5 The Theoretical Framework

5.1 The Cosmological Principle

The observational evidence discussed in Chap. 2, particularly the isotropy of the Cosmic Microwave Background Radiation, indicates that the natural starting point for the construction of cosmological models is to assume that, to first approximation, the Universe is isotropic and homogeneous at the present epoch. This is precisely what Einstein assumed in developing his static model of 1917, the first fully self-consistent model of the Universe, derived long before the large-scale isotropy of our Universe was established (Einstein, 1917). Likewise, Friedman's discovery of what were to become the standard models for the large-scale dynamics of the Universe predated the discovery of the expansion of the Universe. The Friedman models were based upon expanding solutions of Einstein's equations, following clues provided by de Sitter and Lanczos.[1]

One of the problems facing the pioneers of relativistic cosmology was the interpretation of the space and time coordinates to be used in these calculations. For example, de Sitter's solution for an empty universe could be written in apparently stationary form, or as an exponentially expanding solution. By 1935, the problem was solved independently by Robertson and Walker (Robertson, 1935; Walker, 1936). They derived the metric of space–time for *all* isotropic, homogeneous, uniformly expanding models of the Universe. This form of the metric is independent of the assumption that the large-scale dynamics of the Universe are described by Einstein's General Theory of Relativity – whatever the physics of the expansion, the space–time metric must be of *Robertson–Walker* form, because of the assumptions of isotropy and homogeneity.

A key step in the development of these models was the introduction by Hermann Weyl in 1923 of what is known as *Weyl's postulate* (Weyl, 1923). To eliminate the arbitrariness in the choice of coordinate frames, Weyl introduced the idea that, in the words of Hermann Bondi (Bondi, 1960):

> The particles of the substratum (representing the nebulae) lie in space–time on a bundle of geodesics diverging from a point in the (finite or infinite) past.

[1] For details of the historical development of the standard world models, see my book *The Cosmic Century* (Longair, 2006)

The most important aspect of this statement is the postulate that the geodesics, which represent the world lines of galaxies, do not intersect, except at a singular point in the finite, or infinite, past. Again, it is remarkable that Weyl introduced this postulate *before* Hubble's discovery of the recession of the nebulae. By the term 'substratum', Bondi meant an imaginary medium which can be thought of as a fluid which defines the overall kinematics of the system of galaxies. A consequence of Weyl's postulate is that there is only one geodesic passing through each point in space–time, except at the origin. Once this postulate is adopted, it becomes possible to assign a notional observer to each world line and these are known as *fundamental observers*. Each fundamental observer carries a standard clock and time measured on that clock from the singular point is called *cosmic time*.

One further assumption is needed before we can derive the framework for the standard models. This is the assumption known as the *cosmological principle* and it can be stated:

We are not located at any special location in the Universe.

A corollary of this statement is that we are located at a *typical* position in the Universe and that any other fundamental observer located anywhere in the Universe at the same cosmic epoch would observe the same large-scale features which we observe. Thus, we assert that every fundamental observer at the same cosmic epoch observes the same Hubble expansion of the distribution of galaxies, the same isotropic Cosmic Microwave Background Radiation, the same large-scale spongy structure in the distribution of galaxies and voids, and so on. As we showed in Sect. 2.3, the combination of Hubble's law and the isotropy of the Universe implies that the system of galaxies as a whole is expanding uniformly and every observer on every galaxy partaking in the uniform expansion observes the same Hubble flow at the same epoch – all of them correctly believe that they are at the centre of a uniformly expanding Universe. The isotropy of the background radiation, the evidence of the scaling of the two-point correlation function with apparent magnitude and the ubiquity of the sponge-like structure of the distribution of galaxies suggest that the cosmological principle is a sensible starting point for the construction of cosmological models.

The specific features of the observable Universe we need in what follows are its overall isotropy and homogeneity, as well as Hubble's law. The combination of these with the Minkowski metric of special relativity results in the *Robertson–Walker metric* for any isotropic, uniformly expanding world model.

5.2 Isotropic Curved Spaces

During the late eighteenth century, non-Euclidean spaces began to be taken seriously by mathematicians who realised that Euclid's fifth postulate, that parallel lines meet only at infinity, might not be essential for the construction of self-consistent geometries. The first suggestions that the global geometry of space might not be Euclidean were discussed by Lambert and Saccheri. In 1786, Lambert noted that, if space were hyperbolic rather than flat, the radius of curvature of space could be used

as an absolute measure of distance. In 1816, Gauss repeated this proposal in a letter to Gerling and was well aware of the fact that a test of the local geometry of space could be carried out by measuring the sum of the angles of a triangle between three high mountain peaks (Longair, 2006).

The fathers of non-Euclidean geometry were Nikolai Ivanovich Lobachevsky in Russia and János Bolyai in Transylvania (Lobachevsky, 1829, 1830; Bolyai, 1832). In his papers, *On the Principles of Geometry* of 1829 and 1930, Lobachevsky at last solved the problem of the existence of non-Euclidean geometries and showed that Euclid's fifth postulate could not be deduced from the other postulates. Non-Euclidean geometry was placed on a firm theoretical basis by the studies of Bernhard Riemann and the English-speaking world was introduced to these ideas through the works of Clifford and Cayley.

Einstein's monumental achievement was to combine special relativity and the theory of gravity through the use of Riemannian geometry and tensor calculus to create the General Theory of Relativity (see Chap. 6). Within a couple of years of formulating the theory, Einstein realised that he now had the tools with which fully self-consistent models for the Universe as a whole could be constructed. In Einstein's model, which we discuss in Sect. 7.3, the Universe is static, closed and has isotropic, spherical geometry. The Friedman solutions, published in 1922 and 1924, were also isotropic models but they were expanding solutions and included geometries that were both spherical and hyperbolic (Friedman, 1922, 1924).

It turns out that it is not necessary to become enmeshed in the details of Riemannian geometry to appreciate the geometrical properties of isotropic curved spaces. We can demonstrate simply why the only isotropic curved spaces are those in which the two-dimensional curvature of any space section κ is constant throughout the space and can only take positive, zero or negative values. The essence of the following argument was first shown to me by my colleague, the late Peter Scheuer.

Let us consider first of all the simplest two-dimensional curved geometry, the surface of a sphere (Fig. 5.1). In the diagram, a triangle is shown consisting of two lines drawn from the north pole down to the equator, the angle between them being 90°; the triangle is completed by the line drawn along the equator of the sphere. The three sides of this triangle are all segments of great circles on the sphere and so are the shortest distances between the three corners of the triangle. The three lines are *geodesics* in the curved geometry.

We need a procedure for working out how non-Euclidean the curved geometry is. The way this is done in general is by the procedure known as the *parallel displacement* or *parallel transport* of a vector on making a complete circuit around a closed figure such as the triangle in Fig. 5.1. Suppose we start with a little vector perpendicular to AC at the pole and lying in the surface of the sphere. We then transport that vector from A to C, keeping it perpendicular to AC. At C, we rotate the vector through 90° so that it is now perpendicular to CB. We then transport the vector, keeping it perpendicular to CB to the corner B. We make a further rotation through 90° to rotate the vector perpendicular to BA and then transport it back to A. At that point, we make a final rotation through 90° to bring the vector back to its original direction. Thus, the total rotation of the vector is 270°. Clearly, the surface

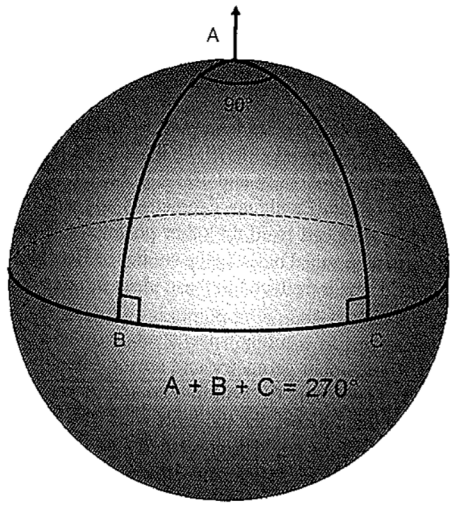

Fig. 5.1. Illustrating the sum of the angles of a triangle on the surface of a sphere.

of the sphere is a non-Euclidean space. This procedure illustrates how we can work out the geometrical properties of any two-space, entirely by making measurements within the two-space, in this case, on the surface of the sphere.

Another simple calculation illustrates an important feature of parallel transport on the surface of a sphere. Suppose the angle at A is not 90° but some arbitrary angle θ. Then, if the radius of the sphere is R_c, the surface area of the triangle ABC is $A = \theta R_c^2$. Thus, if $\theta = 90°$, the area is $\pi R_c^2/2$ and the sum of the angles of the triangle is 270°; if $\theta = 0°$, the area is zero and the sum of the angles of the triangle is 180°. Evidently, the difference of the sum of the angles of the triangle from 180° is proportional to the area of the triangle, that is

$$(\text{Sum of angles of triangle} - 180°) \propto (\text{Area of triangle}) \,. \tag{5.1}$$

This result is a general property of isotropic curved spaces.

Let us now work out the sum of the angles round a closed figure in an isotropic curved space. The procedure is shown schematically in Fig. 5.2a which shows two geodesics from the origin at O being crossed by another pair of geodesics at distances r and $r + \Delta x$ from the origin. The angle $d\theta$ between the geodesics at O is assumed to be small. In Euclidean space, the length of the segment of the geodesic AB would be $\xi = r\,d\theta$. However, this is no longer true in non-Euclidean space and instead, we write

$$\xi(r) = f(r)\,d\theta \,. \tag{5.2}$$

It is straightforward to work out the angle between the diverging geodesics at distance r from the origin. From Fig. 5.2a, it can be seen that the angle between the geodesics is

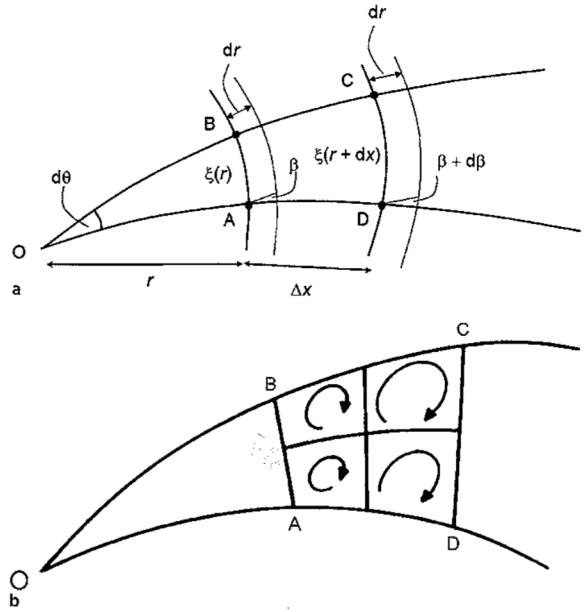

Fig. 5.2. a A schematic diagram illustrating the change in angle β between the geodesics from O over the distance interval Δx. **b** Illustrating how the sum of the rotations around the subloops add up linearly to the total rotation $d\beta$ round the large loop and hence that the total rotation is proportional to the area enclosed by the loop

desics is

$$\beta = \frac{\xi(r + dr) - \xi(r)}{dr} = \frac{d\xi(r)}{dr} = d\theta \frac{d\,f(r)}{dr} \,. \tag{5.3}$$

Let us now move a distance Δx further along the geodesics. The change in the angle β, $\Delta\beta$ is

$$\Delta\beta = \frac{d\xi(r + \Delta x)}{dr} - \frac{d\xi(r)}{dr} = \frac{d^2\xi(r)}{dr^2}\,\Delta x = \frac{d^2 f(r)}{dr^2}\,\Delta x\, d\theta \,. \tag{5.4}$$

Let us check that this result makes sense. In Euclidean space, $\xi(r) = f(r)\,d\theta = r\,d\theta$, $f(r) = r$ and hence (5.3) becomes $\beta = d\theta$. Furthermore, in Euclidean space, $d^2 f(r)/dr^2 = 0$ and so $\Delta\beta = 0$, in other words, $\beta = d\theta$ remains true for all values of r.

Now, the rotation of the vector $d\beta$ depends upon the area of the quadrilateral ABCD. In the case of an isotropic space, we should obtain the same rotation wherever we place the loop in the two-space. Furthermore, if we were to split the loop up into a number of subloops, the rotations around the separate subloops must add

up linearly to the total rotation $d\beta$ (Fig. 5.2b). Thus, in an isotropic two-space, the rotation $d\beta$ should be proportional to the area of the loop ABCD and must be a constant everywhere in the two-space, just as we found in the particular case of a spherical surface in Fig. 5.1.

The area of the loop is $dA = \xi(r)\Delta x = f(r)\Delta x\, d\theta$, and so we can write

$$\frac{d^2 f(r)}{dr^2} = -\kappa\, f(r)\,, \tag{5.5}$$

where κ is a constant, the minus sign being chosen for convenience. This is the equation of simple harmonic motion which has solution

$$f(r) = A \sin \kappa^{1/2} r\,. \tag{5.6}$$

We can find the value of A from the expression for $\xi(r)$ for very small values of r, which must reduce to the Euclidean expression $d\theta = \xi/r$. Therefore, $A = \kappa^{-1/2}$ and

$$f(r) = \frac{\sin \kappa^{1/2} r}{\kappa^{1/2}}\,. \tag{5.7}$$

κ is the *curvature* of the two-space and can be positive, negative or zero. If it is negative, we can write $\kappa = -\kappa'$, where κ' is positive and then the circular functions become hyperbolic functions

$$f(r) = \frac{\sinh \kappa'^{1/2} r}{\kappa'^{1/2}}\,. \tag{5.8}$$

As we showed above, in the Euclidean case, $d^2 f(r)/dr^2 = 0$ and so $\kappa = 0$.

The results we have derived include all possible isotropic curved two-spaces. The constant κ can be positive, negative or zero corresponding to spherical, hyperbolic and flat spaces respectively. In geometric terms, $R_c = \kappa^{-1/2}$ is the radius of curvature of a two-dimensional section through the isotropic curved space and has the same value at all points and in all orientations within the plane. It is often convenient to write the expression for $f(r)$ in the form

$$f(r) = R_c \sin \frac{r}{R_c}\,, \tag{5.9}$$

where R_c is real for closed spherical geometries, imaginary for open hyperbolic geometries and infinite for the case of Euclidean geometry.

The simplest examples of such spaces are the spherical geometries in which R_c is just the radius of the sphere as illustrated in Fig. 5.1. The hyperbolic spaces are more difficult to envisage. The fact that R_c is imaginary can be interpreted in terms of the principal radii of curvature of the surface having opposite sign. The geometry of a hyperbolic two-sphere can be represented by a saddle-shaped figure (Fig. 5.3), just as a two-sphere provides an visualisation of the properties of a spherical two-space.

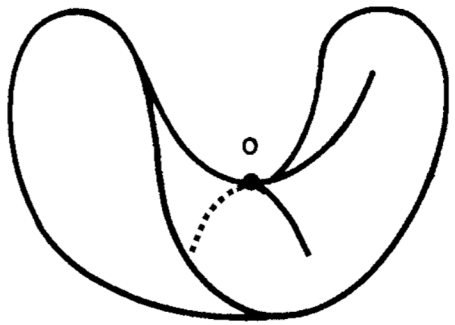

Fig. 5.3. Illustrating the geometry of an isotropic hyperbolic two-space. The principal radii of curvature of the surface are equal in magnitude but have opposite signs in orthogonal directions

5.3 The Space–Time Metric for Isotropic Curved Spaces

In flat space, the distance between two points separated by dx, dy, dz is

$$dl^2 = dx^2 + dy^2 + dz^2. \tag{5.10}$$

Let us now consider the simplest example of an isotropic *two-dimensional* curved space, namely the surface of a sphere which we discussed in Sect. 5.2. We can set up an orthogonal frame of reference at each point locally on the surface of the sphere. It is convenient to work in spherical polar coordinates to describe positions on the surface of the sphere as indicated in Fig. 5.4. In this case, the orthogonal coordinates are the angular coordinates θ and ϕ, and the expression for the increment of distance dl between two neighbouring points on the surface can be written

$$dl^2 = R_c^2\, d\theta^2 + R_c^2 \sin^2\theta\, d\phi^2, \tag{5.11}$$

where R_c is the radius of curvature of the two-space, which in this case is just the radius of the sphere.

The expression (5.11) is known as the *metric* of the two-dimensional surface and can be written more generally in tensor form

$$dl^2 = g_{\mu\nu}\, dx^\mu dx^\nu. \tag{5.12}$$

It is a fundamental result of differential geometry that the *metric tensor* $g_{\mu\nu}$ contains all the information about the intrinsic geometry of the space. The problem is that we can set up a variety of different coordinate systems to define the coordinates of a point on any two-dimensional surface. For example, in the case of a Euclidean plane, we could use rectangular *Cartesian coordinates* so that

$$dl^2 = dx^2 + dy^2, \tag{5.13}$$

or we could use *polar coordinates* in which

$$dl^2 = dr^2 + r^2 d\phi^2. \tag{5.14}$$

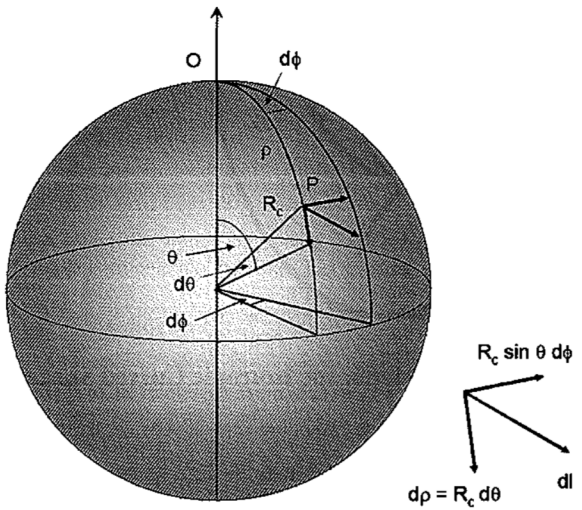

Fig. 5.4. The surface of a sphere as the simplest example of a two-dimensional curved space

How can we determine the *intrinsic curvature* of the space simply in terms of the $g_{\mu\nu}$ of the metric tensor? Gauss first showed how it is possible to do this (Weinberg, 1972; Berry, 1989). For the case of two-dimensional metric tensors which can be reduced to diagonal form, as in the cases of the metrics (5.11), (5.13) and (5.14), the intrinsic curvature of the space is given by the quantity

$$\kappa = \frac{1}{2g_{11}g_{22}}\left\{-\frac{\partial^2 g_{11}}{\partial x_2^2}-\frac{\partial^2 g_{22}}{\partial x_1^2} + \frac{1}{2g_{11}}\left[\frac{\partial g_{11}}{\partial x_1}\frac{\partial g_{22}}{\partial x_1}+\left(\frac{\partial g_{11}}{\partial x_2}\right)^2\right] + \frac{1}{2g_{22}}\left[\frac{\partial g_{11}}{\partial x_2}\frac{\partial g_{22}}{\partial x_2}+\left(\frac{\partial g_{22}}{\partial x_1}\right)^2\right]\right\} . \tag{5.15}$$

It is a useful exercise to use (5.15) to show that both metrics (5.13) and (5.14) have zero curvature and that, for the surface of a sphere, the metric (5.11) corresponds to a space of positive curvature with $\kappa = R_c^{-2}$ at all points on the sphere. κ is known as the *Gaussian curvature* of the two-space and is the same as the definition of the curvature introduced in Sect. 5.2. In general curved spaces, the curvature κ varies from point to point in the space. The extension to isotropic three-spaces is straightforward if we remember that any two-dimensional section through an isotropic three-space must be an isotropic two-space and we already know the metric tensor for this case.

We have already worked out the length of the distance increment dl (5.11). The natural system of coordinates for an isotropic two-space is a spherical polar system in which a radial distance ϱ round the sphere is measured from the pole and the angle ϕ measures angular displacements at the pole. From Fig. 5.4, the distance ϱ round the arc of a great circle from the point O to P is $\varrho = \theta R_c$ and so the metric can be written

$$dl^2 = d\varrho^2 + R_c^2 \sin^2\left(\frac{\varrho}{R_c}\right) d\phi^2 . \tag{5.16}$$

The distance ϱ is the shortest distance between O and P on the surface of the sphere since it is part of a great circle and is therefore the *geodesic distance* between O and P in the isotropic curved space. Geodesics play the role of straight lines in curved space.

We can write the metric in an alternative form if we introduce a distance measure

$$x = R_c \sin\left(\frac{\varrho}{R_c}\right) . \tag{5.17}$$

Differentiating and squaring, we find

$$dx^2 = \left[1-\sin^2\left(\frac{\varrho}{R_c}\right)\right] d\varrho^2 \qquad d\varrho^2 = \frac{dx^2}{1-\kappa x^2} , \tag{5.18}$$

where $\kappa = 1/R_c^2$ is the curvature of the two-space.

Therefore, we can rewrite the metric in the form

$$dl^2 = \frac{dx^2}{1-\kappa x^2} + x^2 d\phi^2 . \tag{5.19}$$

Notice the interpretation of the distance measure x. It can be seen from the metric (5.19) that $dl = x\,d\phi$ is a *proper dimension* perpendicular to the radial coordinate ϱ and that it is the correct expression for the length of a line segment which subtends the angle $d\phi$ at geodesic distance ϱ from O. It is therefore what is known as an *angular diameter distance* since it is guaranteed to give the correct answer for the length of a line segment perpendicular to the line of sight. We can use either ϱ or x in our metric but notice that, if we use x, the increment of geodesic distance is $d\varrho = dx/(1-\kappa x^2)^{1/2}$. We recall that the curvature $\kappa = 1/R_c^2$ can be *positive* as in the spherical two-space discussed above, *zero* in which case we recover flat Euclidean space ($R_c \to \infty$) and *negative* in which case the geometry becomes *hyperbolic* rather than spherical.

We can now write down the expression for the spatial increment in any isotropic, three-dimensional curved space. As mentioned above, the trick is that any two-dimensional section through an isotropic three-space must be an isotropic two-space for which the metric is (5.16) or (5.19). We note that, in spherical polar coordinates, the general angular displacement perpendicular to the radial direction is

$$d\Phi^2 = d\theta^2 + \sin^2\theta\, d\phi^2 , \tag{5.20}$$

and can be found by rotating the coordinate system about the radial direction. Note that the θs and ϕs in (5.20) are different from those used in Fig. 5.4. Thus, by a straightforward extension of the formalism we have derived already, we can write the spatial increment

$$\mathrm{d}l^2 = \mathrm{d}\varrho^2 + R_c^2 \sin^2\left(\frac{\varrho}{R_c}\right)\left[\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2\right] , \tag{5.21}$$

in terms of the three-dimensional spherical polar coordinates $(\varrho,\ \theta,\ \phi)$. An exactly equivalent form is obtained if we write the spatial increment in terms of $x,\ \theta,\ \phi$ in which case we find

$$\mathrm{d}l^2 = \frac{\mathrm{d}x^2}{1-\kappa x^2} + x^2\left[\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2\right] . \tag{5.22}$$

We are now in a position to write down the *Minkowski metric* in any isotropic three-space. It is given by

$$\mathrm{d}s^2 = \mathrm{d}t^2 - \frac{1}{c^2}\mathrm{d}l^2 , \tag{5.23}$$

where $\mathrm{d}l$ is given by either of the above forms of the spatial increment, (5.21) or (5.22). Notice that we have to be careful about the meanings of the distance coordinates $-x$ and ϱ are equivalent but physically quite distinct distance measures. We can now proceed to derive from this metric the *Robertson–Walker metric.*

5.4 The Robertson–Walker Metric

In order to apply the metric (5.23) to isotropic, homogeneous world models, we need the *cosmological principle* and the concepts of *fundamental observers* and *cosmic time* which were introduced in Sect. 5.1. For uniform, isotropic world models, we define a set of *fundamental observers*, who move in such a way that the Universe always appears to be isotropic to them. Each of them has a clock and proper time measured by that clock is called *cosmic time*. There are no problems of synchronisation of the clocks carried by the fundamental observers because, according to Weyl's postulate, the geodesics of all observers meet at one point in the past and cosmic time can be measured from that reference epoch.

We can now write down the metric for such Universes from the considerations of Sect. 5.3. From (5.21) and (5.23), the metric can be written in the form

$$\mathrm{d}s^2 = \mathrm{d}t^2 - \frac{1}{c^2}\left[\mathrm{d}\varrho^2 + R_c^2\sin^2(\varrho/R_c)\left(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2\right)\right] . \tag{5.24}$$

t is cosmic time and $\mathrm{d}\varrho$ is an increment of proper distance in the radial direction.

There is a problem in applying this metric to the expanding Universe as is illustrated by the space–time diagram shown in Fig. 5.5. Since light travels at a finite

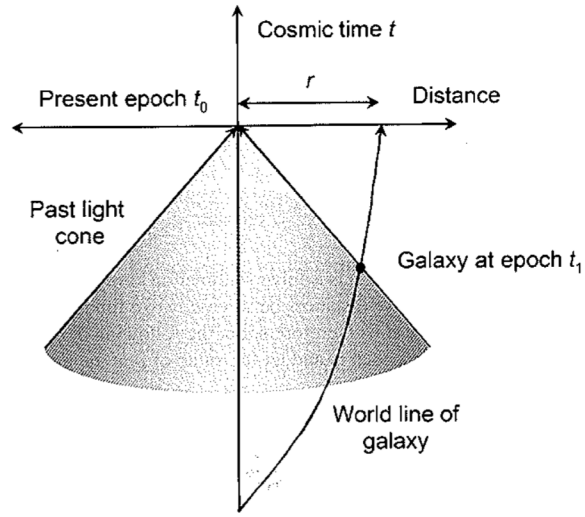

Fig. 5.5. A simple space–time diagram illustrating the definition of the comoving radial coordinate distance

speed, we observe all astronomical objects along a *past light cone* which is centred on the Earth at the present epoch t_0. Therefore, when we observe distant objects, we do not observe them at the present epoch but rather at an earlier epoch t_1 when the Universe was still homogeneous and isotropic but the distances between fundamental observers were smaller and the spatial curvature different. The problem is that we can only apply the metric (5.24) to an isotropic curved space defined *at a single epoch.*

To resolve this problem, we perform the following thought experiment. To measure a proper distance which can be included in the metric (5.24), we line up a set of fundamental observers between the Earth and the galaxy whose distance we wish to measure. The observers are all instructed to measure the distance $\mathrm{d}\varrho$ to the next fundamental observer at a particular cosmic time t which they read on their own clocks. By adding together all the $\mathrm{d}\varrho$s, we can find a proper distance ϱ which is measured *at a single epoch* and which can be used in the metric (5.24). Notice that ϱ is a *fictitious distance* in that we cannot actually measure distances in this way. We observe distant galaxies as they were at some epoch earlier than the present and we do not know how to project their positions relative to us forward to the present epoch until we know the kinematics of the expanding Universe. Thus, *the distance measure ϱ depends upon the choice of cosmological model.*

Let us work out how the ϱ coordinates of galaxies change in a uniformly expanding Universe. The definition of a uniform expansion is that between two cosmic

epochs, t_1 and t_2, the distances of any two fundamental observers, i and j, change such that

$$\frac{\varrho_i(t_1)}{\varrho_j(t_1)} = \frac{\varrho_i(t_2)}{\varrho_j(t_2)} = \text{constant} , \tag{5.25}$$

that is,

$$\frac{\varrho_i(t_1)}{\varrho_i(t_2)} = \frac{\varrho_j(t_1)}{\varrho_j(t_2)} = \ldots = \text{constant} = \frac{a(t_1)}{a(t_2)} . \tag{5.26}$$

For isotropic world models, $a(t)$ is a universal function known as the *scale factor* which describes how the relative distances between *any* two fundamental observers change with cosmic time t. Let us therefore adopt the following definitions. We set $a(t)$ equal to 1 at the present epoch t_0 and let the value of ϱ at the present epoch be r, that is, we can rewrite (5.26) as

$$\varrho(t) = a(t) r . \tag{5.27}$$

The term r thus becomes a *distance label* which is attached to a galaxy or fundamental observer for all time and the variation in proper distance in the expanding Universe is taken care of by the scale factor $a(t)$; r is called the *comoving radial distance coordinate*.

Proper distances perpendicular to the line of sight must also change by a factor a between the epochs t and t_0 because of the isotropy and homogeneity of the world model,

$$\frac{\Delta l(t)}{\Delta l(t_0)} = a(t) . \tag{5.28}$$

From the metric (5.24),

$$a(t) = \frac{R_c(t) \sin\left[\varrho / R_c(t)\right] \mathrm{d}\theta}{R_c(t_0) \sin\left[r / R_c(t_0)\right] \mathrm{d}\theta} . \tag{5.29}$$

Reorganising this equation and using (5.27), we see that

$$\frac{R_c(t)}{a(t)} \sin\left[\frac{a(t) r}{R_c(t)}\right] = R_c(t_0) \sin\left[\frac{r}{R_c(t_0)}\right] . \tag{5.30}$$

This is only true if

$$R_c(t) = a(t)\, R_c(t_0) , \tag{5.31}$$

that is, the radius of curvature of the spatial sections is proportional to the scale factor $a(t)$. Thus, in order to preserve isotropy and homogeneity, *the curvature of space changes as the Universe expands as* $\kappa = R_c^{-2} \propto a^{-2}$. Notice that κ cannot change sign and so, if the geometry of the Universe was once, say, hyperbolic, it will always remain so.

Let us call the value of $R_c(t_0)$, that is, the radius of curvature of the spatial geometry at the present epoch, $\mathfrak{R}$. Then

$$R_c(t) = a(t)\, \mathfrak{R} . \tag{5.32}$$

Substituting (5.27) and (5.32) into the metric (5.24), we obtain

$$\mathrm{d}s^2 = \mathrm{d}t^2 - \frac{a^2(t)}{c^2}[\mathrm{d}r^2 + \mathfrak{R}^2 \sin^2(r/\mathfrak{R})(\mathrm{d}\theta^2 + \sin^2\theta\, \mathrm{d}\phi^2)] . \tag{5.33}$$

This is the *Robertson–Walker metric* in the form we will use in much of our future analysis. Notice that it contains one unknown function $a(t)$, the scale factor, which describes the dynamics of the Universe and an unknown constant $\mathfrak{R}$ which describes the spatial curvature of the Universe at the present epoch.

It is possible to rewrite this metric in different ways. For example, if we use a *comoving angular diameter distance* $r_1 = \mathfrak{R} \sin(r/\mathfrak{R})$, the metric becomes

$$\mathrm{d}s^2 = \mathrm{d}t^2 - \frac{a^2(t)}{c^2}\left[\frac{\mathrm{d}r_1^2}{1 - \kappa r_1^2} + r_1^2(\mathrm{d}\theta^2 + \sin^2\theta\, \mathrm{d}\phi^2)\right] , \tag{5.34}$$

where $\kappa = 1/\mathfrak{R}^2$. By a suitable rescaling of the r_1 coordinate $\kappa r_1^2 = r_2^2$, the metric can equally well be written

$$\mathrm{d}s^2 = \mathrm{d}t^2 - \frac{R_1^2(t)}{c^2}\left[\frac{\mathrm{d}r_2^2}{1 - k r_2^2} + r_2^2(\mathrm{d}\theta^2 + \sin^2\theta\, \mathrm{d}\phi^2)\right] , \tag{5.35}$$

with $k = +1, 0$ and -1 for universes with spherical, flat and hyperbolic geometries respectively. Notice that, in this rescaling, the value of $R_1(t) = R_c(t_0) a = \mathfrak{R} a$ and so the value of $R_1(t)$ at the present epoch is $\mathfrak{R}$ rather than unity. This is a popular form for the metric, but I will normally use (5.33) because the r coordinate has an obvious and important physical meaning.

The importance of the metrics (5.33), (5.34) and (5.35) is that they enable us to define the invariant interval $\mathrm{d}s^2$ between events at any epoch or location in the expanding Universe. Let us recall the meanings of the various components and variables in the metric (5.33):

- The term t is cosmic time, that is, time as measured by a clock carried by a fundamental observer.
- The term r is the *comoving radial distance coordinate* which is fixed to a galaxy for all time and which is the proper distance the galaxy would have if its world line were projected forward to the present epoch t_0 and its distance measured at that time.
- The term $a(t)\, \mathrm{d}r$ is the element of proper (or geodesic) distance in the radial direction at the epoch t.
- The term $a(t)\, [\mathfrak{R} \sin(r/\mathfrak{R})]\, \mathrm{d}\theta = a(t)\, r_1\, \mathrm{d}\theta$ is the element of proper distance perpendicular to the radial direction subtended by the angle $\mathrm{d}\theta$ at the origin.

– Similarly, $a(t)\,[\Re \sin(r/\Re)]\sin\theta\,\mathrm{d}\phi = a(t)\,r_1 \sin\theta\,\mathrm{d}\phi$ is the element of proper distance in the ϕ-direction.

Notice that so far we have specified nothing about the physics which determines the rate of expansion of the Universe; this has all been absorbed into the function $a(t)$. Note the key point that, whatever the physics which determines the function $a(t)$, only the three types of isotropic geometry described by the Robertson–Walker metric are allowed and these types are fixed for all time, although the curvature changes as $a^{-2}(t)$.

5.5 Observations in Cosmology

Many of the most important results which relate the intrinsic properties of distant objects to their observed properties are independent of the specific cosmological model. It is therefore useful to produce a catalogue of results which describe how the observed properties of objects are related to their intrinsic properties and which are independent of the particular form of $a(t)$. First of all, let us elucidate the real meaning of redshift in cosmology.

5.5.1 The Cosmological Redshift

By cosmological redshift, we mean the shift of spectral lines to longer wavelengths associated with the isotropic expansion of the system of galaxies. If λ_e is the wavelength of the line as emitted and λ_0 the observed wavelength, the redshift z is defined to be

$$z = \frac{\lambda_0 - \lambda_\mathrm{e}}{\lambda_\mathrm{e}} \,. \tag{5.36}$$

If the redshift z were interpreted as the recession velocity v of a galaxy, these would be related by the Newtonian Doppler shift formula

$$v = cz \,. \tag{5.37}$$

This is the type of velocity which Hubble used in deriving the velocity–distance relation, $v = H_0 r$. As discussed in Sect. 2.3 and elaborated in Chap. 12, it is incorrect to use the special relativistic Doppler shift formula

$$1 + z = \left(\frac{1 + v/c}{1 - v/c}\right)^{1/2} \,, \tag{5.38}$$

at large redshifts. Rather, because of the requirements of isotropy and homogeneity, the relation $v \propto r$ applies at all comoving radial distances, including those at which the recession velocity would exceed the speed of light.

The key point is that the redshift has a much deeper meaning in cosmology, which we can demonstrate from an analysis of the Robertson–Walker metric. Consider a wave packet of frequency ν_1 emitted between cosmic times t_1 and $t_1 + \Delta t_1$ from a distant galaxy. This wave packet is received by an observer at the present epoch in the interval of cosmic time t_0 to $t_0 + \Delta t_0$. The signal propagates along null cones, $\mathrm{d}s^2 = 0$, and so, considering radial propagation from source to observer, $\mathrm{d}\theta = 0$ and $\mathrm{d}\phi = 0$, the metric (5.33) gives us the relation

$$\mathrm{d}t = -\frac{a(t)}{c}\,\mathrm{d}r \qquad \frac{c\,\mathrm{d}t}{a(t)} = -\mathrm{d}r \,. \tag{5.39}$$

Notice that $a(t)\,\mathrm{d}r$ is simply the interval of proper distance at cosmic time t. The minus sign appears because the origin of the r coordinate is the observer at $t = t_0$. Considering first the leading edge of the wave packet, the integral of (5.39) is

$$\int_{t_1}^{t_0} \frac{c\,\mathrm{d}t}{a(t)} = -\int_r^0 \mathrm{d}r \,. \tag{5.40}$$

The end of the wave packet must travel the same distance in units of comoving distance coordinate since the r coordinate is fixed to the galaxy for all time. Therefore,

$$\int_{t_1+\Delta t_1}^{t_0+\Delta t_0} \frac{c\,\mathrm{d}t}{a(t)} = -\int_r^0 \mathrm{d}r \,, \tag{5.41}$$

that is,

$$\int_{t_1}^{t_0} \frac{c\,\mathrm{d}t}{a(t)} + \frac{c\,\Delta t_0}{a(t_0)} - \frac{c\,\Delta t_1}{a(t_1)} = \int_{t_1}^{t_0} \frac{c\,\mathrm{d}t}{a(t)} \,. \tag{5.42}$$

Since $a(t_0) = 1$, we find that

$$\Delta t_0 = \frac{\Delta t_1}{a(t_1)} \,. \tag{5.43}$$

This is the cosmological expression for the phenomenon of *time dilation*. Distant galaxies are observed at some earlier cosmic time t_1 when $a(t_1) < 1$ and so phenomena are observed to take longer in our frame of reference than they do in that of the source. The phenomenon is precisely the same as time dilation in special relativity, whereby, for example, relativistic muons, created at the top of the atmosphere, are observed to have longer lifetimes in the observer's frame as compared with their proper lifetimes.

The result (5.43) provides us with an expression for *redshift*. If $\Delta t_1 = \nu_1^{-1}$ is the period of the emitted waves and $\Delta t_0 = \nu_0^{-1}$ the observed period, then

$$\nu_0 = \nu_1 a(t_1) \,. \tag{5.44}$$

Rewriting this result in terms of redshift z,

$$z = \frac{\lambda_0 - \lambda_e}{\lambda_e} = \frac{\lambda_0}{\lambda_e} - 1 = \frac{\nu_1}{\nu_0} - 1 \,, \tag{5.45}$$

that is,

$$a(t_1) = \frac{1}{1+z} \,. \tag{5.46}$$

This is one of the most important relations in cosmology and displays the real meaning of the redshifts of galaxies. *Redshift is a measure of the scale factor of the Universe when the radiation was emitted by the source.* When we observe a galaxy with redshift $z = 1$, the scale factor of the Universe when the light was emitted was $a(t) = 0.5$, that is, the distances between fundamental observers (or galaxies) were half their present values. Note, however, that we obtain no information about *when* the light was emitted. If we did, we could determine directly from observation the function $a(t)$. Understanding of the astrophysical evolution of galaxies is improving all the time and it may eventually be possible to determine $a(t)$ in this way.

One important consequence of this calculation is that we can now derive an expression for the comoving radial distance coordinate r. Equation (5.40) can be written

$$r = \int_{t_1}^{t_0} \frac{c\,\mathrm{d}t}{a(t)} \,. \tag{5.47}$$

Thus, once we know $a(t)$, we can immediately find r by integration. This integral emphasises the point that r is an artificial distance which depends upon how the Universe has expanded between the emission and reception of the radiation.

The expression (5.43) for the time dilation as a function of redshift provides a direct test of the Robertson–Walker formalism. The discovery that supernovae of Type 1a have a narrow dispersion in their absolute magnitudes and have exactly the same light curves, that is, the time-variation of their luminosities throughout the supernova outburst, has made these objects particularly important cosmological tools. Their properties and their use in determining cosmological parameters are described in more detail in Sect. 8.5.3, These standard properties become even more precisely defined when account is taken of a correlation between the maximum luminosity and the width of the light curve (Fig. 5.6). These supernovae have such great luminosities at maximum light that they can be observed at large redshifts. Figure 5.7a shows a plot of the width w of the light curves for a large sample of supernovae from the Calán-Tololo and Supernova Cosmology Program projects as a function of redshift z, or, more precisely, $(1 + z)$ (Goldhaber et al., 2001). In the lower panel (Fig. 5.7b), the observed light curve width w has been divided by $(1+z)$ for each supernova. It can be seen that the observations are in excellent agreement with the expectations of (5.43).

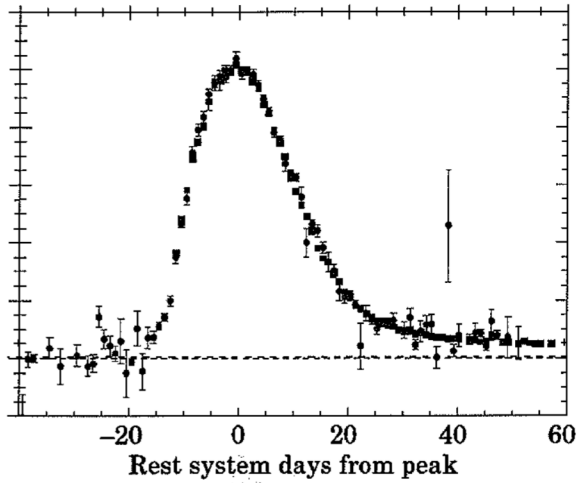

Fig. 5.6. The average time variation of the brightness of a Type 1a supernova from a large sample of supernovae observed in the Calán-Tololo and Supernova Cosmology Program projects. The light curves have been corrected for the effects of time dilation and the luminosity–width correlation (Goldhaber et al., 2001)

Another way of testing the time dilation relation using the remarkably standard properties of the Type 1a supernovae is to use their spectral evolution as a clock to compare the time evolution of low and high redshift supernovae. This test has been carried out for the supernova SN 1997ex, which had redshift 0.361, by members of the Supernova Cosmology Program team (Foley et al., 2005). The time between the first two spectra was 24.88 days and between the first and third spectra 30.95 days. The amount of aging in the supernova rest frame should be a factor of $1/(1 + z)$ smaller corresponding to ages of 18.28 and 22.74 days. The spectral feature age technique applied to the Keck spectra observed for the supernova showed that the corresponding elapsed times in the supernova rest frame were 16.97 ± 2.75 and 18.01 ± 3.14 days, respectively, in excellent agreement with the expectations of cosmological time dilation. Similar results are found from the ESSENCE programme which involves a large consortium of the key players in the Type 1a supernova area (Wood-Vasey et al., 2007).

5.5.2 Hubble's Law

In terms of proper distances, Hubble's law can be written $v = H\varrho$ and so

$$\frac{\mathrm{d}\varrho}{\mathrm{d}t} = H\varrho \,. \tag{5.48}$$

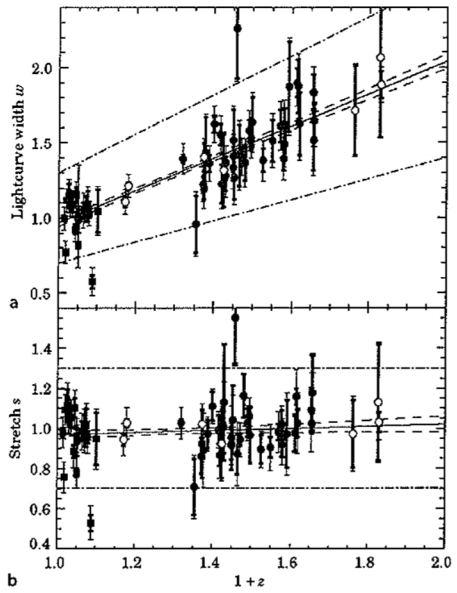

Fig. 5.7. a The observed width w of the light curves of Type 1a supernovae plotted against $(1+z)$. The *low redshift squares* are from the Calán-Tololo supernova programme and the *high redshift circles* are for a subset of 35 supernovae from the Supernova Cosmology Program (SCP). The *open circles* are for the remainder of the 42 SCP fully analyzed supernovae. The band delineated by the *dash-dotted lines* corresponds to stretch values 0.7 to 1.3 which encompass the bulk of the data, except for two outliers. The best-fitting linear relation and the 1σ limits are also shown. **b** The stretch s plotted against $(1+z)$. Stretch is defined as the observed light curve width w divided by $(1+z)$ for each supernova. The notation is the same as in **a** (Goldhaber et al., 2001)

We have written H rather than H_0 in Hubble's law since a 'Hubble's constant' H can be defined at any epoch as we show below. Substituting $\varrho = a(t)r$, we find that

$$r\frac{\mathrm{d}a(t)}{\mathrm{d}t} = Ha(t)r\,, \tag{5.49}$$

that is,

$$H = \dot{a}/a\,. \tag{5.50}$$

Since we measure Hubble's constant H_0 at the present epoch, $t = t_0$, $a = 1$, we find

$$H_0 = (\dot{a})_{t_0}\,. \tag{5.51}$$

Thus, Hubble's constant H_0 defines the present expansion rate of the Universe. Notice that we can define a value of Hubble's constant at any epoch through the more general relation

$$H(t) = \dot{a}/a\,. \tag{5.52}$$

5.5.3 Angular Diameters

The great simplification which results from the use of the Robertson–Walker metric in the form (5.33) is apparent in working out the angular size of an object of proper length d perpendicular to the radial coordinate at redshift z. The relevant spatial component of the metric (5.33) is the term in $\mathrm{d}\theta$. The proper length d of an object at redshift z, corresponding to scale factor $a(t)$, is given by the increment of proper length perpendicular to the radial direction in the metric (5.33), that is,

$$d = a(t)\,\Re \sin\left(\frac{r}{\Re}\right)\Delta\theta = a(t)D\,\Delta\theta = \frac{D\Delta\theta}{(1+z)}\,; \tag{5.53}$$

$$\Delta\theta = \frac{d(1+z)}{D}\,, \tag{5.54}$$

where we have introduced a *distance measure* $D = \Re \sin(r/\Re)$. For small redshifts, $z \ll 1$, $r \ll \Re$, (5.54) reduces to the Euclidean relation $d = r\Delta\theta$.

The expression (5.54) can also be written in the form

$$\Delta\theta = \frac{d}{D_\mathrm{A}}\,, \tag{5.55}$$

so that the relation between d and $\Delta\theta$ looks like the standard Euclidean relation. To achieve this, we have to introduce another distance measure $D_\mathrm{A} = D/(1+z)$ which is known as the *angular diameter distance* and which is often used in the literature.

Another useful calculation is the angular diameter of an object which continues to partake in the expansion of the Universe. This is the case for infinitesimal perturbations in the expanding Universe. A good example is the angular diameter which large-scale structures present in the Universe today would have subtended at an earlier epoch, say, the epoch of recombination, if they had simply expanded with the Universe. This calculation is used to work out physical sizes today corresponding to the angular scales of the fluctuations observed in the Cosmic Microwave Background Radiation. If the physical size of the object is $d(t_0)$ now and it expanded with the Universe, its physical size at redshift z was $d(t_0)a(t) = d(t_0)/(1+z)$. Therefore, the object subtended an angle

$$\Delta\theta = \frac{d(t_0)}{D}\,. \tag{5.56}$$

Notice that in this case the $(1+z)$ factor has disappeared from (5.53).

5.5.4 Apparent Intensities

Suppose a source at redshift z has luminosity $L(\nu_1)$ (measured in $\mathrm{W\,Hz^{-1}}$), that is, the total energy emitted over 4π steradians per unit time per unit frequency interval. What is the flux density $S(\nu_0)$ of the source at the observing frequency ν_0, that is, the energy received per unit time, per unit area and per unit bandwidth ($\mathrm{W\,m^{-2}\,Hz^{-1}}$) where $\nu_0 = a(t_1)\nu_1 = \nu_1/(1+z)$? Suppose the source emits $N(\nu_1)$ photons of energy $h\nu_1$ in the bandwidth ν_1 to $\nu_1 + \Delta\nu_1$ in the proper time interval Δt_1. Then the luminosity $L(\nu_1)$ of the source is

$$L(\nu_1) = \frac{N(\nu_1)\, h\nu_1}{\Delta\nu_1 \Delta t_1} \,. \tag{5.57}$$

These photons are distributed over a 'sphere' centred on the source at epoch t_1 and, when the 'shell' of photons arrives at the observer at the epoch t_0, a certain fraction of them is intercepted by the telescope. The photons are observed at the present epoch t_0 with frequency $\nu_0 = a(t_1)\nu_1$, in a proper time interval $\Delta t_0 = \Delta t_1/a(t_1)$ and in the waveband $\Delta\nu_0 = a(t_1)\Delta\nu_1$.

We also need to know how the photons spread out over a sphere between the epochs t_1 and t_0, that is, we must relate the diameter of our telescope Δl to the angular diameter $\Delta\theta$ which it subtends at the source at epoch t_1. The metric (5.33) provides an elegant answer. The proper distance Δl refers to the present epoch at which $R(t) = 1$ and hence

$$\Delta l = D\Delta\theta \,, \tag{5.58}$$

where $\Delta\theta$ is the angle measured by a fundamental observer, located at the source.

We can also understand this result by considering how the photons emitted by the source spread out over solid angle $\mathrm{d}\Omega$, as observed from the source in the curved geometry. If the Universe were not expanding, the surface area over which the photons would be observed at a time t after their emission would be

$$\mathrm{d}A = R_\mathrm{c}^2 \sin^2\frac{x}{R_\mathrm{c}}\,\mathrm{d}\Omega \,, \tag{5.59}$$

where $x = ct$. In the expanding Universe, R_c changes as the Universe expands and so, in place of the expression x/R_c, we should write

$$\frac{1}{\Re}\int_{t_1}^{t_0}\frac{c\,\mathrm{d}t}{a} = \frac{r}{\Re} \,, \tag{5.60}$$

where r is the comoving radial distance coordinate. Thus,

$$\mathrm{d}A = \Re^2 \sin^2\frac{r}{\Re}\,\mathrm{d}\Omega \,. \tag{5.61}$$

Therefore, the diameter of the telescope as observed from the source is $\Delta l = D\Delta\theta$. Notice how the use of the comoving radial distance coordinate takes account of the changing geometry of the Universe in this calculation. Notice also the difference between (5.54) and (5.58). They correspond to angular diameters measured in opposite directions along the light cone. The factor of $(1+z)$ difference between them is part of a more general relation concerning angular diameter measures along light cones which is known as the *reciprocity theorem*.

Therefore, the surface area of the telescope is $\pi\Delta l^2/4$ and the solid angle subtended by this area at the source is $\Delta\Omega = \pi\Delta\theta^2/4$. The number of photons incident upon the telescope in time Δt_0 is therefore

$$N(\nu_1)\Delta\Omega/4\pi \,, \tag{5.62}$$

but they are now observed with frequency ν_0. Therefore, the flux density of the source, that is, the energy received per unit time, per unit area and per unit bandwidth is

$$S(\nu_0) = \frac{N(\nu_1)\, h\nu_0\, \Delta\Omega}{4\pi\Delta t_0\, \Delta\nu_0\, (\pi/4)\Delta l^2} \,. \tag{5.63}$$

We can now relate the quantities in (5.63) to the properties of the source, using (5.43) and (5.44)

$$S(\nu_0) = \frac{L(\nu_1)a(t_1)}{4\pi D^2} = \frac{L(\nu_1)}{4\pi D^2(1+z)} \,. \tag{5.64}$$

If the spectra of the sources are of power law form, $L(\nu) \propto \nu^{-\alpha}$, this relation becomes

$$S(\nu_0) = \frac{L(\nu_0)}{4\pi D^2(1+z)^{1+\alpha}} \,. \tag{5.65}$$

We can repeat the analysis for *bolometric* luminosities and flux densities. In this case, we consider the total energy emitted in a finite bandwidth $\Delta\nu_1$ which is received in the bandwidth $\Delta\nu_0$, that is

$$\begin{aligned} L_\mathrm{bol} = L(\nu_1)\Delta\nu_1 &= 4\pi D^2 S(\nu_0)(1+z) \times \Delta\nu_0(1+z) \\ &= 4\pi D^2(1+z)^2 S_\mathrm{bol} \,, \end{aligned} \tag{5.66}$$

where the bolometric flux density is $S_\mathrm{bol} = S(\nu_0)\Delta\nu_0$. Therefore,

$$S_\mathrm{bol} = \frac{L_\mathrm{bol}}{4\pi D^2(1+z)^2} = \frac{L_\mathrm{bol}}{4\pi D_\mathrm{L}^2} \,. \tag{5.67}$$

The quantity $D_\mathrm{L} = D(1+z)$ is called the *luminosity distance* of the source since this definition makes the relation between S_bol and L_bol look like an inverse square law. The bolometric luminosity can be integrated over any suitable bandwidth so long as the corresponding redshifted bandwidth is used to measure the bolometric flux density at the present epoch,

$$\sum_{\nu_0} S(\nu_0)\Delta\nu_0 = \frac{\sum_{\nu_1} L(\nu_1)\Delta\nu_1}{4\pi D^2(1+z)^2} = \frac{\sum_{\nu_1} L(\nu_1)\Delta\nu_1}{4\pi D_\mathrm{L}^2} \,. \tag{5.68}$$

The formula (5.64) is the best expression for relating the observed intensity $S(\nu_0)$ to the intrinsic luminosity of the source $L(\nu_1)$. We can also write (5.67) in terms of the luminosity of the source at the observing frequency ν_0 as

$$S(\nu_0) = \frac{L(\nu_0)}{4\pi D_{\rm L}^2}\left[\frac{L(\nu_1)}{L(\nu_0)}(1+z)\right], \tag{5.69}$$

but this now requires knowledge of the spectrum of the source $L(\nu)$. The last term in square brackets is a form of what is known as the *K-correction*. K-corrections were introduced by the pioneer optical cosmologists in the 1930s in order to 'correct' the apparent magnitude of distant galaxies for the effects of redshifting their spectra when observations are made through standard filters with a fixed mean observing frequency ν_0 (Sandage, 1961b). Taking logarithms and multiplying by -2.5, we can write (5.69) in terms of absolute (M) and apparent (m) magnitudes through the relations $M = \text{constant} - 2.5\log_{10} L(\nu_0)$ and $m = \text{constant} - 2.5\log_{10} S(\nu_0)$. We find

$$M = m - 5\log_{10}(D_{\rm L}) - K(z) - 2.5\log_{10}(4\pi), \tag{5.70}$$

where

$$K(z) = -2.5\log_{10}\left[\frac{L(\nu_1)}{L(\nu_0)}(1+z)\right]. \tag{5.71}$$

This form of K-correction is correct for *monochromatic* flux densities and luminosities. In the case of observations in the optical waveband, apparent magnitudes are measured through standard filters which usually have quite wide pass-bands. Therefore, to determine the appropriate K-corrections, the spectral energy distribution of the galaxy has to be convolved with the transmission function of the filter in the rest frame and at the redshift of the galaxy. This is a straightforward calculation once the spectrum of the object is known.

Although I prefer to work directly with (5.64) and take appropriate averages, K-corrections are rather firmly established in the literature and it is often convenient to use the term to describe the effects of shifting the emitted spectrum into the observing wavelength window.

5.5.5 Number Densities

We often need to know the number of objects in a particular redshift interval, z to $z + \mathrm{d}z$. Since there is a one-to-one relation between r and z, the problem is straightforward because, by definition, r is a radial proper distance coordinate defined *at the present epoch*. Therefore, the number of objects in the interval of comoving radial coordinate distance r to $r + \mathrm{d}r$ is given by results already obtained in Sect. 5.3. The space–time diagram shown in Fig. 5.5 illustrates how we can evaluate the numbers of objects in the comoving distance interval r to $r + \mathrm{d}r$ entirely by working in terms of *comoving volumes* at the present epoch. At the present epoch, the radius of curvature of the spatial geometry is $\Re$ and so the volume of a spherical shell of thickness $\mathrm{d}r$ at comoving distance coordinate r is

$$\mathrm{d}V = 4\pi\Re^2 \sin^2(r/\Re)\,\mathrm{d}r = 4\pi D^2\,\mathrm{d}r\,. \tag{5.72}$$

Therefore, if N_0 is the present space density of objects and their number is conserved as the Universe expands,

$$\mathrm{d}N = N(z)\,\mathrm{d}z = 4\pi N_0 D^2\,\mathrm{d}r\,. \tag{5.73}$$

The definition of comoving coordinates automatically takes care of the expansion of the Universe. Another way of expressing this result is to state that (5.73) gives the number density of objects in the redshift interval z to $z + \mathrm{d}z$, assuming the *comoving number density* of the objects is unchanged with cosmic epoch. If, for some reason, the comoving number density of objects changes with cosmic epoch as, say, $f(z)$ with $f(z = 0) = 1$, then the number of objects expected in the redshift interval $\mathrm{d}z$ is

$$\mathrm{d}N = N(z)\,\mathrm{d}z = 4\pi N_0\, f(z)\, D^2\,\mathrm{d}r\,. \tag{5.74}$$

5.5.6 The Age of the Universe

Finally, let us work out an expression for the age of the Universe, T_0, from a rearranged version of (5.39). The basic differential relation is

$$-\frac{c\,\mathrm{d}t}{a(t)} = \mathrm{d}r\,, \tag{5.75}$$

and hence

$$T_0 = \int_0^{t_0} \mathrm{d}t = \int_0^{r_{\rm max}} \frac{a(t)\,\mathrm{d}r}{c}\,, \tag{5.76}$$

where $r_{\rm max}$ is the comoving distance coordinate corresponding to $a = 0$, $z = \infty$.

5.6 Summary

The results we have derived can be used to work out the relations between intrinsic properties of objects and observables for any isotropic, homogeneous world model. Let us summarise the procedures described above:

1. First work out from theory, or otherwise, the function $a(t)$ and the curvature of space at the present epoch $\kappa = \Re^{-2}$. Once we know $a(t)$, we know the redshift–cosmic time relation.

2. Now work out the *comoving radial distance coordinate* r from the integral

$$r = \int_{t_1}^{t_0} \frac{c\,\mathrm{d}t}{a(t)} \,. \tag{5.77}$$

Recall what this expression means: the proper distance interval $c\,\mathrm{d}t$ at epoch t is projected forward to the present epoch t_0 by the scale factor $a(t)$. This integration yields an expression for r as a function of redshift z.

3. Next, work out the *distance measure* D from

$$D = \mathfrak{R} \sin \frac{r}{\mathfrak{R}} \,. \tag{5.78}$$

This relation determines D as a function of redshift z.

4. If so desired, the *angular diameter distance* $D_\mathrm{A} = D/(1+z)$ and the *luminosity distance* $D_\mathrm{L} = D(1+z)$ can be introduced to relate physical sizes and luminosities to angular diameters and flux densities respectively.
5. The number of objects $\mathrm{d}N$ in the redshift interval $\mathrm{d}z$ and solid angle Ω can be found from the expression

$$\mathrm{d}N = \Omega N_0 D^2\,\mathrm{d}r \,, \tag{5.79}$$

where N_0 is the number density of objects at the present epoch which are assumed to be conserved as the Universe expands.

We will develop some explicit solutions for these functions in Chap. 7.

6 An Introduction to Relativistic Gravity

The standard world models which are used as the framework for astrophysical cosmology and for studying the problems of galaxy formation are based upon Einstein's General Theory of Relativity. General Relativity is a beautiful theory but it requires a thorough understanding of tensor calculus in four-dimensional non-Euclidean spaces to appreciate fully Einstein's epoch-making achievement. This is beyond the scope of the present text and so Sects. 6.1 to 6.5 are intended to provide some flavour of the full theory and to introduce some key ideas which will be needed later.[1] In Sect. 6.6, the experimental and observational status of General Relativity is reviewed and it is shown that it has triumphantly survived the many critical tests of the theory which have been devised since its inception in 1915.[2] If you are happy to accept General Relativity at its face value, you may advance to Chap. 7.

6.1 The Principle of Equivalence

As Einstein expressed it many years later:

> I was sitting in a chair in the patent office in Bern when all of a sudden a thought occurred to me: 'If a person falls freely he will not feel his own weight'. I was startled. This simple thought made a deep impression upon me. It impelled me towards a theory of gravitation.

Expressed in more technical terms, a key consideration which led Einstein to the General Theory was the null result of the Eötvös experiment, which showed rather precisely that gravitational mass m_g is proportional to inertial mass m_I. Following Will's exposition (Will, 2006), the deviations from linearity can be written

$$m_\mathrm{g} = m_\mathrm{I} + \sum_\mathrm{A} \frac{\eta^\mathrm{A} E^\mathrm{A}}{c^2} \,. \tag{6.1}$$

[1] I have given a more extended introduction to the theory in Chap. 17 of my book *Theoretical Concepts in Physics* (Longair, 2003). For an up-to-date physical exposition of General Relativity, the book *General Relativity: An Introduction for Physicists* by Hobson, Efstathiou and Lasenby can be recommended (Hobson et al., 2006).

[2] This assessment is based upon the superb article by Will published in *Living Reviews in Relativity* (Will, 2006).

$E^{\rm A}$ is the internal energy of the body generated by interaction A, η is a dimensionless parameter that measures the strength of the violation of the linearity of the relation between $m_{\rm g}$ and $m_{\rm I}$ induced by that interaction, and c is the speed of light. The internal energy terms include all the mass-energy terms which can contribute to the inertial mass of the body, for example, the body's rest energy, its kinetic energy, its electromagnetic energy, weak-interaction energy, binding energy and so on. If the inertial and gravitational masses were not exactly linearly proportional to each other, there would be a finite value $\eta^{\rm A}$ which would be exhibited as an difference in the accelerations of bodies of the same inertial mass composed of different materials. A measurement of, or limit to, the fractional difference in accelerations between two bodies yields the quantity known as the 'Eötvös ratio',

$$\eta = 2\frac{|a_1 - a_2|}{a_1 + a_2} = \sum_{\rm A} \eta^{\rm A}\left(\frac{E_1^{\rm A}}{m_1 c^2} - \frac{E_2^{\rm A}}{m_2 c^2}\right), \tag{6.2}$$

where the subscript 'I' has been dropped from the inertial masses.

A summary of the limits to η from the time of Eötvös to the present day is shown in Fig. 6.1 which includes the famous torsion balance experiments of Eötvös, Dicke, Braginsky and their collaborators. From the late 1980s onwards, numerous experiments were carried out primarily to search for a "fifth force" but their null results also provided limits to the Eötvös ratio. For example, as discussed by Will, the "Eöt-Wash" experiments carried out at the University of Washington used a sophisticated torsion balance tray to compare the accelerations of various materials toward local topographical features on Earth, movable laboratory masses, the Sun and the Galaxy and provided limits of $\eta \leq 3 \times 10^{-13}$. Thus, gravitational mass $m_{\rm g}$ is proportional to inertial mass $m_{\rm I}$ to better than one part in 3×10^{12}.

The principle of equivalence asserts that the gravitational field $\mathbf{g}$ at any point in space can be precisely replaced by an accelerated frame of reference $\mathbf{a}$. In Newton's terminology, 'mass' is proportional to 'weight'. The statement that, locally inertial and gravitational mass are the same is known as the *weak equivalence principle*. Einstein's version of the principle is much stronger. In his own words:

> All local, freely falling, non-rotating laboratories are fully equivalent for the performance of all physical experiments.

By *free-fall*, we mean a frame of reference which is accelerated at the local gravitational acceleration at that point in space, $\mathbf{a} = \mathbf{g}$. This statement formally identifies inertial and gravitational mass, since the force acting on a particle in a gravitational field depends upon the particle's *gravitational mass*, whereas the acceleration depends upon its *inertial mass*.

A more transparent statement of the principle is given by Will who clarifies exactly what is assumed in what he calls the *Einstein equivalence principle*. In Will's words:

> The Einstein equivalence principle (EEP) is a more powerful and far-reaching concept; it states that:
>
> 1. The weak equivalence principle is valid.

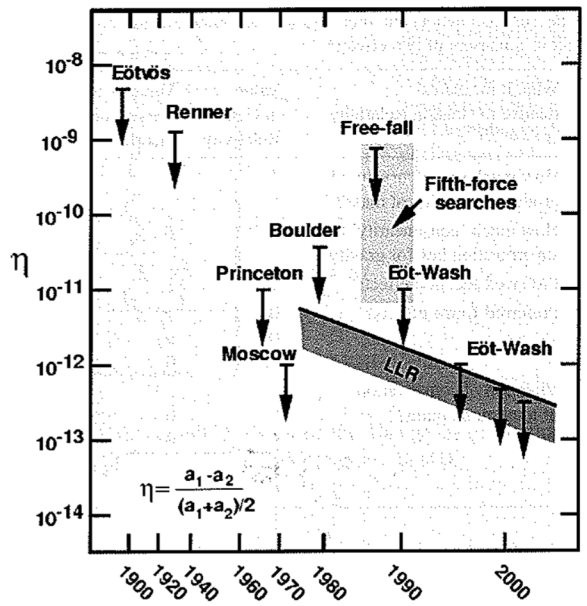

Fig. 6.1. Selected tests of the weak equivalence principle, showing bounds on η, which measures the fractional differences in accelerations of different materials or bodies. The free-fall and Eöt-Wash experiments were originally performed to search for a fifth force (*light grey region*, representing many experiments). The *dark grey band* shows bounds on η for gravitating bodies from lunar laser ranging (LLR) (Will, 2006)

> 2. The outcome of any local non-gravitational experiment is independent of the velocity of the freely falling reference frame in which it is performed.
> 3. The outcome of any local non-gravitational experiment is independent of where and when in the universe it is performed.
>
> The second piece of EEP is called local Lorentz invariance (LLI), and the third piece is called local position invariance (LPI).

The importance of this description of the assumptions behind the General Theory of Relativity is that it makes clear the scope for developing alternative theories of relativistic gravity. Thus, if either (2) or (3) were to be relaxed, a much wider range of possible theories of relativistic gravity could be developed and these are illustrated by the range of additional parameters listed in Table 6.1. For our modest ambitions in this chapter, we simply use the parameterised post-Newtonian (PPN) coefficients listed in Table 6.1 as measures of the success of standard General Relativity.

Table 6.1. The PPN parameters and their significance (note that α_3 has been shown twice to indicate that it is a measure of two effects)

Parameter	What it measures relative to General Relativity	Value in General Relativity	Value in semi-conservative theories	Value in fully conservative theories
γ	How much space-curvature is produced by unit rest mass?	1	γ	γ
β	How much 'non-linearity' in the superposition law for gravity?	1	β	β
ξ	Preferred-location effects?	0	ξ	ξ
α_1	Preferred-frame effects?	0	α_1	0
α_2		0	α_2	0
α_3		0	0	0
α_3	Violation of conservation			
ζ_1	of total momentum?	0	0	0
ζ_2		0	0	0
ζ_3		0	0	0
ζ_4		0	0	0

The principle of equivalence has profound consequences for our understanding of the nature of space and time in a gravitational field. Let us illustrate some of these by two elementary examples.

6.2 The Gravitational Redshift

In the first example, we replace a stationary frame of reference located in a uniform gravitational field $\mathbf{g}$ by a frame of reference which is accelerated in the opposite direction. Consider a light wave of frequency ν propagating from the ceiling to the floor of a lift in a gravitational field $\mathbf{g} = -\mathbf{a}$ (Fig. 6.2).

We assume that the acceleration is small. If the height of the lift is h, a light signal travels from the ceiling to the floor in a time $t = h/c$. According to the principle of equivalence, we can replace the gravitational field by an accelerated frame of reference and so, after time t, the floor is accelerated to a speed $u = at = |\mathbf{g}|t$. Hence,

$$u = \frac{|\mathbf{g}|h}{c} . \tag{6.3}$$

Therefore, the light wave is observed with a higher frequency when it arrives at the floor of the lift because of the Doppler effect. To first order in u/c, the observed frequency ν' is

$$\nu' = \nu\left(1 + \frac{u}{c}\right) = \nu\left(1 + \frac{|\mathbf{g}|h}{c^2}\right) . \tag{6.4}$$

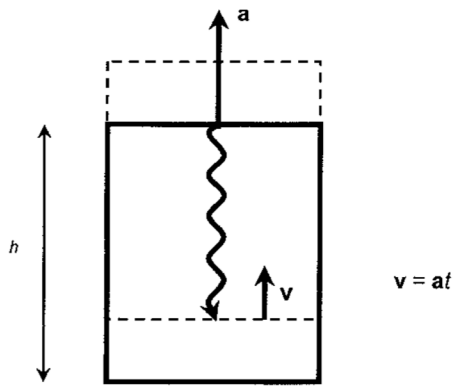

Fig. 6.2. Illustrating the gravitational redshift/blueshift of an electromagnetic wave propagating from the ceiling to the floor of a stationary lift in a gravitational field according to the principle of equivalence

Let us now express this result in terms of the change in gravitational potential between the ceiling and floor of the lift. Since $\mathbf{g} = -\text{grad}\,\phi$

$$|\mathbf{g}| = -\frac{\Delta\phi}{h} . \tag{6.5}$$

Notice that, because of the attractive nature of the gravitational force, ϕ is more negative at $h = 0$ than at the ceiling. Therefore,

$$\nu' = \nu\left(1 - \frac{\Delta\phi}{c^2}\right) . \tag{6.6}$$

This is the formula for the *gravitational redshift* z_g in the 'Newtonian' limit. Recalling the definition of redshift,

$$z = \frac{\lambda_{obs} - \lambda_{em}}{\lambda_{em}} = \frac{\nu - \nu'}{\nu} , \tag{6.7}$$

we find that

$$z_g = \frac{\Delta\phi}{c^2} . \tag{6.8}$$

In this simple example, since $\Delta\phi$ is negative, z_g is also negative corresponding to a gravitational blueshift rather than redshift. If the light waves propagated from the floor to the ceiling, we would obtain a redshift of the same magnitude. Thus, the frequency of the waves depends upon to the *gravitational potential* in which the light waves are propagated.

A test of the expression (6.8) for the gravitational redshift was proposed by Eddington in 1924. He estimated that the gravitational redshift of the lines in the spectrum of the white dwarf star Sirius B should be $cz_g = 20\,\text{km}\,\text{s}^{-1}$. The value measured by Adams in 1925 was $19\,\text{km}\,\text{s}^{-1}$. Eddington was jubilant (Eddington, 1926):

> Prof. Adams has thus killed two birds with one stone. He has carried out a new test of Einstein's theory of General Relativity, and has shown that matter at least 2000 times denser than platinum is not only possible, but actually exists in the stellar universe.

Laboratory experiments to measure the gravitational redshift were carried out by Pound, Rebka and Snider who measured the difference in redshift of γ-ray photons moving up then down a tower 22.5 m high at Harvard University using the Mössbauer effect (Pound and Rebka, 1960; Pound and Snider, 1965). In this effect, the recoil effects of the emission and absorption of the γ-ray photons are zero since the momentum is absorbed by the whole atomic lattice. The γ-ray resonance is therefore very sharp indeed and only tiny Doppler shifts are needed to move off resonance absorption. In the Harvard experiment, the difference in redshifts for γ-ray photons moving up and down the tower was:

$$z_{\text{up}} - z_{\text{down}} = \frac{2gh}{c^2} = 4.905 \times 10^{-15} . \tag{6.9}$$

The measured value was $(4.900 \pm 0.037) \times 10^{-15}$, a precision of about 1%. Notice the key point that the gravitational redshift is incompatible with special relativity, according to which the observers at the top and bottom of the tower are at rest in the same inertial frame of reference.

Suppose we now write (6.6) in terms of the period of the waves T. Then,

$$T' = T\left(1 + \frac{\Delta\phi}{c^2}\right) . \tag{6.10}$$

This expression is exactly the same as the time dilation formula between inertial frames of reference in special relativity, only now the expression refers to different locations in the gravitational field. This expression for time dilation is exactly what would be evaluated for any time interval and so we can write in general

$$\mathrm{d}t' = \mathrm{d}t\left(1 + \frac{\Delta\phi}{c^2}\right) . \tag{6.11}$$

Let us now take the gravitational potential to be zero at infinity and measure the gravitational potential at any point in the field relative to that value. We assume that we are in the weak field limit in which changes in the gravitational potential are small. Then, at any point in the gravitational field, we can write

$$\mathrm{d}t'^2 = \mathrm{d}t^2\left[1 + \frac{\phi(r)}{c^2}\right]^2 , \tag{6.12}$$

where $\mathrm{d}t$ is the time interval measured at $\phi = 0$, that is, at $r = \infty$. Since $\phi(r)/c^2$ is small, we can write this expression as

$$\mathrm{d}t'^2 = \mathrm{d}t^2\left[1 + \frac{2\phi(r)}{c^2}\right] . \tag{6.13}$$

If we now adopt the Newtonian expression for the gravitational potential for a point mass M,

$$\phi(r) = -\frac{GM}{r} , \tag{6.14}$$

we find

$$\mathrm{d}t'^2 = \mathrm{d}t^2\left(1 - \frac{2GM}{rc^2}\right) . \tag{6.15}$$

Let us now introduce this expression for the time interval into the standard Minkowski metric of special relativity,

$$\mathrm{d}s^2 = \mathrm{d}t'^2 - \frac{1}{c^2}\mathrm{d}l^2 , \tag{6.16}$$

where $\mathrm{d}l$ is the differential element of proper distance. The metric of space–time about the point mass can therefore be written as

$$\mathrm{d}s^2 = \mathrm{d}t^2\left(1 - \frac{2GM}{rc^2}\right) - \frac{1}{c^2}\mathrm{d}l^2 . \tag{6.17}$$

This calculation shows how the metric coefficients become more complicated than those of Minkowski space–time when we attempt to derive a relativistic theory of gravity. Notice how careful we have to be about keeping track of time in General Relativity. The time interval measured by an observer at a point in the gravitational field is $\mathrm{d}t'$; the interval $\mathrm{d}t$ is a time interval at infinity. The gravitational redshift relates these differences in time keeping. Notice further that both of these are different from the time measured by an observer in free-fall in the gravitational field.

6.3 The Bending of Light Rays

Let us show how the expression for $\mathrm{d}l$ has to be changed as well. Consider the propagation of light rays in our lift but now travelling perpendicular to the gravitational acceleration. We again use the principle of equivalence to replace the stationary lift in a gravitational field by an accelerated lift in free space (Fig. 6.3).

In the time the light ray propagates across the lift, a distance l, the lift moves upwards a distance $\frac{1}{2}|\mathbf{g}|t^2$. Therefore, in the frame of reference of the accelerated lift, and also in the stationary frame in the gravitational field, the light ray follows a parabolic path as illustrated in Figs. 6.3a–c. Let us approximate the light path by a circular arc of radius R. The length of the chord d across the circle is then

$$d^2 = \tfrac{1}{4}|\mathbf{g}|^2 t^4 + l^2 . \tag{6.18}$$

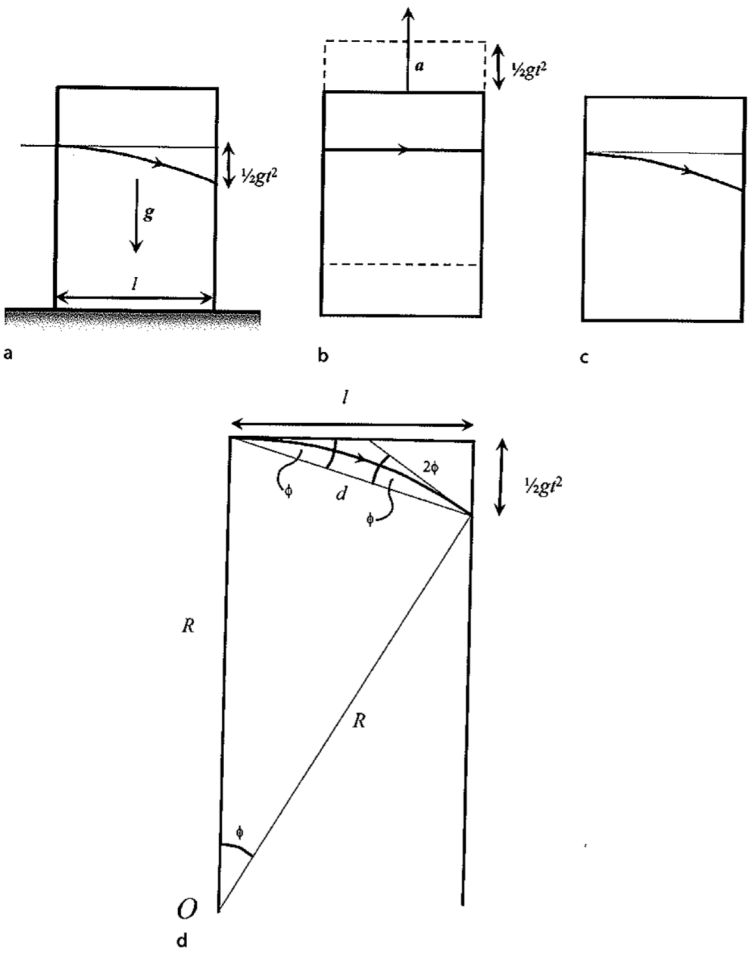

Fig. 6.3a–d. Illustrating the application of the principle of equivalence to the propagation of a light ray in a gravitational field and in a uniformly accelerated lift. In the equivalent accelerated frame of reference, the light ray travels along a curved path

Now, from the geometry of the diagram, it can be seen that $\phi = |\mathbf{g}|t^2/2l$. Hence, since $R\phi = d$,

$$R^2 = \frac{d^2}{\phi^2} = l^2 + \frac{4l^4}{|\mathbf{g}|^2 t^4} \,. \tag{6.19}$$

Now, $\frac{1}{2}|\mathbf{g}|t^2 \ll l$, $l = ct$ and so the second term in (6.19) is much larger than the first. Therefore,

$$R = \frac{2l^2}{|\mathbf{g}|t^2} = \frac{2c^2}{|\mathbf{g}|} \,. \tag{6.20}$$

Thus, the radius of curvature of the path of the light ray depends only upon the local gravitational acceleration $|\mathbf{g}|$. Since $\mathbf{g}$ is determined by the gradient of the gravitational potential, it follows that the *curvature of the paths of light rays depends upon the mass distribution.*

6.4 Further Complications

The consequences of these two elementary calculations are that the rate at which clocks tick depends upon the gravitational potential in which they are located and the paths of light rays are bent by the gravitational influence of the mass-energy distribution. In other words, not only is space curved but, more generally, *space–time* is curved. Neither the space nor time coordinates take the simple 'Euclidean' values which appear in the Minkowski metric, which can be written in polar coordinates

$$\mathrm{d}s^2 = \mathrm{d}t^2 - \frac{1}{c^2}\left[\mathrm{d}r^2 + r^2(\mathrm{d}\theta^2 + \sin^2\theta^2 \mathrm{d}\phi^2)\right] \,. \tag{6.21}$$

It must be emphasised that the arguments of Sects. 6.2 and 6.3 are *illustrative* of how Newtonian gravity has to be modified to incorporate the principle of equivalence and the rules of special relativity and many unsatisfactory steps were involved.

To complicate matters further, any relativistic theory of gravity must be non-linear. This follows from Einstein's mass-energy relation $E = mc^2$ as applied to the gravitational field. The gravitational field due to some mass distribution has a certain local energy density at each point in space. Since $E = mc^2$, it follows that there is a certain inertial mass density in the gravitational field which is itself a source of gravitational field. This property contrasts with that of, say, an electric field distribution. This possesses a certain amount of electromagnetic field energy and a corresponding inertial mass density but this does not generate additional electrostatic charge. Thus, relativistic gravity is intrinsically a non-linear theory and this accounts for a great deal of its complexity.

This feature of relativistic gravity was recognised by Einstein in 1912. From his student days, he vaguely remembered Gauss's theory of surfaces and consulted his old school friend, the mathematician Marcel Grossmann, about the most general forms of transformation between frames of reference for metrics of the form

$$\mathrm{d}s^2 = g_{\mu\nu}\,\mathrm{d}x^\mu\,\mathrm{d}x^\nu. \tag{6.22}$$

Although outside Grossmann's field of expertise, he soon came back with the answer that the most general transformation formulae were the Riemannian geometries, but that they had the 'bad feature' that they are non-linear. Einstein instantly recognised that, on the contrary, this was a great advantage since any satisfactory theory of relativistic gravity must be non-linear.

Finally, although we can eliminate the *acceleration* due to gravity at a particular point in space, we cannot eliminate completely the effects of gravity in the vicinity of that point. This is most easily seen by considering the gravitational field at distance r from a point mass M (Fig. 6.4). It is apparent that we need different freely falling lifts at different points in space in order to eliminate gravity everywhere. Even over very limited regions of space, if we make very precise measurements, neighbouring particles will be observed to begin to move under the influence of the quadrupole field which cannot be eliminated by transforming to a single accelerated reference frame. As an example, consider a standard Euclidean (x, y, z) coordinate frame inside an orbiting Space Station, the z-coordinate being taken in the radial direction. It is a useful exercise to show that, if two test particles are released from rest, with

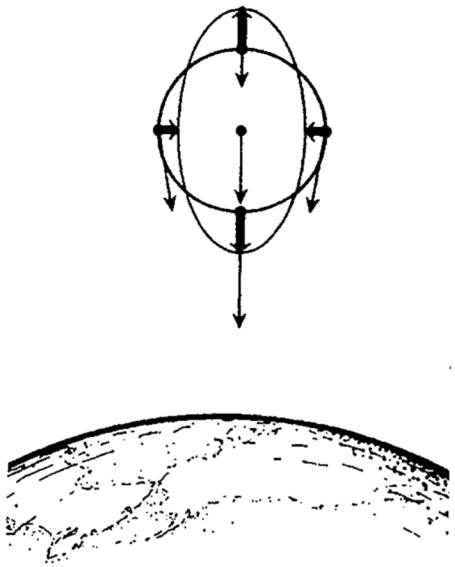

Fig. 6.4. Illustrating the 'tidal forces' which cannot be eliminated when the acceleration due to gravity **g** is replaced by an accelerated reference frame at a particular point in space. In this example, the observer is in centrifugal equilibrium in the field of a point mass. Initially, test masses are located on a sphere about the observer. At a later time, the sphere is distorted into an ellipsoid because of the tidal forces which cannot be eliminated when transforming away the gravitational field at the observer (Penrose, 1997)

an initial separation vector ξ, this separation vector varies with time as

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\begin{bmatrix}\xi^x\\ \xi^y\\ \xi^z\end{bmatrix} = \begin{bmatrix}-GM/r^3 & 0 & 0\\ 0 & -GM/r^3 & 0\\ 0 & 0 & +2GM/r^3\end{bmatrix}\begin{bmatrix}\xi^x\\ \xi^y\\ \xi^z\end{bmatrix}. \tag{6.23}$$

The pleasant aspect of this analysis is that it can be seen that the uncompensated forces depend upon r^{-3}. This is the part of the gravitational field which cannot be eliminated by transforming to a single freely falling frame. Notice that it has the form of a 'tidal force', which depends upon r^{-3}, of exactly the same type which causes Earth–Moon and Earth–Sun tides. We therefore need a theory which reduces locally to Einstein's special relativity in a freely falling frame and which transforms correctly into another freely falling reference frame when we move to a different point in space. There is no such thing as a global Lorentz frame in the presence of a non-uniform gravitational field.

Einstein's General Relativity enables us to find the metric of space–time in the presence of mass-energy. The simplest example is the metric of space–time about a point mass of mass M, the *Schwarzschild metric*, which can be written

$$\mathrm{d}s^2 = \mathrm{d}t^2\left(1-\frac{2GM}{rc^2}\right) - \frac{1}{c^2}\left[\frac{\mathrm{d}r^2}{\left(1-\frac{2GM}{rc^2}\right)} + r^2\left(\mathrm{d}\theta^2 + \sin^2\theta^2\mathrm{d}\phi^2\right)\right], \tag{6.24}$$

where the metric has been written in spherical polar coordinates. The Schwarzschild metric is *exact* for a stationary point mass in General Relativity. Some elements of the Schwarzschild metric are similar to those which were derived in our approximate analyses. For example, the increment of *proper time* is

$$\mathrm{d}t' = \mathrm{d}t\left(1-\frac{2GM}{rc^2}\right)^{1/2}, \tag{6.25}$$

and has the same properties which we derived above, namely, the *coordinate time* t keeps track of how clocks measure time at infinity. Clocks closer to the origin run slower relative to clocks at infinity by the factor $(1-2GM/rc^2)^{1/2}$. This enables us to derive the general expression for *gravitational redshift*. The period of the light waves changes by precisely this factor as the light ray propagates from radius r from the point mass to infinity. Therefore, the change of frequency is

$$\nu' = \nu\left(1-\frac{2GM}{rc^2}\right)^{-1/2}, \tag{6.26}$$

where ν is the frequency measured at infinity. Thus, the redshift z of the radiation is

$$z_g = \frac{\lambda_{\mathrm{obs}}-\lambda_{\mathrm{em}}}{\lambda_{\mathrm{em}}} = \frac{\lambda_{\mathrm{obs}}}{\lambda_{\mathrm{em}}} - 1 = \left(1-\frac{2GM}{rc^2}\right)^{-1/2} - 1, \tag{6.27}$$

or

$$1 + z_g = \left(1 - \frac{2GM}{rc^2}\right)^{-1/2} . \tag{6.28}$$

Light rays emitted from the *Schwarzschild radius* $r_g = 2GM/c^2$ are shifted to infinite wavelengths.[3]

6.5 The Route to General Relativity

Einstein's great achievement was to understand how the features discussed in Sects. 6.1 to 6.4 could be incorporated into a self-consistent theory of relativistic gravity. The remarkable story of Einstein's struggles to discover the theory is told in some detail by Abraham Pais in his splendid scientific biography of Einstein *Subtle is the Lord ...* (Pais, 1982). This is not the place to go into the technical details of what Einstein did. In summary, his thinking was guided by four ideas:

- The influence of gravity on light
- The principle of equivalence
- Riemannian space–time
- The principle of covariance

My recommended approach would be to begin with Rindler's excellent introductory text *Relativity: Special, General, and Cosmological* (Rindler, 2001), and then proceed to either Weinberg's *Gravitation and Cosmology* or Hobson, Efstathiou and Lasenby's *General Relativity: An Introduction for Physicists*, which both describe clearly why General Relativity has to be as complex as it is. In both books the physical content of the theory and the mathematics are elucidated at each stage (Weinberg, 1972; Hobson et al., 2006). Another useful recommendation is d'Inverno's *Introducing Einstein's Relativity* which is particularly clear on the geometric aspects of the theory (d'Inverno, 1992). The understanding of the theory requires considerable effort. Let me outline some of the key steps in its formal development.

6.5.1 Four-Tensors in Relativity

In formulating the Special Theory of Relativity, Einstein realised that all the laws of physics, with the exception of gravity, could be written in Lorentz-invariant form. By this, we mean that the equations are *form-invariant* under Lorentz transformations:

[3] I have given further examples of the use of the Schwarzschild metric to derive expressions for light deflection by point masses and the advance of the perihelion of planetary orbits, as well as an introduction to Schwarzschild black holes, in Chap. 17 of my book *Theoretical Concepts in Physics* (Longair, 2003).

this is often called *Lorentz covariance*. The simplest example is the introduction of *four-vectors* into special relativity; these are designed to be objects which are form-invariant under Lorentz transformations. Just for this subsection, we use the notation used by professional relativists in which the velocity v is measured in units of the speed of light, equivalent to setting the value of $c = 1$. Then, the time coordinate $x^0 = t$ and the spatial components are $x = x^1$, $y = x^2$ and $z = x^3$. As a result, we can write the transformation of a four-vector V^α between two inertial frames of reference in standard configuration in the form

$$V^\alpha \rightarrow V'^\alpha = \Lambda^\alpha_\beta V^\beta , \tag{6.29}$$

where the matrix Λ^α_β is the standard Lorentz transformation

$$\Lambda^\alpha_\beta = \begin{bmatrix} \gamma & -\gamma v & 0 & 0 \\ -\gamma v & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} , \tag{6.30}$$

and $\gamma = (1 - v^2/c^2)^{-1/2}$. The convention of summing over identical indices is adopted in (6.29). Some familiar examples of four-vectors are listed in Table 6.2, along with their translations into more familiar quantities.

Many physical quantities are naturally described in terms of tensors rather than vectors. The natural extension of the concept of four-vectors is then to *four-tensors* which are objects which transform according to the rule

$$T^{\alpha\gamma} \rightarrow T'^{\alpha\gamma} = \Lambda^\alpha_\beta \Lambda^\gamma_\delta T^{\beta\delta} . \tag{6.31}$$

For example, relativists call matter without any internal pressure 'dust' and the *energy–momentum tensor* for dust is $T^{\alpha\beta} = \varrho_0 u^\alpha u^\beta$, where ϱ_0 is the proper mass density of the dust, meaning the density measured by an observer moving with the flow, or a comoving observer; u^α is the velocity four-vector. Writing out the

Table 6.2. Examples of common four-vectors. The second column gives the four-vector notation used in this section. The third column translates the components of the four-vectors into more familiar quantities. In all cases, $c = 1$ and $\gamma = (1 - v^2/c^2)^{-1/2}$

Displacement four-vector	$[x^0, x^1, x^2, x^3]$	$[t, x, y, z]$
Velocity four-vector	$[v^0, v^1, v^2, v^3]$	$[\gamma, \gamma v_x, \gamma v_y, \gamma v_z]$
Momentum four-vector	$[p^0, p^1, p^2, p^3]$	$[\gamma m_0, \gamma m_0 v_x, \gamma m_0 v_y, \gamma m_0 v_z]$
Acceleration four-vector	$[a^0, a^1, a^2, a^3]$	$\left[\gamma \frac{\mathrm{d}\gamma}{\mathrm{d}t}, \gamma \frac{\mathrm{d}(\gamma v_x)}{\mathrm{d}t}, \gamma \frac{\mathrm{d}(\gamma v_y)}{\mathrm{d}t}, \gamma \frac{\mathrm{d}(\gamma v_z)}{\mathrm{d}t}\right]$
Frequency four-vector	$[k^0, k^1, k^2, k^3]$	$[\omega, k_x, k_y, k_z]$
Four-momentum of photon	$[p^0, p^1, p^2, p^3]$	$[\hbar\omega, \hbar k_x, \hbar k_y, \hbar k_z]$

components of $T^{\alpha\beta}$ in terms of the quantities in the third column of Table 6.2, we find

$$T^{\alpha\beta} = \gamma^2 \varrho_0 \begin{bmatrix} 1 & u_x & u_y & u_z \\ u_x & u_x^2 & u_y u_x & u_z u_x \\ u_y & u_x u_y & u_y^2 & u_z u_y \\ u_z & u_x u_z & u_y u_z & u_z^2 \end{bmatrix} , \tag{6.32}$$

where (u_x, u_y, u_z) are the components of the three-velocity measured in the chosen reference frame. It is instructive to note the form of the T^{00} component of this four-tensor, $T^{00} = \gamma^2 \varrho_0$, which corresponds to the total energy density. This quantity has a natural interpretation in special relativity. The observed density of the dust ϱ as it moves with the flow is increased by two powers of the Lorentz factor γ over the proper value ϱ_0. One of these is associated with the formula for the relativistic three-momentum of the dust, $p = \gamma m \mathbf{u}$, and the other with length contraction in the direction of motion of the dust, $l = l_0/\gamma$.

When the pressure cannot be neglected, the energy–momentum tensor becomes

$$T^{\alpha\beta} = (\varrho_0 + p) u^\alpha u^\beta - p g^{\alpha\beta} , \tag{6.33}$$

where $g^{\alpha\beta}$ is the metric tensor, which in the case of special relativity is the matrix

$$g^{\alpha\beta} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} . \tag{6.34}$$

Then, it is a pleasant exercise to show that the equation

$$\partial_\beta T^{\alpha\beta} = 0 , \tag{6.35}$$

expresses the laws of conservation of momentum and energy in relativity, where ∂_β means partial differentiation of the tensor components with respect to β and so the operator ∂_β has the form

$$[\partial/\partial x_0, \partial/\partial x_1, \partial/\partial x_2, \partial/\partial x_3] . \tag{6.36}$$

Maxwell's equations in a vacuum can be written in compact form in terms of the antisymmetric electromagnetic field tensor $F^{\alpha\beta}$

$$F^{\alpha\beta} = \begin{bmatrix} 0 & E_x & E_y & E_z \\ -E_x & 0 & B_z & -B_y \\ -E_y & -B_z & 0 & B_x \\ -E_z & B_y & -B_x & 0 \end{bmatrix} , \tag{6.37}$$

and the current density four-vector $j^\alpha = [\varrho_e, \mathbf{j}]$. I apologise for deviating from my normal practice of using strictly SI units. This form of Maxwell's equations is written

in Heaviside–Lorentz units with $c = 1$. The equation of continuity becomes

$$\partial_\alpha j^\alpha = 0 . \tag{6.38}$$

Maxwell's equations for the relations between electric and magnetic fields and their sources become

$$\partial_\beta F^{\alpha\beta} = j^\alpha . \tag{6.39}$$

Thus, four-tensors provide the natural language for expressing the laws of physics in a form which guarantees that they transform correctly according to the Lorentz transformations.

6.5.2 What Einstein Did

The elementary considerations of Sects. 6.1 to 6.4 indicate that the aim of General Relativity is to incorporate the influence of the mass-energy distribution upon space–time into the metric coefficients $g_{\mu\nu}$. The metric of space–time locally has to reduce to the standard Minkowski metric

$$\mathrm{d}s^2 = \mathrm{d}t^2 - \frac{1}{c^2} \mathrm{d}l^2 . \tag{6.40}$$

Therefore, the natural starting point for the development of general transformations between arbitrary four-dimensional spaces is the *Riemannian metric* of form

$$\mathrm{d}s^2 = \sum_{\mu,\nu} g_{\mu\nu} \, \mathrm{d}x^\mu \mathrm{d}x^\nu = g_{\mu\nu} \, \mathrm{d}x^\mu \mathrm{d}x^\nu , \tag{6.41}$$

where the coordinates x^μ and x^ν define points in four-dimensional space and the interval $\mathrm{d}s^2$ is given by a homogeneous quadratic differential form in these coordinates. The components of the *metric tensor* $g_{\mu\nu}$ vary from point-to-point in space–time and define its local curvature. Since the local curvature defines the properties of the gravitational field, the $g_{\mu\nu}$ can be thought of as being analogous to gravitational potentials.

We need to develop a way of relating the $g_{\mu\nu}$ to the mass-energy distribution, that is, to find the analogue of Poisson's equation in Newtonian gravity which involves second-order partial differential equations. As an illustrative example, in deriving (6.13), we rationalised that g_{00} should have the form

$$g_{00} = \left(1 + \frac{2\phi}{c^2}\right) , \tag{6.42}$$

(see also the Schwarzschild metric (6.24)). Poisson's equation for gravity is

$$\nabla^2 \phi = 4\pi G \varrho , \tag{6.43}$$

and hence, from (6.42) and (6.43), we find that

$$\nabla^2 g_{00} = \frac{8\pi G}{c^2} T_{00} \,. \tag{6.44}$$

This is a crude calculation but it shows why it is reasonable to expect a close relation between the derivatives of $g_{\mu\nu}$ and the corresponding components of the energy–momentum tensor $T_{\mu\nu}$.

The tensor equivalent of this analysis involves the differentiation of tensors and this is where the complications begin: partial differentiation of tensors does not generally yield other tensors. Thus, the definitions of the equivalent vector operations of grad, div and curl are correspondingly more complicated for tensors as compared with vectors. Furthermore, the analysis can no longer be carried out in Minkowski space–time since space–time is necessarily curved. How this problem was solved and the components of the metric tensor $g_{\mu\nu}$ are related to the energy–momentum tensor $T_{\mu\nu}$ was Einstein's extraordinary achievement of the years 1912 to 1915.

What is needed is a tensor which involves the metric tensor $g_{\mu\nu}$ and its first and second derivatives and which is linear in its second derivatives. It turns out that there is a unique answer to this problem: it is the fourth-rank tensor $R^{\lambda}_{\mu\nu\kappa}$ which is known as the *Riemann–Christoffel tensor*. Other tensors can be formed from this tensor by contraction, the most important of these being the *Ricci tensor*

$$R_{\mu\kappa} = R^{\lambda}_{\mu\lambda\kappa} \,, \tag{6.45}$$

and the *curvature scalar*

$$R = g^{\mu\kappa} R_{\mu\kappa} \,. \tag{6.46}$$

Einstein's stroke of genius was to propose that these tensors are related to the energy–momentum tensor in the following way

$$R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R = -\frac{8\pi G}{c^2} T_{\mu\nu} \,. \tag{6.47}$$

This is the key relation which shows how the components of the metric tensor $g_{\mu\nu}$ are related to the mass-energy distribution $T_{\mu\nu}$ in the Universe.

We will go no further along this route, except to note that Einstein realised that he could add an additional term to the left-hand side of (6.47). This is the origin of the famous cosmological constant Λ and was originally introduced in order to construct a static closed model for the Universe. Equation (6.47) then becomes

$$R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R + \Lambda g_{\mu\nu} = -\frac{8\pi G}{c^2} T_{\mu\nu} \,. \tag{6.48}$$

In the discussion of Chap. 7, we will use the Newtonian equivalents of these equations but it must be appreciated that we can only do this with the reassurance that the complete Einstein equations give fully self-consistent world models, without any need to introduce ad hoc assumptions. Our Newtonian equivalences can however provide intuitive impressions of the physical content of the theory.

6.6 Experimental and Observational Tests of General Relativity

In Einstein's exposition of the General Theory of Relativity, three tests of the theory were proposed, the gravitational redshift, the advance of the perihelion of planetary orbits and the deflection of light rays by the Sun. In 1964, Irwin Shapiro proposed a fourth test, the time delay of electromagnetic waves due to the distortion of space–time in the gravitational field of the Sun (Shapiro, 1964). The history of these tests and their status up to 1993 were comprehensively described by Will in his excellent book *Theory and Experiment in Gravitational Physics* (Will, 1993). More recently, he has updated the status of these tests and many other approaches to validating General Relativity (Will, 2006). Let us first review the four classic tests of the theory and then look briefly at the current status of possible modifications to General Relativity.

6.6.1 The Four Tests of General Relativity

Traditionally, there are four tests of the theory. The *first* is the measurement of the *gravitational redshift* of electromagnetic waves in a gravitational field which was discussed in Sect. 6.2. There, we described the use of the Mössbauer effect to measure the redshift of γ-ray photons in terrestrial experiments and the observation of the gravitational redshift of the emission lines in white dwarfs. More recent versions of the test have involved placing hydrogen masers in rocket payloads and measuring very precisely the change in frequency with altitude. These experiments have demonstrated directly the gravitational redshift of light. In the rocket experiments, the gravitational redshift was measured with a precision of about 5 parts in 10^5. Nowadays, it is preferable to regard this as a test of the conservation of energy in a gravitational field.

The *second* and oldest test, and the first great triumph of General Relativity, was the explanation of the *perihelion shift* of the orbit of the planet Mercury. Mercury's orbit has ellipticity $e = 0.2$ and, in 1859, Le Verrier found that, once account is taken of the influence of the other planets in the Solar System, there remained a small but significant advance of the perihelion of its orbit which amounted to about $\dot{\omega} \approx 43$ arcsec per century (Le Verrier, 1859). The origin of this perihelion shift remained a mystery, possible explanations including the presence of a hitherto unknown planet close to the Sun, oblateness of the solar interior, deviations from the inverse square law of gravity near the Sun and so on. Continued observations of Mercury by radar ranging have established the advance of the perihelion of its orbit to about 0.1% precision with the result $\dot{\omega} = 42.98(1 \pm 0.001)$ arcsec per century, once the perturbing effects of the other planets had been taken into account (Shapiro, 1990). Einstein's theory of General Relativity predicts a value of $\dot{\omega} = 42.98$ arcsec per century, in remarkable agreement with the observed value.

There has been some debate as to whether or not the agreement really is as good as this comparison suggests because there might be a contribution to the perihelion advance if the core of the Sun were rapidly rotating and so possessed a finite quadrupole moment. Observations of the vibrational modes of the Sun,

or *helioseismology*, have shown that the core of the Sun is not rotating sufficiently rapidly to upset the excellent agreement between the predictions of General Relativity and the observed perihelion advance. Specifically, the quadrupole moment of the Sun has now been measured to be $J_2 = (2.2 \pm 0.1) \times 10^{-7}$ and so its contribution to the perihelion advance is less than 0.1% of the predicted advance. In Will's recent assessment, he quotes the limits in terms of the values of PPN coefficients γ and β (see Sect. 6.6.3),

$$\dot{\omega} = 42.98 \left[\frac{1}{3}(2 + 2\gamma - \beta) + 3 \times 10^{-4} \frac{J_2}{10^{-7}} \right] \text{ arcsec per century} . \qquad (6.49)$$

Adopting the above value of J_2, the limit of 0.1% accuracy for $\dot{\omega}$ corresponds to $(2\gamma - \beta - 1) < 3 \times 10^{-3}$.

The *third* test was the measurement of the deflection of light by the Sun. For light rays just grazing the limb of the Sun, the deflection amounts to $\Delta\theta_{GR} = 4GM/R_\odot c^2 = 1.75$ arcsec, where $R_\odot$ is the radius of the Sun. Historically, this was a very important result. According to Newtonian theory, if we assume that the photon has a momentum $p = h\nu/c$ and then use the Rutherford scattering formula to work out the deviation of the light path, we find that the Newtonian deflection amounts to half the prediction of General Relativity, $\Delta\theta_{Newton} = 2GM/R_\odot c^2$. This prediction led to the famous eclipse expeditions of 1919 led by Eddington and Crommelin to measure precisely the angular deflections of the positions of stars observed close to the limb of the Sun during a solar eclipse. One expedition went to Sobral in Northern Brazil and the other to the island of Principe, off the coast of West Africa. The Sobral result was 1.98 ± 0.012 arcsec and the Principe result 1.61 ± 0.3 arcsec. These were technically demanding observations and there has been some controversy about the reliability of the results (Coles, 2001).

The modern version of the test originally involved measuring very precisely the angular separations between compact radio sources as they are observed close to the Sun. By means of Very Long Baseline Interferometry (VLBI), an angular precision of 100 microarcsec has now been achieved. In recent experiments, the VLBI technique has been used to measure deflections by the Sun over the whole sky. For example, at 90° to the direction of the Sun, the deflection of the radio waves still amounts to 4 milliarcsec, which is readily measurable by VLBI techniques.

The evolution of the precision of the light deflection test from the early optical studies to the most recent VLBI experiments is shown in the upper panel of Fig. 6.5. Transcontinental and intercontinental VLBI observations of quasars and radio galaxies have been used to monitor the Earth's rotation and these are sensitive to the deflection of light over almost the entire celestial sphere. An analysis of almost 2 million VLBI observations of 541 radio sources made by 87 VLBI sites over the period 1979 to 1999 yielded the following value for the parameter γ: $(\gamma - 1) = (-1.7 \pm 4.5) \times 10^{-4}$, or equivalently, $(1 + \gamma)/2 = 0.99992 \pm 0.00023$ (Shapiro et al., 2004). Notice also the limits shown in Fig. 6.5 obtained from the very precise positions of stars measured by the ESA *Hipparcos* astrometric satellite. General relativistic corrections had to be made for stars over the whole sky in order to obtain the quoted accuracy of about one milliarcsecond for the stars in the *Hipparcos* catalogue.

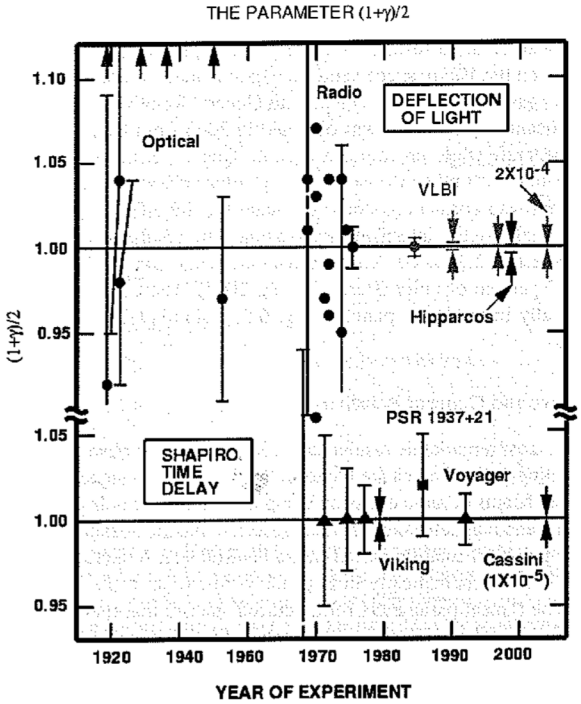

Fig. 6.5. Measurements of the quantity $(1 + \gamma)/2$ from light deflection and time delay experiments. The value of γ according to General Relativity is unity. The arrows at the top of the diagram denote anomalously large values from early eclipse expeditions. The time delay measurements from the Cassini spacecraft yielded agreement with General Relativity at the level of 10^{-3} percent. VLBI radio deflection measurements have reached 0.02 percent accuracy. The *Hipparcos* limits were derived from precise measurements of the positions stars over the whole sky and resulted in a precision of 0.1 percent in the measurement of γ (Will, 2006)

The forthcoming GAIA mission of ESA will measure the positions of about a billion stars in the Galaxy with a precision of about ten microarcsecond for stars brighter than 15th magnitude and so further estimates of the value of γ can be expected.

The *fourth* of the traditional tests is closely related to the deflection of light by the Sun and concerns the time delay expected when an electromagnetic wave propagates through a varying gravitational potential. In 1964, Shapiro realised that the gravitational redshift of radio signals passing close to the Sun causes a small time delay which can be measured by very precise timing of signals which are reflected from planets or space vehicles as they are about to be occulted by the Sun (Shapiro, 1964). Originally, the radio signals were reflected from the surface of the planets,

but later experiments used transponders on space vehicles which passed behind the Sun. The most accurate results from the early experiments were obtained using transponders on the *Viking* space vehicles which landed on Mars. These 'anchored' transponders gave results in agreement with General Relativity to within 0.1%.

A significant improvement was obtained in 2003 from Doppler tracking of the Cassini spacecraft while en route to Saturn. This experiment had the advantage of carrying out the timing measurements at two different radio frequencies and so much improved corrections could be made for the effects of the dispersion of the radio signals by the interplanetary plasma. The result of this experiment was $(\gamma - 1) = (2.1 \pm 2.3) \times 10^{-5}$. Hence the coefficient $\frac{1}{2}(1 + \gamma)$ must be within at most 0.0012 percent of unity (Bertotti et al., 2003). These results are summarised diagramatically in the lower panel of Fig. 6.5, which is taken from Will's review (Will, 2006).

6.6.2 Pulsars and General Relativity

Some of the most remarkable results have come from radio observations of *pulsars*. These pulsating radio sources are identified with rotating, magnetised neutron stars and they emit beams of radio emission along their magnetic poles. It is assumed that the rotational and magnetic axes are misaligned so that the distant observer normally detects one pulse per rotation period of the neutron star. A sketch of this model for a pulsar is shown schematically in Fig. 6.6 for the case of the binary pulsar PSR 1913+16. The typical parameters for a neutron star are that their masses are about $1.4M_\odot$, their radii about 10 km and their magnetic flux densities range from 10^4 to 10^9 T. Observations by Joseph Taylor and his colleagues using the Arecibo radio telescope have demonstrated that these are among the most stable clocks we know of in the Universe (Taylor, 1992).

The most intriguing systems are those pulsars which are members of binary systems, particularly those which are referred to as *relativistic binaries* in which both members of the binaries are neutron stars and their binary periods are less than a day. The first of these to be discovered was the binary pulsar PSR 1913+16 (Hulse and Taylor, 1975). The system has a binary period of only 7.75 hours and its orbital eccentricity is large, $e = 0.617$. This system is a pure gift for the relativist. To test General Relativity, we need a perfect clock in a rotating frame of reference and systems such as PSR 1913+16 are ideal for this purpose. The neutron stars are so inert and compact that the binary system is very 'clean' and so can be used for some of the most sensitive tests of General Relativity yet devised.

Precise timing of the arrival times of the pulses enables many independent parameters of the binary system to be determined and these depend upon the masses of the neutron stars. In Fig. 6.7, the most accurately determined three parameters are used to estimate the masses of the neutron stars in the binary system PSR 1913+16, assuming that General Relativity is the correct theory of gravity. It can be seen that the different loci intersect very precisely at a single point in the m_1/m_2 plane. Some measure of the precision with which the theory is known to be correct can be obtained from the accuracy with which the masses of the neutron stars are known:

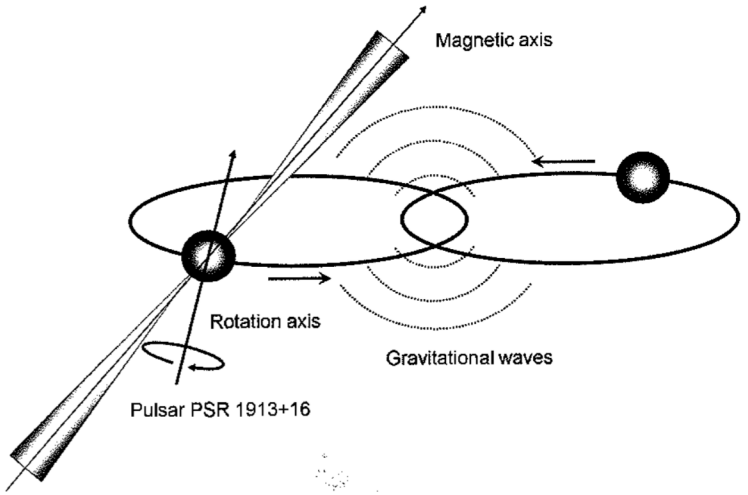

Fig. 6.6. A schematic diagram of the orbit of the binary pulsar PSR 1913+16. The pulsar is one of a pair of neutron stars in binary orbits about their common centre of mass. There is a displacement between the axis of the magnetic dipole and the rotation axis of the neutron star. The radio pulses are assumed to be due to beams of radio emission from the poles of the magnetic field distribution and are associated with the passage of the beam across the line of sight to the observer. As a result of the ability to measure precisely many parameters of the binary orbit from ultraprecise pulsar timing, the masses of the two neutron stars have been measured with very high precision (Taylor, 1992; Will, 2006)

$m_1 = 1.4414 \pm 0.0002\ M_\odot$ and $m_2 = 1.3867 \pm 0.0002\ M_\odot$. These are the most accurately known masses for any extrasolar system object.

Four other neutron star–neutron star binaries are known, including the system J0737-3039 in which both neutron stars are observed as pulsars (Lyne et al., 2004). This system is of the greatest interest since ultimately even better estimates of the orbital parameters of the system can be found than is the case for PSR 1913+16. It has not, however, yet been observed over as long a time period as PSR 1913+16.

A second remarkable measurement has been the rate of loss of orbital rotational energy by the *emission of gravitational waves*. A binary star system loses energy by the emission of gravitational radiation and the rate at which energy is lost can be precisely predicted once the masses of the neutron stars and the parameters of the binary orbit are known.[4] The rate of change of the angular frequency Ω of the orbit due to the emission of gravitational radiation is precisely known, $\mathrm{d}\Omega/\mathrm{d}t \propto \Omega^5$.

[4] I have given a simple heuristic derivation of the formula for the rate of loss of energy of a binary system by gravitational radiation in the Explanatory Supplement to Chap. 8 of my book *The Cosmic Century* (Longair, 2006).

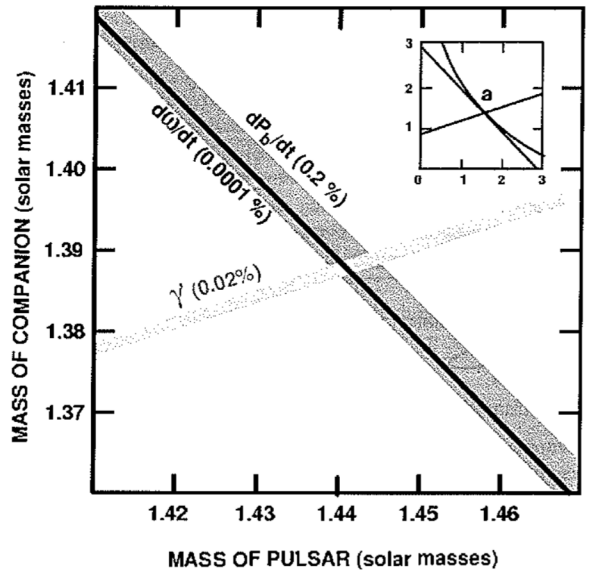

Fig. 6.7. Constraints on the masses of the pulsar PSR 1913+16 and its invisible companion from precise timing data, assuming General Relativity to be the correct theory of gravity. The width of each strip in the plane reflects the observational uncertainties, shown as a percentage. The *inset* shows the same three most accurate constraints on the full mass plane; the intersection region has been magnified 400 times in the large figure (Will, 2006)

The change in orbital phase of the binary pulsar PSR 1913+16 has been observed over a period of 30 years and General Relativity is in precise agreement with the observed changes over that period (Fig. 6.8). Thus, although the gravitational waves themselves have not been detected, exactly the correct energy loss rate from the system has been measured; it is generally assumed that this is convincing evidence for the existence of gravitational waves and this observation acts as a spur to their direct detection by future generations of gravitational wave detectors.

This is a very important result for the theory of gravitation since it enables a range of alternative theories of gravity to be excluded. For example, since General Relativity predicts only quadrupole emission of gravitational radiation, any theory which, say, involved the dipole emission of gravitational waves can potentially be excluded. The only problem with this argument for the system PSR 1913+16 is that the masses of the two neutron stars are almost exactly the same and so it possesses a rather small dipole moment. It turns out that at the moment the solar system tests provide better constraints on theories of relativistic gravity.

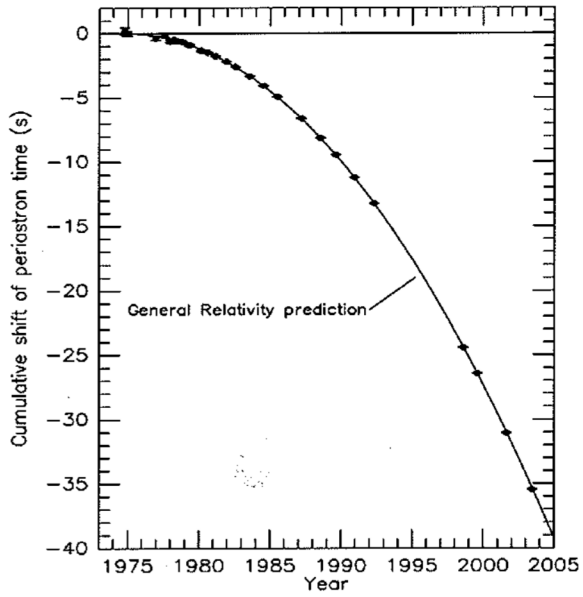

Fig. 6.8. The change of orbital phase as a function of time for the binary neutron star system PSR 1913+16 compared with the expected changes due to gravitational radiation energy loss by the binary system (Taylor, 1992; Will, 2006)

6.6.3 Parameterised Post-Newtonian Models

We have already introduced the parameters β and γ without explaining precisely what they mean. These quantities are found in what are called *parameterised post-Newtonian (PPN)* models for theories of relativistic gravity. To understand this approach to comparing the theories with observation, it is simplest to quote the words of Will (Will, 2006):

> The comparison of metric theories of gravity with each other and with experiment becomes particularly simple when one takes the slow-motion, weak-field limit. This approximation, known as the post-Newtonian limit, is sufficiently accurate to encompass most solar-system tests that can be performed in the foreseeable future. It turns out that, in this limit, the space–time metric predicted by nearly every metric theory of gravity has the same structure. It can be written as an expansion about the Minkowski metric in terms of dimensionless gravitational potentials of varying degrees of smallness.

The spirit of this approach is to relax the powerful constraints implied by the Einstein equivalence principle discussed in Sect. 6.1 and so allow a wider range of possible theories of relativistic gravity. To give some impression of what is involved in this approach, Table 6.1, taken from Will's survey, shows a list of the various parameters involved in these theories. Notice that the entries for ξ and for α_1, α_2 and α_3 correspond to relaxing the second and third conditions involved in the Einstein equivalence principle.

These modifications would change the metric coefficients from the values they take in General Relativity. Thus, quoting Will, the metric coefficients would become:

$$\begin{aligned} g_{00} = &-1 + 2U - 2\beta U^2 - 2\xi\Phi_W + (2\gamma + 2 + \alpha_3 + \zeta_1 - 2\xi)\Phi_1 \\ &+2(3\gamma - 2\beta + 1 + \zeta_2 + \xi)\Phi_2 + 2(1 + \zeta_3)\Phi_3 + 2(3\gamma + 3\zeta_4 - 2\xi)\Phi_4 \\ &-(\zeta_1 - 2\xi)\mathcal{A} - (\alpha_1 - \alpha_2 - \alpha_3)w^2 U - \alpha_2 w^i w^j U_{ij} \\ &+(2\alpha_3 - \alpha_1)w^i V_i + \mathcal{O}(\epsilon^3) \end{aligned} \tag{6.50}$$

$$\begin{aligned} g_{0i} = &-\tfrac{1}{2}(4\gamma + 3 + \alpha_1 - \alpha_2 + \zeta_1 - 2\xi)V_i - \tfrac{1}{2}(1 + \alpha_2 - \zeta_1 + 2\xi)W_i \\ &-\tfrac{1}{2}(\alpha_1 - 2\alpha_2)w^i U - \alpha_2 w^j U_{ij} + \mathcal{O}(\epsilon^{5/2}) \end{aligned} \tag{6.51}$$

$$g_{ij} = (1 + 2\gamma U)\delta_{ij} + \mathcal{O}(\epsilon^2)\,. \tag{6.52}$$

The quantities U, U_{ij}, Φ_W, Φ_1, Φ_2, Φ_3, Φ_4, $\mathcal{A}$, V_i, W_i are various metric potentials which can be interpreted in terms of Newtonian gravity. Thus, U, defined by

$$U = \int \frac{\varrho'}{|x - x'|}\,\mathrm{d}^3x'\,, \tag{6.53}$$

is just the Newtonian gravitational potential.

The expressions (6.50) to (6.52) look rather forbidding at first sight, but the important point is that it is possible to test theories in which the Einstein equivalence principle is relaxed and provide further constraints upon acceptable theories. As Will points out in his review, some of the theories may appear somewhat unphysical within the realms of known physics, but in some extensions of the Standard Model of particle physics, for example, even some of our most cherished theories, such as Lorentz invariance, might have to be sacrificed. Another example involves scalar-tensor modifications of General Relativity which are involved in unification schemes such as string theory, and in cosmological model building.

Some of these post-Newtonian 'corrections' have quite obvious meanings. For example, inspection of the first three terms of (6.50) shows that, for a point mass, the first two are just the familiar metric coefficient $(1 + \phi/c^2)$ in our notation and the third is a non-linear term in the square of the potential $\beta(\phi/c^2)^2$. In the same way, inspection of (6.52) shows that γ describes how much space-curvature is produced by unit mass, reducing to the standard result if $\gamma = 1$. The limits which can be set to possible deviations from the Einstein equivalence principle are listed in Table 6.3. Will's review should be consulted for more details about these possibilities.

Table 6.3. Current limits on the PPN parameters. Here η_N is a combination of other parameters given by $\eta_N = 4\beta - \gamma - 3 - 10\xi/3 - \alpha_1 + 2\alpha_2/3 - 2\zeta_1/3 - \zeta_2/3$

Parameter	Effect	Limit	Remarks
$\gamma - 1$	Time delay	2.3×10^{-5}	Cassini tracking
	Light deflection	4×10^{-4}	VLBI
$\beta - 1$	Perihelion shift	3×10^{-3}	$J_2 = 10^{-7}$ from helioseismology
	Nordtvedt effect	2.3×10^{-4}	$\eta_N = 4\beta - \gamma - 3$ assumed
ξ	Earth tides	10^{-3}	Gravimeter data
α_1	Orbital polarisation	10^{-4}	Lunar laser ranging
		2×10^{-4}	PSR J2317+1439
α_2	Spin precession	4×10^{-7}	Solar alignment with ecliptic
α_3	Pulsar acceleration	4×10^{-20}	Pulsar $\dot{P}$ statistics
η_N	Nordtvedt effect	9×10^{-4}	Lunar laser ranging
ζ_1	–	2×10^{-2}	Combined PPN bounds
ζ_2	Binary acceleration	4×10^{-5}	$\ddot{P}_p$ for PSR 1913+16
ζ_3	Newtons 3rd law	10^{-8}	Lunar acceleration
ζ_4	–	–	Not independent ($6\zeta_4 = 3\alpha_3 + 2\zeta_1 - 3\zeta_3$)

Table 6.4. Constancy of the gravitational constant G. For binary pulsar data, the bounds are dependent upon the theory of gravity in the strong-field regime and on the neutron star equation of state. Big Bang nucleosynthesis bounds assume a specific form for the time dependence of G

Method	$(\dot{G}/G)/10^{-13}$ year^{-1}	Reference
Lunar laser ranging	4 ± 9	(Williams et al., 2004)
Binary pulsar PSR 1913+16	40 ± 50	(Kaspi et al., 1994)
Helioseismology	0 ± 16	(Guenther et al., 1998)
Big Bang nucleosynthesis	0 ± 4	(Copi et al., 2004)

6.6.4 Variation of the Gravitational Constant with Cosmic Epoch

An important question for cosmology is whether or not the gravitational constant G has varied with time. A summary of recent results is shown in Table 6.4 (Will, 2006).

- The technique of lunar laser ranging has provided the strongest limit to date. In the analysis of these data, evidence is sought for steady changes in the lunar orbit which could be attributed to changes in the gravitational constant with time.
- The techniques of accurate pulsar timing have been used to determine whether or not there is any evidence for steady changes in the pulsar's orbital period due to steady variations in the gravitational constant G with time. This test is somewhat dependent upon the equation of state used to describe the interior of the neutron star.

- The helioseismology limit is derived from the remarkable success of the standard astrophysical model of the solar interior in accounting for its internal structure. The frequency spacings between the p-modes of different azimuthal and radial order are very sensitive to the sound speed in the central regions of the star. If the gravitational constant had varied with cosmic epoch, the chemical composition and hence the speed of sound and frequency separations in the central regions would have been significantly different from the observed values.
- The primordial nucleosynthesis argument follows from the fact that, if the gravitational constant were greater in the past, the early evolution of the Universe would have been more rapid than in the standard model of its early stages and so helium would have been dramatically overproduced relative to its observed cosmic value. We will return to this argument in Chap. 10.

The sense of the data listed in Table 6.4 is that there can have been little change in the value of the gravitational constant over typical cosmological time-scales which are about 10^{10} years.

6.7 Summary

General Relativity has passed every observational and experimental test which has been made of the theory and so we can have confidence that it is the correct starting point for the development of models of the large-scale dynamical structure of our Universe.

7 The Friedman World Models

7.1 Einstein's Field Equations

Einstein realised that, in General Relativity, he had discovered a theory which enabled fully self-consistent models for the Universe as a whole to be constructed. The standard models contain three essential ingredients:

- The *cosmological principle*, which, combined with the observations that the Universe is isotropic, homogeneous and uniformly expanding on a large scale, leads to the Robertson–Walker metric (5.33).
- *Weyl's postulate*, according to which the world lines of particles meet at a singular point in the finite or infinite past. This means that there is a unique world line passing through every point in space–time. The fluid moves along streamlines in the universal expansion and so behaves like a perfect fluid for which the energy–momentum tensor is given by the $T^{\alpha\beta}$ of (6.33).
- General Relativity, which enables us to relate the energy–momentum tensor to the geometrical properties of space–time through (6.47) or (6.48).

The assumptions of isotropy and homogeneity result in enormous simplifications of Einstein's field equations which reduce to the following pair of equations:

$$\ddot{a} = -\frac{4\pi G}{3} a \left(\varrho + \frac{3p}{c^2} \right) + \tfrac{1}{3} \Lambda a \, ; \tag{7.1}$$

$$\dot{a}^2 = \frac{8\pi G \varrho}{3} a^2 - \frac{c^2}{\Re^2} + \tfrac{1}{3} \Lambda a^2 \, . \tag{7.2}$$

In these equations, a is the scale factor normalised to the value unity at the present epoch t_0, ϱ is the total inertial mass density of the matter and radiation content of the Universe and p the associated total pressure. $\Re$ is the radius of curvature of the geometry of the world model at the present epoch and so the term $-c^2/\Re^2$ is a constant of integration. The *cosmological constant* Λ was introduced by Einstein in 1917 in order to create a static Universe with closed geometry which he hoped would enable Mach's principle to be incorporated into General Relativity (Einstein, 1917).

Let us look more closely at the meanings of (7.1) and (7.2). Equation (7.2) is referred to as *Friedman's equation* and has the form of an energy equation, the