3

From relativity to cosmology

3.1 Historical background

In 1915 Einstein put the finishing touches to the general theory of relativity. The Schwarzschild solution described in Chapter 2 was the first physically significant solution of the field equations of general relativity. It showed how spacetime is curved around a spherically symmetric distribution of matter. The problem solved by Schwarzschild is basically a local problem, in the sense that the distortions of spacetime geometry from the Minkowski geometry of special relativity gradually diminish to zero as we move further and further away from the gravitating sphere. This result can be easily verified from the line element (2.123) by letting the radial coordinate r go to infinity. In technical jargon, a spacetime satisfying this property is called *asymptotically flat*. In general any spacetime geometry generated by a local distribution of matter is expected to have this property. Even from Newtonian gravity we expect an analogous result: that the gravitational field of a local distribution of matter will die away at a large distance from the distribution. Can the universe be approximated by a local distribution of matter?

Einstein felt that the answer to the above question would be in the negative. Rather, he expected the universe to be filled with matter, however far we are able to probe it. A Schwarzschild-type solution cannot therefore provide the correct spacetime geometry of such a distribution of matter. Since we can never get away from gravitating matter, the concept of asymptotic flatness must break down. A new type of solution is therefore needed to describe a universe filled everywhere with matter. Einstein published such a solution in 1917.

Before we consider Einstein's solution, it is worth noting that more than two centuries earlier Newton also had attempted a solution describing a

matter-filled universe of infinite extent. A highly symmetric distribution of matter does lead to a solution in Newtonian gravity. Imagine, for example, a uniform distribution of matter filling the infinite Euclidean space. An observer viewing the universe from any vantage point will find that it looks the same in all directions and that it presents the same aspect from all vantage points. These two properties are known as *isotropy* and *homogeneity*, and they will turn out to play simplifying roles in relativistic cosmology as well. Newton found that such a universe would be static, for, any particle of matter is being attracted equally in all directions, so it should stay put where it is.

On the other hand, homogeneity precludes any pressure gradients in the universe. And we know that any finite distribution of pressure-free matter would tend to shrink under its own gravity. Stars are able to maintain a stationary shape only because they have large enough pressure gradients inside to withstand their own gravity. Clearly, in going from a finite to an infinite universe something new has entered the argument: the boundary conditions at infinity. Considerable ambiguity arises in Newtonian theory when we try to interpret these boundary conditions.

Newton also found his solution to be unstable: any local inhomogeneity would precipitate gravitational contraction that would tend to augment the local inhomogeneity. Newton compared the instability of the solution to that of a set of needles finely balanced on their points.

Nevertheless, in 1934 E. A. Milne and W. H. McCrea showed how some of the problems of Newtonian cosmology can be resolved. The reader interested in this approach may find some properties of Newtonian cosmology outlined in Exercises 1 to 3 at the end of this chapter and also in Chapter 4.

We will now return to Einstein's solution of 1917.

3.2 The Einstein universe

It is evident from the field equations of general relativity derived in Chapter 2 that their solution in the most general form – the solution of an interlinked set of nonlinear partial differential equations – is beyond the present range of techniques available to applied mathematics. It is necessary to impose simplifying symmetry assumptions in order to make any progress towards a solution. Just as Schwarzschild assumed spherical symmetry in his local solution, Einstein assumed homogeneity and isotropy in his cosmological problem. He further assumed, like Schwarzschild, that spacetime is static. This enabled him to choose a time

coordinate t such that the line element of spacetime could be described by

$$ds^2 = c^2\,dt^2 - \alpha_{\mu\nu}\,dx^\mu\,dx^\nu, \tag{3.1}$$

where $\alpha_{\mu\nu}$ are functions of space coordinates $x^\mu(\mu, \nu = 1, 2, 3)$ only.

Note that constraint of homogeneity implies that the coefficient of dt^2 can only be a constant, which we have normalized to c^2. Similarly, the condition of isotropy tells us that there should be no terms of the form $dt\,dx^\mu$ in the line element. This can be seen easily in the following way. If we had terms like $g_{0\mu}\,dt\,dx^\mu$ in the line element, then spatial displacements dx^μ and $-dx^\mu$ would contribute oppositely to ds^2 over a small time interval dt, and such directional variation is forbidden by isotropy. Can we say anything more about $\alpha_{\mu\nu}$?

Einstein believed that the universe has so much matter as to 'close' the space. And this assumption led him to a specific form for $\alpha_{\mu\nu}$. We will now elaborate a little on the notion of closed space and on how to arrive at $\alpha_{\mu\nu}$. Let us begin with examples from lower-dimensional spaces.

As the simplest example of an open space is the Euclidean straight line extending indefinitely in both directions, we can use a real variable r to denote a typical point on the line with $-\infty < r < \infty$. Figure 3.1(a) shows such a straight line. Figure 3.1(b) shows an example of a closed curve Σ_1. It has no boundary, but if we use a real variable r to denote points on the curve then we will find that a finite range of r will suffice. If we go beyond this range we will begin to go over the curve again and again. A familiar simple example of this is the circle S_1 of radius S shown in Figure 3.1(c). If we use the Euclidean measure of distance to locate a point and denote by r the distance of this point from a fixed point N, we find that the range $0 \leq r < 2\pi S$ describes all the points on the circle.

While both the curves in Figure 3.1(b) and 3.1(c) are closed, the circle evidently has more symmetries than the curve Σ_1. This can be demonstrated as follows. If we take a small section (an arc) of the circle and slide it along the circle, it will always lie flush on it. We cannot do the same for the curve Σ_1. We can express this by saying that the circle S_1 describes homogeneous space, while the curve Σ_1 does not.

Figure 3.2 illustrates the corresponding situation in two dimensions. Two coordinates r and ϕ $(0 \leq r < \infty,\ 0 \leq \phi < 2\pi)$ are needed to locate a point on the Euclidean plane of Figure 3.2(a). The surface Σ_2 shown in Figure 3.2(b) and the sphere S_2 of radius S shown in Figure 3.2(c) are closed surfaces, of which S_2 is homogenous but Σ_2 is not. This latter property can be easily verified by our technique of sliding a small section of each surface along itself.

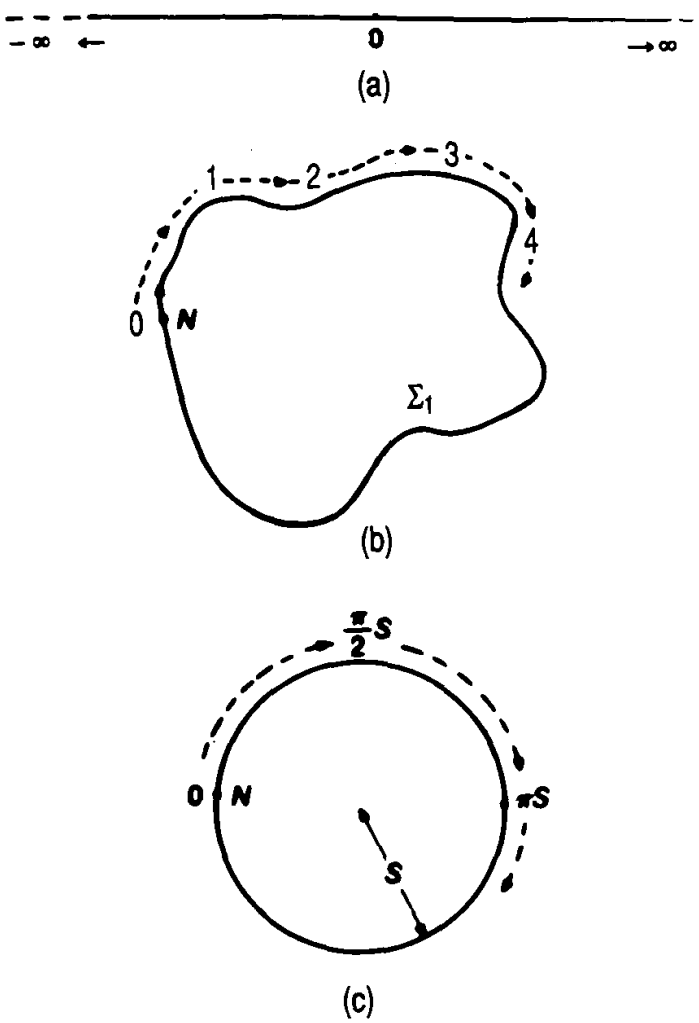

Fig. 3.1 Curves in one-dimensional space. (a) A straight line extending from $-\infty$ to ∞. This is an example of open space. (b) A closed curve Σ_1. Starting from a point N on it as the origin, we can use the length r along the curve to label points on it. If the length of the curve is L, when $r = L$ we come back to the starting point. This is a closed space. (c) A closed space S_1 that is homogeneous: it is a circle. If it has radius S, $L = 2\pi S$.

There is another symmetry inherent in the spherical surface, which can be demonstrated as follows. At any point O on it draw a small arc lying on the surface and then rotate this arc around the point O, trying all the while to keep the arc lying flush on the surface. Again the spherical surface S_2 allows you to do this, but Σ_2 does not. This means that the surface S_2 shows isotropy about O.

We can now see how to construct the homogenous and isotropic closed space of three dimensions that Einstein wanted for his model of the universe. It is S_3, the 3-surface of a four-dimensional hypersphere of radius S. The equation of such a 3-surface is given in Cartesian coordinates x_1, x_2, x_3, and x_4 by

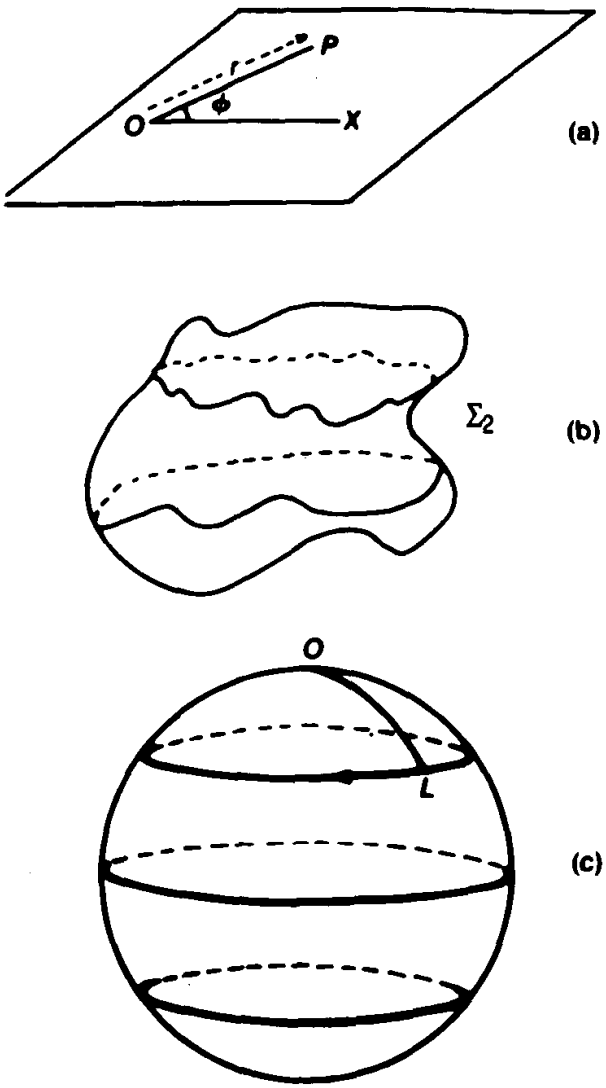

Fig. 3.2 (a) The plane is an open two-dimensional space. From any point O on it draw the straight line OX in any direction in the plane. The coordinates (r, ϕ) in the illustration show how to specify any point P on the plane. (b) An arbitrary closed surface Σ_2. (*c*) A closed surface S_2 that is homogeneous and isotropic. It is a sphere. Take any point O on S_2 and draw a small arc of a great circle OL lying on S_2. As OL is rotated around O, the point L moves along a small circle on S_2 and the arc always stays on S_2. This is an example of isotropy: as seen from O, the surface S_2 shows no preferential direction.

$$(x_1)^2 + (x_2)^2 + (x_3)^2 + (x_4)^2 = S^2. \tag{3.2}$$

To use coordinates *intrinsic* to the surface we define

$$x_4 = S\cos\chi, \qquad x_1 = \sin\chi\cos\theta, \qquad x_2 = S\sin\chi\sin\theta\cos\phi,$$
$$x_3 = S\sin\chi\sin\theta\sin\phi. \tag{3.3}$$

The spatial line element *on* the surface S_3 is therefore given by

$$\begin{aligned} \mathrm{d}\sigma^2 &= (\mathrm{d}x_1)^2 + (\mathrm{d}x_2)^2 + (\mathrm{d}x_3)^2 + (\mathrm{d}x_4)^2 \\ &= S^2[\mathrm{d}\chi^2 + \sin^2\chi(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)]. \end{aligned} \tag{3.4}$$

The ranges of θ, ϕ, and χ are given by

$$0 \leqslant \chi \leqslant \pi, \qquad 0 \leqslant \theta \leqslant \pi, \qquad 0 \leqslant \phi \leqslant 2\pi. \tag{3.5}$$

At this stage it is worth pointing that there are two alternatives open to us. The first is what we have tacitly taken for granted, that χ takes the entire range $0 \leqslant \chi \leqslant \pi$, and this gives us what is commonly known as *spherical space*. If, however, we identify the antipodal points, the space is called *elliptical space*.

Another way to express $d\sigma^2$ is through coordinates r, θ, ϕ, with $r = \sin\chi (0 \leqslant r \leqslant 1)$. In elliptical space r runs through this range once: in spherical space it does so twice:

$$d\sigma^2 = S^2\left[\frac{dr^2}{1-r^2} + r^2(d\theta^2 + \sin^2\theta\, d\phi^2)\right]. \tag{3.6}$$

The constant S is called the 'radius' of the universe. The line element for the Einstein universe is therefore given by

$$\begin{aligned} ds^2 &= c^2\,dt^2 - d\sigma^2 \\ &= c^2\,dt^2 - S^2[d\chi^2 + \sin^2\chi(d\theta^2 + \sin^2\theta\, d\phi^2)] \\ &= c^2\,dt^2 - S^2\left[\frac{dr^2}{1-r^2} + r^2(d\theta^2 + \sin^2\theta\, d\phi^2)\right]. \end{aligned} \tag{3.7}$$

Note that we have derived the line element (3.7) entirely from the various assumptions of symmetry. The field equations have not yet been used. We will now see what happens when we use the above line element to compute the left-hand side of Einstein's equations.

This is easily done with the machinery developed in Chapter 2. We write $x^0 = ct$, $x^1 = r$, $x^2 = \theta$, $x^3 = \phi$, so that

$$g_{00} = 1, \qquad g_{11} = -\frac{S^2}{1-r^2}, \qquad g_{22} = -S^2r^2, \qquad g_{33} = -S^2r^2\sin^2\theta.$$

$$g^{00} = 1, \qquad g^{11} = -\frac{1-r^2}{S^2}, \qquad g^{22} = -\frac{1}{S^2r^2}, \qquad g^{33} = -\frac{1}{S^2r^2\sin^2\theta}.$$

Elementary calculus then tells us that the only nonzero components of Γ^i_{kl} are the following:

$$\Gamma^1_{11} = \frac{r}{1-r^2}, \qquad \Gamma^2_{12} = \Gamma^3_{13} = \frac{1}{r}, \qquad \Gamma^1_{22} = -r(1-r^2),$$

$$\Gamma^1_{33} = -r(1-r^2)\sin^2\theta, \qquad \Gamma^2_{33} = -\sin\theta\cos\theta, \qquad \Gamma^3_{23} = \cot\theta.$$

Next, using the formulae given in the last chapter, we find the following nonzero components of the Einstein tensor:

$$R^0_0 - \tfrac{1}{2}R = -\frac{3}{S^2}, \tag{3.8}$$

$$R_1^1 - \tfrac{1}{2}R = R_2^2 - \tfrac{1}{2}R = R_3^3 - \tfrac{1}{2}R = -\frac{1}{S^2}. \tag{3.9}$$

To complete the field equations, Einstein used the energy tensor for dust derived in (2.83). For dust at rest in the above frame of reference, u^i has only one component, the time component, nonzero. We therefore get

$$\begin{aligned} T_0^0 &= \rho_0 c^2, \\ T_1^1 &= T_2^2 = T_3^3 = 0. \end{aligned} \tag{3.10}$$

Thus the two equations (3.8) and (3.9) lead to two independent equations:

$$-\frac{3}{S^2} = -\frac{8\pi G}{c^2}\rho_0, \qquad -\frac{1}{S^2} = 0. \tag{3.11}$$

Clearly, no sensible solution is possible from these equations, thus suggesting that no static homogeneous isotropic and dense model of the universe is possible under the Einstein equations.

It was his inability to generate such a model that led Einstein to modify his equations (2.98) to (2.102), thus introducing the now famous (or infamous) λ-term. If we introduce this additional constant into the picture, our equations in (3.11) are modified to

$$\lambda - \frac{3}{S^2} = -\frac{8\pi G}{c^2}\rho_0 \tag{3.12}$$

and

$$\lambda - \frac{1}{S^2} = 0. \tag{3.13}$$

We now do have a sensible solution. We get

$$S = \left(\frac{1}{\lambda}\right)^{1/2} = \frac{c}{2(\pi G \rho_0)^{1/2}}. \tag{3.14}$$

Einstein considered this solution as justifying his conjecture that with sufficiently high density it should be possible to 'close' the universe. In (3.14) we have the radius S of the universe as given by the matter density ρ_0, with the result that the larger the value of ρ_0, the smaller is the value of S. However, if λ is a given universal constant like G, both ρ_0 and S are determined in terms of λ (as well as G and c). How big is λ?

In 1917 very little information was available about ρ_0, from which λ could be determined. The value of

$$S \approx 10^{26} - 10^{27} \text{ cm}$$

quoted in those days is therefore only of historical interest. If we take ρ_0

as $\sim 10^{-31}\,\mathrm{g\,cm^{-3}}$ as the rough estimate of mass density in the form of galaxies (see Chapter 9), we get $S \sim 10^{29}$ cm and $\lambda \approx 10^{-58}\,\mathrm{cm^{-2}}$.

The λ-term introduces a force of repulsion between two bodies that increases in proportion to the distance between them. The above value of λ is too small to make any detectable difference from the prediction of standard general relativity (that is, with $\lambda = 0$) in any of the Solar System tests mentioned in Chapter 2. Thus the Einstein universe faced no threat from the local tests of gravity. The model, however, did not survive much longer than a decade, for reasons discussed below.

3.3 The expanding universe

In the late nineteenth century the philosopher and scientist Ernst Mach raised certain conceptual objections to Newton's laws of motion. Mach critically examined the role of a background against which motion is to be measured and argued that unless there is a material background it is not possible to attach any meaning to the concepts of rest or motion. Einstein was greatly influenced by Mach's discussion. The Einstein universe described above includes matter-filled space and thus a background of distant matter against which a local observer can measure motion and formulate laws of mechanics. In fact, as we have just seen, the density of matter determines the precise geometrical nature of spacetime in the Einstein model.

Einstein believed this to be a unique feature of general relativity. He felt that the presence of matter was essential to have a meaningful spacetime geometry. However, his expectation that general relativity can yield only such matter-filled spacetimes as solutions of the field equations was proved wrong shortly after the publication of his paper in 1917. For in 1917 W. de Sitter published another solution of the field equations in (2.102) with the line element given by

$$\mathrm{d}s^2 = c^2\left(1 - \frac{H^2R^2}{c^2}\right)\mathrm{d}t^2 - \frac{\mathrm{d}R^2}{1 - \left(\dfrac{H^2R^2}{c^2}\right)} - R^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2), \tag{3.15}$$

where H is a constant related to λ by

$$\lambda = \frac{3H^2}{c^2}. \tag{3.16}$$

The remarkable feature of the de Sitter universe is that *it is empty*.

Moreover, although the above coordinates give the impression that the universe is static, it is possible to find a new set of coordinates (t, r, θ, ϕ) in terms of which the line element (3.15) takes the form

$$ds^2 = c^2\,dt^2 - e^{2Ht}[dr^2 + r^2(d\theta^2 + \sin^2\theta\,d\phi^2)]. \qquad (3.17)$$

It is easy to verify that test particles with constant values of (r, θ, ϕ) follow timelike geodesics in this model. Thus the proper separation between any two particles measured at a given time t increases with time as e^{Ht}. That is, these particles are all moving apart from one another.

However, these particles have no material status. They have no masses and they do not influence the geometry of spacetime. In the dynamic sense the universe is empty, although in the kinematic sense it is expanding. As Eddington once put it, the de Sitter universe has motion without matter, in contrast to the Einstein universe, which has matter without motion.

The de Sitter universe showed, however, that empty spacetimes could be obtained as solutions of general relativity. For reasons discussed above, a universe of this type fails to meet Mach's criterion that there should be a background of distant matter against which local motion can be measured. Although the property of emptiness of the de Sitter universe was embarrassing, its property of expansion turned out to contain the germ of the truth. For by the end of the third decade of this century, the observations of Hubble and Humason indicated that the universe is not static but is indeed expanding.

Chapter 1 summarized these observations. The phenomenon of nebular redshift observed by Hubble and Humason in the 1920s has now been observed in practically all extragalactic objects. As mentioned in section 1.8, a Newtonian interpretation of such redshifts involves the Doppler effect. How can we express this phenomenon in the language of general relativity? Can we generate models of the universe that combine de Sitter's notion of expansion with Einstein's notion of nonemptiness? The Friedmann models to be discussed in Chapter 4 do just that, and were in fact obtained by Friedmann between 1922 and 1924, five years *before* Hubble's data became well known.

The rest of this chapter outlines the kinematic features of the expanding models of the universe. We will first describe how to generalize the arguments that led Einstein to the static line element (3.7). This generalization will lead us to a nonstatic line element that preserves the properties of homogeneity and isotropy assumed by Einstein, but is potentially capable of explaining Hubble's data.

3.4 Simplifying assumptions of cosmology

Once we decide to generalize from a static to a non-static model of the universe, our task becomes more complicated. Figure 3.3(a) shows a spacetime diagram with a swarm of world lines representing particles moving in arbitrary ways. There is no order in this picture, and where two world lines intersect we have colliding particles. It would indeed be very

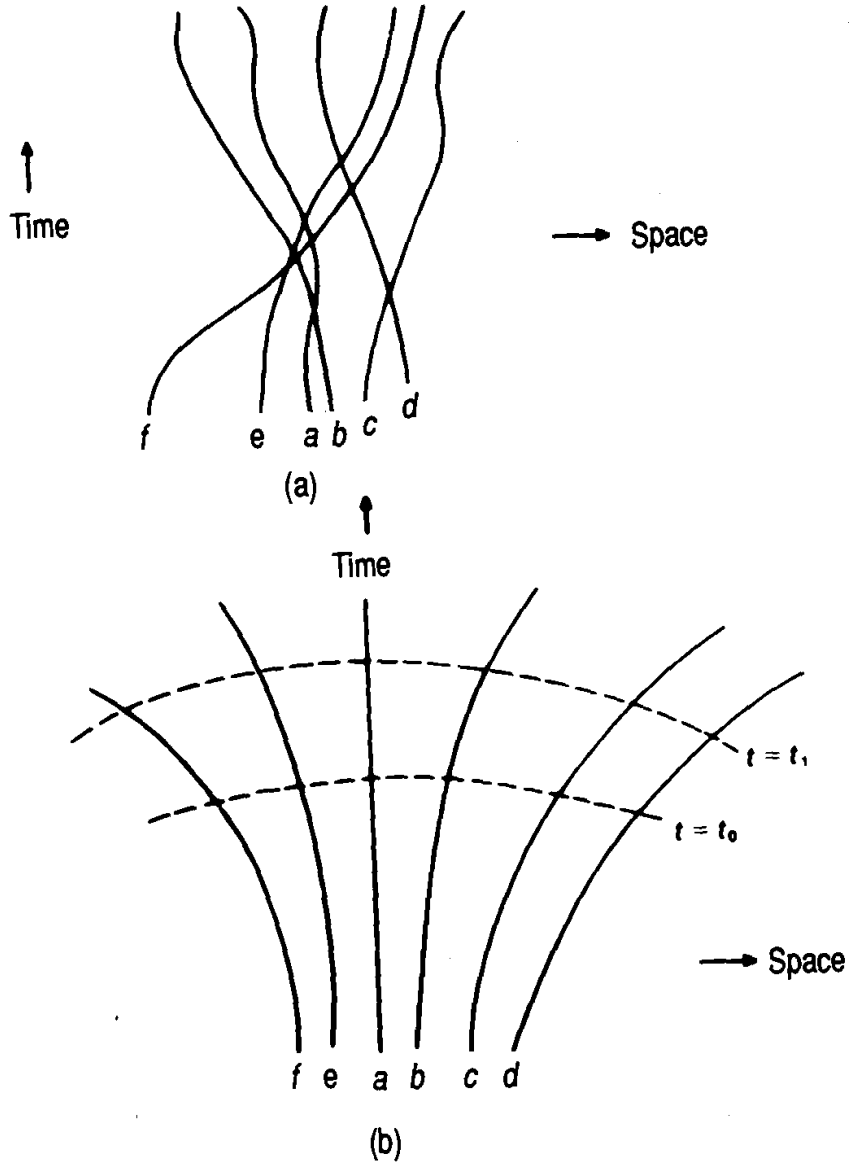

Fig. 3.3 (a) An arbitrary bundle of world lines $a, b, c, \ldots$ describes particles moving haphazardly. Intersecting world lines denote particle collisions. (b) Particles move along nonintersecting world lines $a, b, c, \ldots$ which have no wobbles or irregularities. This is the regularity expressed formally by the Weyl postulate. Note that this regularity enables us to construct a sequence of spacelike hypersurfaces orthogonal to the world lines of the bundle. These are hypersurfaces of constant cosmic time t. Thus the cosmologist can talk of cosmic epochs $t = t_0$, $t = t_1$, and so on in an unambiguous fashion.

difficult to solve the Einstein field equations for such a mess of gravitating matter. Fortunately, the real universe does not appear to be so messy.

Hubble's observations indicate that the universe is (or at least seems to be) an orderly structure in which the galaxies, considered as basic units, are moving apart from one another. Thus Figure 3.3(b) represents a typical spacetime section of the universe in which the world lines represent the histories of galaxies. These world lines, unlike those of Figure 3.3(a), are nonintersecting and form a funnel-like structure in which the separation between any two world lines is steadily increasing.

This intuitive picture of regularity is often expressed formally as the *Weyl postulate*, after the early work of the mathematician Hermann Weyl. The postulate states that the world lines of galaxies designated as *fundamental observers* form a 3-bundle of nonintersecting geodesics orthogonal to a series of spacelike hypersurfaces.

To appreciate the full significance of Weyl's postulate, let us try to express it in terms of coordinates and metric of spacetime. Accordingly we use three spacelike coordinates x^μ ($\mu = 1, 2, 3$) to label a typical world line in the 3-bundle of galaxy world lines. Further, let the coordinate x^0 label a typical member of the series of spacelike hypersurfaces mentioned above. Thus

$$x^0 = \text{constant}$$

is a typical spacelike hypersurface orthogonal to the typical world line given by

$$x^\mu = \text{constant}.$$

Although in practice the galaxies form a discrete set, we can extend the discrete set (x^μ) to a continuum by the *smooth fluid approximation*. This approximation is none other than the widely used device of going over from a discrete distribution of particles to a continuum density distribution. In this case we can treat the quantities x^μ as forming a continuum along with x^0 and use them as the four coordinates x^i to describe space and time.

It is worth emphasizing the importance of the nonintersecting world lines. If two galaxy world lines did intersect, our coordinate system above would break down, for we would than have two different values of x^μ specifying the same spacetime point (the point of intersection). In the next chapter we will, however, encounter an exceptional situation in which all world lines intersect at one singular point!

Let the metric in terms of these coordinates be given by the tensor g_{ik}.

What can we assert about this metric tensor on the basis of the Weyl postulate? The orthogonality condition tells us that

$$g_{0\mu} = 0. \tag{3.18}$$

Further, the fact that the line $x^\mu =$ constant is a *geodesic* tells us that the geodesic equations

$$\frac{\mathrm{d}^2 x^i}{\mathrm{d}x^2} + \Gamma^i_{kl} \frac{\mathrm{d}x^k}{\mathrm{d}s} \frac{\mathrm{d}x^l}{\mathrm{d}s} = 0 \tag{3.19}$$

are satisfied for $x^i =$ constant, $i = 1, 2, 3$. Therefore

$$\Gamma^\mu_{00} = 0, \qquad \mu = 1, 2, 3. \tag{3.20}$$

From (3.18) and (3.20) we therefore get

$$\frac{\partial g_{00}}{\partial x^\mu} = 0, \qquad \mu = 1, 2, 3. \tag{3.21}$$

Thus g_{00} depends on x^0 only. We can therefore replace x^0 by a suitable function of x^0 to make g_{00} constant. Hence we take, without loss of generality,

$$g_{00} = 1. \tag{3.22}$$

The line element therefore becomes

$$\begin{aligned} \mathrm{d}s^2 &= (\mathrm{d}x^0)^2 + g_{\mu\nu}\,\mathrm{d}x^\mu\,\mathrm{d}x^\nu \\ &= c^2\,\mathrm{d}t^2 + g_{\mu\nu}\,\mathrm{d}x^\mu\,\mathrm{d}x^\nu, \end{aligned} \tag{3.23}$$

where we have put $ct = x^0$. This time coordinate is called the *cosmic time*. It is easily seen that the spacelike hypersurfaces in Weyl's postulate are the surfaces of simultaneity with respect to the cosmic time. Moreover, t is the proper time kept by any galaxy.

The second important assumption of cosmology is embodied in the *cosmological principle*. This principle states that at any given cosmic time, the universe is homogeneous and isotropic. That is, the surfaces $t =$ constant exhibit the properties discussed earlier in connection with the Einstein universe. There we saw that the three-dimensional surface S_3 of a hypersphere has the requisite properties of homogeneity and isotropy. But is this the only alternative available?

Einstein, as we saw earlier, selected this alternative because he believed space to be closed. However, if we do not insist on closed space, two more alternatives are available to us, which can be seen in the following way. First let us consider an analogy in lower dimensions.

Figure 3.4 shows three surfaces. Figure 3.4(a) shows a section of the Euclidean plane, Figure 3.4(b) a spherical surface, Figure 3.4(c) a saddle-shaped surface. Suppose we try to cover these surfaces with a plain

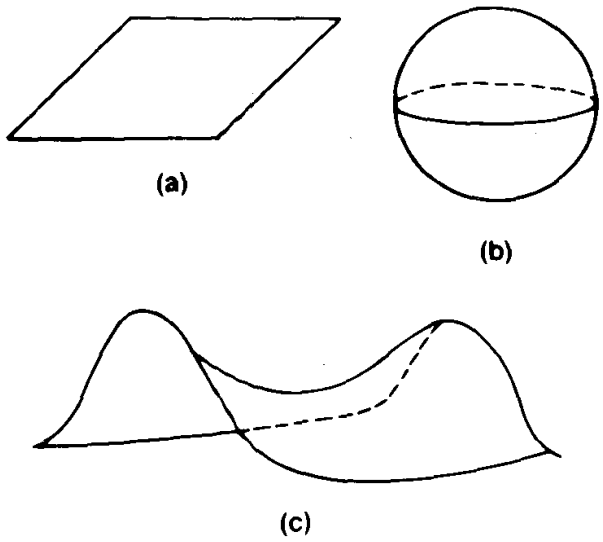

Fig. 3.4 Examples of surfaces of (a) zero curvature (b) positive curvature, and (c) negative curvature.

sheet of paper. We will find that our sheet fits exactly and smoothly on the plane surface. If we try to cover the spherical surface, the sheet of paper develops wrinkles, indicating that the sheet of paper has area in excess of that needed to cover the surface. Similarly, in trying to cover the saddle our paper will be torn, being short of the necessary covering area. These differences can be expressed in differential geometry by the notion of curvature. The plain surface has zero curvature, the spherical surface has positive curvature, and the saddle has negative curvature. Our paper-covering experiment tells us in general whether a given surface has a zero, positive, or negative curvature. These ideas can be extended to higher dimensions as well.

In the Einstein universe the space sections were the 3-surfaces of hyperspheres, and hence they had a constant *positive curvature*. The constancy of curvature is necessary to ensure the properties of homogeneity and isotropy; for if the curvature of space differs from place to place, physical measurements could be devised to detect the differences. We can similarly get other homogeneous and isotropic spaces by considering them as 3-surfaces of *constant negative curvature* or of *zero curvature*.

In terms of the Cartesian coordinates x_1, x_2, x_3, x_4 used earlier, a 3-surface of constant negative curvature is given by an equation of the form

$$x_1^2 + x_2^2 + x_3^2 - x_4^2 = -S^2. \tag{3.24}$$

where S is a constant. The substitution

$$x_1 = S\sinh\chi\cos\theta, \qquad x_2 = S\sinh\chi\sin\theta\cos\phi,$$
$$x_3 = S\sinh\chi\sin\theta\sin\phi, \qquad x_4 = S\cosh\chi \tag{3.25}$$

gives

$$\mathrm{d}x_1^2 + \mathrm{d}x_2^2 + \mathrm{d}x_3^2 - \mathrm{d}x_4^2 = S^2[\mathrm{d}\chi^2 + \sinh^2\chi(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)]. \tag{3.26}$$

Notice the minus sign in front of $\mathrm{d}x_4^2$. It means that we are embedding our 3-surface not in a Euclidean space but in a pseudo-Euclidean space. (In Euclidean space the Pythagoras theorem holds with the line-element given by $\mathrm{d}x^2 = \mathrm{d}x_1^2 + \mathrm{d}x_2^2 + \mathrm{d}x_3^2 + \cdots$. If some of the + signs on the right-hand side are changed to − signs, the result is a pseudo-Euclidean space. Thus Minkowski space is pseudo-Euclidean.) If we further substitute

$$r = \sinh\chi, \tag{3.27}$$

(3.26) becomes

$$\mathrm{d}\sigma^2 = S^2\left[\frac{\mathrm{d}r^2}{1+r^2} + r^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)\right]. \tag{3.28}$$

Compare this with the expression (3.6) for the space of positive curvature:

$$\mathrm{d}\sigma^2 = S^2\left[\frac{\mathrm{d}r^2}{1-r^2} + r^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)\right]. \tag{3.29}$$

Both the expressions can be combined into a single expression by introducing a parameter k that takes values ± 1:

$$\mathrm{d}\sigma^2 = S^2\left[\frac{\mathrm{d}r^2}{1-kr^2} + r^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)\right]. \tag{3.30}$$

Notice that if we set $k = 0$ we get the third alternative – the 3-surface of zero curvature:

$$\mathrm{d}\sigma^2 = S^2[\mathrm{d}r^2 + r^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)]. \tag{3.31}$$

The right-hand side of (3.31) is simply the Euclidean line element scaled by the constant factor S.

The constant S can, however, depend on cosmic time, since we were considering a typical $t =$ constant hypersurface in the above argument. Thus the most general line element satisfying the Weyl postulate and the cosmological principle is given by

$$\mathrm{d}s^2 = c^2\,\mathrm{d}t^2 - S^2(t)\left[\frac{\mathrm{d}r^2}{1-kr^2} + r^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)\right], \tag{3.32}$$

where the 3-spaces $t =$ constant are Euclidean for $k = 0$, closed with positive curvature for $k = \pm 1$, and open with negative curvature for $k = -1$. For reasons that will become clearer later, the scale factor $S(t)$ is often called the *expansion factor*.

The line element (3.32) that we have obtained using partly intuitive and partly heuristic arguments was rigorously derived in the 1930s by H. P. Robertson and A. G. Walker (independently). If is often referred to as the *Robertson–Walker line element*.

The Robertson–Walker line element is sometimes expressed in a slightly different form with the help of the following radial coordinate transformation:

$$\bar{r} = \frac{2r}{1 + (1 - kr^2)^{1/2}}. \tag{3.33}$$

We then get the line element as

$$ds^2 = c^2\,dt^2 - \frac{S^2(t)}{\left(1 + \dfrac{k\bar{r}^2}{4}\right)}[d\bar{r}^2 + \bar{r}^2(d\theta^2 + \sin^2\theta\,d\phi^2)] \tag{3.34}$$

This line element is manifestly isotropic in $\bar{r}$, θ, ϕ. We will, however continue to use (3.32).

Notice how the simplifying postulates of cosmology have reduced the number of unknowns in the metric tensor from 10 to the single function $S(t)$ and the discrete parameter k that characterize the Robertson–Walker metric. The task of the relativist is now simplified to solving an ordinary differential equation in the independent variable t. We will defer the solution of this problem to the next chapter.

We end this chapter with a discussion of some of the important observational features of a typical Robertson–Walker spacetime. These features show how a non-Euclidean geometry can substantially alter conclusions based on naive Euclidean concepts.

3.5 The redshift

Let us first try to understand how the nebular redshift found by Hubble and Humason is accounted for by the Robertson–Walker model. We begin by recalling that the basic units of Weyl's postulate are galaxies with constant coordinates x^μ. We can easily identify the x^μ with the (r, θ, ϕ) of Robertson–Walker spacetime. Thus each galaxy has a constant set of coordinates (r, θ, ϕ). This coordinate frame is often referred to as the *cosmological rest frame*. As observers we are located in our Galaxy, which also has constant (r, θ, ϕ) coordinates. Later on, in Chapter 9, we will show that this remark is only approximately correct, because our Galaxy has a small motion relative to this cosmological frame. Without loss of generality we can take $r = 0$ for our vantage point.

Although this assumption suggests that we are placing ourselves at the centre of the universe, this does not confer any special status on us. Because of the assumption of homogeneity, *any* galaxy could be chosen to have $r = 0$. Our particular choice is simply dictated by convenience.

Consider a galaxy G_1 at (r_1, θ_1, ϕ_1) emitting light waves towards us. Let us denote by t_0 the present epoch of observation. At what time should a light wave leave G_1 in order to arrive at $r = 0$ at time $t = t_0$? To find the answer to this question we need to know the path of the wave from G_1 to us. Since light travels along null geodesics, as described in Chapter 2, we need to calculate the null geodesic from G_1 to us.

From the symmetry of a spacetime we can guess that a null geodesic from $r = 0$ to $r = r_1$ will maintain a constant spatial direction. That is, we expect to have $\theta = \theta_1$, $\phi = \phi_1$ all along the null geodesic. This guess proves to be correct when we substitute these values into the geodesic equations. Accordingly we will assume that only r and t change along the null geodesic. Next we recall that a first integral of the null geodesic equation is simply $ds = 0$. For the Robertson–Walker line element this gives us

$$c\,dt = \pm \frac{S\,dr}{(1 - kr^2)^{1/2}}. \tag{3.35}$$

Since r decreases as t increases along this null geodesic, we should take the minus sign in the above relation. Suppose the null geodesic left G_1 at time t_1. Then we get from the above relation

$$\int_{t_1}^{t_0} \frac{c\,dt}{S(t)} = \int_0^{r_1} \frac{dr}{(1 - kr^2)^{1/2}}. \tag{3.36}$$

Thus if we know $S(t)$ and k, we know the answer to our question.

However, consider what happens to successive wave crests emitted by G_1. Suppose the wave crests were emitted at t_1 and $t_1 + \Delta t_1$ and received by us at t_0 and $t_0 + \Delta t_0$, respectively. Then, comparably to (3.36), we have

$$\int_{t_1+\Delta t_0}^{t_0+\Delta t_1} \frac{c\,dt}{S(t)} = \int_0^{r_1} \frac{dr}{(1 - kr^2)^{1/2}}. \tag{3.37}$$

If $S(t)$ is a slowly varying function so that it effectively remains unchanged over the small intervals Δt_0 and Δt_1, we get by subtraction of (3.36) from (3.37)

$$\frac{c\,\Delta t_0}{S(t_0)} - \frac{c\,\Delta_1}{S(t_1)} = 0,$$

that is,