1

Some basic concepts and an overview of cosmology

In this chapter we present an elementary discussion of some basic concepts in cosmology. Although the mathematical formalism is essential, some of the main ideas underlying the formalism are simple and it helps to have an intuitive and qualitative notion of these ideas.

Cosmology is the study of the large-scale structure and behaviour of the universe, that is, of the universe taken as a whole. The term ‘as a whole’ applied to the universe needs a precise definition, which will emerge in the course of this book. It will be sufficient for the present to note that one of the points that has emerged from cosmological studies in the last few decades is that the universe is not simply a random collection of irregularly distributed matter, but it is a single entity, all parts of which are in some sense in unison with all other parts. This, at any rate, is the view taken in the ‘standard models’ which will be our main concern. We may have to modify these assertions when considering the inflationary models in a later chapter.

When considering the large-scale structure of the universe, the basic constituents can be taken to be galaxies, which are congregations of about 10^{11} stars bound together by their mutual gravitational attraction. Galaxies tend to occur in groups called clusters, each cluster containing anything from a few to a few thousand galaxies. There is some evidence for the existence of clusters of clusters, but not much evidence of clusters of clusters of clusters or higher hierarchies. ‘Superclusters’ and voids (empty regions) have received much attention (see Chapter 5). Observations indicate that on the average galaxies are spread uniformly throughout the universe at any given time. This means that if we consider a portion of the universe which is large compared to the distance between typical nearest galaxies (this is of the order of a million light years), then the number of galaxies in that portion is roughly the same as the number in another portion with the same volume at any given time. This proviso 'at any given time' about the uniform distribution of galaxies is important because, as we shall see, the universe is in a dynamic state and so the number of galaxies in any given volume changes with time. The distribution of galaxies also appears to be isotropic about us, that is, it is the same, on the average, in all directions from us. If we make the assumption that we do not occupy a special position amongst the galaxies, we conclude that the distribution of galaxies is isotropic about any galaxy. It can be shown that if the distribution of galaxies is isotropic about every galaxy, then it is necessarily true that galaxies are spread uniformly throughout the universe.

We adopt here a working definition of the universe as the totality of galaxies causally connected to the galaxies that we observe. We assume that observers in the furthest-known galaxies would see distributions of galaxies around them similar to ours, and the furthest galaxies in their field of vision in the opposite direction to us would have similar distributions of galaxies around them, and so on. The totality of galaxies connected in this manner could be defined to be the universe.

E. P. Hubble discovered around 1930 (see, for example, Hubble (1929, 1936)) that the distant galaxies are moving away from us. The velocity of recession follows Hubble's law, according to which the velocity is proportional to distance. This rule is approximate because it does not hold for galaxies which are very near nor for those which are very far, for the following reasons. In addition to the systematic motion of recession every galaxy has a component of random motion. For nearby galaxies this random motion may be comparable to the systematic motion of recession and so nearby galaxies do not obey Hubble's law. The very distant galaxies also show departures from Hubble's law partly because light from the very distant galaxies was emitted billions of years ago and the systematic motion of galaxies in those epochs may have been significantly different from that of the present epoch. In fact by studying the departure from Hubble's law of the very distant galaxies one can get useful information about the overall structure and evolution of the universe, as we shall see.

Hubble discovered the velocity of recession of distant galaxies by studying their red-shifts, which will be described quantitatively later. The red-shift can be caused by other processes than the velocity of recession of the source. For example, if light is emitted by a source in a strong gravitational field and received by an observer in a weak gravitational field, the observer will see a red-shift. However, it seems unlikely that the red-shift of distant galaxies is gravitational in origin; for one thing these red-shifts are rather large for them to be gravitational and, secondly, it is difficult to understand the systematic increase with faintness on the basis of a gravitational origin. Thus the present consensus is that the red-shift is due to velocity of recession, but an alternative explanation of at least a part of these red-shifts on the basis of either gravitation or some hitherto unknown physical process cannot be completely ruled out.

The universe, as we have seen, appears to be homogeneous and isotropic as far as we can detect. These properties lead us to make an assumption about the model universe that we shall be studying, called the Cosmological Principle. According to this principle the universe is homogeneous everywhere and isotropic about every point in it. This is really an extrapolation from observation. This assumption is very important, and it is remarkable that the universe seems to obey it. This principle asserts what we have mentioned before, that the universe is not a random collection of galaxies, but it is a single entity.

The Cosmological Principle simplifies considerably the study of the large-scale structure of the universe. It implies, amongst other things, that the distance between any two typical galaxies has a universal factor, the same for any pair of galaxies (we will derive this in detail later). Consider any two galaxies A and B which are taking part in the general motion of expansion of the universe. The distance between these galaxies can be written as $f_{AB}R$, where f_{AB} is independent of time and R is a function of time. The constant f_{AB} depends on the galaxies A and B. Similarly, the distance between galaxies C and D is $f_{CD}R$, where the constant f_{CD} depends on the galaxies C and D. Thus if the distance between A and B changes by a certain factor in a definite period of time then the distance between C and D also changes by the same factor in that period of time. The large-scale structure and behaviour of the universe can be described by the single function R of time. One of the major current problems of cosmology is to determine the exact form of $R(t)$. The function $R(t)$ is called the scale factor or the radius of the universe. The latter term is somewhat misleading because, as we shall see, the universe may be infinite in its spatial extent in which case it will not have a finite radius. However, in some models the universe has finite spatial extent, in which case R is related to the maximum distance between two points in the universe.

It is helpful to consider the analogy of a spherical balloon which is expanding and which is uniformly covered on its surface with dots. The dots can be considered to correspond to 'galaxies' in a two-dimensional universe. As the balloon expands, all dots move away from each other and from any given dot all dots appear to move away with speeds which at any given time are proportional to the distance (along the surface). Let the

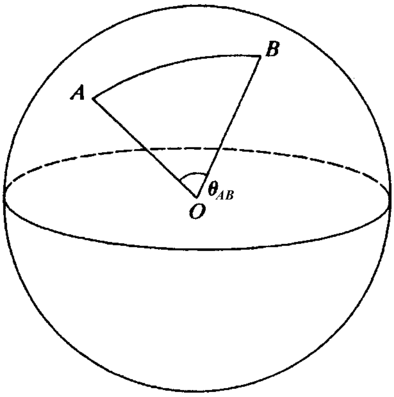

Fig. 1.1. Diagram to illustrate Equation (1.1).

radius of the balloon at time t be denoted by $R'(t)$. Consider two dots which subtend an angle θ_{AB} at the centre, the dots being denoted by A and B (Fig. 1.1). The distance d_{AB} between the dots on a great circle is given by

$$d_{AB} = \theta_{AB} R'(t). \tag{1.1}$$

The speed v_{AB} with which A and B are moving relative to each other is given by

$$v_{AB} = \dot{d}_{AB} = \theta_{AB}\dot{R}' = d_{AB}(\dot{R}'/R'), \quad \dot{R}' \equiv \frac{\mathrm{d}R'}{\mathrm{d}t}, \text{ etc.} \tag{1.2}$$

Thus the relative speed of A and B around a great circle is proportional to the distance around the great circle, the factor of proportionality being $\dot{R}'/R'$, which is the same for any pair of dots. The distance around a great circle between any pair of dots has the same form, for example, $\theta_{CD}R'$, where θ_{CD} is the angle subtended at the centre by dots C and D. Because the expansion of the balloon is uniform, the angles θ_{AB}, θ_{CD}, etc., remain the same for all t. We thus have a close analogy between the model of an expanding universe and the expansion of a uniformly dotted spherical balloon. In the case of galaxies Hubble's law is approximate but for dots on a balloon the corresponding relation is strictly true. From (1.1) it follows that if the distance between A and B changes by a certain factor in any period of time, the distance between *any* pair of dots changes by the same factor in that period of time.

From the rate at which galaxies are receding from each other, it can be deduced that *all* galaxies must have been very close to each other *at the same time* in the past. Considering again the analogy of the balloon, it is

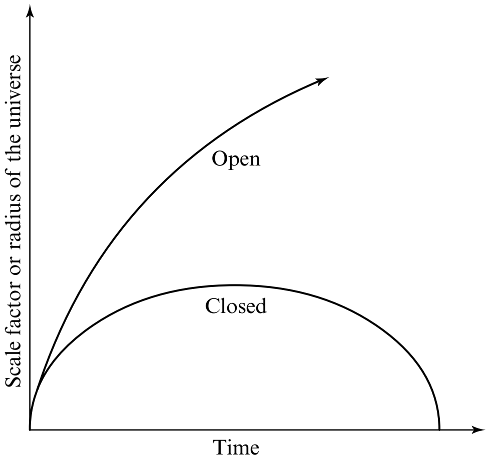

Fig. 1.2. Evolution of the scale factor or radius with time in the open and closed models of the universe.

like saying that the balloon must have started with zero radius and at this initial time all dots must have been on top of each other. For the universe it is believed that at this initial moment (some time between 10 and 20 billion years ago) there was a universal explosion, at every point of the universe, in which matter was thrown asunder violently. This was the ‘big bang’. The explosion could have been at every point of an infinite or a finite universe. In the latter case the universe would have started from zero volume. An infinite universe remains infinite in spatial extent all the time down to the initial moment; as in the case of the finite universe, the matter becomes more and more dense and hot as one traces the history of the universe to the initial moment, which is a ‘space-time singularity’ about which we will learn more later. The universe is expanding now because of the initial explosion. There is not necessarily any force propelling the galaxies apart, but their motion can be explained as a remnant of the initial impetus. The recession is slowing down because of the gravitational attraction of different parts of the universe to each other, at least in the simpler models. This is not necessarily true in models with a cosmological constant, as we shall see later.

The expansion of the universe may continue forever, as in the ‘open’ models, or the expansion may halt at some future time and contraction set in, as in the ‘closed’ models, in which case the universe will collapse at a finite time later into a space-time singularity with infinite or near infinite density. These possibilities are illustrated in Fig. 1.2. In the Friedmann models the open universes have infinite spatial extent whereas the closed

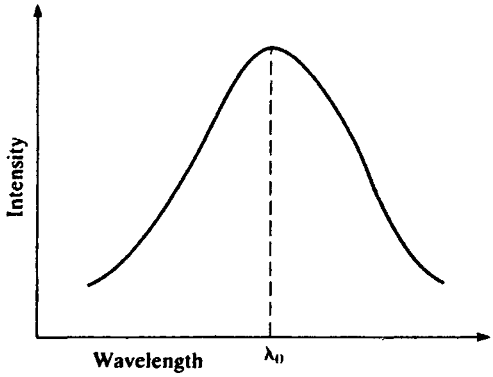

Fig. 1.3. Graph of intensity versus wavelength for black-body radiation. For the cosmic background radiation λ_0 is just under 0.1 cm.

models are finite. This is not necessarily the case for the Lemaître models. Both the Friedmann and Lemaître models will be discussed in detail in later chapters.

There is an important piece of evidence apart from the recession of the galaxies that the contents of the universe in the past must have been in a highly compressed form. This is the 'cosmic background radiation', which was discovered by Penzias and Wilson in 1965 and confirmed by many observations later. The existence of this radiation can be explained as follows. As we trace the history of the universe backwards to higher densities, at some stage galaxies could not have had a separate existence, but must have been merged together to form one great continuous mass. Due to the compression the temperature of the matter must have been very high. There is reason to believe, as we shall see, that there must also have been present a great deal of electromagnetic radiation, which at some stage was in equilibrium with the matter. The spectrum of the radiation would thus correspond to a black body of high temperature. There should be a remnant of this radiation, still with black-body spectrum, but corresponding to a much lower temperature. The cosmic background radiation discovered by Penzias, Wilson and others indeed does have a black-body spectrum (Fig. 1.3) with a temperature of about 2.7 K.

Hubble's law implies arbitrarily large velocities of the galaxies as the distance increases indefinitely. There is thus an apparent contradiction with special relativity which can be resolved as follows. The red-shift z is defined as $z = (\lambda_r - \lambda_i)/\lambda_i$, where λ_i is the original wavelength of the radiation given off by the galaxy and λ_r is the wavelength of this radiation when received

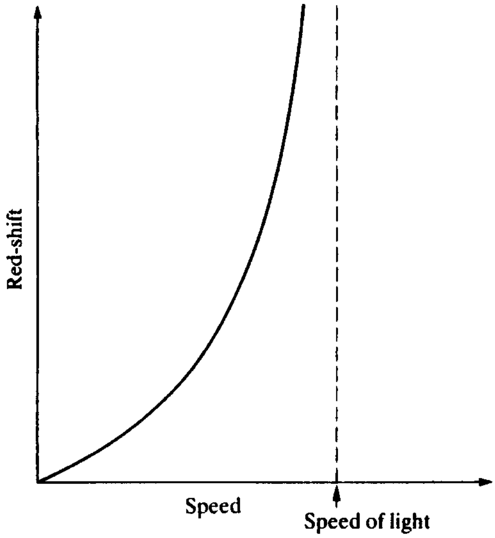

Fig. 1.4. This graph shows the relation between the red-shift (z) and the speed of recession. As z tends to infinity, the speed of recession tends to the speed of light.

by us. As the velocity of the galaxy approaches that of light, z tends towards infinity (Fig. 1.4), so it is not possible to *observe* higher velocities than that of light. The distance at which the red-shift of a galaxy becomes infinite is called the *horizon*. Galaxies beyond the horizon are indicated by Hubble's law to have higher velocities than light, but this does not violate special relativity because the presence of gravitation radically alters the nature of space and time according to general relativity. It is not as if a material particle is going past an observer at a velocity greater than that of light, but it is space which is in some sense expanding faster than the speed of light. This will become clear when we derive the expressions for the velocity, red-shift, etc., analytically later.

As mentioned earlier, in the open model the universe will expand forever whereas in the closed model there will be contraction and collapse in the future. It is not known at present whether the universe is open or closed. There are several interconnecting ways by which this could be determined. One way is to measure the present average density of the universe and compare it with a certain critical density. If the density is above the critical density, the attractive force of different parts of the universe towards each other will be enough to halt the recession eventually and to pull the galaxies together. If the density is below the critical density, the attractive force is insufficient and the expansion will continue forever. The critical density at any time (this will be derived in detail later) is given by

$$\varepsilon_c = 3H^2/8\pi G, \quad H = \dot{R}/R. \tag{1.3}$$

Here G is Newton's gravitational constant and R is the scale factor which is a function of time; it corresponds to $R'(t)$ of (1.1) and represents the 'size' of the universe in a sense which will become clear later. If t_0 denotes the present time, then the present value of H, denoted by H_0, is called Hubble's constant. That is, $H_0 = H(t_0)$. For galaxies which are not too near nor too far, the velocity v is related to the distance d by Hubble's constant:

$$v = H_0 d. \tag{1.4}$$

(Compare (1.2), (1.3) and (1.4).) The present value of the critical density is thus $3H_0^2/8\pi G$, and is dependent on the value of Hubble's constant. There are some uncertainties in the value of the latter, the likely value being between 50 km s^{-1} and 100 km s^{-1} per million parsecs. That is, a galaxy which is 100 million parsecs distant has a velocity away from us of 5000–10000 km s^{-1}. For a value of Hubble's constant given by 50 km s^{-1} per million parsecs, the critical density equals about 5×10^{-30} g cm^{-3}, or about three hydrogen atoms per thousand litres of space.

There are several other related ways of determining if the universe will expand forever. One of these is to measure the rate at which the expansion of the universe is slowing down. This is measured by the deceleration parameter, about which there are also uncertainties. Theoretically in the simpler models, in suitable units, the deceleration parameter is half the ratio of the actual density to the critical density. This ratio is usually denoted by Ω. Thus if $\Omega < 1$, the density is subcritical and the universe will expand forever, the opposite being the case if $\Omega > 1$. The present observed value of Ω is somewhere between 0.1 and 2 (the lower limit could be less). In the simpler models the deceleration parameter, usually denoted by q_0, is thus $\frac{1}{2}\Omega$, so that the universe expands forever in these models if $q_0 < \frac{1}{2}$, the opposite being the case if $q_0 > \frac{1}{2}$.

Another way to find out if the universe will expand forever is to determine the precise age of the universe and compare it with the 'Hubble time'. This is the time elapsed since the big bang until now if the rate of expansion had been the same as at present. In Fig. 1.5 if ON denotes the present time (t_0), then clearly PN is $R(t_0)$. If the tangent at P to the curve $R(t)$ meets the t-axis at T at an angle α, then

$$\tan \alpha = PN/NT = \dot{R}(t_0), \tag{1.5}$$

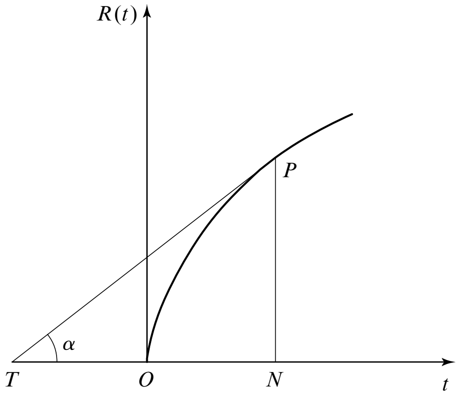

Fig. 1.5. Diagram to define Hubble time.

so that

$$NT = PN/\dot{R}(t_0) = R(t_0)/\dot{R}(t_0) = H_0^{-1}. \tag{1.6}$$

Thus NT, which is, in fact, Hubble's time, is the reciprocal of Hubble's constant in the units considered here. For the value of 50 km s^{-1} per million parsecs of Hubble's constant, the Hubble time is about 20 billion years. Again in the simpler models, if the universe is older than two-thirds of the Hubble time it will expand forever, the opposite being the case if its age is less than two-thirds of the Hubble time.

Whether the universe will expand forever is one of the most important unresolved problems in cosmology, both theoretically and observationally, but all the above methods of ascertaining this contain many uncertainties.

In this book we shall use the term 'open' to mean a model which expands forever, and 'closed' for the opposite. Sometimes the expression 'closed' is used to mean a universe with a finite volume, but, as mentioned earlier, it is only in the Friedmann models that a universe has infinite volume if it expands forever, etc.

The standard big-bang model of the universe has had three major successes. Firstly, it predicts that something like Hubble's law of expansion must hold for the universe. Secondly, it predicts the existence of the microwave background radiation. Thirdly, it predicts successfully the formation of light atomic nuclei from protons and neutrons a few minutes after the big bang. This prediction gives the correct abundance ratio for He^3, D, He^4 and Li^7. (We shall discuss this in detail later.) Heavier elements are thought to have been formed much later in the interior of stars. (See Hoyle, Burbidge and Narlikar (2000) for an alternative point of view.)

Certain problems and puzzles remain in the standard model. One of these is that the universe displays a remarkable degree of large-scale homogeneity. This is most evident in the microwave background radiation which is known to be uniform in temperature to about one part in 1000. (There is, however, a systematic variation of about one part in 3000 attributed to the motion of the Earth in the Galaxy and the motion of the Galaxy in the local group of galaxies, and also a smaller variation in all directions, presumably due to the 'graininess' that existed in the matter at the time the radiation 'decoupled'.) The uniformity that exists is a puzzle because, soon after the big bang, regions which were well separated could not have communicated with each other or known of each other's existence. Roughly speaking, at a time t after the big bang, light could have travelled only a distance ct since the big bang, so regions separated by a distance greater than ct at time t could not have influenced each other. The fact that microwave background radiation received from all directions is uniform implies that there is uniformity in regions whose separation must have been many times the distance ct (the *horizon distance*) a second or so after the big bang. How did these different regions manage to have the same density, etc.? Of course there is no problem if one simply *assumes* that the uniformity persists up to time $t = 0$, but this requires a very special set of initial conditions. This is known as the *horizon problem*.

Another problem is concerned with the fact that a certain amount of inhomogeneity must have existed in the primordial matter to account for the clumping of matter into galaxies and clusters of galaxies, etc., that we observe today. Any small inhomogeneity in the primordial matter rapidly grows into a large one with gravitational self-interaction. Thus one has to assume a considerable smoothness in the primordial matter to account for the inhomogeneity in the scale of galaxies at the present time. The problem becomes acute if one extrapolates to 10^{-45} s after the big bang, when one has to assume an unusual situation of almost perfect smoothness but not quite absolute smoothness in the initial state of matter. This is known as the *smoothness problem*.

A third problem of the standard big-bang model has to do with the present observed density of matter, which we have denoted by the parameter Ω. If Ω were initially equal to unity (this corresponds to a flat universe) it would stay equal to unity forever. On the other hand, if Ω were initially different from unity, its depature from unity would increase with time. The present value of Ω lies somewhere between 0.1 and 2. For this to be the