1

First Principles

In this chapter, our aim is to provide an introduction to the basic mathematical structure of modern cosmological models based on Einstein's theory of gravity, the General Theory of Relativity or general relativity for short. This theory is mathematically challenging, but fortunately we do not really need to use its fully general form. Throughout this chapter we will therefore illustrate the key results with Newtonian analogies. We begin our study with a discussion of the Cosmological Principle, the ingredient that makes relativistic cosmology rather more palatable than it might otherwise be.

1.1 The Cosmological Principle

Whenever science enters a new field and is faced with a dearth of observational or experimental data some guiding principle is usually needed to assist during the first tentative steps towards a theoretical understanding. Such principles are often based on ideas of symmetry which reduce the number of degrees of freedom one has to consider. This general rule proved to be the case in the early years of the 20th century when the first steps were taken, by Einstein and others, towards a scientific theory of the Universe. Little was then known empirically about the distribution of matter in the Universe and Einstein's theory of gravity was found to be too difficult to solve for an arbitrary distribution of matter. In order to make progress the early cosmologists therefore had to content themselves with the construction of simplified models which they hoped might describe some aspects of the Universe in a broad-brush sense. These models were based on an idea called the *Cosmological Principle*. Although the name 'principle' sounds grand, principles are generally introduced into physics when one has no data to go on, and cosmology was no exception to this rule.

The Cosmological Principle is the assertion that, on sufficiently large scales (beyond those traced by the large-scale structure of the distribution of galaxies), the Universe is both homogeneous and isotropic. Homogeneity is the property of being identical everywhere in space, while isotropy is the property of looking the same in every direction. The Universe is clearly not exactly homogeneous, so cosmologists define homogeneity in an average sense: the Universe is taken to be identical in different places when one looks at sufficiently large pieces. A good analogy is that of a patterned carpet which is made of repeating units of some basic design. On the scale of the individual design the structure is clearly inhomogeneous but on scales larger than each unit it is homogeneous.

There is quite good observational evidence that the Universe does have these properties, although this evidence is not completely watertight. One piece of evidence is the observed near-isotropy of the cosmic microwave background radiation. Isotropy, however, does not necessarily imply homogeneity without the additional assumption that the observer is not in a special place: the so-called *Copernican Principle*. One would observe isotropy in any spherically symmetric distribution of matter, but only if one were in the middle of the pattern. A circular carpet bearing a design consisting of a series of concentric rings would look isotropic only to an observer standing in the centre of the pattern. Observed isotropy, together with the Copernican Principle, therefore implies the Cosmological Principle.

The Cosmological Principle was introduced by Einstein and subsequent relativistic cosmologists without any observational justification whatsoever. Indeed, it was not known until the 1920s that the spiral nebulae (now known to be galaxies like our own) were outside our own galaxy, the Milky Way. A term frequently used to describe the entire Universe in those days was metagalaxy, indicating that it was thought that the Milky Way was essentially the entire cosmos. The Galaxy certainly does not look the same in all directions: it presents itself as a prominent band across the night sky.

In advocating the Cosmological Principle, Einstein was particularly motivated by ideas associated with Ernst Mach. Mach's Principle, roughly speaking, is that the laws of physics are determined by the distribution of matter on large scales. For example, the value of the gravitational constant G was thought perhaps to be related to the amount of mass in the Universe. Einstein thought that the only way to put theoretical cosmology on a firm footing was to assume that there was a basic simplicity to the global structure of the Universe enabling a similar simplicity in the local behaviour of matter. The Cosmological Principle achieves this and leads to relatively simple cosmological models, as we shall see shortly.

There are various approaches one can take to this principle. One is philosophical, and is characterised by the work of Milne in the 1930s and later by Bondi, Gold and Hoyle in the 1940s. This line of reasoning is based, to a large extent, on the aesthetic appeal of the Cosmological Principle. Ultimately this appeal stems from the fact that it would indeed be very difficult for us to understand the Universe if physical conditions, or even the laws of physics themselves, were to vary dramatically from place to place. These thoughts have been taken further, leading to the *Perfect Cosmological Principle*, in which the Universe is the same not only in all places and in all directions, but also at all times. This stronger version of the Cosmological Principle was formulated by Bondi and Gold (1948) and it subsequently led Hoyle (1948) and Hoyle and Narlikar (1963, 1964) to develop the steady-state cosmology. This theory implies, amongst other things, the continuous creation of matter to keep the density of the expanding Universe constant. The steady-state universe was abandoned in the 1960s because of the properties of the cosmic microwave background, radio sources and the cosmological helium abundance which are more readily explained in a Big Bang model than in a steady state. Nowadays the latter is only of historical interest (see Chapter 3 later).

Attempts have also been made to justify the Cosmological Principle on more direct physical grounds. As we shall see, homogeneous and isotropic universes described by the theory of general relativity possess what is known as a 'cosmological horizon': regions sufficiently distant from each other cannot have been in causal contact ('have never been inside each other's horizon') at any stage since the Big Bang. The size of the regions whose parts are in causal contact with each other at a given time grows with cosmological epoch; the calculation of the horizon scale is performed in Section 2.7. The problem then arises as to how one explains the observation that the Universe appears homogeneous on scales much larger than the scale one expects to have been in causal contact up to the present time. The mystery is this: if two regions of the Universe have never been able to communicate with each other by means of light signals, how can they even know the physical conditions (density, temperature, etc.) pertaining to each other? If they cannot know this, how is it that they evolve in such a way that these conditions are the same in each of the regions? One either has to suppose that causal physics is not responsible for this homogeneity, or that the calculation of the horizon is not correct. This conundrum is usually called the *Cosmological Horizon Problem* and we shall discuss it in some detail in Chapter 7.

Various attempts have been made to avoid this problem. For example, particular models of the Universe, such as some that are homogeneous but not isotropic, do not possess the required particle horizon. These models can become isotropic in the course of their evolution. A famous example is the 'mix-master' universe of Misner (1968) in which isotropisation is effected by viscous dissipation involving neutrinos in the early universe. Another way to isotropise an initially anisotropic universe is by creating particles at the earliest stage of all, the Planck era (Chapter 6). More recently still, Guth (1981) proposed an idea which could resolve the horizon problem: *the inflationary universe*, which is of great contemporary interest in cosmology, and which we discuss in Chapter 7.

In any case, the most appropriate approach to this problem is an empirical one. We accept the Cosmological Principle because it agrees with observations. We shall describe the observational evidence for this in Chapter 4; data concerning radiogalaxies, clusters of galaxies, quasars and the microwave background all demonstrate that the level of anisotropy of the Universe on large scales is about one part in 10^5.

1.2 Fundamentals of General Relativity

The strongest force of nature on large scales is gravity, so the most important part of a physical description of the Universe is a theory of gravity. The best candidate we have for this is Einstein's General Theory of Relativity. We therefore begin this chapter with a brief introduction to the basics of this theory. Readers familiar with this material can skip Section 1.2 and resume reading at Section 1.3. In fact, about 90% of this book does not require the use of general relativity at all so readers only interested in a Newtonian treatment may turn directly to Section 1.11.

In *Special Relativity*, the invariant *interval* between two events at coordinates (t, x, y, z) and $(t + \mathrm{d}t, x + \mathrm{d}x, y + \mathrm{d}y, z + \mathrm{d}z)$ is defined by

$$\mathrm{d}s^2 = c^2\,\mathrm{d}t^2 - (\mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2), \tag{1.2.1}$$

where $\mathrm{d}s$ is invariant under a change of coordinate system and the path of a light ray is given by $\mathrm{d}s = 0$. The paths of material particles between any two events are such as to give stationary values of $\int_{\rm path} \mathrm{d}s$; this corresponds to the shortest distance between any two points being a straight line. This all applies to the motion of particles under no external forces; actual forces such as gravitation and electromagnetism cause particle tracks to deviate from the straight line.

Gravitation exerts the same force per unit mass on all bodies and the essence of Einstein's theory is to transform it from being a force to being a property of space–time. In his theory, the space–time is not necessarily flat as it is in Minkowski space–time (1.2.1) but may be curved. The interval between two events can be written as

$$\mathrm{d}s^2 = g_{ij}\,\mathrm{d}x^i\,\mathrm{d}x^j, \tag{1.2.2}$$

where repeated suffixes imply summation and i, j both run from 0 to 3; $x^0 = ct$ is the time coordinate and x^1, x^2, x^3 are space coordinates. The tensor g_{ij} is the *metric tensor* describing the space–time geometry; we discuss this in much more detail in Section 1.3. As we mentioned above, particle moves in such a way that the integral along its path is stationary:

$$\delta \int_{\rm path} \mathrm{d}s = 0, \tag{1.2.3}$$

but such tracks are no longer straight because of the effects of gravitation contained in g_{ij}. From Equation (1.2.3), the path of a free particle, which is called a *geodesic*, can be shown to be described by

$$\frac{\mathrm{d}^2x^i}{\mathrm{d}s^2} + \Gamma^i_{kl}\frac{\mathrm{d}x^k}{\mathrm{d}s}\frac{\mathrm{d}x^l}{\mathrm{d}s} = 0, \tag{1.2.4}$$

where the Γs are called *Christoffel symbols*,

$$\Gamma^i_{kl} = \tfrac{1}{2}g^{im}\left[\frac{\partial g_{mk}}{\partial x^l} + \frac{\partial g_{ml}}{\partial x^k} - \frac{\partial g_{kl}}{\partial x^m}\right], \tag{1.2.5}$$

and

$$g^{im}g_{mk} = \delta^i_k \tag{1.2.6}$$

is the Kronecker delta, which is unity when $i = k$ and zero otherwise. Free particles move on geodesics but the metric g_{ij} is itself determined by the matter. The key factor in Einstein's equations is the relationship between the distribution of matter and the metric describing the space–time geometry.

In general relativity all equations are tensor equations. A general tensor is a quantity which transforms as follows when coordinates are changed from x^i to x'^i:

$$A'^{kl\ldots}_{pq\ldots} = \frac{\partial x'^k}{\partial x^m}\frac{\partial x'^l}{\partial x^n}\cdots\frac{\partial x^r}{\partial x'^p}\frac{\partial x^s}{\partial x'^q}\cdots A^{mn\ldots}_{rs\ldots}, \tag{1.2.7}$$

where the upper indices are *contravariant* and the lower are *covariant*. The difference between these types of index can be illustrated by considering a tensor of rank 1 which is simply a vector (the rank of a tensor is the number of indices it carries). A vector will undergo a transformation according to some rules when the coordinate system in which it is expressed is changed. Suppose we have an original coordinate system x^i and we transform it to a new system x'^k. If the vector $\boldsymbol{A}$ transforms in such a way that $\boldsymbol{A}' = \partial x'^k/\partial x^i \boldsymbol{A}$, then the vector $\boldsymbol{A}$ is a contravariant vector and it is written with an upper index, i.e. $\boldsymbol{A} = A^i$. On the other hand, if the vector transforms according to $\boldsymbol{A}' = \partial x^i/\partial x'^k \boldsymbol{A}$, then it is covariant and is written $\boldsymbol{A} = A_i$. The tangent vector to a curve is an example of a contravariant vector; the normal to a surface is a covariant vector. The rule (1.2.7) is a generalisation of these concepts to tensors of arbitrary rank and to tensors of mixed character.

In Newtonian and special-relativistic physics a key role is played by conservation laws of mass, energy and momentum. Our task is now to obtain similar laws for general relativity. With the equivalence of mass and energy brought about by Special Relativity, these laws can be written

$$\frac{\partial T_{ik}}{\partial x^k} = 0. \tag{1.2.8}$$

The energy–momentum tensor T_{ik} describes the matter distribution: for a perfect fluid, with pressure p and energy density ρ, it is

$$T_{ik} = (p + \rho c^2)U_i U_k - p g_{ik}; \tag{1.2.9}$$

the vector U_i is the fluid four-velocity

$$U_i = g_{ik}U^k = g_{ik}\frac{\mathrm{d}x^k}{\mathrm{d}s}, \tag{1.2.10}$$

where $x^k(s)$ is the world line of a fluid element, i.e. the trajectory in space–time followed by the particle. Equation (1.2.10) is a special case of the general rule for raising or lowering suffixes using the metric tensor.

It is easy to see that the Equation (1.2.8) cannot be correct in general relativity since $\partial T^{ik}/\partial x^k$ and $\partial T_{ik}/\partial x^k$ are not tensors. Since

$$T'_{mn} = \frac{\partial x^i}{\partial x'^m}\frac{\partial x^k}{\partial x'^n}T_{ik},$$

it is evident that $\partial T_{mn}/\partial x'^n$ involves terms such as $\partial^2 x^i/\partial x'^m \partial x'^n$, so it will not be a tensor. However, although the ordinary derivative of a tensor is not a tensor, a quantity called the *covariant derivative* can be shown to be one. The covariant derivative of a tensor $\boldsymbol{A}$ is defined by

$$A^{kl...}_{pq...;j} = \frac{\partial A^{kl...}_{pq...}}{\partial x^j} + \Gamma^k_{mj}A^{ml...}_{pq...} + \Gamma^l_{nj}A^{kn...}_{pq...} + \cdots - \Gamma^r_{pj}A^{kl...}_{rq...} - \Gamma^s_{qj}A^{kl...}_{ps...} - \cdots \tag{1.2.11}$$

in an obvious notation. The conservation law can therefore be written in a fully covariant form:

$$T_i{}^k{}_{;k} = 0. \tag{1.2.12}$$

A covariant derivative is usually written as a ';' in the subscript; ordinary derivatives are usually written as a ',' so that Equation (1.2.8) can be written $T_{ik,k} = 0$.

Einstein wished to find a relation between matter and metric and to equate T_{ik} to a tensor obtained from g_{ik}, which contains only the first two derivatives of g_{ik} and has zero covariant derivative. Because, in the appropriate limit, Equation (1.2.12) must reduce to Poisson's equation describing Newtonian gravity

$$\nabla^2\varphi = 4\pi G\rho, \tag{1.2.13}$$

it should be linear in the second derivative of the metric. The properties of curved spaces were well-known when Einstein was working on this theory. For example, it was known that the *Riemann–Christoffel tensor*,

$$R^i_{klm} = \frac{\partial \Gamma^i_{km}}{\partial x^l} - \frac{\partial \Gamma^i_{kl}}{\partial x^m} + \Gamma^i_{\mathrm{nl}}\Gamma^n_{km} - \Gamma^i_{nm}\Gamma^n_{kl}, \tag{1.2.14}$$

could be used to determine whether a given space is curved or flat. (Incidentally, Γ^i_{km} is not a tensor so it is by no means obvious, though it is actually true, that R^i_{klm} is a tensor.) From the Riemann–Christoffel tensor one can form the *Ricci tensor*:

$$R_{ik} = R^l{}_{ilk}. \tag{1.2.15}$$

Finally, one can form a scalar curvature, the *Ricci scalar*:

$$R = g^{ik}R_{ik}. \tag{1.2.16}$$

Now we are in a position to define the *Einstein tensor*

$$G_{ik} \equiv R_{ik} - \tfrac{1}{2}g_{ik}R. \tag{1.2.17}$$

Einstein showed that

$$G_i{}^k{}_{;k} = 0. \tag{1.2.18}$$

The tensor G_{ik} contains second derivatives of g_{ik}, so Einstein proposed as his fundamental equation

$$G_{ik} \equiv R_{ik} - \tfrac{1}{2} g_{ik} R = \frac{8\pi G}{c^4} T_{ik}, \tag{1.2.19}$$

where the quantity $8\pi G/c^4$ (G is Newton's gravitational constant) ensures that Poisson's equation in its standard form (1.2.13) results in the limit of a weak gravitational field. He subsequently proposed the alternative form

$$G_{ik} \equiv R_{ik} - \tfrac{1}{2} g_{ik} R - \Lambda g_{ik} = \frac{8\pi G}{c^4} T_{ik}, \tag{1.2.20}$$

where Λ is called the *cosmological constant*; as $g_i{}^k{}_{;k} = 0$, we still have $T_i{}^k{}_{;k} = 0$. He actually did this in order to ensure that static cosmological solutions could be obtained. We shall return to be the issue of Λ later, in Section 1.12.

1.3 The Robertson–Walker Metric

Having established the idea of the Cosmological Principle, our task is to see if we can construct models of the Universe in which this principle holds. Because general relativity is a geometrical theory, we must begin by investigating the geometrical properties of homogeneous and isotropic spaces. Let us suppose we can regard the Universe as a continuous fluid and assign to each fluid element the three spatial coordinates x^α ($\alpha = 1, 2, 3$). Thus, any point in space–time can be labelled by the coordinates x^α, corresponding to the fluid element which is passing through the point, and a time parameter which we take to be the proper time t measured by a clock moving with the fluid element. The coordinates x^α are called *comoving coordinates*. The geometrical properties of space–time are described by a metric; the meaning of the metric will be divulged just a little later. One can show from simple geometrical considerations only (i.e. without making use of any field equations) that the most general space–time metric describing a universe in which the Cosmological Principle is obeyed is of the form

$$\mathrm{d}s^2 = (c\,\mathrm{d}t)^2 - a(t)^2 \left[\frac{\mathrm{d}r^2}{1 - Kr^2} + r^2 (\mathrm{d}\vartheta^2 + \sin^2\vartheta\,\mathrm{d}\varphi^2) \right], \tag{1.3.1}$$

where we have used spherical polar coordinates: r, ϑ and φ are the comoving coordinates (r is by convention dimensionless); t is the proper time; $a(t)$ is a function to be determined which has the dimensions of a length and is called the *cosmic scale factor* or the *expansion parameter*; the *curvature parameter* K is a constant which can be scaled in such a way that it takes only the values 1, 0 or −1. The metric (1.3.1) is called the *Robertson–Walker metric*.

The significance of the metric of a space–time, or more specifically the metric tensor g_{ik}, which we introduced briefly in Equation (1.2.2),

$$ds^2 = g_{ik}(x)\,dx^i\,dx^k \quad (i,k = 0,1,2,3) \tag{1.3.2}$$

(as usual, repeated indices imply a summation), is such that, in Equation (1.3.2), ds^2 represents the space–time interval between two points labelled by x^j and $x^j + dx^j$. Equation (1.3.1) merely represents a special case of this type of relation. The metric tensor determines all the geometrical properties of the space–time described by the system of coordinates x^j. It may help to think of Equation (1.3.2) as a generalisation of Pythagoras's theorem. If $ds^2 > 0$, then the interval is *timelike* and ds/c would be the time interval measured by a clock which moves freely between x^j and $x^j + dx^j$. If $ds^2 < 0$, then the interval is *spacelike* and $|ds^2|^{1/2}$ represents the length of a ruler with ends at x^j and $x^j + dx^j$ measured by an observer at rest with respect to the ruler. If $ds^2 = 0$, then the interval is *lightlike* or *null*; this type of interval is important because it means that the two points x^j and $x^j + dx^j$ can be connected by a light ray.

If the distribution of matter is uniform, then the space is uniform and isotropic. This, in turn, means that one can define a *universal time* (or *proper time*) such that at any instant the three-dimensional spatial metric

$$dl^2 = \gamma_{\alpha\beta}\,dx^\alpha\,dx^\beta \quad (\alpha,\beta = 1,2,3), \tag{1.3.3}$$

where the interval is now just the spatial distance, is identical in all places and in all directions. Thus, the space–time metric must be of the form

$$ds^2 = (c\,dt)^2 - dl^2 = (c\,dt)^2 - \gamma_{\alpha\beta}\,dx^\alpha\,dx^\beta. \tag{1.3.4}$$

This coordinate system is called the *synchronous gauge* and is the most commonly used way of slicing the four-dimensional space–time into three space dimensions and one time dimension.

To find the three-dimensional (spatial) metric tensor $\gamma_{\alpha\beta}$ let us consider first the simpler case of an isotropic and homogeneous space of only two dimensions. Such a space can be either (i) the usual Cartesian plane (flat Euclidean space with infinite curvature radius), (ii) a spherical surface of radius R (a curved space with positive Gaussian curvature $1/R^2$), or (iii) the surface of a hyperboloid (a curved space with negative Gaussian curvature).

In the first case the metric, in polar coordinates ρ ($0 \leqslant \rho < \infty$) and φ ($0 \leqslant \varphi < 2\pi$), is of the form

$$dl^2 = a^2(dr^2 + r^2\,d\varphi^2); \tag{1.3.5 a}$$

we have introduced the dimensionless coordinate $r = \rho/a$, which lies in the range $0 \leqslant r < \infty$, and the arbitrary constant a, which has the dimensions of a length. On the surface of a sphere of radius R the metric in coordinates ϑ ($0 \leqslant \vartheta \leqslant \pi$) and φ ($0 \leqslant \varphi < 2\pi$) is just

$$dl^2 = a^2(d\vartheta^2 + \sin^2\vartheta\,d\varphi^2) = a^2\left(\frac{dr^2}{1-r^2} + r^2\,d\varphi^2\right), \tag{1.3.5 b}$$

where $a = R$ and the dimensionless variable $r = \sin\vartheta$ lies in the interval $0 \leqslant r \leqslant 1$ ($r = 0$ at the poles and $r = 1$ at the equator). In the hyperboloidal case the metric is given by

$$\mathrm{d}l^2 = a^2(\mathrm{d}\vartheta^2 + \sinh^2\vartheta\,\mathrm{d}\varphi^2) = a^2\left(\frac{\mathrm{d}r^2}{1+r^2} + r^2\,\mathrm{d}\varphi^2\right), \tag{1.3.5 c}$$

where the dimensionless variable $r = \sinh\vartheta$ lies in the range $0 \leqslant r < \infty$.

The Robertson–Walker metric is obtained from (1.3.4), where the spatial part is simply the three-dimensional generalisation of (1.3.5). One finds that for the three-dimensional flat, positively curved and negatively curved spaces one has, respectively,

$$\mathrm{d}l^2 = a^2(\mathrm{d}r^2 + r^2\,\mathrm{d}\Omega^2), \tag{1.3.6 a}$$

$$\mathrm{d}l^2 = a^2(\mathrm{d}\chi^2 + \sin^2\chi\,\mathrm{d}\Omega^2) = a^2\left(\frac{\mathrm{d}r^2}{1-r^2} + r^2\,\mathrm{d}\Omega^2\right), \tag{1.3.6 b}$$

$$\mathrm{d}l^2 = a^2(\mathrm{d}\chi^2 + \sinh^2\chi\,\mathrm{d}\Omega^2) = a^2\left(\frac{\mathrm{d}r^2}{1+r^2} + r^2\,\mathrm{d}\Omega^2\right), \tag{1.3.6 c}$$

where $\mathrm{d}\Omega^2 = \mathrm{d}\vartheta^2 + \sin^2\vartheta\,\mathrm{d}\varphi^2$; $0 \leqslant \chi \leqslant \pi$ in (1.3.6 *b*) and $0 \leqslant \chi < \infty$ in (1.3.6 *c*). The values of $K = 1, 0, -1$ in (1.3.1) correspond, respectively, to the hypersphere, Euclidean space and space of constant negative curvature.

The geometrical properties of Euclidean space ($K = 0$) are well known. On the other hand, the properties of the hypersphere ($K = 1$) are complex. This space is closed, i.e. it has finite volume, but has no boundaries. This property is clear by analogy with the two-dimensional case of a sphere: beginning from a coordinate origin at the pole, the surface inside a radius $r_c(\vartheta) = a\vartheta$ has an area $S(\vartheta) = 2\pi a^2(1-\cos\vartheta)$, which increases with r_c and has a maximum value $S_{\max} = 4\pi a^2$ at $\vartheta = \pi$. The perimeter of this region is $L(\vartheta) = 2\pi a\sin\vartheta = 2\pi a r$, which is maximum at the 'equator' ($\vartheta = \frac{1}{2}\pi$), where it takes the value $2\pi a$, and is zero at the 'antipole' ($\vartheta = \pi$): the sphere is therefore a closed surface, with finite area and no boundary. In the three-dimensional case the volume of the region contained inside a radius

$$r_c(\chi) = a\chi = a\sin^{-1} r \tag{1.3.7}$$

has volume

$$V(\chi) = 2\pi a^3(\chi - \tfrac{1}{2}\sin 2\chi), \tag{1.3.8}$$

which increases and has a maximum value for $\chi = \pi$,

$$V_{\max} = 2\pi^2 a^3, \tag{1.3.9}$$

and area

$$S(\chi) = 4\pi a^2\sin^2\chi, \tag{1.3.10}$$

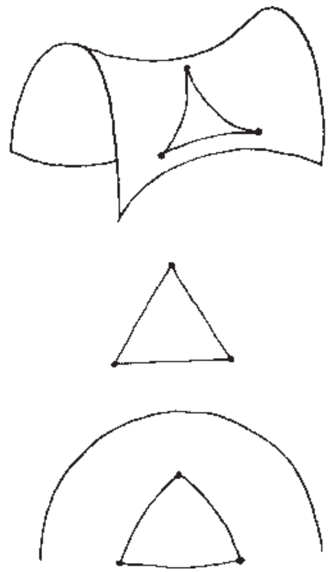

Figure 1.1 Examples of curved spaces in two dimensions: in a space with negative curvature (open), for example, the sum of the internal angles of a triangle is less than 180°, while for a positively curved space (closed) it is greater.

maximum at the 'equator' ($\chi = \frac{1}{2}\pi$), where it takes the value $4\pi a^2$, and is zero at the 'antipole' ($\chi = \pi$). In such a space the value of $S(\chi)$ is more than in Euclidean space, and the sum of the internal angles of a triangle is more than π. The properties of a space of constant negative curvature ($K = -1$) are more similar to those of Euclidean space: the hyperbolic space is open, i.e. infinite. All the relevant formulae for this space can be obtained from those describing the hypersphere by replacing trigonometric functions by hyperbolic functions. One can show, for example, that $S(\chi)$ is less than the Euclidean case, and the sum of the internal angles of a triangle is less than π.

In cases with $K \neq 0$, the parameter a, which appears in (1.3.1), is related to the curvature of space. In fact, the Gaussian curvature is given by $C_G = K/a^2$; as expected it is positive for the closed space and negative for the open space. The Gaussian curvature radius $R_G = C_G^{-1/2} = a/\sqrt{K}$ is, respectively, positive or imaginary in these two cases. In cosmology one uses the term radius of curvature to describe the modulus of R_G; with this convention a always represents the radius of spatial curvature. Of course, in a flat universe the parameter a does not have any geometrical significance.

As we shall see later in this chapter, the Einstein equations of general relativity relate the geometrical properties of space–time with the energy–momentum tensor describing the contents of the Universe. In particular, for a homogeneous and isotropic perfect fluid with rest-mass energy density ρc^2 and pressure p, the solutions of the Einstein equations are the *Friedmann cosmological equations*:

$$\ddot{a} = -\tfrac{4}{3}\pi G\left(\rho + 3\frac{p}{c^2}\right)a, \tag{1.3.11 a}$$

$$\dot{a}^2 + Kc^2 = \tfrac{8}{3}\pi G\rho a^2 \tag{1.3.11 b}$$

(the dot represents a derivative with respect to cosmological proper time t); the time evolution of the expansion parameter a which appears in the Robertson–Walker metric (1.3.1) can be derived from (1.3.11) if one has an equation of state relating p to ρ. From Equation (1.3.11 *b*) one can derive the curvature

$$\frac{K}{a^2} = \frac{1}{c^2}\left(\frac{\dot{a}}{a}\right)^2\left(\frac{\rho}{\rho_c} - 1\right), \tag{1.3.12}$$

where

$$\rho_c = \frac{3}{8\pi G}\left(\frac{\dot{a}}{a}\right)^2 \tag{1.3.13}$$

is called the *critical density*. The space is closed ($K = 1$), flat ($K = 0$) or open ($K = -1$) according to whether the *density parameter*

$$\Omega(t) = \frac{\rho}{\rho_c} \tag{1.3.14}$$

is greater than, equal to, or less than unity.

It will sometimes be useful to change the time variable we use from proper time to *conformal time*:

$$\tau = \int \frac{dt}{a(t)}; \tag{1.3.15}$$

with such a time variable the Robertson–Walker metric becomes

$$ds^2 = a(\tau)^2\left[(c\,d\tau)^2 - \left(\frac{dr^2}{1 - Kr^2} + r^2\,d\Omega^2\right)\right]. \tag{1.3.16}$$

1.4 The Hubble Law

The *proper distance*, d_P, of a point P from another point P_0, which we take to define the origin of a set of polar coordinates r, ϑ and φ, is the distance measured by a chain of rulers held by observers which connect P to P_0 at time t. From the Robertson–Walker metric (1.3.1) with $dt = 0$ this can be seen to be

$$d_P = \int_0^r \frac{a\,dr'}{(1 - Kr'^2)^{1/2}} = af(r), \tag{1.4.1}$$

where the function $f(r)$ is, respectively,

$$f(r) = \sin^{-1} r \qquad (K = 1), \tag{1.4.2 a}$$

$$f(r) = r \qquad (K = 0), \tag{1.4.2 b}$$

$$f(r) = \sinh^{-1} r \qquad (K = -1). \tag{1.4.2 c}$$

Of course this proper distance is of little operational significance because one can never measure simultaneously all the distance elements separating P from P_0. The proper distance at time t is related to that at the present time t_0 by

$$d_{\mathrm{P}}(t_0) = a_0 f(r) = \frac{a_0}{a} d_{\mathrm{P}}(t), \tag{1.4.3}$$

where a_0 is the value of $a(t)$ at $t = t_0$. Instead of the comoving coordinate r one could also define a radial comoving coordinate of P by the quantity

$$d_{\mathrm{c}} = a_0 f(r). \tag{1.4.4}$$

In this case the relation between comoving coordinates and proper coordinates is just

$$d_{\mathrm{c}} = \frac{a_0}{a} d_{\mathrm{P}}. \tag{1.4.5}$$

The proper distance d_{P} of a source may change with time because of the time-dependence of the expansion parameter a. In this case a source at P has a radial velocity with respect to the origin P_0 given by

$$v_{\mathrm{r}} = \dot{a} f(r) = \frac{\dot{a}}{a} d_{\mathrm{P}}. \tag{1.4.6}$$

Equation (1.4.6) is called the *Hubble law* and the quantity

$$H(t) = \dot{a}/a \tag{1.4.7}$$

is called the *Hubble constant* or, more accurately, the *Hubble parameter* (because it is not constant in time). As we shall see, the value of this parameter evaluated at the present time for our Universe, $H(t_0) = H_0$, is not known to any great accuracy. It is believed, however, to have a value around

$$H_0 \simeq 65 \text{ km s}^{-1} \text{ Mpc}^{-1}. \tag{1.4.8}$$

The unit 'Mpc' is defined later on in Section 4.1. It is conventional to take account of the uncertainty in H_0 by defining the dimensionless parameter h to be $H_0/100 \text{ km s}^{-1} \text{ Mpc}^{-1}$ (see Section 4.2). The law (1.4.6) can, in fact, be derived directly from the Cosmological Principle if $v \ll c$. Consider a triangle defined by the three spatial points O, O′ and P. Let the velocity of P and O′ with respect to O be, respectively, $\boldsymbol{v}(\boldsymbol{r})$ and $\boldsymbol{v}(\boldsymbol{d})$. The velocity of P with respect to O′ is

$$\boldsymbol{v}'(\boldsymbol{r}') = \boldsymbol{v}(\boldsymbol{r}) - \boldsymbol{v}(\boldsymbol{d}). \tag{1.4.9}$$

From the Cosmological Principle the functions $\boldsymbol{v}$ and $\boldsymbol{v}'$ must be the same. Therefore

$$\boldsymbol{v}(\boldsymbol{r} - \boldsymbol{d}) = \boldsymbol{v}'(\boldsymbol{r} - \boldsymbol{d}) = \boldsymbol{v}(\boldsymbol{r}) - \boldsymbol{v}(\boldsymbol{d}). \tag{1.4.10}$$

Equation (1.4.10) implies a linear relationship between $\boldsymbol{v}$ and $\boldsymbol{r}$:

$$v_\alpha = H_\alpha{}^\beta x_\beta \quad (\alpha, \beta = 1, 2, 3). \tag{1.4.11}$$

If we impose the condition that the velocity field is *irrotational*,

$$\nabla \times \boldsymbol{v} = \boldsymbol{0}, \tag{1.4.12}$$

which comes from the condition of isotropy, one can deduce that the matrix $H_\alpha{}^\beta$ is symmetric and can therefore be diagonalised by an appropriate coordinate transformation. From isotropy, the velocity field must therefore be of the form

$$v_i = Hx_i, \tag{1.4.13}$$

where H is only a function of time. Equation (1.4.13) is simply the Hubble law (1.4.6).

Another, simpler, way to derive Equation (1.4.6) is the following. The points O, O′ and P are assumed to be sufficiently close to each other that relativistic space–time curvature effects are negligible. If the universe evolves in a homogeneous and isotropic manner, the triangle OO′P must always be similar to the original triangle. This means that the length of all the sides must be multiplied by the same factor a/a_0. Consequently, the distance between any two points must also be multiplied by the same factor. We therefore have

$$l = \frac{a}{a_0} l_0, \tag{1.4.14}$$

where l_0 and l are the lengths of a line segment joining two points at times t_0 and t, respectively. From (1.4.14) we recover immediately the Hubble law (1.4.6).

One property of the Hubble law, which is implicit in the previous reasoning, is that we can treat any spatial position as the origin of a coordinate system. In fact, referring again to the triangle OO′P, we have

$$\boldsymbol{v}_{\mathrm{P}} = \boldsymbol{v}_{\mathrm{O}'} + \boldsymbol{v}'_{\mathrm{P}} = H\boldsymbol{d} + \boldsymbol{v}'_{\mathrm{P}} = H\boldsymbol{r} \tag{1.4.15}$$

and, therefore,

$$\boldsymbol{v}'_{\mathrm{P}} = H(\boldsymbol{r} - \boldsymbol{d}) = H\boldsymbol{r}', \tag{1.4.16}$$

which again is just the Hubble law, this time expressed about the point O′.

1.5 Redshift

It is useful to introduce a new variable related to the expansion parameter a which is more directly observable. We call this variable the *redshift* z and we shall use it extensively from now on in describing the evolution of the Universe because many of the relevant formulae are very simple when expressed in terms of this variable.

We define the redshift of a luminous source, such as a distant galaxy, by the quantity

$$z = \frac{\lambda_0 - \lambda_e}{\lambda_e}, \tag{1.5.1}$$

where λ_0 is the wavelength of radiation from the source observed at O (which we take to be the origin of our coordinate system) at time t_0 and emitted by the source at some (earlier) time t_e; the source is moving with the expansion of the universe and is at a comoving coordinate r. The wavelength of radiation emitted by the source is λ_e. The radiation travels along a light ray (null geodesic) from the source to the observer so that $ds^2 = 0$ and, therefore,

$$\int_{t_e}^{t_0} \frac{c\,dt}{a(t)} = \int_0^r \frac{dr}{(1 - Kr^2)^{1/2}} = f(r). \tag{1.5.2}$$

Light emitted from the source at $t'_e = t_e + \delta t_e$ reaches the observer at $t'_0 = t_0 + \delta t_0$. Given that $f(r)$ does not change, because r is a comoving coordinate and both the source and the observer are moving with the cosmological expansion, we can write

$$\int_{t'}^{t'_0} \frac{c\,dt}{a(t)} = f(r). \tag{1.5.3}$$

If δt and, therefore, δt_0 are small, Equations (1.5.2) and (1.5.3) imply that

$$\frac{\delta t_0}{a_0} = \frac{\delta t}{a}. \tag{1.5.4}$$

If, in particular, $\delta t = 1/\nu_e$ and $\delta t_0 = 1/\nu_0$ (ν_e and ν_0 are the frequencies of the emitted and observed light, respectively), we will have

$$\nu_e a = \nu_0 a_0 \tag{1.5.5}$$

or, equivalently,

$$\frac{a}{\lambda_e} = \frac{a_0}{\lambda_0}, \tag{1.5.6}$$

from which

$$1 + z = \frac{a_0}{a}. \tag{1.5.7}$$

A line of reasoning similar to the previous one can be made to recover the evolution of the velocity $v_p(t)$ of a test particle with respect to a comoving observer. At time $t + dt$ the particle has travelled a distance $dl = v_p(t)\,dt$ and thus finds itself moving with respect to a new reference frame which, because of the expansion of the universe, has an expansion velocity $dv = (\dot{a}/a)\,dl$. The velocity of the particle with respect to the new comoving observer is therefore

$$v_p(t + dt) = v_p(t) - \frac{\dot{a}}{a}\,dl = v_p(t) - \frac{\dot{a}}{a} v_p(t)\,dt, \tag{1.5.8}$$

which, integrated, gives

$$v_{\mathrm{p}} \propto a^{-1}. \tag{1.5.9}$$

The results expressed by Equations (1.5.5) and (1.5.11) are a particular example of the fact that, in a universe described by the Robertson–Walker metric, the momentum q of a free particle (whether relativistic or not) evolves according to $q \propto a^{-1}$.

There is also a simply way to recover Equation (1.5.7), which does not require any knowledge of the metric. Consider two nearby points P and P′, participating in the expansion of the Universe. From the Hubble law we have

$$\mathrm{d}v_{\mathrm{P}'} = H\,\mathrm{d}l = \frac{\dot{a}}{a}\,\mathrm{d}l, \tag{1.5.10}$$

where $\mathrm{d}v_{\mathrm{P}'}$ is the relative velocity of P′ with respect to P and $\mathrm{d}l$ is the (infinitesimal) distance between P and P′. The point P′ sends a light signal at time t and frequency ν which arrives at P with frequency ν' at time $t + \mathrm{d}t = t + (\mathrm{d}l/c)$. Since $\mathrm{d}l$ is infinitesimal, as is $\mathrm{d}v_{\mathrm{P}'}$, we can apply the approximate formula describing the *Doppler effect*:

$$\frac{\nu' - \nu}{\nu} = \frac{\mathrm{d}\nu}{\nu} \simeq -\frac{\mathrm{d}v_{\mathrm{P}'}}{c} = -\frac{\dot{a}}{a}\,\mathrm{d}t = -\frac{\mathrm{d}a}{a}. \tag{1.5.11}$$

The Equation (1.5.11) integrates immediately to give (1.5.5) and therefore (1.5.7).

1.6 The Deceleration Parameter

The Hubble parameter $H(t)$ measures the expansion rate at any particular time t for any model obeying the Cosmological Principle. It does, however, vary with time in a way that depends upon the contents of the Universe. One can express this by expanding the cosmic scale factor for times t close to t_0 in a power series:

$$a(t) = a_0[1 + H_0(t - t_0) - \tfrac{1}{2}q_0 H_0^2 (t - t_0)^2 + \cdots], \tag{1.6.1}$$

where

$$q_0 = -\frac{\ddot{a}(t_0)a_0}{\dot{a}(t_0)^2} \tag{1.6.2}$$

is called the *deceleration parameter*; the suffix '0', as always, refers to the fact that $q_0 = q(t_0)$. Note that while the Hubble parameter has the dimensions of inverse time, q is actually dimensionless.

Putting the redshift, defined by Equation (1.5.7), into Equation (1.6.1) we find that

$$z = H_0(t_0 - t) + (1 + \tfrac{1}{2}q_0)H_0^2(t_0 - t)^2 + \cdots, \tag{1.6.3}$$

which can be inverted to yield

$$t_0 - t = \frac{1}{H_0}[z - (1 + \tfrac{1}{2}q_0)z^2 + \cdots]. \tag{1.6.4}$$

To find r as a function of z one needs to recall that, for a light ray,

$$\int_t^{t_0} \frac{c\,\mathrm{d}t}{a} = \int_0^r \frac{\mathrm{d}r}{(1-Kr^2)^{1/2}}, \tag{1.6.5}$$

which becomes, using Equations (1.5.7) and (1.6.3),

$$\frac{c}{a_0}\int_t^{t_0}[1 + H_0(t_0-t) + (1+\tfrac{1}{2}q_0)H_0^2(t_0-t)^2 + \cdots]\,\mathrm{d}t = r + \mathcal{O}(r^3), \tag{1.6.6}$$

and therefore

$$r = \frac{c}{a_0}[(t_0-t) + \tfrac{1}{2}H_0(t_0-t)^2 + \cdots]. \tag{1.6.7}$$

Substituting Equation (1.6.4) into (1.6.7) we have, finally,

$$r = \frac{c}{a_0 H_0}[z - \tfrac{1}{2}(1+q_0)z^2 + \cdots]. \tag{1.6.8}$$

Expressions of this type are useful because they do not require full solutions of the Einstein equations for $a(t)$; the quantity q_0 is used to parametrise a family of approximate solutions for t close to t_0.

1.7 Cosmological Distances

We have shown how the comoving coordinate system we have adopted relates to *proper distance* (i.e. distances measured in a hypersurface of constant proper time) in spaces described by the Robertson–Walker metric. Obviously, however, we cannot measure proper distances to astronomical objects in any direct way. Distant objects are observed only through the light they emit which takes a finite time to travel to us; we cannot therefore make measurements along a surface of constant proper time, but only along the set of light paths travelling to us from the past – our past *light cone*. One can, however, define operationally other kinds of distance which are, at least in principle, directly measurable.

One such distance is the *luminosity distance* d_{L}. This is defined in such a way as to preserve the Euclidean inverse-square law for the diminution of light with distance from a point source. Let L denote the power emitted by a source at a point P, which is at a coordinate distance r at time t. Let l be the power received per unit area (i.e. the flux) at time t_0 by an observer placed at P_0. We then define

$$d_{\mathrm{L}} = \left(\frac{L}{4\pi l}\right)^{1/2}. \tag{1.7.1}$$

The area of a spherical surface centred on P and passing through P_0 at time t_0 is just $4\pi a_0^2 r^2$. The photons emitted by the source arrive at this surface having been redshifted by the expansion of the universe by a factor a/a_0. Also, as we have seen, photons emitted by the source in a small interval δt arrive at P_0 in an interval $\delta t_0 = (a_0/a)\delta t$ due to a time-dilation effect. We therefore find

$$l = \frac{L}{4\pi a_0^2 r^2}\left(\frac{a}{a_0}\right)^2, \tag{1.7.2}$$

from which

$$d_{\mathrm{L}} = a_0^2 \frac{r}{a}. \tag{1.7.3}$$

Following the same procedure as in Section 1.6, one can show that

$$d_{\mathrm{L}} = \frac{c}{H_0}[z + \tfrac{1}{2}(1 - q_0)z^2 + \cdots], \tag{1.7.4}$$

in contrast with the proper distance, d_{P}, defined by Equation (1.4.1), which has the form $d_{\mathrm{P}} = a_0 r$, with $f(r)$ given by Equations (1.4.2).

Next we define the *angular-diameter distance* d_{A}. Again, this is constructed in such a way as to preserve a geometrical property of Euclidean space, namely the variation of the angular size of an object with its distance from an observer. Let $D_{\mathrm{P}}(t)$ be the (proper) diameter of a source placed at coordinate r at time t. If the angle subtended by D_{P} is denoted $\Delta\vartheta$, then Equation (1.2.1) implies

$$D_{\mathrm{P}} = ar\Delta\vartheta. \tag{1.7.5}$$

We define d_{A} to be the distance

$$d_{\mathrm{A}} = \frac{D_{\mathrm{P}}}{\Delta\vartheta} = ar; \tag{1.7.6}$$

it should be noted that a decreases as r increases for the same D_{P} and, in some models, the angular size of a source can actually increase with its luminosity distance.

Other measures of distance, less often used, are the *parallax distance*

$$d_\mu = a_0 \frac{r}{(1 - Kr^2)^{1/2}}, \tag{1.7.7}$$

and the *proper motion distance*

$$d_{\mathrm{M}} = a_0 r. \tag{1.7.8}$$

Evidently, for $r \to 0$, and therefore for $t \to t_0$, we have

$$d_{\mathrm{p}} \simeq d_{\mathrm{L}} \simeq d_{\mathrm{A}} \simeq d_\mu \simeq d_{\mathrm{M}} \simeq d_{\mathrm{c}}, \tag{1.7.9}$$

so that at small distances we recover the Euclidean behaviour.

1.8 The m–z and N–z Relations

The general relationship we have established between redshift and distance allows us to establish some interesting properties of the Universe which could, in principle, be used to probe its spatial geometry and, in particular, to test the Cosmological Principle. In fact, there are severe complications with the implementation of this idea, as we discuss in Section 4.7. If celestial objects (such as galaxy clusters, galaxies, radio sources, quasars, etc.) are distributed homogeneously and isotropically on large scales, it is interesting to consider two relationships: the m–z relationship between the apparent magnitude of a source and its redshift and the $N(> l)$–z relationship between the number of sources of a given type with apparent luminosity greater than some limit l and redshift less than z. These relations are also important because, in principle, they provide a way of determining the deceleration parameter q_0.

As we have seen previously,

$$d_{\mathrm{L}} = \frac{c}{H_0}[z + \tfrac{1}{2}(1 - q_0)z^2 + \cdots], \tag{1.8.1}$$

from which

$$l = \frac{L}{4\pi d_{\mathrm{L}}^2} = \frac{LH_0^2}{4\pi c^2 z^2}[1 + (q_0 - 1)z + \cdots]. \tag{1.8.2}$$

Astronomers do not usually work with the absolute luminosity L and apparent flux l. Instead they work with quantities related to these: the *absolute magnitude* M and the *apparent magnitude* m (for more details see Section 4.1). The magnitude scale is defined logarithmically by taking a factor of 100 in received flux to be a difference of 5 magnitudes. The zero-point can be fixed in various ways; for historical reasons it is conventional to take Polaris to have an apparent magnitude of 2.12 in visible light but different choices can and have been made. The absolute magnitude is defined to be the apparent magnitude the source would have if it were placed at a distance of 10 parsec. The relationship between the luminosity distance of a source, its apparent magnitude m and its absolute magnitude M is, therefore, just

$$d_{\mathrm{L}} = 10^{1+(m-M)/5}\ \mathrm{pc}. \tag{1.8.3}$$

The quantity

$$m - M = -5 + 5\log d_{\mathrm{L}}(\mathrm{pc}) \tag{1.8.4}$$

is called the *distance modulus*. Using Equation (1.8.2) we find

$$m - M \simeq 25 - 5\log_{10} H_0 + 5\log cz + 1.086(1 - q_0)z + \cdots, \tag{1.8.5}$$

with H_0 in km s^{-1} Mpc^{-1} and c in km s^{-1}. Here one should remember that 1 Mpc = 10^6 pc and the logarithms are always defined to the base 10. The behaviour of $m(z)$ is sensitive to the value of q_0 only for $z > 0.1$. In reality, as we shall see, there are many other factors which intervene in this type of analysis with the result that we can say very little about q_0, or even its sign. In the regime where it is accurate, that is for $z < z_{\max} \simeq 0.2$, Equation (1.8.5) can provide an estimate of H_0, together with a strong confirmation of the validity of the Hubble law and, therefore, of the Cosmological Principle.

Another test of this principle is the so-called *Hubble test*, which relates the number $N(> l)$ of sources of a particular type with apparent luminosity greater than l as a function of l. If the Universe were Euclidean and galaxies all had the same absolute luminosity L, and were distributed uniformly with mean number-density n_0, we would have

$$N(l) = \tfrac{4}{3}\pi n_0 d_l^3, \tag{1.8.6}$$

with d_l given by

$$d_l = \left(\frac{L}{4\pi l}\right)^{1/2}, \tag{1.8.7}$$

from which

$$N(l) \propto l^{-3/2} \tag{1.8.8}$$

and, therefore, introducing the apparent magnitude in the form $m = 2.5 \log_{10} l +$ const.,

$$\log N(l) = 0.6m + \text{const.} \tag{1.8.9}$$

Equation (1.8.9) is also true if the sources have an arbitrary distribution of luminosities around L; in this case all that changes is the value of the constant.

In the non-Euclidean case we have

$$N(l) = 4\pi \int_0^r \frac{n[t(r')]a[t(r')]^3 r'^2}{(1 - Kr'^2)^{1/2}}\, \mathrm{d}r', \tag{1.8.10}$$

where $t(r')$ is the time at which a source at r' emitted a light signal which arrives now at the observer. If the galaxies are neither created nor destroyed in the interval $t(r) < t < t_0$, so that $na^3 = n_0 a_0^3$, we see that, upon expanding as a power series, Equation (1.8.10) leads to

$$N(l) = 4\pi n_0 a_0^3 (\tfrac{1}{3} r^3 + \tfrac{1}{10} K r^5 + \cdots). \tag{1.8.11}$$

Recalling that

$$r = \frac{c}{a_0 H_0}[z - \tfrac{1}{2}(1 + q_0)z^2 + \cdots], \tag{1.8.12}$$

Equation (1.8.11) becomes

$$\log N(l) = 3 \log z - 0.651(1 + q_0)z + \text{const.}, \tag{1.8.13}$$

from which one can, in principle, recover q_0. In practice, however, there are many effects (the most important being various evolutionary phenomena) which effectively mean that the constant terms in the above equations all actually depend on z. Nevertheless, Equation (1.8.13) works well for $z < 0.2$, where the term in q_0 is negligible and the constant is, effectively, constant.

1.9 Olbers' Paradox

Having established the behaviour of light in the expanding relativistic cosmology, it is worth revisiting an idea from the pre-relativistic era. Before the development of relativity, astronomers generally believed the Universe to be infinite, homogeneous, Euclidean and static. This picture was of course shattered by the discovery of the Hubble expansion in 1929, which we discuss in Chapter 4. It is nevertheless interesting to point out that this model, which we might call the Eighteenth Century Universe, gave rise to an interesting puzzle now known as *Olbers' Paradox* (Olbers 1826). As a matter of fact, Olbers' Paradox had previously been analysed by a number of others, including (incorrectly) Halley (1720) and (correctly) Loys de Chéseaux (1744). The argument proceeds from the simple observation that the night sky is quite dark. In an Eighteenth Century Universe, the apparent luminosity l of a star of absolute luminosity L placed at a distance r from an observer is just

$$l = \frac{L}{4\pi r^2} \tag{1.9.1}$$

if one neglects absorption. This is the same as Equation (1.7.1). Let us assume, for simplicity, that all stars have the same absolute luminosity and the (constant) number density of stars per unit volume is n. The radiant energy arriving at the observer from the whole Universe is then

$$l_{\text{tot}} = \int_0^\infty \frac{L}{4\pi r^2} 4\pi r^2 \, \mathrm{d}r = nL \int_0^\infty \mathrm{d}r, \tag{1.9.2}$$

which is infinite. This is the Olbers Paradox. It was thought in the past that this paradox could be resolved by postulating the presence of interstellar absorption, perhaps by dust; such an explanation was actually advanced by Lord Kelvin in the 19th century. What would happen if this were the case would be that, after a sufficient time, the absorbing material would be brought into thermodynamic equilibrium with the radiation and would then emit as much radiation as it absorbed, though perhaps in a different region of the electromagnetic spectrum. To be fair to Kelvin, however, one should mention that at that time it was not known that light and heat were actually different aspects of the same phenomenon, so the argument was reasonable given what was then known about the nature of radiation.

In the modern version of the expanding Universe the conditions necessary for an Olbers Paradox to arise are violated in a number of ways we shall discuss later: the light from a distant star would be redshifted; the spatial geometry is not necessarily flat; the Universe may not be infinite in spatial or temporal extent. In fact, the basic reason why an Olbers Paradox does not arise in modern cosmological theories is much simpler than any of these possibilities. The key fact is that no star can burn for an infinite time: a star of mass M can at most radiate only so long as it takes to radiate away its rest energy Mc^2. As one looks further and further out into space, one must see stars which are older and older. In order for them all, out to infinite distance, to be shining light that we observe now, they must have switched on at different times depending on their distance from us. Such a coordination is not only unnatural, it also requires us to be in a special place. So an Olbers Paradox would only really be expected to happen if the Universe were actually inhomogeneous on large scales and the Copernican Principle were violated. The other effects mentioned above are important in determining the exact amount of radiation received by an observer from the cosmological background, but any cosmology that respects the relativistic notion that $E = mc^2$ (and the Cosmological Principle) is not expected to have an infinitely bright night sky. Exactly how much background light there is in the Universe is an observation that can in principle be used to test cosmological models in much the same way as the number-counts discussed in Section 1.8.

1.10 The Friedmann Equations

So far we have developed much of the language of modern relativistic cosmology without actually using the field Equations (1.2.20). We have managed to discuss many important properties of the universe in terms of geometry or using simple kinematics. To go further we must use general relativity to relate the geometry of space–time, expressed by the metric tensor $g_{ij}(x_k)$, to the matter content of the universe, expressed by the energy–momentum tensor $T_{ij}(x_k)$. The *Einstein equations* (without the cosmological constant; see Section 1.12) are

$$R_{ij} - \tfrac{1}{2} g_{ij} R = \frac{8\pi G}{c^4} T_{ij}, \tag{1.10.1}$$

where R_{ij} and R are the Ricci tensor and Ricci scalar, respectively. A test particle moves along a space–time geodesic, that is a trajectory in a four-dimensional space whose 'length' is stationary with respect to small variations in the trajectory.

In cosmology, the energy–momentum tensor which is of greatest relevance is that of a perfect fluid:

$$T_{ij} = (p + \rho c^2) U_i U_j - p g_{ij}, \tag{1.10.2}$$

where p is the pressure, ρc^2 is the energy density (which includes the rest-mass energy), and U_k is the fluid four-velocity, defined by Equation (1.2.10). If the metric is of Robertson–Walker type, the Einstein equations then yield

$$\ddot{a} = -\frac{4\pi}{3} G \left(\rho + 3\frac{p}{c^2} \right) a, \tag{1.10.3}$$

for the time–time component, and

$$a\ddot{a} + 2\dot{a}^2 + 2Kc^2 = 4\pi G \left(\rho - \frac{p}{c^2} \right) a^2, \tag{1.10.4}$$

for the space–space components. The space–time components merely give $0 = 0$. Eliminating $\ddot{a}$ from (1.10.3) and (1.10.4) we obtain

$$\dot{a}^2 + Kc^2 = \tfrac{8}{3}\pi G \rho a^2. \tag{1.10.5}$$

In reality, as we shall see, Equations (1.10.3) and (1.10.5) – the *Friedmann equations* – are not independent: the second can be recovered from the first if one takes the adiabatic expansion of the universe into account, i.e.

$$d(\rho c^2 a^3) = -p\, da^3. \tag{1.10.6 a}$$

The last equation can also be expressed as

$$\dot{p} a^3 = \frac{d}{dt}[a^3(\rho c^2 + p)] \tag{1.10.6 b}$$

or

$$\dot{\rho} + 3\left(\rho + \frac{p}{c^2}\right)\frac{\dot{a}}{a} = 0. \tag{1.10.6 c}$$

1.11 A Newtonian Approach

Before proceeding further, it is worth demonstrating how one can actually get most of the way towards the Friedmann equations using only Newtonian arguments.

Birkhoff's theorem (1923) proves that a spherically symmetric gravitational field in an empty space is static and is always described by the Schwarzschild exterior metric (i.e. the metric generated in empty space by a point mass). This property is very similar to a result proved by Newton and usually known as *Newton's spherical theorem* which is based on the application of Gauss's theorem to the gravitational field. In the Newtonian version the gravitational field outside a spherically symmetric body is the same as if the body had all its mass concentrated at its centre. Birkhoff's theorem can also be applied to the field inside an empty spherical cavity at the centre of a homogeneous spherical distribution of mass–energy, even if the distribution is not static. In this case the metric inside the cavity is the Minkowski (flat-space) metric: $g_{ij} = \eta_{ij}$ ($\eta_{ij} = -1$ for $i = j = 1, 2, 3$; $\eta_{ij} = 1$ for $i = j = 0$; $\eta_{ij} = 0$ for $i \neq j$). This corollary of Birkhoff's theorem also has a Newtonian analogue: the gravitational field inside a homogeneous spherical shell of matter is always zero. This corollary can also be applied if the space outside the cavity is infinite: the only condition that must be obeyed is that the distribution of mass–energy must be spherically symmetric.

A proof of Birkhoff's theorem is beyond the scope of this book, but we will use its existence to justify a Newtonian approach to the time-evolution of a homogeneous and isotropic distribution of material. Let us consider the evolution of the mass m contained inside a sphere of radius l centred at the point O in such a universe. By Birkhoff's theorem the space inside the sphere is flat. If the radius l is such that

$$\frac{Gm}{lc^2} \ll 1, \tag{1.11.1}$$

one can use Newtonian mechanics to describe the behaviour of the particle. Equation (1.11.1) means in effect that the free-fall time for the sphere, $\tau_{\rm ff} \simeq (G\rho)^{-1/2}$, is much greater than the light-crossing time $\tau \simeq l/c$. Alternatively, Equation (1.11.1) means that the radius of the sphere is much larger than the Schwarzschild radius corresponding to the mass m, $r_S = 2mG/c^2$.

As we have seen in Section 1.4, the Cosmological Principle requires that

$$l = d_c \frac{a}{a_0}, \tag{1.11.2}$$

where a is the expansion parameter of the universe which, according to our conventions, has the dimensions of a length, while the comoving coordinate d_c is dimensionless. One can always pick d_c small enough so that at any instant the inequality (1.11.1) is satisfied. We shall see, however, that this quantity actually disappears from the formulae.

Applying a Newtonian approximation to describe the motion of a unit mass at a point P on the surface of the sphere yields

$$\frac{d^2 l}{dt^2} = -\frac{Gm}{l^2} = -\frac{4\pi}{3} G\rho l, \tag{1.11.3}$$

or, multiplying by $\dot{l}$,

$$\frac{d}{dt}\frac{\dot{l}^2}{2} = -\frac{Gm}{l^2}\dot{l} = \frac{d}{dt}\frac{Gm}{l}, \tag{1.11.4}$$

and, integrating,

$$\dot{l}^2 = \frac{2Gm}{l} + C, \tag{1.11.5}$$

which is nothing more than the law of conservation of energy per unit mass: the constant of integration C is proportional to the total energy. From Equations (1.11.2) and (1.11.5) it is easy to obtain the Equation (1.10.4) in the form

$$\dot{a}^2 + Kc^2 = \tfrac{8}{3}\pi G\rho a^2 \tag{1.11.6}$$

by putting

$$C = -K\left(\frac{d_c c}{a_0}\right)^2. \tag{1.11.7}$$

It is clear that, with an appropriate redefinition of d_c, one can scale K so as to take the values 1, 0 or −1. The case $K = 1$ corresponds to $C < 0$ (negative total energy). In this case the expansion eventually ceases and collapse ensues. In the case $K = -1$ the total energy is positive, so the expansion never ends. The case $K = 0$ corresponds to total energy of exactly zero: this represents the 'escape velocity' situation where the expansion ceases at $t \to \infty$.

Equation (1.11.3) implies that there are no forces due to pressure gradients, which is in accord with our assumption of homogeneity and isotropy. Equation (1.11.6) was obtained under the assumption that the sphere contains only non-relativistic matter ($p \ll \rho c^2$). A result from general relativity shows that, in the presence of relativistic particles, one should replace the density of matter in Equation (1.11.3) by

$$\rho_{\mathrm{eff}} = \rho + 3\frac{p}{c^2}, \tag{1.11.8}$$

where ρ now means the energy density (including the rest-mass energy) divided by c^2. In this way, Equation (1.11.3) becomes

$$\ddot{a} = -\frac{4\pi}{3}G\left(\rho + 3\frac{p}{c^2}\right)a. \tag{1.11.9}$$

It is important to note that, from Equation (1.10.6 a),

$$\mathrm{d}(\rho c^2 a^3 r_0^3) = -p\,\mathrm{d}(a^3 r_0^3); \tag{1.11.10}$$

from (1.11.9) one obtains (1.11.6) in both the non-relativistic ($p \simeq 0$, $\rho = \rho_{\mathrm{m}}$) and ultra-relativistic ($p \simeq \rho c^2$) cases. In fact Equation (1.11.9), after multiplying by $\dot{a}$, gives

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\dot{a}^2 = -\frac{4\pi}{3}G\left(\rho a\dot{a} + 3\frac{p}{c^2}a\dot{a}\right). \tag{1.11.11}$$

From (1.11.10) we have

$$3\frac{p}{c^2}a\dot{a} = -3\rho a\dot{a} - \dot{\rho}a^2, \tag{1.11.12}$$

which, substituted in Equation (1.11.11), yields

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\dot{a}^2 = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{4\pi}{3}G\rho a^2\right). \tag{1.11.13}$$

From Equation (1.11.13), by integration, one obtains Equation (1.11.6).

What this shows is that, with Birkhoff's theorem and a reinterpretation of the quantity ρ to take account of intrinsically relativistic effects, we can derive the Friedmann equations using an essentially Newtonian approach.

1.12 The Cosmological Constant

Einstein formulated his theory of general relativity without a cosmological constant in 1916; at this time it was generally accepted that the Universe was static. We outlined the development of this theory in Section 1.2, and the field equations themselves appear as Equation (1.10.1). A glance at the equation

$$\ddot{a} = -\frac{4\pi}{3}G\left(\rho + 3\frac{p}{c^2}\right)a \tag{1.12.1}$$

shows one that universes evolving according to this theory cannot be static, unless

$$\rho = -3\frac{p}{c^2}; \tag{1.12.2}$$

in other words, either the energy density or the pressure must be negative. Given that this type of fluid does not seem to be physically reasonable, Einstein (1917) modified the Equation (1.10.1) by introducing the cosmological constant term Λ:

$$R_{ij} - \tfrac{1}{2} g_{ij} R - \Lambda g_{ij} = \frac{8\pi G}{c^4} T_{ij}; \tag{1.12.3}$$

as we shall see, with an appropriate choice of Λ, one can obtain a static cosmological model. Equation (1.12.3) represents the most general possible modification of the Einstein equations that still satisfies the condition that T_{ij} is equal to a tensor constructed from the metric g_{ij} and its first and second derivatives, and is linear in the second derivative. This modification does not change the covariant character of the equations, and does not alter the continuity condition (1.2.12). The strongest constraint one can place on Λ from observations is that it should be sufficiently small so as not to change the laws of planetary motion, which are known to be well described by (1.10.1).

The Equation (1.12.3) can be written in a form similar to (1.10.1) by modifying the energy–momentum tensor:

$$R_{ij} - \tfrac{1}{2} g_{ij} R = \frac{8\pi G}{c^4} \tilde{T}_{ij}, \tag{1.12.4}$$

with $\tilde{T}_{ij}$ formally given by

$$\tilde{T}_{ij} = T_{ij} + \frac{\Lambda c^4}{8\pi G} g_{ij} = -\tilde{p} g_{ij} + (\tilde{p} + \tilde{\rho} c^2) U_i U_j, \tag{1.12.5}$$

where the effective pressure $\tilde{p}$ and the effective density $\tilde{\rho}$ are related to the corresponding quantities for a perfect fluid by

$$\tilde{p} = p - \frac{\Lambda c^4}{8\pi G}, \qquad \tilde{\rho} = \rho + \frac{\Lambda c^2}{8\pi G}; \tag{1.12.6}$$

these relations show that $|\Lambda|^{-1/2}$ has the dimensions of a length. One can then show that, for a universe described by the Robertson–Walker metric, we can get equations which are analogous to (1.10.3) and (1.10.5), respectively:

$$\ddot{a} = -\frac{4\pi}{3} G \left(\tilde{\rho} + 3 \frac{\tilde{p}}{c^2} \right) a \tag{1.12.7}$$

and

$$\dot{a}^2 + Kc^2 = \frac{8\pi G}{3} \tilde{\rho} a^2. \tag{1.12.8}$$

These equations admit a static solution for

$$\tilde{\rho} = -3 \frac{\tilde{p}}{c^2} = \frac{3Kc^2}{8\pi G a^2}. \tag{1.12.9}$$

For a 'dust' universe ($p = 0$), which is a good approximation to our Universe at the present time, Equations (1.12.9) and (1.12.6) give

$$\Lambda = \frac{K}{a^2}, \qquad \rho = \frac{Kc^2}{4\pi G a^2}. \tag{1.12.10}$$

Since $\rho > 0$, we must have $K = 1$ and therefore $\Lambda > 0$. The value of Λ which makes the universe static is just

$$\Lambda_{\mathrm{E}} = \frac{4\pi G\rho}{c^2}. \tag{1.12.11}$$

The model we have just described is called the *Einstein universe*. This universe is static (but unfortunately unstable, as one can show), has positive curvature and a curvature radius

$$a_{\mathrm{E}} = \Lambda_{\mathrm{E}}^{-1/2} = \frac{c}{(4\pi G\rho)^{1/2}}. \tag{1.12.12}$$

After the discovery of the expansion of the Universe in the late 1920s there was no longer any reason to seek static solutions to the field equations. The motivation which had led Einstein to introduce his cosmological constant term therefore subsided. Einstein subsequently regarded the Λ-term as the biggest mistake he had made in his life. Since then, however, Λ has not died but has been the subject of much interest and serious study on both conceptual and observational grounds. The situation here is reminiscent of Aladdin and the genie: after he released the genie from the lamp, it took on a life of its own. For more than 60 years the genie lingered, providing neither compelling observational evidence of its existence nor strong theoretical reasons for it to be taken seriously. However, observations do now suggest that it may have been there all along. We shall return to this resurgence of Λ in the next chapter and also in Chapter 7, but in the meantime we shall restrict ourselves to brief comments on two particularly important models involving the cosmological constant, because we shall encounter them again when we discuss inflation.

The *de Sitter universe* (de Sitter 1917) is a cosmological model in which the universe is empty ($p = 0$; $\rho = 0$) and flat ($K = 0$). From Equations (1.12.6) we get

$$\tilde{p} = -\tilde{\rho}c^2 = -\frac{\Lambda c^4}{8\pi G}, \tag{1.12.13}$$

which, on substitution in (1.12.8), gives

$$\dot{a}^2 = \tfrac{1}{3}\Lambda c^2 a^2; \tag{1.12.14}$$

this equation implies that Λ is positive. Equation (1.12.14) has a solution of the form

$$a = A\exp[(\tfrac{1}{3}\Lambda)^{1/2}ct], \tag{1.12.15}$$

corresponding to a Hubble parameter $H = \dot{a}/a = c(\Lambda/3)^{1/2}$, which is actually constant in time. In the de Sitter vacuum universe, test particles move away from each other because of the repulsive gravitational effect of the positive cosmological constant.

The de Sitter model was only of marginal historical interest until the last 20 years or so. In recent years, however, it has been a major component of inflationary universe models in which, for a certain interval of time, the expansion assumes an exponential character of the type (1.12.15). In such a universe the equation of state of the fluid is of the form $p \simeq -\rho c^2$ due to quantum effects which we discuss in Chapter 7.

In the *Lemaître model* (1927) the universe has positive spatial curvature ($K = 1$). One can demonstrate that the expansion parameter in this case is always increasing, but there is a period in which it remains practically constant. This model was invoked around 1970 to explain the apparent concentration of quasars at a redshift of $z \simeq 2$. Subsequent data have, however, shown that this is not the explanation for the redshift evolution of quasars, so this model is again of only marginal historical interest.

1.13 Friedmann Models

Having dealt with a few special cases, we now introduce the standard cosmological models described by the solutions (1.10.3) and (1.10.5). Their name derives from A. Friedmann, who derived their properties in 1922. His work was not well known at that time partly because his models were not static, and the discovery of the Hubble expansion was still some way in the future. His work was in any case not widely circulated in the western scientific literature. Independently, and somewhat later, the Belgian priest George Lemaître obtained essentially the same results and his work achieved more immediate attention, especially in England where he was championed by Eddington. When the work of Lemaître (1927) was published, Hubble's observations were just becoming known, so in the West Lemaître is often credited with being the father of the Big Bang cosmology, although that honour should probably be conferred on Friedmann.

The Friedmann models are so important that we shall devote the next chapter to their behaviour. Here we shall just whet the readers appetite with some basic properties. First, we assume a perfect fluid with some density ρ and pressure p. The form of equation of state giving p as a function of ρ does not concern us for now; we discuss it in Section 2.1. For the moment we also ignore the cosmological constant.

The equations we need to solve are (1.10.3) and (1.10.5), which we rewrite here for completeness:

$$\ddot{a} = -\frac{4\pi}{3}G\left(\rho + 3\frac{p}{c^2}\right)a \tag{1.13.1}$$

and

$$\dot{a}^2 + Kc^2 = \frac{8\pi G}{3}\rho a^2, \tag{1.13.2}$$

as well as the Equation (1.10.6)

$$d(\rho a^3) = -3\frac{p}{c^2}a^2\,da. \tag{1.13.3}$$

The Equations (1.13.1)–(1.13.3) allow one, at least in principle, to calculate the time evolution of $a(t)$ as well as $\rho(t)$ and $p(t)$ if we know the equation of state.

Let us focus for now on Equation (1.13.3), which can be rewritten in a convenient form for $a = a_0$:

$$\left(\frac{\dot a}{a_0}\right)^2 - \frac{8\pi}{3}G\rho\left(\frac{a}{a_0}\right)^2 = H_0^2\left(1-\frac{\rho_0}{\rho_{0c}}\right) = H_0^2(1-\Omega_0) = -\frac{Kc^2}{a_0^2}, \tag{1.13.4}$$

where $H_0 = \dot a_0/a_0$, Ω_0 is the (present) *density parameter* and

$$\rho_{0c} = \frac{3H_0^2}{8\pi G}. \tag{1.13.5}$$

The suffix '0' refers here to a generic reference time t_0 which is also used in the particular case where t is the present time. Equation (1.13.5) is a reminder of the importance of ρ_{0c}: if $\rho_0 < \rho_{0c}$, then $K = -1$, while if $\rho_0 > \rho_{0c}$, $K = 1$; $K = 0$ corresponds to the 'critical' case when $\rho_0 = \rho_{0c}$.

Let us now include the cosmological constant term Λ. In Section 1.12 we showed how one can treat the cosmological constant as a form of fluid with a strange equation of state, as well as a modification of the law of gravity. In that sense, Λ can be thought of as belonging either on the left-hand or right-hand side of the Einstein equations. Either way, the upshot is that Equations (1.13.1) and (1.13.2) become

$$\ddot a = -\frac{4\pi}{3}G\left(\rho + 3\frac{p}{c^2}\right)a + \frac{\Lambda c^2 a}{3} \tag{1.13.6}$$

and

$$\dot a^2 + Kc^2 = \frac{8\pi G}{3}\rho a^2 + \frac{\Lambda c^2 a^2}{3}, \tag{1.13.7}$$

respectively. If we ignore the original terms in p and ρ we can see that Equation (1.13.7) can be written in a form similar to Equation (1.13.4):

$$\left(\frac{\dot a}{a_0}\right)^2 - \frac{\Lambda c^2}{3} = H_0^2\left(1-\frac{\Lambda}{\Lambda_c}\right) = H_0^2(1-\Omega_{0\Lambda}) = -\frac{Kc^2}{a_0^2}. \tag{1.13.8}$$

In this case the 'critical' value is

$$\Lambda_c = \frac{3H_0^2}{c^2}, \tag{1.13.9}$$

so that $\Omega_{0\Lambda} = \Lambda c^2/3H_0^2$.

If we now reinstate the 'ordinary' matter we began with, we can see that the curvature is zero as long as $\Omega_{0\Lambda} + \Omega_0 = 1$. The cosmological constant therefore breaks the relationship between the matter density and curvature. Even if $\Omega_0 < 1$, a suitably chosen value of $\Omega_{0\Lambda} = 1 - \Omega_0$ can be invoked to ensure flat space sections.

Bibliographic Notes on Chapter 1

The classic papers of Einstein (1917), de Sitter (1917), Friedmann (1922) and Lemaître (1927) are all well worth reading for historical insights. A particularly erudite overview of the role of observation in expanding world models is given by Sandage (1988). More detailed discussions of the basic background, including the role of general relativity in cosmology, can be found in Berry (1989), Harrison (1981), Kenyon (1990), Landau and Lifshitz (1975), Milne (1935), Misner *et al*. (1972), Narlikar (1993), Peebles (1993), Peacock (1999), Raychaudhuri (1979), Roos (1994), Wald (1984), Weinberg (1972) and Zel'dovich and Novikov (1983).

Problems

1. Suppose that it is discovered that Newton's law of gravitation is incorrect, and that the force F on a test particle of mass m due to a body of mass M has an additional term that does not depend on M and is proportional to the separation r:

$$F = -\frac{GMm}{r^2} + \frac{Amr}{3}.$$

   Assuming that Newton's sphere theorem continues to hold, derive the appropriate form of the Friedmann equation in this case and comment on your result.

2. The most general form of a space–time four-metric in the synchronous gauge is

$$\mathrm{d}s^2 = c^2\,\mathrm{d}t^2 - g_{\alpha\beta}\,\mathrm{d}x^\alpha\,\mathrm{d}x^\beta = c^2\,\mathrm{d}t^2 - \mathrm{d}l^2,$$

   where $g_{\alpha\beta}$ is the three-metric of the spatial hypersurfaces. By writing the equation of the three-space as that of a constrained surface in four dimensions, show that the most general form of the three-metric compatible with homogeneity and isotropy is given by the Robertson–Walker form.

3. Show that the special-relativistic formula for the Doppler shift,

$$1 + z = \sqrt{\frac{1 + v/c}{1 - v/c}},$$

   reduces to $z \simeq v/c$ in the limit of small velocities. Invert the formula to give v/c in terms of z. Calculate the recession velocity of a galaxy at $z = 5$ using the special-relativistic formula.

4. A model is constructed with $\Omega_0 < 1$, $\Lambda \neq 0$ and $k = 0$. Show in this case that

$$q_0 = \tfrac{3}{2}\Omega_0 - 1.$$

5. An object has luminosity distance d_{L} and angular-diameter distance d_{A}. Show that

$$\frac{d_{\mathrm{A}}}{d_{\mathrm{L}}} = \frac{1}{(1+z)^2},$$

   independent of cosmology.