmetric tensor, which also makes sense; the xy component (for example) describes a rotation about the z axis, while the yx component describes a rotation in the opposite sense around the same axis.

The evolution of our congruence is described by the covariant derivative of these quantities along the path, $D/d\tau = U^\sigma \nabla_\sigma$. We can straightforwardly calculate this for the entire tensor $B_{\mu\nu}$, and then take the appropriate decomposition. We have

$$\begin{aligned}\frac{DB_{\mu\nu}}{d\tau} \equiv U^\sigma \nabla_\sigma B_{\mu\nu} &= U^\sigma \nabla_\sigma \nabla_\nu U_\mu \\ &= U^\sigma \nabla_\nu \nabla_\sigma U_\mu + U^\sigma R^\lambda{}_{\mu\nu\sigma} U_\lambda \\ &= \nabla_\nu (U^\sigma \nabla_\sigma U_\mu) - (\nabla_\nu U^\sigma)(\nabla_\sigma U_\mu) - R_{\lambda\mu\nu\sigma} U^\sigma U^\lambda \\ &= -B^\sigma{}_\nu B_{\mu\sigma} - R_{\lambda\mu\nu\sigma} U^\sigma U^\lambda. \end{aligned} \tag{F.10}$$

Taking the trace of this equation yields an evolution equation for the expansion,

$$\boxed{\frac{d\theta}{d\tau} = -\frac{1}{3}\theta^2 - \sigma_{\mu\nu}\sigma^{\mu\nu} + \omega_{\mu\nu}\omega^{\mu\nu} - R_{\mu\nu}U^\mu U^\nu.} \tag{F.11}$$

This is **Raychaudhuri's equation**, and plays a crucial role in the proofs of the singularity theorems. [Sometimes the demand that the congruence obey the geodesic equation is dropped; this simply adds a term $\nabla_\mu(U^\nu \nabla_\nu U^\mu)$ to the right-hand side.] Similarly, the symmetric trace-free part of (F.10) is

$$\begin{aligned}\frac{D\sigma_{\mu\nu}}{d\tau} = &-\frac{2}{3}\theta\sigma_{\mu\nu} - \sigma_{\mu\alpha}\sigma^\alpha{}_\nu - \omega_{\mu\rho}\omega^\rho{}_\nu + \frac{1}{3}P_{\mu\nu}(\sigma_{\alpha\beta}\sigma^{\alpha\beta} - \omega_{\alpha\beta}\omega^{\alpha\beta}) \\ &+ C_{\alpha\nu\mu\beta}U^\alpha U^\beta + \frac{1}{2}\bar{R}_{\mu\nu}, \end{aligned} \tag{F.12}$$

where $\bar{R}_{\mu\nu}$ is the spatially-projected trace-free part of $R_{\mu\nu}$,

$$\bar{R}_{\mu\nu} = P^\alpha{}_\mu P^\beta{}_\nu R_{\alpha\beta} - \tfrac{1}{3}P_{\mu\nu}P^{\alpha\beta}R_{\alpha\beta}, \tag{F.13}$$

and the antisymmetric part of (F.10) is

$$\frac{D\omega_{\mu\nu}}{d\tau} = -\frac{2}{3}\theta\omega_{\mu\nu} + \sigma_\mu{}^\alpha\omega_{\nu\alpha} - \sigma_\nu{}^\alpha\omega_{\mu\alpha}. \tag{F.14}$$

These equations do not get used as frequently as Raychaudhuri's equation, but they're nice to have around.

Let's give a brief example of the way in which Raychaudhuri's equation gets used. First notice that, since the shear and rotation are both "spatial" tensors, we have

$$\sigma_{\mu\nu}\sigma^{\mu\nu} \geq 0, \quad \omega_{\mu\nu}\omega^{\mu\nu} \geq 0. \tag{F.15}$$

Next, notice that the last term in (F.11) is just what appears if we combine Einstein's equation with the Strong Energy Condition; from Einstein's equation we know

$$R_{\mu\nu}U^\mu U^\nu = 8\pi G\left(T_{\mu\nu} - \tfrac{1}{2}Tg_{\mu\nu}\right)U^\mu U^\nu, \tag{F.16}$$

and the SEC demands that the right-hand side of this expression be nonnegative for any timelike U^μ. We therefore have

$$R_{\mu\nu}U^\mu U^\nu \geq 0 \tag{F.17}$$

if the SEC holds. Finally, note that $\omega_{\mu\nu} = 0$ if and only if the vector field U^μ is orthogonal to a family of hypersurfaces. This follows straightforwardly from the facts that the rotation is a spatial tensor ($U^\mu\omega_{\mu\nu} = 0$), and by Frobenius's theorem a necessary and sufficient condition for a vector field U^μ to be hypersurface-orthogonal is $U_{[\mu}\nabla_\nu U_{\rho]}$; the details are left for you to check. Therefore, if we have a congruence whose tangent field is hypersurface-orthogonal, in a spacetime obeying Einstein's equations and the SEC, Raychaudhuri's equation implies

$$\frac{d\theta}{d\tau} \leq -\frac{1}{3}\theta^2. \tag{F.18}$$

This equation is easily integrated to obtain

$$\theta^{-1}(\tau) \geq \theta_0^{-1} + \tfrac{1}{3}\tau. \tag{F.19}$$

Consider a hypersurface-orthogonal congruence, which is initially converging ($\theta_0 < 0$) rather than expanding. Then (F.19) tells us convergence will continue, and we must hit a caustic (a place where geodesics cross) in a finite proper time $\tau \leq -3\theta_0^{-1}$. In other words, matter obeying the SEC can never begin to push geodesics apart, it can only increase the rate at which they are converging. Of course, this result only applies to some arbitrarily-chosen congruence, and the appearance of caustics certainly doesn't indicate any singularity in the spacetime (geodesics cross all the time, even in flat spacetime). But many of the proofs of the singularity theorems take advantage of this property of the Raychaudhuri equation to show that spacetime must be geodesically incomplete in some way.

We turn next to the behavior of congruences of null geodesics. These are trickier, essentially because our starting point (studying the evolution of vectors in a three-dimensional subspace normal to the tangent field) doesn't make as much sense, since the tangent vector of a null curve is normal to itself. Instead, in the null case what we care about is the evolution of vectors in a *two-dimensional* subspace of "spatial" vectors normal to the null tangent vector field $k^\mu = dx^\mu/d\lambda$.

Unfortunately, there is no unique way to define this subspace, as observers in different Lorentz frames will have different notions of what constitutes a spatial vector. Faced with this dilemma, we have two sensible approaches. A slick approach would be to define an abstract two-dimensional vector space by starting with the three-dimensional space of vectors orthogonal to k^μ, and then taking equivalence classes where two vectors are equivalent if they differ by a multiple of k^μ. The grungier approach, which we will follow, is simply to choose a second "auxiliary" null vector l^μ, which (in some frame) points in the opposite spatial direction to k^μ, normalized such that

$$l^\mu l_\mu = 0, \quad l^\mu k_\mu = -1. \tag{F.20}$$

We furthermore demand that the auxiliary vector be parallel-transported,

$$k^\mu \nabla_\mu l^\nu = 0, \tag{F.21}$$

which is compatible with (F.20) because parallel transport preserves inner products. The auxiliary null vector l^μ is by no means unique, since as we've just noted the idea of pointing in opposite spatial directions is frame-dependent. Nevertheless, we can make a choice and hope that important quantities are independent of the arbitrary choice. Having done so, the two-dimensional space of normal vectors we are interested in, called $T_\perp$, consists simply of those vectors V^μ that are orthogonal to both k^μ and l^μ,

$$T_\perp = \{V^\mu | V^\mu k_\mu = 0, V^\mu l_\mu = 0\}. \tag{F.22}$$

Our task now is to follow the evolution of deviation vectors living in this subspace, which represent a family of neighboring null geodesics.

Projecting into the normal subspace $T_\perp$ requires a slightly modified definition of the projection tensor; it turns out that

$$Q_{\mu\nu} = g_{\mu\nu} + k_\mu l_\nu + k_\nu l_\mu \tag{F.23}$$

does the trick. Namely, $Q_{\mu\nu}$ will act like the metric when acting on vectors V^μ, W^μ in $T_\perp$, while annihilating anything proportional to k^μ or l^μ. Some useful properties of this projection tensor include

$$\begin{aligned} Q_{\mu\nu} V^\nu W^\nu &= g_{\mu\nu} V^\mu W^\nu \\ Q^\mu{}_\nu V^\nu &= V^\mu \\ Q^\mu{}_\nu k^\nu &= 0 \\ Q^\mu{}_\nu l^\nu &= 0 \\ Q^\mu{}_\nu Q^\nu{}_\sigma &= Q^\mu{}_\sigma \\ k^\sigma \nabla_\sigma Q^\mu{}_\nu &= 0. \end{aligned} \tag{F.24}$$

Just as for timelike geodesics, the failure of a normal deviation vector V^μ to be parallel-propagated is governed by the tensor $B^\mu{}_\nu = \nabla_\nu k^\mu$, in the sense that

$$\frac{DV^\mu}{d\lambda} \equiv k^\nu \nabla_\nu V^\mu = B^\mu{}_\nu V^\nu. \tag{F.25}$$

However, in the null case the tensor $B_{\mu\nu}$ is actually more than we need; the relevant information is completely contained in the projected version,

$$\widehat{B}^\mu{}_\nu = Q^\mu{}_\alpha Q^\beta{}_\nu B^\alpha{}_\beta. \tag{F.26}$$

To see this, we simply play around with (F.25), using the various properties in (F.24):

$$\begin{aligned}
\frac{DV^\mu}{d\lambda} &= k^\nu \nabla_\nu V^\mu \\
&= k^\nu \nabla_\nu (Q^\mu{}_\rho V^\rho) \\
&= Q^\mu{}_\rho k^\nu \nabla_\nu V^\rho \\
&= Q^\mu{}_\rho B^\rho{}_\nu V^\nu \\
&= Q^\mu{}_\rho B^\rho{}_\nu Q^\nu{}_\sigma V^\sigma \\
&= \widehat{B}^\mu{}_\sigma V^\sigma.
\end{aligned} \tag{F.27}$$

So we only have to keep track of the evolution of this projected tensor, not the full $B_{\mu\nu}$.

To understand that evolution, we again decompose into the expansion, shear, and rotation:

$$\widehat{B}_{\mu\nu} = \tfrac{1}{2}\theta Q_{\mu\nu} + \widehat{\sigma}_{\mu\nu} + \widehat{\omega}_{\mu\nu}, \tag{F.28}$$

where

$$\begin{aligned}
\theta &= Q^{\mu\nu}\widehat{B}_{\mu\nu} = \widehat{B}^\mu{}_\mu \\
\widehat{\sigma}_{\mu\nu} &= \widehat{B}_{(\mu\nu)} - \tfrac{1}{2}\theta Q_{\mu\nu} \\
\widehat{\omega}_{\mu\nu} &= \widehat{B}_{[\mu\nu]}.
\end{aligned} \tag{F.29}$$

We find factors of $\frac{1}{2}$ rather than $\frac{1}{3}$ because our normal space $T_\perp$ is two-dimensional, reflected in the fact that $Q^{\mu\nu}Q_{\mu\nu} = 2$. As in the timelike case, $\widehat{\omega}_{\mu\nu} = 0$ is a necessary and sufficient condition for the congruence to be hypersurface-orthogonal. The evolution of $\widehat{B}_{\mu\nu}$ along the path is given by

$$\begin{aligned}
\frac{D\widehat{B}_{\mu\nu}}{d\lambda} \equiv k^\sigma \nabla_\sigma \widehat{B}_{\mu\nu} &= k^\sigma \nabla_\sigma (Q^\alpha{}_\mu Q^\beta{}_\nu \nabla_\alpha k_\beta) \\
&= Q^\alpha{}_\mu Q^\beta{}_\nu k^\sigma \nabla_\sigma \nabla_\alpha k_\beta
\end{aligned}$$

$$\begin{aligned} &= -Q^{\alpha}{}_{\mu} Q^{\beta}{}_{\nu} (B_{\alpha}{}^{\sigma} B_{\beta\sigma} + R_{\alpha\lambda\beta\sigma} k^{\lambda} k^{\sigma}) \\ &= -\widehat{B}_{\mu}{}^{\sigma} \widehat{B}_{\nu\sigma} - Q^{\alpha}{}_{\mu} Q^{\beta}{}_{\nu} R_{\mu\lambda\nu\sigma} k^{\lambda} k^{\sigma}. \end{aligned} \tag{F.30}$$

Continuing to follow our previous logic, we can take the trace of this equation to find an evolution equation for the expansion of null geodesics,

$$\frac{d\theta}{d\lambda} = -\frac{1}{2}\theta^2 - \widehat{\sigma}_{\mu\nu}\widehat{\sigma}^{\mu\nu} + \widehat{\omega}_{\mu\nu}\widehat{\omega}^{\mu\nu} - R_{\mu\nu}k^{\mu}k^{\nu}. \tag{F.31}$$

Happily, this equation turns out to be completely independent of our arbitrarily chosen auxiliary vector l^{μ}. First, the expansion itself is independent of l^{μ}, as we easily verify:

$$\begin{aligned} \theta &= Q^{\mu\nu}\widehat{B}_{\mu\nu} \\ &= Q^{\mu\nu}B_{\mu\nu} \\ &= g^{\mu\nu}B_{\mu\nu}, \end{aligned} \tag{F.32}$$

where the second line follows from $Q^{\mu\nu}Q^{\alpha}{}_{\nu} = Q^{\mu\alpha}$, and the third from $k^{\mu}B_{\mu\nu} = k^{\nu}B_{\mu\nu} = 0$. (This is why we never put a hat on θ to begin with.) Second, both $\widehat{\sigma}_{\mu\nu}\widehat{\sigma}^{\mu\nu}$ and $\widehat{\omega}_{\mu\nu}\widehat{\omega}^{\mu\nu}$ are likewise independent of l^{μ} (as you are welcome to verify), even though $\widehat{\sigma}_{\mu\nu}$ and $\widehat{\omega}_{\mu\nu}$ themselves are not. Finally, the projection tensors dropped out of the curvature-tensor piece when we took the trace. We therefore have a well-defined notion of the evolution of the expansion, independent of any arbitrary choices we made. Notice that, because k^{μ} is null, Einstein's equation implies

$$\begin{aligned} R_{\mu\nu}k^{\mu}k^{\nu} &= 8\pi G\left(T_{\mu\nu} - \tfrac{1}{2}T g_{\mu\nu}\right)k^{\mu}k^{\nu} \\ &= 8\pi G T_{\mu\nu}k^{\mu}k^{\nu}. \end{aligned} \tag{F.33}$$

For this to be nonnegative, we need only invoke the Null Energy Condition, which is the least restrictive of all the energy conditions we discussed in Chapter 3. Thus, null geodesics tend to converge to caustics under more general circumstances than timelike ones.

We can continue on to get evolution equations for the shear,

$$\frac{D\widehat{\sigma}_{\mu\nu}}{d\lambda} = -\theta\widehat{\sigma}_{\mu\nu} - Q^{\alpha}{}_{\mu}Q^{\beta}{}_{\nu}C_{\mu\lambda\nu\sigma}k^{\lambda}k^{\sigma}, \tag{F.34}$$

and for the rotation,

$$\frac{D\widehat{\omega}_{\mu\nu}}{d\lambda} = -\theta\widehat{\omega}_{\mu\nu}. \tag{F.35}$$

These equations are less natural than the one for the expansion, since the shear and rotation do depend on our choice of l^{μ}; nevertheless, they can be useful in specific circumstances.

APPENDIX

G Conformal Transformations

A **conformal transformation** is essentially a local change of scale. Since distances are measured by the metric, such transformations are implemented by multiplying the metric by a spacetime-dependent (nonvanishing) function:

$$\tilde{g}_{\mu\nu} = \omega^2(x) g_{\mu\nu}, \tag{G.1}$$

or equivalently

$$\tilde{ds}^2 = \omega^2(x) ds^2, \tag{G.2}$$

for some nonvanishing function $\omega(x)$. (Here x is used to denote the collection of spacetime coordinates x^μ.) Note that the inverse conformal transformation is trivial: $g_{\mu\nu} = \omega^{-2}\tilde{g}_{\mu\nu}$. Transformations of this sort have a number of uses in GR; our favorite purposes will be to change dynamical variables in scalar-tensor theories (as in Section 4.8), and to remap spacetimes into convenient conformal diagrams (as in the following Appendix).

We first mention one critical fact: *null curves are left invariant by conformal transformations*. By this we mean simply that, if $x^\mu(\lambda)$ is a curve that is null with respect to $g_{\mu\nu}$, it will also be null with respect to $\tilde{g}_{\mu\nu}$. This follows immediately once we understand that a curve $x^\mu(\lambda)$ is null if and only if its tangent vector $dx^\mu/d\lambda$ is null,

$$g_{\mu\nu}\frac{dx^\mu}{d\lambda}\frac{dx^\nu}{d\lambda} = 0. \tag{G.3}$$

Then in the conformally-related metric we have

$$\tilde{g}_{\mu\nu}\frac{dx^\mu}{d\lambda}\frac{dx^\nu}{d\lambda} = \omega^2(x) g_{\mu\nu}\frac{dx^\mu}{d\lambda}\frac{dx^\nu}{d\lambda} = 0. \tag{G.4}$$

Thus, curves that are null as defined by one metric will also be null as defined by any conformally-related metric. We may say that "conformal transformations leave light cones invariant." (Indeed, you can check that they leave angles between any two four-vectors invariant, a feature that our conformal transformations share with the familiar conformal transformations of complex analysis.)

Let us next consider how geometrical quantities change under conformal transformations. A conformal transformation is not a change of coordinates, but an actual change of the geometry—timelike geodesics of $\tilde{g}_{\mu\nu}$, for example, will generally differ from timelike geodesics of $g_{\mu\nu}$. However, we can use conformal transformations to change our dynamical variables: anything that is a function of $g_{\mu\nu}$

can be equally well thought of as a function of $\tilde{g}_{\mu\nu}$ and $\omega(x)$. We then say that the quantities are expressed in the **conformal frame**. In this Appendix we collect some expressions for how quantities in the original metric $g_{\mu\nu}$ are related to those in the conformal metric $\tilde{g}_{\mu\nu}$.

We begin by considering the Christoffel symbols. Because the connection coefficients are linear in derivatives of the metric and also linear in the inverse metric, the conformally-transformed connection takes the form

$$\widetilde{\Gamma}^{\rho}_{\mu\nu} = \Gamma^{\rho}_{\mu\nu} + C^{\rho}{}_{\mu\nu}. \tag{G.5}$$

$C^{\rho}{}_{\mu\nu}$ is clearly a tensor, as it is the difference of two connections. An explicit calculation reveals it to be given by

$$C^{\rho}{}_{\mu\nu} = \omega^{-1}\left(\delta^{\rho}_{\mu}\nabla_{\nu}\omega + \delta^{\rho}_{\nu}\nabla_{\mu}\omega - g_{\mu\nu}g^{\rho\lambda}\nabla_{\lambda}\omega\right). \tag{G.6}$$

This formula immediately becomes useful when we consider how the Riemann tensor behaves under conformal transformations. In fact under any change of connection of the form (G.5), we have

$$\widetilde{R}^{\rho}{}_{\sigma\mu\nu} = R^{\rho}{}_{\sigma\mu\nu} + \nabla_{\mu}C^{\rho}{}_{\nu\sigma} - \nabla_{\nu}C^{\rho}{}_{\mu\sigma} + C^{\rho}{}_{\mu\lambda}C^{\lambda}{}_{\nu\sigma} - C^{\rho}{}_{\nu\lambda}C^{\lambda}{}_{\mu\sigma}. \tag{G.7}$$

Thus it is a matter of simply plugging in and grinding away to get

$$\begin{aligned}\widetilde{R}^{\rho}{}_{\sigma\mu\nu} = R^{\rho}{}_{\sigma\mu\nu} &- 2\left(\delta^{\rho}_{[\mu}\delta^{\alpha}_{\nu]}\delta^{\beta}_{\sigma} - g_{\sigma[\mu}\delta^{\alpha}_{\nu]}g^{\rho\beta}\right)\omega^{-1}(\nabla_{\alpha}\nabla_{\beta}\omega)\\ &+ 2\left(2\delta^{\rho}_{[\mu}\delta^{\alpha}_{\nu]}\delta^{\beta}_{\sigma} - 2g_{\sigma[\mu}\delta^{\alpha}_{\nu]}g^{\rho\beta} + g_{\sigma[\mu}\delta^{\rho}_{\nu]}g^{\alpha\beta}\right)\omega^{-2}(\nabla_{\alpha}\omega)(\nabla_{\beta}\omega).\end{aligned} \tag{G.8}$$

Contracting the first and third indices yields the Ricci tensor,

$$\begin{aligned}\widetilde{R}_{\sigma\nu} = R_{\sigma\nu} &- \left[(n-2)\delta^{\alpha}_{\sigma}\delta^{\beta}_{\nu} + g_{\sigma\nu}g^{\alpha\beta}\right]\omega^{-1}(\nabla_{\alpha}\nabla_{\beta}\omega)\\ &+ \left[2(n-2)\delta^{\alpha}_{\sigma}\delta^{\beta}_{\nu} - (n-3)g_{\sigma\nu}g^{\alpha\beta}\right]\omega^{-2}(\nabla_{\alpha}\omega)(\nabla_{\beta}\omega),\end{aligned} \tag{G.9}$$

where n is the number of dimensions. Raising an index (with $\tilde{g}^{\mu\nu} = \omega^{-2}g^{\mu\nu}$) and contracting again gets us the curvature scalar,

$$\widetilde{R} = \omega^{-2}R - 2(n-1)g^{\alpha\beta}\omega^{-3}(\nabla_{\alpha}\nabla_{\beta}\omega) - (n-1)(n-4)g^{\alpha\beta}\omega^{-4}(\nabla_{\alpha}\omega)(\nabla_{\beta}\omega). \tag{G.10}$$

Another useful quantity is the covariant derivative of a scalar field ϕ. The first covariant derivative is equal in the original or conformal frame, since they are both equal to the partial derivative:

$$\widetilde{\nabla}_{\mu}\phi = \nabla_{\mu}\phi = \partial_{\mu}\phi. \tag{G.11}$$

The second derivative, however, involves the Christoffel symbol, and therefore has a nontrivial transformation:

$$\widetilde{\nabla}_\mu \widetilde{\nabla}_\nu \phi = \nabla_\mu \nabla_\nu \phi - \left(\delta^\alpha_\mu \delta^\beta_\nu + \delta^\beta_\mu \delta^\alpha_\nu - g_{\mu\nu} g^{\alpha\beta}\right) \omega^{-1} (\nabla_\alpha \omega)(\nabla_\beta \phi). \tag{G.12}$$

We can contract this with $\tilde{g}^{\mu\nu}$ to obtain the D'Alembertian,

$$\widetilde{\Box}\phi = \omega^{-2}\Box\phi + (n-2) g^{\alpha\beta} \omega^{-3} (\nabla_\alpha \omega)(\nabla_\beta \phi). \tag{G.13}$$

Finally, we may want to go backward, and express quantities in the original metric in terms of the conformal metric. This is simply a matter of tedious computation, the answers to which are reproduced here for convenience. The curvature tensor and its contractions are

$$\begin{aligned} R^\rho{}_{\sigma\mu\nu} = \widetilde{R}^\rho{}_{\sigma\mu\nu} &+ 2\left(\delta^\rho_{[\mu}\delta^\alpha_{\nu]}\delta^\beta_\sigma - \tilde{g}_{\sigma[\mu}\delta^\alpha_{\nu]}\tilde{g}^{\rho\beta}\right)\omega^{-1}(\widetilde{\nabla}_\alpha \widetilde{\nabla}_\beta \omega) \\ &+ 2\tilde{g}_{\sigma[\mu}\delta^\rho_{\nu]}\tilde{g}^{\alpha\beta}\omega^{-2}(\widetilde{\nabla}_\alpha\omega)(\widetilde{\nabla}_\beta\omega), \end{aligned} \tag{G.14}$$

$$\begin{aligned} R_{\sigma\nu} = \widetilde{R}_{\sigma\nu} &+ \left[(n-2)\delta^\alpha_\sigma\delta^\beta_\nu + \tilde{g}_{\sigma\nu}\tilde{g}^{\alpha\beta}\right]\omega^{-1}(\widetilde{\nabla}_\alpha\widetilde{\nabla}_\beta\omega) \\ &- (n-1)\tilde{g}_{\sigma\nu}\tilde{g}^{\alpha\beta}\omega^{-2}(\widetilde{\nabla}_\alpha\omega)(\widetilde{\nabla}_\beta\omega), \end{aligned} \tag{G.15}$$

and

$$R = \omega^2\widetilde{R} + 2(n-1)\tilde{g}^{\alpha\beta}\omega(\widetilde{\nabla}_\alpha\widetilde{\nabla}_\beta\omega) - n(n-1)\tilde{g}^{\alpha\beta}(\widetilde{\nabla}_\alpha\omega)(\widetilde{\nabla}_\beta\omega), \tag{G.16}$$

while the covariant derivatives of a scalar field are given by

$$\nabla_\mu\nabla_\nu\phi = \widetilde{\nabla}_\mu\widetilde{\nabla}_\nu\phi + \left(\delta^\alpha_\mu\delta^\beta_\nu + \delta^\beta_\mu\delta^\alpha_\nu - \tilde{g}_{\mu\nu}\tilde{g}^{\alpha\beta}\right)\omega^{-1}(\widetilde{\nabla}_\alpha\omega)(\widetilde{\nabla}_\beta\phi) \tag{G.17}$$

and

$$\Box\phi = \omega^2\widetilde{\Box}\phi - (n-2)\tilde{g}^{\alpha\beta}\omega(\widetilde{\nabla}_\alpha\omega)(\widetilde{\nabla}_\beta\phi). \tag{G.18}$$

G.1 ■ EXERCISES

1. Show that conformal transformations leave null geodesics invariant, that is, that the null geodesics of $g_{\mu\nu}$ are the same as those of $\omega^2 g_{\mu\nu}$. (We already know that they leave null *curves* invariant; you have to show that the transformed curves still are geodesics.) What is the relationship between the affine parameters in the original and conformal metrics?

9. Show that in two dimensions, a conformal transformation can always be found (provided that the operator $\nabla^\mu\nabla_\mu$ is invertible) such that the curvature of the transformed metric vanishes, at least in some coordinate chart. (It can't in general be done simultaneously over the entire manifold.) This means that any two-dimensional metric can be written locally as a flat metric multiplied by a conformal factor.

10. Suppose that two metrics are related by an overall conformal transformation of the form

$$\tilde{g}_{\mu\nu} = e^{\alpha(x)} g_{\mu\nu}. \tag{G.19}$$

(a) Show that if ξ^μ is a Killing vector for the metric $g_{\mu\nu}$, then it is a conformal Killing vector for the metric $\widetilde{g}_{\mu\nu}$. A **conformal Killing vector** obeys the equation

$$\nabla_\mu \xi_\nu + \nabla_\nu \xi_\mu = (\nabla_\lambda \alpha)\xi^\lambda g_{\mu\nu}. \tag{G.20}$$

(b) Show that $\xi_\mu k^\mu$ is constant along photon geodesics in $\widetilde{g}_{\mu\nu}$. Here k^μ is the photon's 4-momentum.

(c) Show that the conformal time $\eta = \int dt/R(t)$ is associated with a conformal Killing vector $\xi = \partial_\eta$.

(d) Use part (c) to rederive the relationship between the scale factor and redshift.

APPENDIX

H Conformal Diagrams

Curved spacetime manifolds can in principle be impossibly complex; fortunately, we may often approximate physically realistic situations by manifolds with high degrees of symmetry (especially spherical symmetry). Even symmetric spacetimes, however, can pose formidable challenges to our powers of visualization, if we try to imagine the global structure of such manifolds. It is therefore useful to be able to draw standardized representations of spacetime diagrams that capture the global properties and causal structure of sufficiently symmetric spacetimes. (By "causal structure" we mean the relationship between the past and future of different events, as defined by their light cones.) An elegant fulfillment of this wish is provided by **conformal diagrams** (or Carter–Penrose, or just Penrose diagrams).

A conformal diagram is simply an ordinary spacetime diagram for a metric on which we have performed a particularly clever coordinate transformation. Since our goal is to portray the causal structure of the spacetime, which is defined by its light cones, "clever" means that the new coordinates $x^{\mu'}$ have a "timelike" coordinate and a "radial" one, with the feature that radial light cones can be consistently portrayed at 45° on a spacetime diagram. In addition, we aim for coordinates in which "infinity" is only a finite coordinate value away, so that the structure of the entire spacetime is immediately apparent.

As explained in the previous Appendix, conformal transformations leave light cones invariant. Since we would like to find coordinates in which light cones are at 45°, we need only find coordinates in which the metric of interest is conformally related to a different metric for which we know that the light cones are at 45°. (Of course the angle at which our light cones are drawn depends on our units, or equivalently how we draw our axes; what we really mean is a set of coordinates T, R in which radial null rays satisfy $dT/dR = \pm 1$.)

Let's begin with Minkowski space to see how the technique works. The Minkowski metric in polar coordinates is

$$ds^2 = -\mathrm{d}t^2 + \mathrm{d}r^2 + r^2 d\Omega^2, \tag{H.1}$$

where $d\Omega^2 = \mathrm{d}\theta^2 + \sin^2\theta \mathrm{d}\phi^2$ is the metric on a unit two-sphere. Here it is already true that we can draw light cones at 45° everywhere (the trajectories $t = \pm r$ are null), but we would like to make the causal structure of the entire spacetime more transparent by switching to coordinates with finite ranges. Nothing unusual will happen to the θ, ϕ coordinates, but we will want to keep careful track of the ranges

of the other two coordinates. To start with of course we have

$$-\infty < t < \infty, \quad 0 \leq r < \infty. \tag{H.2}$$

Technically, the worldline $r = 0$ represents a coordinate singularity and should be covered by a different patch, but we all know what is going on so we'll just act like $r = 0$ is well-behaved.

A first guess (which turns out not to work) might be simply to rescale the timelike and radial coordinates so that they cover a finite range. A good candidate is to use the arctangent, portrayed in Figure H.1, and define $\bar{t} = \arctan t$, $\bar{r} = \arctan r$. The metric then would take the form [using $d \tan x = (1/\cos^2 x)dx$]

$$ds^2 = -\frac{1}{\cos^4 \bar{t}}\, d\bar{t}^2 + \frac{1}{\cos^4 \bar{r}}\, d\bar{r}^2 + \tan^2 \bar{r}\, d\Omega^2, \tag{H.3}$$

with

$$-\frac{\pi}{2} < \bar{t} < \frac{\pi}{2}$$
$$0 \leq \bar{r} < \frac{\pi}{2}. \tag{H.4}$$

The good news is that the new coordinates have finite ranges; the bad news is that the slope of the light cones (given by $d\bar{t}/d\bar{r} = \pm\cos^2 \bar{t}/\cos^2 \bar{r}$) is not equal to ± 1, as we wished. If we were to draw the appropriate spacetime diagram (which you might want to do, just for fun), it would not be clear where null rays traveled, especially at the edges of the spacetime.

The way out of this cul-de-sac is, instead of straightforwardly manipulating the original coordinates t and r, to be even more clever and switch to null coordinates:

$$u = t - r$$
$$v = t + r, \tag{H.5}$$

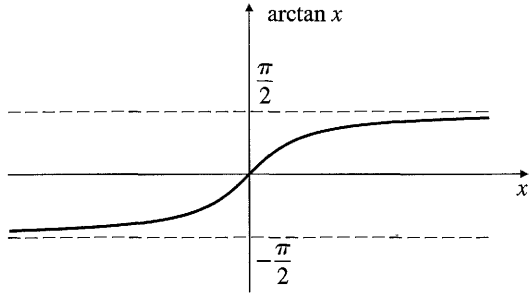

FIGURE H.1 The arctangent maps the real line to a finite interval.

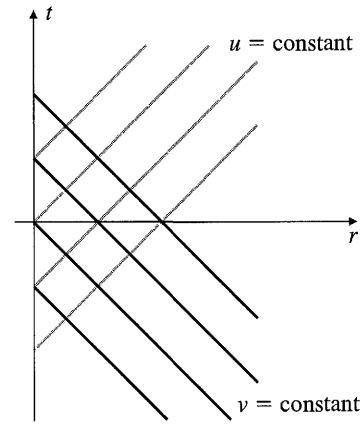

FIGURE H.2 Null radial coordinates on Minkowski space.

with corresponding ranges given by

$$-\infty < u < \infty, \quad -\infty < v < \infty, \quad u \leq v. \tag{H.6}$$

These coordinates are as portrayed in Figure H.2, on which each point represents a 2-sphere of radius $r = \frac{1}{2}(v - u)$. The Minkowski metric in null coordinates is given by

$$ds^2 = -\tfrac{1}{2}(\mathrm{d}u\mathrm{d}v + \mathrm{d}v\mathrm{d}u) + \tfrac{1}{4}(v - u)^2 d\Omega^2. \tag{H.7}$$

Now we use the arctangent to bring infinity into a finite coordinate value, letting

$$\begin{aligned} U &= \arctan u \\ V &= \arctan v, \end{aligned} \tag{H.8}$$

with ranges

$$-\pi/2 < U < \pi/2, \quad -\pi/2 < V < \pi/2, \quad U \leq V. \tag{H.9}$$

We then have

$$\mathrm{d}u\,\mathrm{d}v + \mathrm{d}v\,\mathrm{d}u = \frac{1}{\cos^2 U \cos^2 V}(\mathrm{d}U\,\mathrm{d}V + \mathrm{d}V\,\mathrm{d}U), \tag{H.10}$$

and

$$\begin{aligned} (v - u)^2 = (\tan V - \tan U)^2 &= \frac{1}{\cos^2 U \cos^2 V}(\sin V \cos U - \cos V \sin U)^2 \\ &= \frac{1}{\cos^2 U \cos^2 V} \sin^2(V - U), \end{aligned} \tag{H.11}$$

so that the metric (H.7) in these coordinates is

$$ds^2 = \frac{1}{4\cos^2 U \cos^2 V}\left[-2(\mathrm{d}U\mathrm{d}V + \mathrm{d}V\mathrm{d}U) + \sin^2(V-U)\,d\Omega^2\right]. \quad \text{(H.12)}$$

This form has a certain appeal, since the metric appears as a fairly simple expression multiplied by an overall factor. We can make it even better by transforming back to a timelike coordinate T and a radial coordinate R, via

$$T = V + U, \quad R = V - U, \quad \text{(H.13)}$$

with ranges

$$0 \le R < \pi, \quad |T| + R < \pi. \quad \text{(H.14)}$$

Now the metric is

$$ds^2 = \omega^{-2}(T, R)\left(-\mathrm{d}T^2 + \mathrm{d}R^2 + \sin^2 R\, d\Omega^2\right), \quad \text{(H.15)}$$

where

$$\begin{aligned}\omega &= 2\cos U \cos V \\ &= 2\cos\left[\tfrac{1}{2}(T-R)\right]\cos\left[\tfrac{1}{2}(T+R)\right] \\ &= \cos T + \cos R. \end{aligned} \quad \text{(H.16)}$$

The original Minkowski metric, which we denoted ds^2, may therefore be thought of as related by a conformal transformation to the "unphysical" metric

$$\begin{aligned}\tilde{ds}^2 &= \omega^2(T, R) ds^2 \\ &= -\mathrm{d}T^2 + \mathrm{d}R^2 + \sin^2 R\, d\Omega^2. \end{aligned} \quad \text{(H.17)}$$

This describes the manifold $\mathbf{R} \times S^3$, where the 3-sphere is purely spacelike, perfectly round, and unchanging in time. There is curvature in this metric, unlike in Minkowski spacetime. This shouldn't bother us, since it is unphysical; the true physical metric, obtained by a conformal transformation, is simply flat spacetime, no matter what coordinates we choose. In fact the metric (H.17) is that of the "Einstein static universe," a static solution to Einstein's equation with a perfect fluid and a cosmological constant (Figure H.3). Of course, the full range of coordinates on $\mathbf{R} \times S^3$ would usually be $-\infty < T < \infty$, $0 \le R \le \pi$, while Minkowski space is mapped into the subspace defined by (H.14). The entire $\mathbf{R} \times S^3$ can be drawn as a cylinder, in which each circle of constant T represents a 3-sphere. The shaded region represents Minkowski space. We can unroll the shaded region to portray Minkowski space as a triangle, as shown in Figure H.4. This is the conformal diagram. Each point represents a two-sphere.

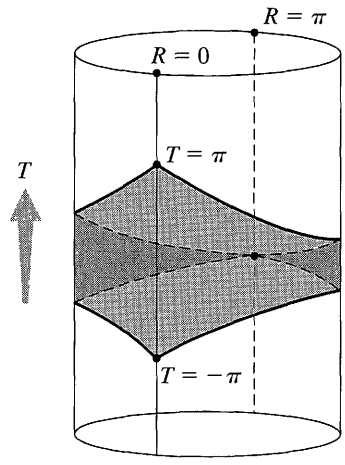

FIGURE H.3 The Einstein static universe, $\mathbf{R} \times S^3$, portrayed as a cylinder. The shaded region is conformally related to Minkowski space.

In fact Minkowski space is only the *interior* of the above diagram (including $R = 0$); the boundaries are not part of the original spacetime. The boundaries are referred to as **conformal infinity**, and the union of the original spacetime with conformal infinity is the **conformal compactification**, which is a manifold with boundary. The structure of the conformal diagram allows us to subdivide conformal infinity into a few different regions:

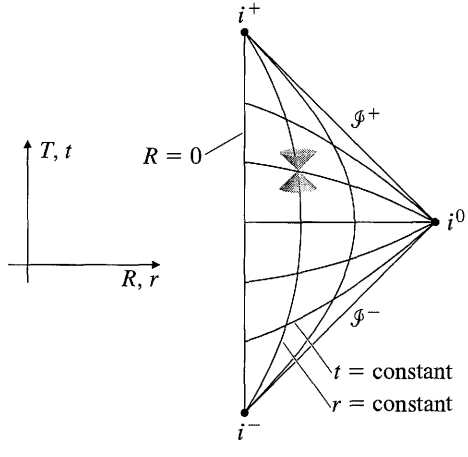

FIGURE H.4 The conformal diagram of Minkowski space. Light cones are at $\pm 45°$ throughout the diagram.

$$i^+ = \text{future timelike infinity } (T = \pi,\ R = 0)$$
$$i^0 = \text{spatial infinity } (T = 0,\ R = \pi)$$
$$i^- = \text{past timelike infinity } (T = -\pi,\ R = 0)$$
$$\mathscr{I}^+ = \text{future null infinity } (T = \pi - R,\ 0 < R < \pi)$$
$$\mathscr{I}^- = \text{past null infinity } (T = -\pi + R,\ 0 < R < \pi)$$

($\mathscr{I}^+$ and $\mathscr{I}^-$ are pronounced as "scri-plus" and "scri-minus," respectively.) Note that i^+, i^0, and i^- are actually *points*, since $R = 0$ and $R = \pi$ are the north and south poles of S^3. Meanwhile $\mathscr{I}^+$ and $\mathscr{I}^-$ are actually null surfaces, with the topology of $\mathbf{R} \times S^2$.

The conformal diagram for Minkowski spacetime contains a number of important features. Radial null geodesics are at $\pm 45°$ in the diagram. All timelike geodesics begin at i^- and end at i^+; all null geodesics begin at $\mathscr{I}^-$ and end at $\mathscr{I}^+$; all spacelike geodesics both begin and end at i^0. On the other hand, there can be nongeodesic timelike curves that end at null infinity, if they become "asymptotically null."

It is nice to be able to fit all of Minkowski space on a small piece of paper, but we don't really learn much that we didn't already know. Conformal diagrams are more useful when we want to represent slightly more complicated spacetimes, such as those for black holes. As discussed in Chapter 6, asymptotically flat spacetimes (or regions of a spacetime) are those that share the structure of $\mathscr{I}^+$, i^0, and $\mathscr{I}^-$ with Minkowski space. Equally importantly, the conformal diagram gives us an idea of the causal structure of the spacetime, for example, whether the past or future light cones of two specified points intersect. In Minkowski space this is always true for any two points, but curved spacetimes can be more interesting, as we saw for the case of an expanding universe in Chapter 2.

Let's consider the conformal diagram for the cosmological spacetime introduced in Chapter 2, which provides a vivid illustration of the usefulness of this technique. When we put polar coordinates on space, the metric becomes

$$ds^2 = -\mathrm{d}t^2 + t^{2q}\left(\mathrm{d}r^2 + r^2 d\Omega^2\right), \tag{H.18}$$

where we have chosen to consider power-law behavior for the scale factor, $a(t) = t^q$, and $0 < q < 1$. A crucial difference between this metric and that of Minkowski space is the singularity at $t = 0$, which restricts the range of our coordinates:

$$0 < t < \infty \tag{H.19}$$
$$0 \leq r < \infty. \tag{H.20}$$

Other than this restricted coordinate range, our analysis follows almost precisely that of the case of flat spacetime. This is because we can bring the metric (H.18) to the form of flat spacetime times a conformal factor; once done, we need only to reproduce our previous coordinate transformations to express our expanding-universe metric as a conformal factor times the Einstein static universe.

We begin by choosing a new time coordinate η, sometimes called **conformal time**, which satisfies

$$dt^2 = t^{2q} d\eta^2, \tag{H.21}$$

or

$$\eta = \frac{1}{1-q} t^{1-q}. \tag{H.22}$$

This simple choice allows us to bring out the scale factor as an overall conformal factor,

$$ds^2 = [(1-q)\eta]^{2q/(1-q)} \left(-d\eta^2 + dr^2 + r^2 d\Omega^2\right). \tag{H.23}$$

The range of η is the same as that of t,

$$0 < \eta < \infty. \tag{H.24}$$

Note that η is a timelike coordinate [in the sense that the vector ∂_η is timelike, $ds^2(\partial_\eta, \partial_\eta) < 0$], but it does not measure the proper time of a comoving clock (one with constant spatial coordinates). If we consider a trajectory $x^\mu(\lambda) = (\eta(\lambda), 0, 0, 0)$, and calculated the proper time $\tau(\eta)$, we would find that it was equal to our previous time coordinate but not our new one: $\tau \propto t \propto \eta^{1/(1-q)}$. So η is a timelike coordinate, but not the time that anyone would measure. This is perfectly okay, and simply serves as an illustration of the independence of the notions of observable quantities and spacetime coordinates.

Now that we have our expanding-universe metric in the form of a conformal factor times Minkowski, we can perform the same sequence of coordinate transformations—(H.5), (H.8), and (H.13)—where we allow η to take the place of t. These changes transform our coordinates from (η, r) to (T, R), where the ranges are now

$$0 \leq R, \quad 0 < T, \quad T + R < \pi. \tag{H.25}$$

The metric (H.23) becomes

$$ds^2 = \omega^{-2}(T, R) \left(-dT^2 + dR^2 + \sin^2 R \, d\Omega^2\right), \tag{H.26}$$

where some heroic use of trigonometric identities reveals that the conformal factor is of the form

$$\omega(T, R) = \left(\frac{\cos T + \cos R}{2 \sin T}\right)^{2q} (\cos T + \cos R). \tag{H.27}$$

The precise form of the conformal factor is actually not of primary importance; the crucial feature is that we have once again expressed our metric as a conformal factor times that of the Einstein static universe. The important distinction between

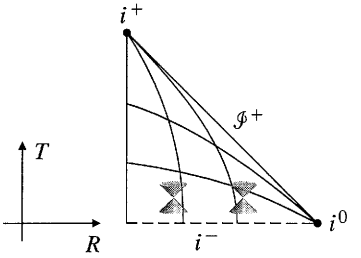

FIGURE H.5 Conformal diagram for a Robertson–Walker universe with $a(t) \propto t^q$ for $0 < q < 1$. The dashed line represents the singularity at $t = 0$ (which also corresponds to $T = 0$).

this case and that of flat spacetime is that the timelike coordinate ends at the singularity at $T = 0$; otherwise the spacetime diagram is identical. We therefore have the conformal diagram of Figure H.5, which resembles the upper half of the Minkowski diagram (Figure H.4). Once again, light cones appear at 45°. We see how the conformal diagram makes the causal structure apparent; it is straightforward to choose two events in the spacetime with the property that their past light cones will hit the singularity before they intersect (while future light cones will always overlap). For more complicated geometries, this convenient way of representing a spacetime will be even more useful.

APPENDIX

I The Parallel Propagator

The idea of parallel-transporting a tensor along a curve is obviously of central importance in GR. For a vector V^μ being transported down a path $x^\mu(\lambda)$, the equation of parallel transport is

$$\frac{dx^\mu}{d\lambda}\nabla_\mu V^\nu \equiv \frac{dx^\mu}{d\lambda}\partial_\mu V^\nu + \frac{dx^\mu}{d\lambda}\Gamma^\nu_{\mu\sigma}V^\sigma = 0. \tag{I.1}$$

It turns out to be possible to write down an explicit and general solution to this equation; it's somewhat formal, but interesting both in its own right and for its connections to techniques in quantum field theory.

We begin by noticing that for some path $\gamma : \lambda \to x^\sigma(\lambda)$, solving the parallel transport equation for a vector V^μ amounts to finding a matrix $P^\mu{}_\rho(\lambda, \lambda_0)$, which relates the vector at its initial value $V^\mu(\lambda_0)$ to its value somewhere later down the path:

$$V^\mu(\lambda) = P^\mu{}_\rho(\lambda, \lambda_0)V^\rho(\lambda_0). \tag{I.2}$$

Of course the matrix $P^\mu{}_\rho(\lambda, \lambda_0)$, known as the **parallel propagator**, depends on the path γ (although it's hard to find a notation that indicates this without making γ look like an index). If we define

$$A^\mu{}_\rho(\lambda) = -\Gamma^\mu_{\sigma\rho}\frac{dx^\sigma}{d\lambda}, \tag{I.3}$$

where the quantities on the right-hand side are evaluated at $x^\nu(\lambda)$, then the parallel transport equation becomes

$$\frac{d}{d\lambda}V^\mu = A^\mu{}_\rho V^\rho. \tag{I.4}$$

Since the parallel propagator must work for any vector, substituting (I.2) into (I.4) shows that $P^\mu{}_\rho(\lambda, \lambda_0)$ also obeys this equation:

$$\frac{d}{d\lambda}P^\mu{}_\rho(\lambda, \lambda_0) = A^\mu{}_\sigma(\lambda)P^\sigma{}_\rho(\lambda, \lambda_0). \tag{I.5}$$

To solve this equation, first integrate both sides:

$$P^\mu{}_\rho(\lambda, \lambda_0) = \delta^\mu_\rho + \int_{\lambda_0}^{\lambda} A^\mu{}_\sigma(\eta)P^\sigma{}_\rho(\eta, \lambda_0)\,d\eta. \tag{I.6}$$

The Kronecker delta, it is easy to see, provides the correct normalization for $\lambda = \lambda_0$.

We can solve (I.6) by iteration, taking the right-hand side and plugging it into itself repeatedly, giving

$$P^{\mu}{}_{\rho}(\lambda, \lambda_0) = \delta^{\mu}_{\rho} + \int_{\lambda_0}^{\lambda} A^{\mu}{}_{\rho}(\eta)\, d\eta + \int_{\lambda_0}^{\lambda} \int_{\lambda_0}^{\eta} A^{\mu}{}_{\sigma}(\eta) A^{\sigma}{}_{\rho}(\eta')\, d\eta' d\eta + \cdots. \tag{I.7}$$

The nth term in this series is an integral over an n-dimensional right triangle, or n-simplex:

$$\int_{\lambda_0}^{\lambda} A(\eta_1)\, d\eta_1$$

$$\int_{\lambda_0}^{\lambda} \int_{\lambda_0}^{\eta_2} A(\eta_2) A(\eta_1)\, d\eta_1 d\eta_2$$

$$\int_{\lambda_0}^{\lambda} \int_{\lambda_0}^{\eta_3} \int_{\lambda_0}^{\eta_2} A(\eta_3) A(\eta_2) A(\eta_1)\, d^3\eta.$$

See Figure I.1.

It would simplify things if we could consider such an integral to be over an n-cube instead of an n-simplex. Is there some way to do this? There are $n!$ such simplices in each cube, so we would have to multiply by $1/n!$ to compensate for this extra volume. But we also want to get the integrand right; using matrix notation, the integrand at nth order is $A(\eta_n)A(\eta_{n-1}) \cdots A(\eta_1)$, but with the special property that $\eta_n \geq \eta_{n-1} \geq \cdots \geq \eta_1$. We therefore define the **path-ordering symbol**, $\mathcal{P}$, to ensure that this condition holds. In other words, the expression

$$\mathcal{P}[A(\eta_n)A(\eta_{n-1}) \cdots A(\eta_1)] \tag{I.8}$$

stands for the product of the n matrices $A(\eta_i)$, ordered in such a way that the largest value of η_i is on the left, and each subsequent value of η_i is less than or

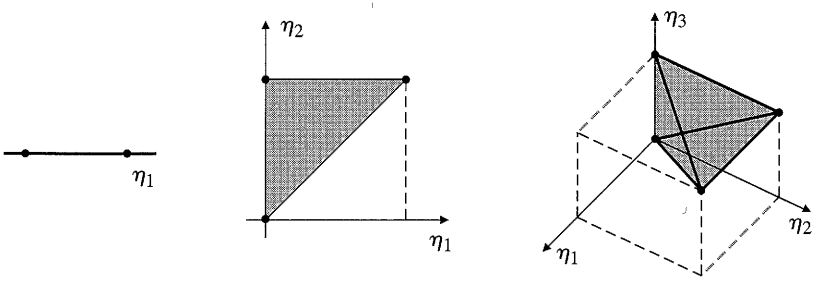

FIGURE I.1 n-simplices (n-dimensional right triangles) for $n = 1, 2, 3$.

equal to the previous one. We then can express the nth-order term in (I.7) as

$$\int_{\lambda_0}^{\lambda}\int_{\lambda_0}^{\eta_n}\cdots\int_{\lambda_0}^{\eta_2} A(\eta_n)A(\eta_{n-1})\cdots A(\eta_1)\,d^n\eta$$
$$= \frac{1}{n!}\int_{\lambda_0}^{\lambda}\int_{\lambda_0}^{\lambda}\cdots\int_{\lambda_0}^{\lambda} \mathcal{P}[A(\eta_n)A(\eta_{n-1})\cdots A(\eta_1)]\,d^n\eta. \tag{I.9}$$

This expression contains no substantive statement about the matrices $A(\eta_i)$; it is just notation. But we can now write (I.7) in matrix form as

$$P(\lambda,\lambda_0) = \mathbf{1} + \sum_{n=1}^{\infty}\frac{1}{n!}\int_{\lambda_0}^{\lambda}\mathcal{P}[A(\eta_n)A(\eta_{n-1})\cdots A(\eta_1)]\,d^n\eta. \tag{I.10}$$

This formula is just the series expression for an exponential; we therefore say that the parallel propagator is given by the path-ordered exponential

$$P(\lambda,\lambda_0) = \mathcal{P}\exp\left(\int_{\lambda_0}^{\lambda} A(\eta)\,d\eta\right), \tag{I.11}$$

where once again this is just notation; the path-ordered exponential is defined to be the right-hand side of (I.10). We can write it more explicitly as

$$\boxed{P^{\mu}{}_{\nu}(\lambda,\lambda_0) = \mathcal{P}\exp\left(-\int_{\lambda_0}^{\lambda}\Gamma^{\mu}_{\sigma\nu}\frac{dx^{\sigma}}{d\eta}\,d\eta\right).} \tag{I.12}$$

It's nice to have an explicit formula, even if it is rather abstract. The same kind of expression appears in quantum field theory as "Dyson's Formula," where it arises because the Schrödinger equation for the time-evolution operator has the same form as (I.5).

An especially interesting example of the parallel propagator occurs when the path is a loop, starting and ending at the same point. Then if the connection is metric-compatible, the resulting matrix will just be a Lorentz transformation on the tangent space at the point. This transformation is known as the "holonomy" of the loop. If you know the holonomy of every possible loop, that turns out to be equivalent to knowing the metric. One can then examine general relativity in the "loop representation," where the fundamental variables are holonomies rather than the explicit metric. A program called "loop quantum gravity" attempts to directly quantize general relativity in these variables (as opposed to something like string theory, in which GR falls out in some limit). A great deal of mathematical progress has been made in this direction, but fundamental obstacles remain.[1]

[1]For a review of this approach, see C. Rovelli, "Loop quantum gravity," *Living Rev. Rel.* **1**, 1 (1998) `http://arxiv.org/gr-qc/9710008`.

APPENDIX

J Noncoordinate Bases

Early on in our study of manifolds, we made a decision to choose bases for our tangent spaces that were adapted to coordinates. For both aesthetic and pragmatic reasons, we should consider once again the formalism of connections and curvature, but this time using sets of basis vectors in the tangent space that are *not* derived from any coordinate system. It will turn out that this slight change in emphasis reveals a different point of view on the connection and curvature, one in which the relationship to gauge theories of particle physics is much more transparent. In fact the concepts to be introduced are very straightforward, but the subject is a notational nightmare, so it looks more difficult than it really is.

Until now we have been taking advantage of the fact that a natural basis for the tangent space T_p at a point p is given by the partial derivatives with respect to the coordinates at that point, $\hat{e}_{(\mu)} = \partial_\mu$. Similarly, a basis for the cotangent space T_p^* is given by the gradients of the coordinate functions, $\hat{\theta}^{(\mu)} = \mathrm{d}x^\mu$. Nothing stops us, however, from setting up any bases we like. Let us therefore imagine that at each point in the manifold we introduce a set of basis vectors $\hat{e}_{(a)}$ (indexed by a Latin letter rather than Greek, to remind us that they are not related to any coordinate system). We will choose these basis vectors to be "orthonormal," in a sense that is appropriate to the signature of the manifold on which we are working. That is, if the canonical form of the metric is written η_{ab}, we demand that the inner product of our basis vectors be

$$g(\hat{e}_{(a)}, \hat{e}_{(b)}) = \eta_{ab}, \tag{J.1}$$

where $g(\ ,\ )$ is the usual metric tensor. Thus, in a Lorentzian spacetime η_{ab} represents the Minkowski metric, while in a space with positive-definite metric it would represent the Euclidean metric. The set of vectors comprising an orthonormal basis is sometimes known as a **tetrad** (from Greek *tetras*, "a group of four") or **vielbein** (from the German for "many legs"). In different numbers of dimensions it occasionally becomes a *vierbein* (four), *dreibein* (three), *zweibein* (two), and so on. Just as we cannot in general find coordinate charts that cover the entire manifold, we will often not be able to find a single set of smooth basis vector fields that are defined everywhere. As usual, we can overcome this problem by working in different patches and making sure things are well-behaved on the overlaps.

The point of having a basis is that any vector can be expressed as a linear combination of basis vectors. Specifically, we can express our old basis vectors

$\hat{e}_{(\mu)} = \partial_\mu$ in terms of the new ones:

$$\hat{e}_{(\mu)} = e_\mu{}^a \hat{e}_{(a)}. \tag{J.2}$$

The components $e_\mu{}^a$ form an $n \times n$ invertible matrix. (In accord with our usual practice of blurring the distinction between objects and their components, we will refer to the $e_\mu{}^a$ as the tetrad or vielbein, and often in the plural as "vielbeins.") We denote their inverse by switching indices to obtain $e^\mu{}_a$, which satisfy

$$e^\mu{}_a e_\nu{}^a = \delta^\mu_\nu, \quad e_\mu{}^a e^\mu{}_b = \delta^a_b. \tag{J.3}$$

These serve as the components of the vectors $\hat{e}_{(a)}$ in the coordinate basis:

$$\hat{e}_{(a)} = e^\mu{}_a \hat{e}_{(\mu)}. \tag{J.4}$$

In terms of the inverse vielbeins, (J.1) becomes

$$g_{\mu\nu} e^\mu{}_a e^\nu{}_b = \eta_{ab}, \tag{J.5}$$

or equivalently

$$g_{\mu\nu} = e_\mu{}^a e_\nu{}^b \eta_{ab}. \tag{J.6}$$

This last equation sometimes leads people to say that the vielbeins are the "square root" of the metric.

We can similarly set up an orthonormal basis of one-forms in T_p, which we denote $\hat{\theta}^{(a)}$. They may be chosen to be compatible with the basis vectors, in the sense that

$$\hat{\theta}^{(a)}(\hat{e}_{(b)}) = \delta^a_b. \tag{J.7}$$

An immediate consequence is that the orthonormal one-forms are related to their coordinate-based cousins $\hat{\theta}^{(\mu)} = \mathrm{d}x^\mu$ by

$$\hat{\theta}^{(\mu)} = e^\mu{}_a \hat{\theta}^{(a)} \tag{J.8}$$

and

$$\hat{\theta}^{(a)} = e_\mu{}^a \hat{\theta}^{(\mu)}. \tag{J.9}$$

The vielbeins $e_\mu{}^a$ thus serve double duty as the components of the coordinate basis vectors in terms of the orthonormal basis vectors, and as components of the orthonormal basis one-forms in terms of the coordinate basis one-forms; while the inverse vielbeins serve as the components of the orthonormal basis vectors in terms of the coordinate basis, and as components of the coordinate basis one-forms in terms of the orthonormal basis.

Any other vector can be expressed in terms of its components in the orthonormal basis. If a vector V is written in the coordinate basis as $V^{\mu}\hat{e}_{(\mu)}$ and in the orthonormal basis as $V^{a}\hat{e}_{(a)}$, the sets of components will be related by

$$V^{a} = e_{\mu}{}^{a}V^{\mu}. \tag{J.10}$$

So the vielbeins allow us to "switch from Latin to Greek indices and back." The nice property of tensors, that there is usually only one sensible thing to do based on index placement, is of great help here. We can go on to refer to multi-index tensors in either basis, or even in terms of mixed components:

$$V^{a}{}_{b} = e_{\mu}{}^{a}V^{\mu}{}_{b} = e^{\nu}{}_{b}V^{a}{}_{\nu} = e_{\mu}{}^{a}e^{\nu}{}_{b}V^{\mu}{}_{\nu}. \tag{J.11}$$

Looking back at (J.5), we see that the components of the metric tensor in the orthonormal basis are just those of the flat metric, η_{ab}. (For this reason the Greek indices are sometimes referred to as "curved" and the Latin ones as "flat.") In fact we can go so far as to raise and lower the Latin indices using the flat metric and its inverse η^{ab}. You can check for yourself that everything works (for example, that the lowering an index with the metric commutes with changing from orthonormal to coordinate bases). In particular, our definition of the inverse vielbeins is consistent with our usual notion of raising and lowering indices,

$$e^{\mu}{}_{a} = g^{\mu\nu}\eta_{ab}e_{\nu}{}^{b}. \tag{J.12}$$

We have introduced the vielbeins $e_{\nu}{}^{a}$ as components of a set of basis vectors, evaluated in a different basis. This is equivalent to thinking of them as the components of a (1, 1) tensor,

$$e = e_{\nu}{}^{a}\mathrm{d}x^{\nu} \otimes \hat{e}_{(a)}. \tag{J.13}$$

But this is actually a tensor we already know and love: the identity map. If we act this tensor on a vector, we get back the same vector, just in a different basis; that's the content of (J.10). Likewise, if we use the inverse vielbein e_{a}^{μ} to convert the Latin index on $e_{\nu}{}^{a}$ to a Greek index, according to (J.3) we get the Kronecker delta δ_{ν}^{μ}, which of course is the identity map on vectors (or one-forms). This point is worth emphasizing because we could also choose to interpret $e_{\nu}{}^{a}$ as a set of vector components (and some references do so), in which case the covariant derivative would look different. By introducing a new set of basis vectors and one-forms, we necessitate a return to our favorite topic of transformation properties. We've been careful all along to emphasize that the tensor transformation law was only an indirect outcome of a coordinate transformation; the real issue was a change of basis. Now that we have noncoordinate bases, these bases can be changed independently of the coordinates. The only restriction is that the orthonormality property (J.1) be preserved. But we know what kind of transformations preserve the flat metric—in a Euclidean signature metric they are orthogonal transforma-

tions, while in a Lorentzian signature metric they are Lorentz transformations. We therefore consider changes of basis of the form

$$\hat{e}_{(a)} \rightarrow \hat{e}_{(a')} = \Lambda^{a}{}_{a'}(x)\hat{e}_{(a)}, \tag{J.14}$$

where the matrices $\Lambda^{a}{}_{a'}(x)$ represent position-dependent transformations which (at each point) leave the canonical form of the metric unaltered:

$$\Lambda^{a}{}_{a'}\Lambda^{b}{}_{b'}\eta_{ab} = \eta_{a'b'}. \tag{J.15}$$

In fact these matrices correspond to what in flat space we called the inverse Lorentz transformations (which operate on basis vectors); as before we also have ordinary Lorentz transformations $\Lambda^{a'}{}_{a}$, which transform the basis one-forms. As far as components are concerned, as before we transform upper indices with $\Lambda^{a'}{}_{a}$ and lower indices with $\Lambda^{a}{}_{a'}$.

So we now have the freedom to perform a Lorentz transformation (or an ordinary Euclidean rotation, depending on the signature) at every point in space. These transformations are therefore called **local Lorentz transformations**, or LLT's. We still have our usual freedom to make changes in coordinates, which are called **general coordinate transformations**, or GCT's. Both can happen at the same time, resulting in a mixed tensor transformation law:

$$T^{a'\mu'}{}_{b'\nu'} = \Lambda^{a'}{}_{a}\frac{\partial x^{\mu'}}{\partial x^{\mu}}\Lambda^{b}{}_{b'}\frac{\partial x^{\nu}}{\partial x^{\nu'}}T^{a\mu}{}_{b\nu}. \tag{J.16}$$

Translating what we know about tensors into noncoordinate bases is for the most part merely a matter of sticking vielbeins in the right places. The crucial exception comes when we begin to differentiate things. In our ordinary formalism, the covariant derivative of a tensor is given by its partial derivative plus correction terms, one for each index, involving the tensor and the connection coefficients. The same procedure will continue to be true for the noncoordinate basis, but we replace the ordinary connection coefficients $\Gamma^{\lambda}_{\mu\nu}$ by the **spin connection**, denoted $\omega_{\mu}{}^{a}{}_{b}$. Each Latin index gets a factor of the spin connection in the usual way:

$$\nabla_{\mu}X^{a}{}_{b} = \partial_{\mu}X^{a}{}_{b} + \omega_{\mu}{}^{a}{}_{c}X^{c}{}_{b} - \omega_{\mu}{}^{c}{}_{b}X^{a}{}_{c}. \tag{J.17}$$

(The name "spin connection" comes from the fact that this can be used to take covariant derivatives of spinors, which is actually impossible using the conventional connection coefficients.) In the presence of mixed Latin and Greek indices we get terms of both kinds.

The usual demand that a tensor be independent of the way it is written allows us to derive a relationship between the spin connection, the vielbeins, and the $\Gamma^{\nu}_{\mu\lambda}$'s. Consider the covariant derivative of a vector X, first in a purely coordinate basis:

$$\begin{aligned}\nabla X &= (\nabla_{\mu}X^{\nu})\mathrm{d}x^{\mu} \otimes \partial_{\nu} \\ &= (\partial_{\mu}X^{\nu} + \Gamma^{\nu}_{\mu\lambda}X^{\lambda})\mathrm{d}x^{\mu} \otimes \partial_{\nu}.\end{aligned} \tag{J.18}$$

Now find the same object in a mixed basis, and convert into the coordinate basis:

$$\begin{aligned}\nabla X &= (\nabla_\mu X^a)\mathrm{d}x^\mu \otimes \hat{e}_{(a)} \\ &= (\partial_\mu X^a + \omega_\mu{}^a{}_b X^b)\mathrm{d}x^\mu \otimes \hat{e}_{(a)} \\ &= (\partial_\mu (e_\nu{}^a X^\nu) + \omega_\mu{}^a{}_b e_\lambda{}^b X^\lambda)\mathrm{d}x^\mu \otimes (e^\sigma{}_a \partial_\sigma) \\ &= e^\sigma{}_a (e_\nu{}^a \partial_\mu X^\nu + X^\nu \partial_\mu e_\nu{}^a + \omega_\mu{}^a{}_b e_\lambda{}^b X^\lambda)\mathrm{d}x^\mu \otimes \partial_\sigma \\ &= (\partial_\mu X^\nu + e^\nu{}_a \partial_\mu e_\lambda{}^a X^\lambda + e^\nu{}_a e_\lambda{}^b \omega_\mu{}^a{}_b X^\lambda)\mathrm{d}x^\mu \otimes \partial_\nu. \end{aligned} \tag{J.19}$$

Comparison with (J.18) reveals

$$\Gamma^\nu_{\mu\lambda} = e^\nu{}_a \partial_\mu e_\lambda{}^a + e^\nu{}_a e_\lambda{}^b \omega_\mu{}^a{}_b, \tag{J.20}$$

or equivalently

$$\omega_\mu{}^a{}_b = e_\nu{}^a e^\lambda{}_b \Gamma^\nu_{\mu\lambda} - e^\lambda{}_b \partial_\mu e_\lambda{}^a. \tag{J.21}$$

A bit of manipulation allows us to write this relation as the vanishing of the covariant derivative of the vielbein,

$$\begin{aligned}\nabla_\mu e_\nu{}^a &= \partial_\mu e_\nu{}^a - \Gamma^\lambda_{\mu\nu} e_\lambda{}^a + \omega_\mu{}^a{}_b e_\nu{}^b \\ &= 0, \end{aligned} \tag{J.22}$$

which is sometimes known as the "tetrad postulate." Note that this is always true; we did not need to assume anything about the connection in order to derive it. Specifically, we did not need to assume that the connection was metric compatible or torsion free. We did, however, implicitly take $e_\nu{}^a$ to represent the (1, 1) tensor (J.13); since this tensor is the identity map, it is no surprise that its covariant derivative vanishes. (Not all references have this philosophy, so be careful.)

Since the connection may be thought of as something we need to introduce in order to fix up the transformation law of the covariant derivative, it should come as no surprise that the spin connection does not itself obey the tensor transformation law. Actually, under GCT's the one lower Greek index does transform in the right way, as a one-form. But under LLT's the spin connection transforms inhomogeneously, as

$$\omega_\mu{}^{a'}{}_{b'} = \Lambda^{a'}{}_a \Lambda^b{}_{b'} \omega_\mu{}^a{}_b - \Lambda^c{}_{b'} \partial_\mu \Lambda^{a'}{}_c. \tag{J.23}$$

You are encouraged to check for yourself that this results in the proper transformation of the covariant derivative.

So far we have done nothing but empty formalism, translating things we already knew into a new notation. But the work we are doing does buy us two things. The first, which we already alluded to, is the ability to describe spinor fields on spacetime and take their covariant derivatives; we won't explore this further here. The second is a change in viewpoint, in which we can think of various tensors as tensor-valued differential forms. For example, an object like $X_\mu{}^a$, which we

think of as a (1, 1) tensor written with mixed indices, can also be thought of as a "vector-valued one-form." It has one lower Greek index, so we think of it as a one-form, but for each value of the lower index it is a vector. Similarly a tensor $A_{\mu\nu}{}^{a}{}_{b}$, antisymmetric in μ and ν, can be thought of as a "(1, 1)-tensor-valued two-form." Thus, any tensor with some number of antisymmetric lower Greek indices and some number of Latin indices can be thought of as a differential form, but taking values in the tensor bundle. (Ordinary differential forms are simply scalar-valued forms.) The usefulness of this viewpoint comes when we consider exterior derivatives. If we want to think of $X_{\mu}{}^{a}$ as a vector-valued one-form, we are tempted to take its exterior derivative:

$$(\mathrm{d}X)_{\mu\nu}{}^{a} = \partial_{\mu} X_{\nu}{}^{a} - \partial_{\nu} X_{\mu}{}^{a}. \tag{J.24}$$

It is easy to check that this object transforms like a two-form [that is, according to the transformation law for (0, 2) tensors] under GCT's, but not as a vector under LLT's (the Lorentz transformations depend on position, which introduces an inhomogeneous term into the transformation law). But we can fix this by judicious use of the spin connection, which can be thought of as a one-form, but not a tensor-valued one-form, due to the nontensorial transformation law (J.23). Thus, the object

$$(\mathrm{d}X)_{\mu\nu}{}^{a} + (\omega \wedge X)_{\mu\nu}{}^{a} = \partial_{\mu} X_{\nu}{}^{a} - \partial_{\nu} X_{\mu}{}^{a} + \omega_{\mu}{}^{a}{}_{b} X_{\nu}{}^{b} - \omega_{\nu}{}^{a}{}_{b} X_{\mu}{}^{b}, \tag{J.25}$$

as you can verify, transforms as a proper tensor.

An immediate application of this formalism is to the expressions for the torsion and curvature, the two tensors that characterize any given connection. The torsion, with two antisymmetric lower indices, can be thought of as a vector-valued two-form $T_{\mu\nu}{}^{a}$. The curvature, which is always antisymmetric in its last two indices, is a (1, 1)-tensor-valued two-form, $R^{a}{}_{b\mu\nu}$. Using our freedom to suppress indices on differential forms, we can express these in terms of the basis one-forms

$$e^{a} = e_{\mu}{}^{a} \mathrm{d}x^{\mu} \tag{J.26}$$

and the spin-connection one-forms

$$\omega^{a}{}_{b} = \omega_{\mu}{}^{a}{}_{b} \mathrm{d}x^{\mu}. \tag{J.27}$$

Notice that we have switched notations, defining $e^{a} \equiv \hat{\theta}^{(a)}$. This is fairly conventional, as well as cleaner. The defining relations for the torsion and curvature are then

$$T^{a} = \mathrm{d}e^{a} + \omega^{a}{}_{b} \wedge e^{b} \tag{J.28}$$

and

$$R^{a}{}_{b} = \mathrm{d}\omega^{a}{}_{b} + \omega^{a}{}_{c} \wedge \omega^{c}{}_{b}. \tag{J.29}$$

Keep in mind that $R^a{}_b$ represents the entire Riemann tensor, with Greek indices suppressed; don't confuse it with the Ricci tensor. These are known as the **Cartan structure equations**. They are equivalent to the usual definitions; let's go through the exercise of showing this for the torsion, and you can check the curvature for yourself. We have

$$\begin{aligned} T_{\mu\nu}{}^{\lambda} &= e^{\lambda}{}_a T_{\mu\nu}{}^a \\ &= e^{\lambda}{}_a(\partial_\mu e_\nu{}^a - \partial_\nu e_\mu{}^a + \omega_\mu{}^a{}_b e_\nu{}^b - \omega_\nu{}^a{}_b e_\mu{}^b) \\ &= \Gamma^{\lambda}_{\mu\nu} - \Gamma^{\lambda}_{\nu\mu}, \end{aligned} \tag{J.30}$$

which is just the original definition we gave. Here we have used (J.20), the expression for the $\Gamma^{\lambda}_{\mu\nu}$'s in terms of the vielbeins and spin connection. We can also express identities obeyed by these tensors as

$$\mathrm{d}T^a + \omega^a{}_b \wedge T^b = R^a{}_b \wedge e^b \tag{J.31}$$

and

$$\mathrm{d}R^a{}_b + \omega^a{}_c \wedge R^c{}_b - R^a{}_c \wedge \omega^c{}_b = 0. \tag{J.32}$$

The first of these is the generalization of $R^{\rho}{}_{[\sigma\mu\nu]} = 0$, while the second is the Bianchi identity $\nabla_{[\lambda|} R^{\rho}{}_{\sigma|\mu\nu]} = 0$. (Sometimes both equations are called Bianchi identities.)

The form of these expressions leads to an almost irresistible temptation to define a "covariant-exterior derivative," which acts on a tensor-valued form by taking the ordinary exterior derivative and then adding appropriate terms with the spin connection, one for each Latin index. Although we won't do that here, it is okay to give in to this temptation, and in fact the right-hand side of (J.28) and the left-hand sides of (J.31) and (J.32) can be thought of as just such covariant-exterior derivatives. But be careful, since (J.29) cannot be; you can't take any sort of covariant derivative of the spin connection, since it's not a tensor.

So far our equations have been true for general connections; let's see what we get for the Christoffel connection. The torsion-free requirement is just that (J.28) vanish; this does not lead immediately to any simple statement about the coefficients of the spin connection. Metric compatibility is expressed as the vanishing of the covariant derivative of the metric: $\nabla g = 0$. We can see what this leads to when we express the metric in the orthonormal basis, where its components are simply η_{ab}:

$$\begin{aligned} \nabla_\mu \eta_{ab} &= \partial_\mu \eta_{ab} - \omega_\mu{}^c{}_a \eta_{cb} - \omega_\mu{}^c{}_b \eta_{ac} \\ &= -\omega_{\mu ab} - \omega_{\mu ba}. \end{aligned} \tag{J.33}$$

Then setting this equal to zero implies

$$\omega_{\mu ab} = -\omega_{\mu ba}. \tag{J.34}$$

Thus, metric compatibility is equivalent to the antisymmetry of the spin connection in its Latin indices. (As before, such a statement is only sensible if both indices are either upstairs or downstairs.) These two conditions together allow us to express the spin connection in terms of the vielbeins. An explicit formula expresses this solution, but in practice it is easier to simply solve the torsion-free condition

$$\omega^a{}_b \wedge e^b = -\mathrm{d}e^a, \tag{J.35}$$

using the asymmetry of the spin connection, to find the individual components.

One of the best reasons for thinking about noncoordinate bases is that they actually lead to great simplifications in certain cases, including the calculation of the curvature tensor. Let's see how this works in a simple example, a spatially flat expanding universe, with metric

$$ds^2 = -\mathrm{d}t^2 + a^2(t)\delta_{ij}\mathrm{d}x^i\mathrm{d}x^j. \tag{J.36}$$

We will use the differential-forms notation of (J.26) and (J.27); calculations such as this are good evidence that this language is practically useful as well as elegant. The metric is thus written (for any geometry)

$$ds^2 = \eta_{ab}e^a \otimes e^b. \tag{J.37}$$

We need to choose basis one-forms e^a such that this matches our metric (J.36). There are many choices (related by local Lorentz transformations), but one obvious one:

$$\begin{aligned} e^0 &= \mathrm{d}t \\ e^i &= a\mathrm{d}x^i. \end{aligned} \tag{J.38}$$

We would now like to solve for the spin connection using (J.35). The good news is that we basically can do it by guessing. First, by appropriately raising and lowering indices (with η^{ab} and η_{ab}) we derive the consequences of the antisymmetry of ω_{ab}:

$$\begin{aligned} \omega^0{}_0 &= 0 \\ \omega^0{}_j &= \omega^j{}_0 \\ \omega^i{}_j &= -\omega^j{}_i. \end{aligned} \tag{J.39}$$

We next calculate the right-hand side of (J.35),

$$\begin{aligned} \mathrm{d}e^0 &= 0 \\ \mathrm{d}e^i &= \mathrm{d}a \wedge \mathrm{d}x^i = \dot{a}\mathrm{d}t \wedge \mathrm{d}x^i, \end{aligned} \tag{J.40}$$

and then the left,

$$\omega^0{}_b \wedge e^b = \omega^0{}_j \wedge e^j = a\omega^0{}_j \wedge \mathrm{d}x^j$$
$$\omega^i{}_b \wedge e^b = \omega^i{}_0 \wedge e^0 + \omega^i{}_j \wedge e^j = \omega^i{}_0 \wedge \mathrm{d}t + \omega^i{}_j \wedge \mathrm{d}x^j. \tag{J.41}$$

Plugging into (J.35) yields

$$\omega^0{}_j \wedge \mathrm{d}x^j = 0$$
$$\omega^i{}_0 \wedge \mathrm{d}t + \omega^i{}_j \wedge \mathrm{d}x^j = -\dot{a}\mathrm{d}t \wedge \mathrm{d}x^i. \tag{J.42}$$

We would like to solve these equations for $\omega^a{}_b$. It is tempting to guess $\omega^0{}_j = 0$; but then to solve the second equation we would require $\omega^i{}_j = -\dot{a}\delta^i_j \mathrm{d}t$, which is incompatible with $\omega^i{}_j = -\omega^j{}_i$ from (J.39). But we can solve the first equation by setting $\omega^0{}_j$ proportional to $\mathrm{d}x^j$ (due to the antisymmetry of the wedge product). Indeed, if we choose

$$\omega^0{}_j = \dot{a}\mathrm{d}x^j, \quad \omega^i{}_0 = \dot{a}\mathrm{d}x^i, \tag{J.43}$$

we find that both equations in (J.42) are solved by setting

$$\omega^i{}_j = 0. \tag{J.44}$$

Now that we know the spin connection, we can easily get the curvature through

$$R^a{}_b = \mathrm{d}\omega^a{}_b + \omega^a{}_c \wedge \omega^c{}_b. \tag{J.45}$$

We first calculate the exterior derivative of the spin connection forms,

$$\mathrm{d}\omega^i{}_0 = \ddot{a}\mathrm{d}t \wedge \mathrm{d}x^i$$
$$\mathrm{d}\omega^0{}_j = \ddot{a}\mathrm{d}t \wedge \mathrm{d}x^j$$
$$\mathrm{d}\omega^i{}_j = 0, \tag{J.46}$$

and then the wedge products,

$$\omega^0{}_c \wedge \omega^c{}_0 = 0$$
$$\omega^i{}_c \wedge \omega^c{}_0 = 0$$
$$\omega^i{}_c \wedge \omega^c{}_j = \dot{a}^2 \mathrm{d}x^i \wedge \mathrm{d}x^j. \tag{J.47}$$

We therefore obtain the curvature two-form,

$$R^0{}_0 = 0$$
$$R^0{}_j = \ddot{a}\mathrm{d}t \wedge \mathrm{d}x^j$$
$$R^i{}_0 = \ddot{a}\mathrm{d}t \wedge \mathrm{d}x^i$$
$$R^i{}_j = \dot{a}^2 \mathrm{d}x^i \wedge \mathrm{d}x^j. \tag{J.48}$$

For purposes of comparison, we can use vielbeins to convert $R^a{}_{b\mu\nu}$ to our conventional expression $R^\rho{}_{\sigma\mu\nu}$, using

$$R^\rho{}_{\sigma\mu\nu} = e^\rho{}_a e_\sigma{}^b R^a{}_{b\mu\nu}. \tag{J.49}$$

In component form the vielbeins (J.38) and their inverse are

$$e_\mu{}^a = \begin{pmatrix} 1 & & & \\ & a & & \\ & & a & \\ & & & a \end{pmatrix}, \quad e^\nu{}_b = \begin{pmatrix} 1 & & & \\ & a^{-1} & & \\ & & a^{-1} & \\ & & & a^{-1} \end{pmatrix}. \tag{J.50}$$

We will also need to evaluate the components of the wedge products of basis forms, which is straightforward enough,

$$(\mathrm{d}x^\alpha \wedge \mathrm{d}x^\beta)_{\mu\nu} = \delta^\alpha_\mu \delta^\beta_\nu - \delta^\alpha_\nu \delta^\beta_\mu. \tag{J.51}$$

Putting it all together yields the components $R^\rho{}_{\sigma\mu\nu}$,

$$\begin{aligned} R^0{}_{j0l} &= a\ddot{a}\delta_{jl} \\ R^i{}_{0k0} &= -\frac{\ddot{a}}{a}\delta^i_k \\ R^i{}_{jkl} &= \dot{a}^2(\delta^i_k\delta_{jl} - \delta^i_l\delta_{jk}), \end{aligned} \tag{J.52}$$

as well as ones obtained by antisymmetry in the last two indices. We may contract to get the components of the Ricci tensor $R_{\sigma\nu} = R^\lambda{}_{\sigma\lambda\nu}$,

$$\begin{aligned} R_{00} &= -3\frac{\ddot{a}}{a} \\ R_{i0} &= 0 \\ R_{ij} &= (a\ddot{a} + 2\dot{a}^2)\delta_{ij}. \end{aligned} \tag{J.53}$$

You can check that this agrees with our results from Chapter 8. Already in this simple example, the tetrad method was computationally simpler than the coordinate-basis method; in more complicated metrics the comparative advantage continues to grow.

In the language of noncoordinate bases, it is possible to compare the formalism of connections and curvature in Riemannian geometry to that of gauge theories in particle physics. In both situations, the fields of interest live in vector spaces that are assigned to each point in spacetime. In Riemannian geometry the vector spaces include the tangent space, the cotangent space, and the higher tensor spaces constructed from these. In gauge theories, on the other hand, we are concerned with "internal" vector spaces. The distinction is that the tangent space and its relatives are intimately associated with the manifold itself, and are naturally defined once the manifold is set up; the tangent space, for example, can be thought of as the space of directional derivatives at a point. In contrast, an internal vector

space can be of any dimension we like, and has to be defined as an independent addition to the manifold. In math jargon, the union of the base manifold with the internal vector spaces (defined at each point) is a **fiber bundle**, and each copy of the vector space is called the "fiber" (in accord with our definition of the tangent bundle).

Besides the base manifold (for us, spacetime) and the fibers, the other important ingredient in the definition of a fiber bundle is the "structure group," a Lie group that acts on the fibers to describe how they are sewn together on overlapping coordinate patches. Without going into details, the structure group for the tangent bundle in a four-dimensional spacetime is generally GL(4, **R**), the group of real invertible 4×4 matrices; if we have a Lorentzian metric, this may be reduced to the Lorentz group SO(3, 1). Now imagine that we introduce an internal three-dimensional vector space, and sew the fibers together with ordinary rotations; the structure group of this new bundle is then SO(3). A field that lives in this bundle might be denoted $\phi^A(x^\mu)$, where A runs from one to three; it is a three-vector (an internal one, unrelated to spacetime) for each point on the manifold. We have freedom to choose the basis in the fibers in any way we wish; this means that "physical quantities" should be left invariant under local SO(3) transformations such as

$$\phi^A(x^\mu) \rightarrow \phi^{A'}(x^\mu) = O^{A'}{}_A(x^\mu)\phi^A(x^\mu), \tag{J.54}$$

where $O^{A'}{}_A(x^\mu)$ is a matrix in SO(3) that depends on spacetime. Such transformations are known as **gauge transformations**, and theories invariant under them are called "gauge theories."

For the most part it is not hard to arrange things such that physical quantities are invariant under gauge transformations. The one difficulty arises when we consider partial derivatives, $\partial_\mu \phi^A$. Because the matrix $O^{A'}{}_A(x^\mu)$ depends on spacetime, it will contribute an unwanted term to the transformation of the partial derivative. By now you should be able to guess the solution: introduce a connection to correct for the inhomogeneous term in the transformation law. We therefore define a connection on the fiber bundle to be an object $A_\mu{}^A{}_B$, with two "group indices" and one spacetime index. Under GCT's it transforms as a one-form, while under gauge transformations it transforms as

$$A_\mu{}^{A'}{}_{B'} = O^{A'}{}_A O^B{}_{B'} A_\mu{}^A{}_B - O^C{}_{B'} \partial_\mu O^{A'}{}_C. \tag{J.55}$$

(Beware: our conventions are different from those in the particle physics literature.) With this transformation law, the "gauge covariant derivative"

$$D_\mu \phi^A = \partial_\mu \phi^A + A_\mu{}^A{}_B \phi^B \tag{J.56}$$

transforms "tensorially" under gauge transformations, as you are welcome to check. [In ordinary electromagnetism the connection is just the conventional vector potential. No indices are necessary, because the structure group U(1) is one-dimensional.]

It is clear that this notion of a connection on an internal fiber bundle is very closely related to the connection on the tangent bundle, especially in the orthonormal-frame picture we have been discussing. The transformation law (J.55), for example, is exactly the same as the transformation law (J.23) for the spin connection. We can also define a curvature or "field strength" tensor which is a two-form

$$F^A{}_B = \mathrm{d}A^A{}_B + A^A{}_C \wedge A^C{}_B, \tag{J.57}$$

in exact correspondence with (J.29). We can parallel transport things along paths, and there is a construction analogous to the parallel propagator; the trace of the matrix obtained by parallel transporting a vector around a closed curve is called a "Wilson loop."

We could go on in the development of the relationship between the tangent bundle and internal vector bundles, but that would be another book. Let us instead finish by emphasizing the important *difference* between the two constructions. The difference stems from the fact that the tangent bundle is closely related to the base manifold, while other fiber bundles are tacked on after the fact. It makes sense to say that a vector in the tangent space at p "points along a path" through p; but this makes no sense for an internal vector bundle. There is therefore no analogue of the coordinate basis for an internal space—partial derivatives along curves have nothing to do with internal vectors. It follows in turn that there is nothing like the vielbeins, which relate orthonormal bases to coordinate bases. The torsion tensor, in particular, is only defined for a connection on the tangent bundle, not for any gauge theory connections; it can be thought of as the covariant exterior derivative of the vielbein, and no such construction is available on an internal bundle. You should appreciate the relationship between the different uses of the notion of a connection, without getting carried away.

J.1 ■ EXERCISES

1. In (J.37) we mention that the metric in an orthonormal basis can be written

$$ds^2 = \eta_{ab} e^a \otimes e^b. \tag{J.58}$$

How can this possibly be? If the components of the metric are η_{ab} everywhere, how can we know what the geometry is?

2. Calculate the connection one-forms, curvature two-forms, and hence the components of the Riemann tensor for the Mixmaster universe. The metric is given by

$$ds^2 = -\mathrm{d}t \otimes \mathrm{d}t + \alpha^2 \sigma^1 \otimes \sigma^1 + \beta^2 \sigma^2 \otimes \sigma^2 + \gamma^2 \sigma^3 \otimes \sigma^3.$$

Here α, β, γ are functions of t only and the one-forms σ^i are given by

$$\sigma^1 = \cos\psi \, \mathrm{d}\theta + \sin\psi \sin\theta \, \mathrm{d}\phi$$

$$\sigma^2 = \sin\psi \, \mathrm{d}\theta - \cos\psi \sin\theta \, \mathrm{d}\phi$$

$$\sigma^3 = \mathrm{d}\psi + \cos\theta \, \mathrm{d}\phi.$$

Bibliography

I have made no attempt to provide careful citations to the original literature in the text. In keeping with the philosophy of focusing on pedagogy, I have included references to recent review articles where appropriate. Here I list a number of books that might be useful supplements to the one you are reading; the list is not meant to be comprehensive, and focuses on books that are in print and with which I happen to be familiar.

There are two websites that are invaluable resources for keeping up with recent work in gravitational physics. The first is the ArXiv e-print server for general relativity and quantum cosmology:

`http://arxiv.org/form/gr-qc/`

This is where researchers all over the world put their most recent papers, which can then be easily downloaded. There are similar servers for other areas of physics. The other website is for Living Reviews in Relativity:

`http://www.livingreviews.org/`

Living Reviews is an on-line journal specializing in review articles in all areas of gravitational physics. It is an excellent starting point for anyone interested in exploring recent work in a topic of current interest.

Special Relativity

E. Taylor and J. Wheeler, *Spacetime Physics* (Freeman, 1992). A very nice introduction to special relativity, making a great effort to explain away the "paradoxes" this subject seems to engender.

A.P. French, *Special Relativity* (W.W. Norton, 1968). Somewhat less colorful than Taylor and Wheeler, but a straightforward introduction to special relativity.

Undergraduate General Relativity

B.F. Schutz, *A First Course in General Relativity* (Cambridge, 1985). This is a very nice introductory text, making a real effort to bridge the transition from common topics in undergraduate physics to the language and results of GR.

J.B. Hartle, *Gravity: An Introduction to Einstein's General Relativity* (Addison-Wesley, 2002). Eases the exploration of GR by concentrating on examples of

curved spacetimes and the behavior of particles in them, putting physics before formalism whenever possible.

E.F. Taylor and J.A. Wheeler, *Exploring Black Holes: An Introduction to General Relativity* (Benjamin Cummings, 2002). Uses black holes as a way to introduce physical principles of GR.

Graduate General Relativity

R. Wald, *General Relativity* (Chicago, 1984). Thorough discussions of a number of advanced topics, including black holes, global structure, and spinors. An invaluable reference, this is the book to turn to if you need the right answer to a well-posed GR question.

C. Misner, K. Thorne and J. Wheeler, *Gravitation* (Freeman, 1973). The book that educated at least two generations of researchers in gravitational physics. Comprehensive and encyclopedic, the book is written in an often-idiosyncratic style that you will either like or not.

S. Weinberg, *Gravitation and Cosmology* (Wiley, 1972). A great book at what it does, especially strong on astrophysics, cosmology, and experimental tests. However, it takes an unusual non-geometric approach to the material, and doesn't discuss black holes. Weinberg is much better than most of us at cranking through impressive calculations.

R. D'Inverno, *Introducing Einstein's Relativity* (Oxford, 1992). A sensible and lucid introduction to general relativity, with solid coverage of the major topics necessary in a modern GR course.

A.P. Lightman, W.H. Press, R.H. Price, and S.A. Teukolsky, *Problem Book in Relativity and Gravitation* (Princeton, 1975). A sizeable collection of problems in all areas of GR, with fully worked solutions, making it all the more difficult for instructors to invent problems the students can't easily find the answers to.

Advanced General Relativity

S. Hawking and G. Ellis, *The Large-Scale Structure of Space-Time* (Cambridge, 1973). An advanced book that emphasizes global techniques, differential topology, and singularity theorems; a classic.

F. de Felice and C. Clarke, *Relativity on Curved Manifolds* (Cambridge, 1990). A mathematical approach, but with an excellent emphasis on physically measurable quantities.

R. Sachs and H. Wu, *General Relativity for Mathematicians* (Springer-Verlag, 1977). Just what the title says, although the typically dry mathematics prose style is here enlivened by frequent opinionated asides about both physics and mathematics (and the state of the world).

J. Stewart, *Advanced General Relativity* (Cambridge, 2003). A short but sweet introduction to some advanced topics, especially spinors, asymptotic structure, and the characteristic initial-value problem.

Mathematical Background

B. Schutz, *Geometrical Methods of Mathematical Physics* (Cambridge, 1980). Another good book by Schutz, this one covering some mathematical points that are left out of the GR book (but at a very accessible level). Included are discussions of Lie derivatives, differential forms, and applications to physics other than GR.

T. Frankel, *The Geometry of Physics: An Introduction* (Cambridge, 2001). A rich, readable book on topics in geometry that are of real use to physics, including manifolds, bundles, curvature, Lie groups, and algebraic topology.

M. Nakahara, *Geometry, Topology and Physics* (Institute of Physics, 2003). An accessible introduction to differential geometry and topology, with an emphasis on topics of interest to physicists.

F.W. Warner, *Foundations of Differentiable Manifolds and Lie Groups* (Springer-Verlag, 1983). A standard text in the field, includes basic topics such as manifolds and tensor fields as well as more advanced subjects.

Specialized Topics

J.D. Jackson, *Classical Electrodynamics* (Wiley, 1999). The classic reference for graduate-level electromagnetism. The problems have left indelible marks on generations of graduate students.

H. Goldstein et al., *Classical Mechanics* (Prentice-Hall, 2002). The classic reference for graduate-level mechanics. An updated edition adds more discussion of nonlinear dynamics.

V.I. Arnold, *Mathematical Methods of Classical Mechanics* (Springer-Verlag, 1989). A scary book for some physicists, but an inspiring treatment of classical mechanics from a mathematically sophisticated point of view. A lot of good differential geometry here.

E.W. Kolb and M.S. Turner, *The Early Universe* (Perseus, 1994). Has become a standard reference for early-universe cosmology, including dark matter, phase transitions, and inflation.

A.R. Liddle and D. Lyth, *Cosmological Inflation and Large-Scale Structure* (Cambridge, 2000). Focusing on inflation and its implications for large-scale structure, gives a careful treatment of cosmological perturbation theory.

B.S. Ryden, *Introduction to Cosmology* (Addison-Wesley, 2002). A very modern and physical introduction to topics in contemporary cosmology, aimed at advanced undergraduates or beginning graduate students.

S. Dodelson, *Modern Cosmology* (Academic Press, 2003). A graduate-level introduction to cosmology, emphasizing cosmological perturbations, large-scale structure, and the cosmic microwave background.

C.M. Will, *Theory and Experiment in Gravitational Physics* (Cambridge, 1993). A useful compendium of alternatives to GR and the experimental constraints on them, including a discussion of the parameterized post-Newtonian formalism.

S.L. Shapiro and S.A. Teukolsky, *Black Holes, White Dwarfs and Neutron Stars: The Physics of Compact Objects* (Wiley, 1983). A self-contained introduction to the physics and astrophysics of compact stars and black holes.

M.E. Peskin and D.V. Schroeder, *An Introduction to Quantum Field Theory* (Westview Press, 1995). Has quickly become the standard textbook in quantum field theory.

J. Polchinski, *String Theory* (Cambridge, 1998). The standard two-volume introduction to modern string theory, including discussions of D-branes and string duality.

C.V. Johnson, *D-Branes* (Cambridge, 2003). A detailed introduction to the extended objects called D-branes, which have become an indispensable part of string theory; prior knowledge of string theory itself is not required.

E.E. Falco, P. Schneider, and J. Ehlers, *Gravitational Lenses* (Springer Verlag, 1999). A thorough introduction to the theory and applications of gravitational lensing.

N.D. Birrell and P.C. Davies, *Quantum Fields in Curved Spacetime* (Cambridge, 1984). The standard book for those who want a practical introduction to quantum field theory in curved spacetime, including the Hawking effect.

R.M. Wald, *Quantum Field Theory in Curved Spacetime and Black Hole Thermodynamics* (Chicago, 1994). A careful and mathematically rigorous exposition of quantum fields in curved spacetimes; if you really want to know what a vacuum state is, look here.

Popular Books

K.S. Thorne, *Black Holes and Time Warps: Einstein's Outrageous Legacy* (W.W. Norton, 1994). Thorne is one of the world's leading researchers in gravitational physics of all kinds, and he offers both a history of work in GR and an introduction to very up-to-date research topics.

R. Geroch, *General Relativity from A to B* (Chicago, 1981). A truly beautiful exposition of the workings of spacetime.

B. Greene, *The Elegant Universe: Superstrings, Hidden Dimensions, and the Quest for the Ultimate Theory* (W.W. Norton, 1999). A timely and personal introduction to the physics of string theory. Not afraid to discuss quite advanced concepts, but aims at a general audience all along; very well written.

A.H. Guth, *The Inflationary Universe: The Quest for a New Theory of Cosmic Origins* (Perseus, 1998). A thorough and lucid introduction to all of modern cosmology, focusing on inflation.

L. Smolin, *Three Roads to Quantum Gravity* (Basic Books, 2002). The "three roads" are string theory, loop quantum gravity, and something more profound; Smolin is a partisan for loop quantum gravity, but the discussion should be interesting for everyone.

G. Kane, *Supersymmetry: Unveiling the Ultimate Laws of Nature* (Perseus, 2001). A nice introduction to supersymmetry, a hypothetical symmetry between

bosons and fermions that may be within the reach of particle accelerators soon.

A. Einstein, H.A. Lorentz, H. Weyl, and H. Minkowski, *The Principle of Relativity* (Dover, 1924). Actually not a "popular" book at all; rather, a collection of the original research articles on special and general relativity, translated into English.

A. Pais, *Subtle Is the Lord: The Science and the Life of Albert Einstein* (Oxford, 1983). A scientific biography of Einstein, complete with equations.

Index