1

Modeling the early universe

Cosmology is the science of the universe treated as a whole. It is the science that tries to understand the origin (if there was one), the shape, and the evolution of the universe, including predictions for its future. We will start with a condensed overview of how the universe can be modeled on the largest scales accessible to observations, using general relativity. Given the scope of this book, we will put an emphasis on features that are relevant for understanding the early universe.

1.1 The standard cosmological model

1.1.1 The cosmological principle

The universe cannot be reproduced in the laboratory, and moreover we are severely limited in our ability to travel across the universe. In fact, all our knowledge about the universe must be inferred from the information that arrives here on Earth now. This, however, is not as limiting as it may sound, because of the finiteness of the speed of light. Light originating from distant places, but which we are observing now, shows us these distant places not as they are now, but as they were when the light was emitted. In this way, when looking further into the universe, we see further into the past, and this allows us to reconstruct the history of our universe (up to a point). Of course, we do not directly observe *our* past, but rather the past of increasingly distant regions. This strategy then allows us to reconstruct our history to the extent that the universe is evolving in the same way in different places, i.e., to the extent that the universe is spatially *homogeneous*.

Observations of the distribution of galaxies in the universe show that these are distributed along filaments, with huge voids of over 100 million light-years in diameter in between; see Fig. 1.1. These observations also

https://doi.org/10.1017/9781009516815.002 Published online by Cambridge University Press

show that on even bigger scales, namely bigger than about 300 million light-years, the universe becomes very similar in all directions. In other words, on these scales the universe becomes *isotropic*.

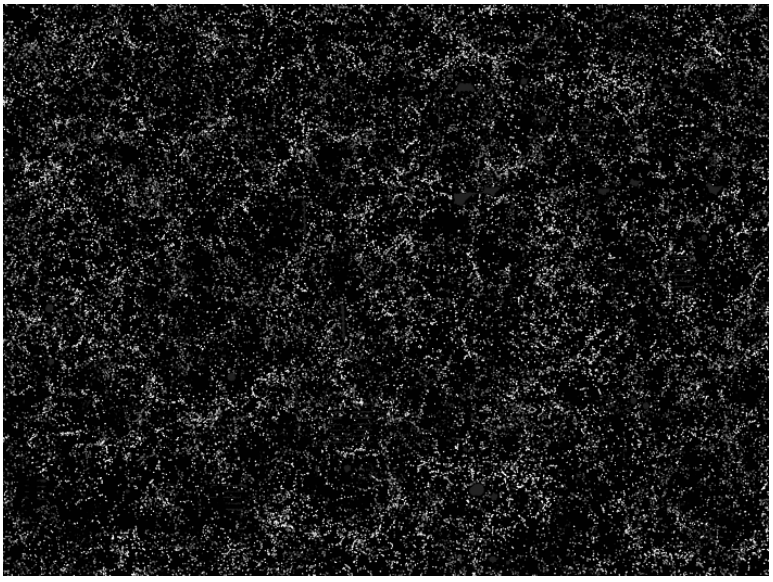


Figure 1.1 The distribution of galaxies in a slice of the universe 6 billion light-years away. The slice is 6 billion light-years wide, 4.5 billion light-years high, and half a billion light-years thick. Each dot represents the position of one galaxy, with brighter dots indicating a closer distance to us. Small gray patches indicate regions without data. One of the goals of cosmology is to understand the statistical properties of the distribution of galaxies in the universe. *Credit: Daniel Eisenstein and the SDSS-III collaboration.*

Thus, motivated by observations, we are led to postulate the *cosmological principle*: the universe is spatially *homogeneous* and *isotropic* on the largest observable scales. Some comments:

- If the cosmological principle were not at least approximately true, we would know very little about the distant universe. Only the realization that the universe appears to obey the same laws and have the same rough properties over large distances has allowed us to gain knowledge about the behavior of the distant universe.
- The cosmological principle certainly breaks down on small scales, but it may very well also break down on scales much larger than the currently observable part of the universe.
- The cosmological principle treats space and time on different footings. This is because the universe is only homogeneous and isotropic in a frame

https://doi.org/10.1017/9781009516815.002 Published online by Cambridge University Press

that is *comoving* with typical galaxies. In other words, in the universe there is a preferred frame, and the universe only appears isotropic to the extent that one is at rest with respect to that frame. A further subtlety is that the existence of a preferred frame does not imply that far-away galaxies must be approximately at rest with respect to each other – on the contrary, there is overwhelming evidence that the universe is expanding. The preferred frame can then be imagined to follow (and largely to be defined by) this expansion.

- Despite its name, the cosmological principle is simply an observed feature of our universe and does not stem from fundamental physical principles. It is one of the main aims of early universe cosmology to find an *explanation* for why the universe is so homogeneous and isotropic.

1.1.2 Robertson–Walker universes

We can use general relativity to describe the evolution of the universe, as general relativity is the (classical) dynamical theory of space, time, and matter. We will later incorporate quantum effects and see that they too likely played an important role in shaping the universe.

At first, we will assume exact spatial homogeneity and isotropy. This is of course an idealization, but the cosmological principle implies that this should be a good approximation in describing the universe on large scales. This assumption is very restrictive, and in fact one can prove that it allows for only three distinct spatial geometries, namely flat 3-dimensional space, a 3-sphere (with constant positive curvature), or a hyperbolic 3-space (with constant negative curvature). We denote the 3-dimensional line element by $\mathrm{d}l^2$. Then, using cartesian coordinates x, y, z, the line element of flat space is simply given by

$$\mathrm{d}l^2_{\mathrm{flat}} = \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2. \tag{1.1}$$

In polar coordinates, defined via $x = r\sin\theta\cos\phi, y = r\sin\theta\sin\phi, z = r\cos\theta$, with $r \geq 0, 0 \leq \theta \leq \pi, 0 \leq \phi < 2\pi$, the line element becomes

$$\mathrm{d}l^2_{\mathrm{flat}} = \mathrm{d}r^2 + r^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2). \tag{1.2}$$

We can obtain the metric for the 3-sphere by using the fact that it can be embedded in 4-dimensional flat space. If we denote the fourth spatial coordinate by w, then the line element for 4-dimensional Euclidean space is $\mathrm{d}l^2_4 = \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2 + \mathrm{d}w^2$, and within this space the unit 3-sphere is defined via the equation

$$x^2 + y^2 + z^2 + w^2 = 1. \tag{1.3}$$

https://doi.org/10.1017/9781009516815.002 Published online by Cambridge University Press

Differentiating, we learn that $x\,\mathrm{d}x + y\,\mathrm{d}y + z\,\mathrm{d}z + w\,\mathrm{d}w = 0$, or

$$\mathrm{d}w = \pm\frac{x\,\mathrm{d}x + y\,\mathrm{d}y + z\,\mathrm{d}z}{\sqrt{1 - x^2 - y^2 - z^2}}. \tag{1.4}$$

We can now obtain the metric for the 3-sphere by plugging this expression for $\mathrm{d}w$ into the 4-dimensional line element, thus restricting the line element to the 3-sphere. The result is

$$\mathrm{d}l^2_{\text{3-sphere}} = \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2 + \frac{(x\,\mathrm{d}x + y\,\mathrm{d}y + z\,\mathrm{d}z)^2}{1 - x^2 - y^2 - z^2}. \tag{1.5}$$

In a similar way one can obtain the metric of 3-dimensional hyperbolic space by embedding the hyperboloid $x^2 + y^2 + z^2 - w^2 = -1$ into 4-dimensional Minkowski space $\mathrm{d}l^2_4 = \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2 - \mathrm{d}w^2$. All three metrics can be conveniently summarized by (where Latin indices run over the spatial coordinates)

$$\mathrm{d}l^2_3 = \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2 + \frac{K(x\,\mathrm{d}x + y\,\mathrm{d}y + z\,\mathrm{d}z)^2}{1 - K(x^2 + y^2 + z^2)} \equiv g_{(3)ij}\mathrm{d}x^i\mathrm{d}x^j, \tag{1.6}$$

or, in polar coordinates,

$$\mathrm{d}l^2_3 = \frac{1}{1 - Kr^2}\mathrm{d}r^2 + r^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2), \tag{1.7}$$

where $K = 0, +1, -1$ for the flat, spherical, and hyperbolic spaces, respectively.[1]

Above, we derived the metric on the 3-sphere only for the case where it has unit radius. We can allow for an arbitrary, and in general time-dependent, size by multiplying the line element by a factor $a^2(t)$, and likewise for the hyperbolic space. In the case of flat space, the multiplicative factor $a^2(t)$ has no absolute meaning, and only relative values at different times are meaningful. The full 4-dimensional *Robertson–Walker* (sometimes *Friedmann–Lemaître–Robertson–Walker*) metric can then be obtained by adding a time direction, yielding

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + a^2(t)\left[\frac{1}{1 - Kr^2}\mathrm{d}r^2 + r^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)\right] \quad \text{(FLRW)}. \tag{1.8}$$

The function $a(t)$ is known as the *scale factor* of the universe. Its meaning can be clarified by using the metric above to calculate the spatial distance between an observer at the origin $r = 0$ and one at $r = R$ (at fixed time t,

[1] In the 2-dimensional analogs of these spaces, for a triangle we have that $\sum \text{angles} > \pi$ for $K = +1$, whereas $\sum \text{angles} = \pi$ for $K = 0$, and $\sum \text{angles} < \pi$ for $K = -1$.

https://doi.org/10.1017/9781009516815.002 Published online by Cambridge University Press

and with the angular coordinates also being fixed), which is given by

$$\int_{r=0}^{r=R} \mathrm{d}s = \int_{r=0}^{r=R} a(t) \frac{1}{\sqrt{1-r^2}} \mathrm{d}r = a(t) \begin{cases} \operatorname{arcsinh} R & (K=-1)\,, \\ R & (K=0)\,, \\ \arcsin R & (K=+1)\,. \end{cases} \tag{1.9}$$

Thus observers at fixed r, θ, ϕ only change their relative distance via the global, space-independent scale factor. In other words, the scale factor describes the overall expansion or contraction of the universe. The coordinates r, θ, ϕ are so-called *comoving* coordinates, and the result above implies that the proper distance between two comoving observers evolves as $a(t)$. Moreover, for a comoving observer, proper time is measured according to $\int \sqrt{-\mathrm{d}s^2} = \int \mathrm{d}t = \Delta t$, which shows that the time coordinate t is simply the *proper time* measured by comoving observers.

Sometimes, it is convenient to define a *conformal time* τ via $\mathrm{d}t = a\mathrm{d}\tau$, so that the metric has an overall factor of a^2:

$$\mathrm{d}s^2 = a^2(\tau) \left[-\mathrm{d}\tau^2 + \frac{1}{1-Kr^2}\mathrm{d}r^2 + r^2(\mathrm{d}\theta^2 + \sin^2\theta\, \mathrm{d}\phi^2) \right]. \tag{1.10}$$

The general form of the Robertson–Walker metric allows us to understand a crucial effect in cosmology, namely the shifting of spectral lines observed in the light coming from distant galaxies. When the scale factor is nonconstant, the frequency of light is in general different at the time of emission than at the time of observation. This can be seen as follows: we assume that we are observing (at time t_0) the light originating (at time t_R) from a galaxy located at $r = R$. Our own position can be chosen to be at $r = 0$. Since light rays propagate according to $\mathrm{d}s^2 = 0$, we obtain from Eq. (1.8) that

$$\mathrm{d}t = -a(t) \frac{\mathrm{d}r}{\sqrt{1-Kr^2}}, \tag{1.11}$$

where the minus sign was chosen because the light is coming towards us along decreasing values of r. Hence the times of emission and observation are related to the coordinate separation R by the first equality below,

$$\int_{t_0}^{t_R} \frac{\mathrm{d}t}{a(t)} = \int_0^R \frac{\mathrm{d}r}{\sqrt{1-Kr^2}} = \int_{t_0+\Delta t_0}^{t_R+\Delta t_R} \frac{\mathrm{d}t}{a(t)}\,. \tag{1.12}$$

Subsequent wave crests of the light wave are emitted at time intervals Δt, if the frequency of the light is $\nu = 1/\Delta t$. So now consider the emission of the next crest, at time $t_R + \Delta t_R$, and its absorption at $t_0 + \Delta t_0$. This is still equal to the same coordinate separation, hence the second equality above holds too. Subtracting the left- from the right-hand side, and working to

https://doi.org/10.1017/9781009516815.002 Published online by Cambridge University Press

leading order in Δt, one gets

$$\frac{\Delta t_R}{a(t_R)} = \frac{\Delta t_0}{a(t_0)}. \tag{1.13}$$

In other words, the frequency at observation is related to the frequency at emission via

$$\nu_0 = \nu_R \frac{a(t_R)}{a(t_0)}. \tag{1.14}$$

Equivalently, the wavelength at observation λ_0 is shifted from the wavelength at emission λ_R via

$$\lambda_0 = \lambda_R \frac{a(t_0)}{a(t_R)} \equiv \lambda_R(1+z)\,, \tag{1.15}$$

where z (not to be confused with the third spatial coordinate) is conventionally called the *redshift*. The wavelength of light simply changes according to the change of the space it lives in – see also Fig. 1.2. In an expanding universe, the wavelength at observation is longer than at emission, $z > 0$, and the light is indeed *redshifted*. In a contracting universe, light decreases in wavelength, $z < 0$, and in that case light is *blueshifted*.

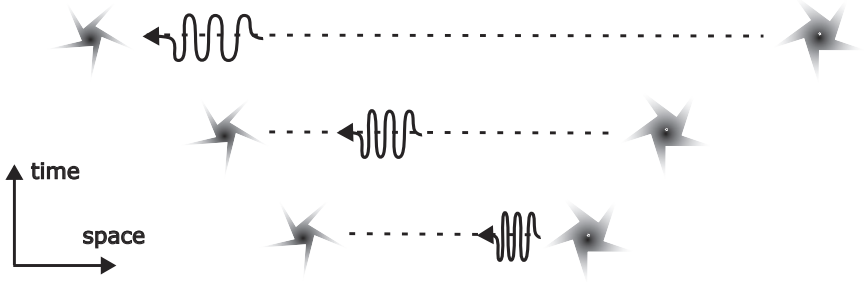


Figure 1.2 In an expanding universe, the space in between well-separated galaxies grows over time (though the space in the vicinity of the galaxies is not expanding). Light emitted from one galaxy is redshifted on the way because its wavelength grows in proportion to the expansion of space.

For nearby sources, we can expand $a(t) = a(t_0)[1 + H_0(t - t_0) + \cdots]$, where

$$H \equiv \frac{\dot{a}}{a}\,. \tag{1.16}$$

Here H is the *Hubble rate* and the subscript 0 customarily refers to a quantity evaluated at the present time. Then, keeping in mind that the speed of light is set to unity, we can identify $t - t_0$ with the proper distance D of the

https://doi.org/10.1017/9781009516815.002 Published online by Cambridge University Press

source, finding that to a first approximation the redshift increases linearly with distance,

$$z \approx H_0\, D \qquad \text{(Hubble)}\,. \tag{1.17}$$

This is the famous *Hubble law*. Astronomical observations, first carried out by Vesto Slipher in the 1920s and to the required accuracy by Edwin Hubble in the 1930s, show that the light from almost all galaxies is redshifted, and that the redshift–distance relation is indeed approximately linear for galaxies that are neither too close nor too far. This provides crucial evidence that our universe is expanding. Only the light from a few nearby galaxies (such as the Andromeda galaxy) is blueshifted, which can be understood as the result of the strong gravitational interactions between these galaxies and the Milky Way, causing them to have significant *peculiar velocities* superimposed on the comoving *Hubble flow*. And at redshifts larger than about $z \approx 0.1$, nonlinear terms in the expression for the Hubble parameter become important.

The observational value of the present-day Hubble rate is a matter of ongoing research and debate. Current observational values fall in the range

$$67\,\mathrm{km\,s^{-1}\,Mpc^{-1}} \lesssim H_0 \lesssim 74\,\mathrm{km\,s^{-1}\,Mpc^{-1}}\,, \tag{1.18}$$

where one parsec is $1\mathrm{pc} \approx 3.26$ light-years. Smaller values tend to arise in measurements involving the cosmic microwave background radiation (explained later), while larger values come from more local measurements, for instance using type Ia supernovae as *standard candles*, i.e., as sources of light with known luminosity (in this case explosions occurring after a critical amount of matter accretion onto a white dwarf). The best current measurements claim an uncertainty of around $1\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$. The discrepancy between different observations is referred to as the Hubble tension, and is currently unsolved.

We should stress that the Robertson–Walker metric is only intended as a zeroth-order approximation to the coarse-grained structure of the universe. In thinking about the expansion/contraction of space, it is important to keep in mind that near localized sources (such as stars, or even entire galaxies) the metric is locally not of Robertson–Walker form, but more closely approximates the Schwarzschild form

$$\mathrm{d}s^2 = \left(1 - \frac{2M}{r}\right)\mathrm{d}t^2 + \frac{\mathrm{d}r^2}{\left(1 - \frac{2M}{r}\right)} + r^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)\,, \tag{1.19}$$

where M is the mass of the source and r is a radial coordinate originating at the center of mass. In the Schwarzschild metric, when $r > 2M$ space is not expanding and, as a consequence, in our solar system space is neither

https://doi.org/10.1017/9781009516815.002 Published online by Cambridge University Press

expanding nor contracting. As an improved approximation, one could think of the entire universe as a kind of "Swiss cheese," with the cheese being well described by the Robertson–Walker metric, and the (small) holes by the Schwarzschild metric. This avoids some standard misconceptions about the expansion of space.

1.1.3 The Friedmann equations

General relativity specifies how spacetime is affected by the matter and energy content of the universe. More concretely, it relates the spacetime curvature, in the form of the Einstein tensor $G_{\mu\nu}$, to the matter content of the universe via its energy–momentum tensor $T_{\mu\nu}$, according to the Einstein equations

$$G_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu} \qquad \text{(Einstein)}\,, \tag{1.20}$$

where Greek indices run over all spacetime coordinates. We will mostly use natural units where $8\pi G = 1$, $c = 1$. In addition, the Einstein tensor obeys the Bianchi identity, from which it follows by consistency that the energy–momentum tensor is covariantly conserved,

$$\nabla^\mu G_{\mu\nu} = 0 \quad \text{(Bianchi)}\,, \qquad \nabla^\mu T_{\mu\nu} = 0 \quad \text{(conservation)}\,. \tag{1.21}$$

We will need to derive appropriate forms for both $G_{\mu\nu}$ and $T_{\mu\nu}$ in the cosmological context.

Let us start with the spacetime. Here the Einstein tensor is defined in terms of the Ricci curvature tensor $R_{\mu\nu}$ according to

$$G_{\mu\nu} \equiv R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R\,. \tag{1.22}$$

The cosmological principle led to the Robertson–Walker form of the metric, Eq. (1.8). In Euclidean coordinates this metric can also be written as

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + a^2(t) g_{(3)ij} \mathrm{d}x^i \mathrm{d}x^j, \tag{1.23}$$

where the spatial 3-metric $g_{(3)}$ was defined in Eq. (1.6). It is a straightforward exercise to derive the associated connections and curvature tensors,

$$\Gamma^0_{ij} = \tfrac{\dot{a}}{a} g_{ij} = a\dot{a} g_{(3)ij}\,, \tag{1.24}$$

$$\Gamma^k_{0j} = \tfrac{\dot{a}}{a} \delta^k_j\,, \tag{1.25}$$

$$\Gamma^k_{ij} = \Gamma^k_{(3)ij} = K g_{(3)ij} x^k\,, \tag{1.26}$$

$$\Gamma^0_{00} = \Gamma^0_{0j} = \Gamma^k_{00} = 0\,, \tag{1.27}$$

https://doi.org/10.1017/9781009516815.002 Published online by Cambridge University Press

$$R_{00} = -3\frac{\ddot{a}}{a}\,, \tag{1.28}$$

$$R_{ij} = (\frac{\ddot{a}}{a} + 2\frac{\dot{a}^2}{a^2} + 2\frac{K}{a^2})g_{ij} = (a\ddot{a} + 2\dot{a}^2 + 2K)g_{(3)ij}\,, \tag{1.29}$$

$$R_{0j} = 0\,, \tag{1.30}$$

$$R = 6(\frac{\ddot{a}}{a} + \frac{\dot{a}^2}{a^2} + \frac{K}{a^2})\,. \tag{1.31}$$

The assumptions of spatial isotropy and homogeneity also lead to strong restrictions on the form of the energy–momentum tensor on large scales. Isotropy, i.e., invariance under spatial rotations, implies that (at a comoving point) the spatial part T_{ij} of the energy–momentum tensor is proportional to the metric g_{ij}, and that the mixed, vectorial component T_{0i} vanishes (otherwise a particular direction would be singled out). Homogeneity means that we must impose these conditions everywhere in space, implying that the function of proportionality in T_{ij} can depend on time alone (otherwise, different spatial locations would not be equivalent). Likewise, T_{00} can depend on time alone. The conventional definitions are

$$T_{00} \equiv \rho(t)\,, \qquad T_{0i} = 0\,, \qquad T_{ij} \equiv p(t)g_{ij}\,, \tag{1.32}$$

where $\rho(t)$ is the proper *energy density* and $p(t)$ the *pressure*. This particular form of the energy–momentum tensor, imposed on us by the assumptions of isotropy and homogeneity, corresponds to that of an ideal fluid. The energy density and the pressure are related to each other via the *equation of state*

$$p = w\rho\,. \tag{1.33}$$

In many cases of interest, w is constant, and we will assume this to be the case unless otherwise noted.

What are the consequences of the covariant conservation $T^{\mu\nu}{}_{;\nu} \equiv \nabla_\nu T^{\mu\nu} = 0$ of the energy–momentum tensor? The $\mu = i$ component of this equation corresponds to momentum conservation, and it is automatically satisfied in a Robertson–Walker background. The $\mu = 0$ component leads to

$$\dot{\rho} + 3H(\rho + p) = 0\,. \tag{1.34}$$

This equation is known as the *equation of continuity*. Note that it shows that in general, in a dynamical universe, the energy density is *not* conserved. For a constant equation of state, the equation of continuity can be integrated to yield

$$\rho \propto \frac{1}{a^{3(1+w)}}. \tag{1.35}$$

Thus, if we know the equation of state of a certain matter component, then we know how its energy density (and thus its relative importance) scales

https://doi.org/10.1017/9781009516815.002 Published online by Cambridge University Press

as the universe contracts or expands. The most commonly considered ideal fluids are

- Pressure-free matter/dust, $w = 0$. Ordinary baryonic matter as well as dark matter fall into this category. Their energy density simply scales inversely to the volume of a given region of space, $\rho \propto a^{-3}$, as expected.
- Relativistic particles/radiation, $w = \frac{1}{3}$. The energy density of radiation or a gas of relativistic particles scales as a^{-4}. This means that in an expanding universe, the energy density of radiation falls off faster than the volume of space increases. Heuristically one can understand the exponent -4 as follows: The number of photons in a given co-moving volume scales with the volume, as a^{-3}. However, the wavelength of a photon also scales with the scale factor (this is the redshift), and thus its frequency/energy scales as a^{-1}. This additional factor of a^{-1} then leads to the overall scaling as a^{-4}.
- Cosmological constant, $w = -1$. The energy density of a cosmological constant is, as the name suggests, constant over time and unaffected by cosmic expansion/contraction. For this reason it is also often described as vacuum energy. Note that $p = -\rho$ implies that the energy–momentum tensor is proportional to the metric $T_{\mu\nu} \propto g_{\mu\nu}$, which is consistent with the Lorentz invariance of the vacuum. Another name that is used to describe a cosmological constant, or in fact any type of matter with an equation of state that is close to $w = -1$, is *dark energy*. While observations suggest that dark energy currently provides the dominant contribution to the overall energy density of the universe, this type of matter remains by far the least understood.

There is one important omission in this list: scalar fields, which are often used in modeling the dynamics of the universe. We will treat these in detail in Chapter 2.

We are finally ready to derive the *Friedmann equations*, which arise by inserting the homogeneous and isotropic expressions for curvature and energy–momentum into the Einstein equations. Using the expressions just derived, one gets

$$H^2 + \frac{K}{a^2} = \tfrac{1}{3}\rho \qquad \text{(Friedmann)}\,, \tag{1.36}$$

$$\frac{\ddot{a}}{a} = -\tfrac{1}{6}(\rho + 3p) = -\tfrac{1}{6}\rho(1 + 3w) \quad \text{(acceleration)}\,. \tag{1.37}$$

The first of these equations is sometimes referred to as simply the *Friedmann equation*, while the second one is often called the *acceleration equation*. These

https://doi.org/10.1017/9781009516815.002 Published online by Cambridge University Press

equations determine the evolution of the universe on the largest observable scales.

One should think of ρ and p above as denoting the total energy density and pressure – these then consist of a sum over all the individual matter types that are present in the universe. Before considering a mixture of ingredients, though, it is useful to look at a few solutions to these equations when only a single matter component is present. We will consider the examples of matter types enumerated above, for the case of a spatially flat universe ($K = 0$):

- Pressure-free matter/dust, $w = 0$. The Friedmann equation then implies that the scale factor evolves as $a(t) \propto t^{\frac{2}{3}}$, while the acceleration equation shows that the universe is decelerating. (Including integration constants, the solution can be written as $a(t) = \frac{(t-t_\star)^{2/3}}{(t_0-t_\star)^{2/3}}$, where t_0 is the time today, and $t_\star$ the time of the big bang, with $a_0 = 1$.)
- Relativistic particles/radiation, $w = \frac{1}{3}$. In this case the scale factor evolves as $a(t) \propto t^{\frac{1}{2}}$, and this also corresponds to decelerated expansion.
- Cosmological constant, $w = -1$. The Friedmann Eq. (1.36) implies that $\rho > 0$ since $K = 0$. In this case the scale factor evolves exponentially, $a(t) \propto e^{Ht}$, and the Hubble parameter is constant. This solution is known as the *de Sitter* universe. According to Eq. (1.37), in the de Sitter solution the expansion is accelerating. Above, we mentioned that matter with equation of state close to -1 is also sometimes called dark energy. Here, we can refine that definition. The acceleration equation implies that, for a positive energy density, the universe is accelerating whenever $w < -\frac{1}{3}$. We have presently entered such an era of accelerated expansion.

The solutions for $K = 0$ are those that are used most often in cosmology, as current observations indicate that our universe is very close to spatially flat. Indeed, whether the universe has positive or negative spatial curvature can be deduced by comparing its density to the *critical density*, cf. Eq. (1.36),

$$\rho_{\text{crit}} = 3H^2 \,. \tag{1.38}$$

And observations show that our universe's energy density is very close to the critical density, which is given by

$$\rho_{\text{crit},0} = 3H_0^2 \approx \frac{3}{8\pi G}(70\,\text{km}\,\text{s}^{-1}\,\text{Mpc}^{-1})^2 \approx 10^{-26}\,\text{kg}\,\text{m}^{-3}\,, \tag{1.39}$$

which amounts to only about one hydrogen atom per cubic meter of volume. As a side remark, let us stress that even spatially flat universes ($K = 0$) are curved in a 4-dimensional sense – their curvature arises from the time evolution of the scale factor.

https://doi.org/10.1017/9781009516815.002 Published online by Cambridge University Press

But even in a universe that appears spatially flat on large scales, the fact that the matter in the universe is distributed inhomogeneously on small scales indicates that on those scales the average curvature must necessarily be positive or negative. Hence it is certainly also important to look at solutions with $K = \pm 1$. The corresponding solutions are more easily found using conformal time τ. In terms of conformal time, the Friedmann equations read

$$a'^2 + Ka^2 = \frac{1}{3}\rho a^4 \,, \tag{1.40}$$

$$a'' + Ka = \frac{1}{6}(\rho - 3p)a^3 \,, \tag{1.41}$$

where a prime denotes a derivative with respect to conformal time.

For $K = +1$, the spatial sections of the universe are spheres, and this is known as the *closed* model of the universe. The Friedmann Eq. (1.40) shows that the expansion can come to a halt, even when the energy density is positive. After such a halt, the universe re-collapses, as can be seen in the solutions obtained in the presence of dark matter or radiation

$$a(\tau) = \begin{cases} 1 - \cos\tau & w = 0 \quad 0 < \tau < 2\pi \,, \\ \sin\tau & w = \frac{1}{3} \quad 0 < \tau < \pi. \end{cases} \tag{1.42}$$

These solutions are used in describing gravitational collapse of overdense regions.

For $K = -1$, the universe is called *open*, and the corresponding solutions are

$$a(\tau) = \begin{cases} \cosh\tau - 1 & w = 0 \quad 0 < \tau < \infty \,, \\ \sinh\tau & w = \frac{1}{3} \quad 0 < \tau < \infty. \end{cases} \tag{1.43}$$

For those solutions, the universe expands forever. Open-universe solutions can serve as useful approximations to the void regions of the universe.

Before proceeding with a more quantitative analysis of the Friedmann equations, let us highlight their main philosophical consequence. They imply that the universe is not static. Space and time are evolving quantities, even on the largest observable scales. This marks a major shift in our thinking about nature: up until about 100 years ago, people were mostly convinced (on philosophical or religious grounds) that the universe on the whole had to be static, immutable, and infinite. The discovery of the expansion of the universe, going hand in hand with the elaboration and vindication of general relativity, has brought with it a major conceptual paradigm shift, as the universe is now seen as an evolving object in itself. The universe has a history, which cosmologists are trying to unravel.

https://doi.org/10.1017/9781009516815.002 Published online by Cambridge University Press

1.2 A thermal history

To obtain a realistic large-scale model of the universe, we must include all of the relevant matter/energy constituents. For the standard, so-called ΛCDM model, these are radiation (mostly photons, but can also include massive particles moving at relativistic speeds), pressure-free matter (cold dark matter and ordinary baryons), and a cosmological constant Λ. We should note that there are good theoretical reasons for why dark energy might not be a simple cosmological constant, but current observations are still compatible with a cosmological constant and hence, for simplicity, we will stick with Λ for now. The Friedmann equation (1.36) then reads

$$3H^2 = \frac{\rho_{\mathrm{r},0}}{a^4} + \frac{\rho_{\mathrm{m},0}}{a^3} - \frac{3K}{a^2} + \Lambda\,. \tag{1.44}$$

If we choose units in which the current scale factor is $a_0 = 1$, then $\rho_{\mathrm{r},0}$ and $\rho_{\mathrm{m},0}$ denote the current energy densities in radiation and pressure-free matter.

As the universe expands, it is clear that matter types that scale with a less negative power of the scale factor eventually dominate over those that scale with more negative powers of a. This leads to a series of epochs, during each of which a different energy component dominates and drives the dynamics. A useful way to rewrite Eq. (1.44) is obtained by dividing by the current critical density $3H_0^2$, resulting in the two equivalent expressions, either in terms of scale factor a or redshift z,

$$\left(\frac{H}{H_0}\right)^2 = \frac{\Omega_{\mathrm{r}}}{a^4} + \frac{\Omega_{\mathrm{m}}}{a^3} + \frac{\Omega_{\mathrm{K}}}{a^2} + \Omega_\Lambda \tag{1.45}$$

$$= \Omega_{\mathrm{r}}(1+z)^4 + \Omega_{\mathrm{m}}(1+z)^3 + \Omega_{\mathrm{K}}(1+z)^2 + \Omega_\Lambda\,, \tag{1.46}$$

where the current fractional energy densities are conventionally defined as

$$\Omega_{\mathrm{r}} = \frac{\rho_{\mathrm{r},0}}{3H_0^2}, \quad \Omega_{\mathrm{m}} = \frac{\rho_{\mathrm{m},0}}{3H_0^2}, \quad \Omega_{\mathrm{K}} = \frac{-K}{H_0^2}, \quad \Omega_\Lambda = \frac{\Lambda}{3H_0^2}\,. \tag{1.47}$$

By definition these sum to unity,

$$\Omega_{\mathrm{r}} + \Omega_{\mathrm{m}} + \Omega_{\mathrm{K}} + \Omega_\Lambda = 1\,. \tag{1.48}$$

It is useful to highlight how theory and observations are checked against each other in practice. Unlike for most laboratory experiments, it is not the case here that the theory makes a number of predictions, then an experiment is set up in order to check them. Rather, in a work of constant refinement, theory and observations are matched to each other. It is in this way that, on the one hand, the current and past abundances of various matter types

https://doi.org/10.1017/9781009516815.002 Published online by Cambridge University Press

are determined, and on the other hand the Einstein/Friedmann equations are verified. Consider for instance Fig. 1.3, which illustrates this process. The figure contains measurements of the apparent magnitude (brightness) of type Ia supernovae, whose intrinsic brightness is well-enough understood to act as reference light sources, at many redshifts. The apparent magnitude however depends on the expansion history $H(z)$ of the universe. And this history can then be fitted from those arising by assuming general relativity and various matter contributions, as exemplified in the figure.

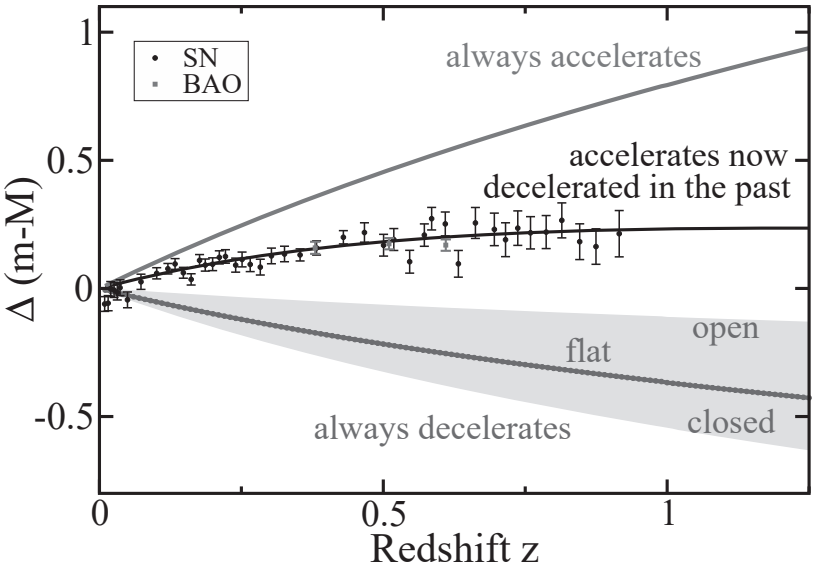


Figure 1.3 A plot of the change in apparent magnitude of cosmological sources (type Ia supernovae from Scolnic et al. (2015) in black, and baryon acoustic oscillations from BOSS Collaboration (2017) in gray) versus redshift. This figure does not contain the most up-to-date data, but illustrates well how the underlying cosmological model is fitted to the data. The reference model is an open empty universe with $\Omega_K = 1$. Models without dark energy ($\Omega_\Lambda = 0$) and various amounts of matter ($0.3 \leq \Omega_m \leq 1.5$) are shown in the shaded area, and are seen to be ruled out. A pure vacuum energy dominated universe with $\Omega_\Lambda = 1$ is shown by the top curve and is also ruled out. Favored is a model with $\Omega_m = 0.3$ and $\Omega_\Lambda = 0.7$, depicted here by the black middle line. Figure reproduced from Huterer and Shafer (2018). c *IOP Publishing. Reproduced with permission. All rights reserved.*

We should stress that this particular measurement is just one of many different probes of both the expansion history and the matter content of the universe. These measurements are remarkably consistent with each other (they tend to agree on the level of a few percent), and they lead to the

https://doi.org/10.1017/9781009516815.002 Published online by Cambridge University Press

following current composition (inferred here from 2018 Planck satellite measurements combined with baryon acoustic oscillation (BAO) measurements, with 1σ errors – see Planck Collaboration; Aghanim et al. (2020a)):

$$\Omega_\Lambda = 0.689 \pm 0.006\,, \quad H_0 \equiv 100 \cdot h = (67.7 \pm 0.4)\,\mathrm{km\,s^{-1}\,Mpc^{-1}}\,,$$
$$\Omega_\mathrm{m} = 0.311 \pm 0.006\,, \quad \Omega_\mathrm{r} = (9.18 \pm 0.17) \times 10^{-5}\,. \tag{1.49}$$

In this fit, it was assumed that the universe is spatially flat, $\Omega_\mathrm{K} = 0$. When the spatial curvature is allowed to vary, one finds what amounts to an upper bound,

$$\Omega_\mathrm{K} = 0.000\,7 \pm 0.001\,9\,. \tag{1.50}$$

Let us take stock then. The main ingredient in today's universe is dark energy, causing the universe to accelerate its expansion. In fact, from Eq. (1.37) we can easily determine the redshift at which acceleration began, when $\ddot{a} = 0$, giving

$$z_\mathrm{acc} \approx 0.64\,. \tag{1.51}$$

This was when the universe had about 61% of its current (linear) size, and was about 7.6 billion years old (we will see below how to calculate ages).

Before the dark energy phase, there was a phase of matter domination, during which the expansion decelerated. For completeness we should point out that out of the current pressure-free matter only about a fifth is thought to be made of ordinary baryons, making up gas and dust clouds, stars, planets, and us. The rest is dark matter, whose effects are seen via gravity (e.g., due to gravitational lensing of light emitted by distant galaxies, and bent by intervening dark matter), but which has not been detected directly. Understanding dark energy and dark matter better are major goals of cosmology, but they have affected the very early universe less strongly, since the matter phase was preceded by a radiation-dominated phase.

The switch between matter and radiation domination occurred at the so-called time of equality, when the size of the universe was

$$a_\mathrm{eq} = \frac{1}{1 + z_\mathrm{eq}} = \frac{\Omega_\mathrm{r}}{\Omega_\mathrm{m}} \approx \frac{1}{3\,388}\,, \tag{1.52}$$

compared to its current size, and its age was about 51 000 years. The radiation density is composed mostly of the cosmic microwave background (discussed below), but also contains contributions from neutrinos, which are very light particles and hence behave relativistically. The main physical characteristic during the radiation phase is the temperature. Recall that photons

https://doi.org/10.1017/9781009516815.002 Published online by Cambridge University Press

have energy $E = h\nu = hc/\lambda = k_B\mathcal{T}$, where ν is the frequency, λ the wavelength, k_B Boltzmann's constant, and $\mathcal{T}$ the absolute temperature, where we have momentarily restored units. Thus the wavelength can be expressed as $\lambda = hc/(k_B\mathcal{T})$. As the universe expands or contracts, the wavelength changes with the scale factor $\lambda \propto a(t)$, and thus the temperature evolves as

$$\mathcal{T}(t) = \mathcal{T}_0 \frac{a_0}{a(t)}. \tag{1.53}$$

This relation implies that the universe was *hotter* in the past, hence the name *hot big bang model*. In the very early universe, photons and other matter particles were tightly coupled, implying that in fact all matter constituents were hot. For this epoch, there exists a useful relation between temperature and time; the Friedmann equation implies that

$$H^2 = \frac{1}{4t^2} = \frac{1}{3}\rho \approx \mathcal{T}^4, \tag{1.54}$$

where we have neglected a factor of order unity relating the energy density in radiation to the fourth power of temperature. The relation above, namely $\mathcal{T} \approx 1/\sqrt{t}$, holds in natural units, and if we restore units, we get the formula

$$\mathcal{T}_{\mathrm{MeV}} \approx \frac{1}{\sqrt{t_{\mathrm{sec}}}}, \tag{1.55}$$

where the temperature is expressed in MeV and the time in seconds (counting from a putative big bang). For completeness we should mention that, going to earlier times and hence hotter temperatures, more and more massive particles start behaving relativistically and hence the number of species contributing to the radiation density goes up as we approach the big bang.

In fact we are now in a position to calculate the age of the universe or, more precisely, the time that has elapsed since the big bang. The Friedmann Eq. (1.45) can be recast as the differential

$$\mathrm{d}t = \frac{\mathrm{d}a}{H_0\sqrt{\Omega_{\mathrm{r}}a^{-2} + \Omega_{\mathrm{m}}a^{-1} + \Omega_{\mathrm{K}} + \Omega_\Lambda a^2}}. \tag{1.56}$$

Integrating from $a = 0$ to $a = a_0 = 1$ gives us the time since the big bang, namely

$$t_{\mathrm{BB}} \approx 0.954\frac{1}{H_0} \approx 13.8 \times 10^9\,\mathrm{years} \qquad (\text{age of universe}). \tag{1.57}$$

This may be compared with the age of the solar system, which formed about 4.65 billion years ago. Integrating (1.56) up to the scale factors a_{eq} and a_{acc} gives the ages mentioned above. Note that, for quick order-of-magnitude estimates, it is useful to memorize that the time since the big bang and the

https://doi.org/10.1017/9781009516815.002 Published online by Cambridge University Press

size of the observable universe (on the order of 10 billion years and 10 billion light-years) are approximately given by 10^{60} Planck times and 10^{60} Planck lengths, respectively.

The high temperatures in the early universe had a marked effect on the behavior of matter. Switching from our backwards-looking perspective to now counting the time forward from the (putative) origin at $t = 0, a = 0$, we may list some of the major events in the history of the universe:

$t < 10^{-14}\,\mathrm{s}$, $\mathcal{T} > \mathrm{TeV}$

We still know very little with certainty about this epoch, as the corresponding particle physics currently cannot be reproduced in accelerator experiments. From our current laws of physics, we can assume that the electroweak symmetry was unbroken during this early phase, and in fact the gauge group describing all matter interactions could have been much bigger (such as $SO(10)$, allowing a grand unified description of particle physics). Also, supersymmetry may have played a role at those early times, but it remains too early to tell. Much of the rest of this book will be about this epoch.

$10^{-14}\,\mathrm{s} < t < 10^{-10}\,\mathrm{s}$, $100\,\mathrm{GeV} < \mathcal{T} < 10\,\mathrm{TeV}$

During this time, the electroweak symmetry $SU(2) \times U(1)$ was broken, and the W, Z bosons became massive. The particle physics that took place at these energies is currently being probed by the Large Hadron Collider at CERN, and is, to a large extent, well understood.

$t \sim 10^{-5}\,\mathrm{s}$, $\mathcal{T} \sim 200\,\mathrm{MeV}$

As the universe cooled, free quarks and gluons bound into baryons (particles with three quarks, such as protons and neutrons) and mesons (particles with a quark and an antiquark).

$t \sim 0.2\,\mathrm{s}$, $\mathcal{T} \sim 1 - 2\,\mathrm{MeV}$

The cross sections of the weak interactions became very small as the temperature dropped to these scales, and thus the neutrinos decoupled from the rest of the matter in the universe. This neutrino background evolved independently from that time on. In the long-term future, one may hope that it will be possible to build a neutrino telescope (perhaps on the far side of the Moon) that can measure these primordial neutrinos directly. This would give us direct information about the physical conditions at these times close to the big bang.

As a byproduct, the ratio of neutrons to protons also *froze out* around this time. Up until this time, protons could easily convert into neutrons,

https://doi.org/10.1017/9781009516815.002 Published online by Cambridge University Press

and vice versa, via the exchange of neutrinos, electrons, and positrons,

$$p + e^- \leftrightarrow n + \nu, \qquad p + \bar{\nu} \leftrightarrow n + e^+. \tag{1.58}$$

The neutron is heavier than the proton by an amount $Q = m_n - m_p \approx 1.3\,\mathrm{MeV}$, where $m_{n,p}$ denote the masses of a neutron and a proton, respectively. Standard thermodynamics then implies that, as long as the two species are still in equilibrium, i.e., as long as the reactions above are efficient, the number densities of neutrons (n_n) and protons (n_p) are related by

$$\frac{n_n}{n_p} \approx e^{-Q/\mathcal{T}}. \tag{1.59}$$

The reactions above require an energy $m_n - m_p - m_e \approx 1.3 - 0.5 \approx 0.8\,\mathrm{MeV}$ to be efficient, where m_e denotes the mass of an electron. Below this temperature, the ratio of neutrons to protons gets frozen in, with $n_n/n_p \approx e^{-1.3/0.8} \approx 1/5$. This has important implications for big bang *nucleosynthesis*, as we will see shortly.

$t \sim 1\,\mathrm{s}$, $\mathcal{T} \sim 0.5\,\mathrm{MeV}$

When the temperature drops to the value of the rest mass of electrons and positrons, these start annihilating each other into photons, while the converse process, spontaneous creation of electron–positron pairs out of photons, becomes rare. The photons that are produced in this way are in thermal equilibrium, and their temperature is slightly higher than that of the neutrinos, which already decoupled earlier. The end result of electron–positron annihilation is that about one electron was left over for each billion photons. Thus there must have been a very slight excess of matter over antimatter in the early universe. This asymmetry remains largely unexplained.

$t < 5\,\mathrm{min}$, $\mathcal{T} \sim 0.05\,\mathrm{MeV}$

Above, we saw that the ratio of neutrons to protons froze out at a value of about 1/5. However, neutrons are only stable when they are locked inside an atomic nucleus. Free neutrons decay into a proton, an electron, and a neutrino with a half life of about 15 min. Over the next few minutes, these decays reduced the neutron-to-proton number density fraction to

$$\frac{n_n}{n_p} \approx \frac{1}{7}. \tag{1.60}$$

Meanwhile, neutrons can combine with protons to form deuterons

$$p + n \to d + \gamma, \tag{1.61}$$

https://doi.org/10.1017/9781009516815.002 Published online by Cambridge University Press

while, in turn, deuterons can combine via the following chain to end up as helium nuclei

$$d + d \to {}^3He + n, \qquad {}^3He + d \to {}^4He + p. \tag{1.62}$$

Assuming that all neutrons end up in helium atoms (which is in fact a good approximation), we would deduce that the total mass fraction in helium should be

$$Y = \frac{n_{He} m_{He}}{n_n m_n + n_p m_p} \approx \frac{(\frac{1}{2} n_n)(4 m_n)}{(n_n + n_p) m_n} = \frac{2 n_n / n_p}{1 + n_n / n_p} \approx \frac{1}{4}. \tag{1.63}$$

Helium is also produced via nuclear fusion in stars. However, looking at stars that are further and further away (and thus looking further into the past) observations show that helium levels decrease and level off around a value of 1/4. Measurements of other light elements, such as deuterium and lithium, are also in reasonably good agreement with theoretical estimates of their production via primordial nucleosynthesis. This is striking evidence that the universe was once hot and dense, and provides strong support for the hot big bang model.

$t \approx 10^5$ years, $\mathcal{T} \sim$ eV

After a time of about 51 000 years (determined by integrating (1.56) up to $a_{\rm eq} = 1/(1 + z_{\rm eq})$), matter came to dominate over the radiation energy density, and the universe started expanding according to $a \propto t^{2/3}$. By the time the universe grew by a further factor of 3 in each direction, the helium nuclei and protons combined with the electrons to form helium and hydrogen atoms. This event is known as *recombination* and will be discussed in more detail below. The name recombination is somewhat misleading, as this is the first time after the big bang that atoms formed.

$t \approx 10^9$ years, $\mathcal{T} \sim$ meV

As we will discuss, small fluctuations in the distribution of matter and radiation were present already at early times in the history of the universe. Under the influence of gravity, small initial overdensities grew and eventually collapsed into stars and galaxies.

About 8 billion years after the big bang, the dark energy came to dominate the energy density in the universe, and the expansion started accelerating (at $z = z_{\rm acc}$). The future dynamics of the universe depends crucially on the nature of dark energy. We will speculate about possible futures at various stages in this book.

https://doi.org/10.1017/9781009516815.002 Published online by Cambridge University Press

1.2.1 A relic from the big bang: the cosmic microwave background

At the time of radiation–matter equality, the universe can still be described as a mixture of mostly photons, electrons, protons, and helium ions. Via Thomson scattering, the photons constantly bounce off the free electrons (the scattering off protons occurs much more rarely due to the significantly larger mass of the proton), which means that they cannot travel far in a straight line and hence the universe is not transparent at that time – the term *primordial soup* is sometimes used to describe this state of affairs.

As the universe cools, at some point it becomes energetically favorable for the electrons to bind with the free protons and helium ions to form hydrogen and helium atoms, respectively. In fact, because helium has a larger ionization potential, helium atoms form first. Afterwards, there are still many free electrons left over, and these can combine with protons to form neutral hydrogen atoms. Since the binding energy of hydrogen is about 13.6 eV, a first guess is that hydrogen production ("recombination") will occur when the universe has reached that temperature. However, it actually occurs a little later, when the universe has cooled off a little more. This is because there are so many more photons than electrons (about a billion times as many), so that a small fraction of atypically energetic photons is enough to keep hydrogen ionized. In the tail of the Planck distribution of the photons, there are about $e^{-B/\mathcal{T}}$ photons with energy larger than B when the temperature is $\mathcal{T}$. Hence, for a binding energy $B \sim 13.6\,\mathrm{eV}$, there are enough energetic photons as long as the temperature is above $\mathcal{T} \approx -B/(\ln 10^{-9}) \approx 3\,000\,\mathrm{K}$. This corresponds to a redshift of $z_* \approx 1\,090$.

At that time, electrons and protons combine into neutral hydrogen atoms. Photons do not interact much with hydrogen atoms, because as a bound system hydrogen is neutral. But this means that, at this particular moment, photons can travel unhindered over large distances for the first time and the universe becomes transparent! The newly freed photons pervade the universe, flying in all directions – this is the cosmic background radiation. Since the photons, electrons, and protons were in thermal equilibrium up to this time, the newly released radiation obeys a blackbody distribution with temperature $\mathcal{T} \approx 3\,000\,\mathrm{K}$.

As the universe expands, this radiation cools, but the blackbody spectrum is preserved. In equilibrium, the energy density of the radiation is given by

$$\rho = 2\int_0^\infty \frac{\mathrm{d}^3 p}{(2\pi)^3} f_{\mathrm{B-E}} E\,, \tag{1.64}$$

https://doi.org/10.1017/9781009516815.002 Published online by Cambridge University Press

where p denotes the momentum of a photon, $f_{\mathrm{B-E}} = 1/(e^{E/\mathcal{T}} - 1)$ denotes the Bose–Einstein distribution appropriate for bosons (it gives the average occupation number at energy E), E is the energy of a photon and is simply $E = p$, and the overall factor of 2 arises because the photon has two helicities. This gives

$$\rho = 2 \int_0^\infty \frac{\mathrm{d}^3 p}{(2\pi)^3} \frac{p}{e^{p/\mathcal{T}} - 1} = \frac{\pi^2}{15} \mathcal{T}^4 \,. \tag{1.65}$$

But radiation scales as $\rho \propto a^{-4}$, and hence we find once more that the temperature of the cosmic microwave background (and of all matter types that it is in equilibrium with during the early universe) scales as

$$\mathcal{T}(t) \propto \frac{1}{a(t)} \,. \tag{1.66}$$

Inside the integral above, the momentum p and the temperature $\mathcal{T}$ scale in such a way with the expansion of the universe that the functional form remains preserved, in particular $p/\mathcal{T}$ remains constant. Thus, the spectrum remains that of a blackbody, with a temperature decreasing inversely to the expansion of the universe.

When released, the cosmic background radiation was at a temperature of about 3 000 K (cf. the Sun's surface temperature of 6 000 K). By today, the wavelength of the emitted background radiation is bigger by a factor $1 + z_* \approx 1090$ and its temperature has correspondingly shrunk to 2.73 K. When emitted, the typical wavelength was on the order of a micrometer. Now it is on the order of a millimeter, and thus the radiation is in the microwave frequency region. This is the origin of the name *cosmic microwave background* (CMB). Using the relations above, we can also infer the number density of photons, which amounts to

$$n_\gamma = 2 \int_0^\infty \frac{\mathrm{d}^3 p}{(2\pi)^3} f_{\mathrm{B-E}} = \frac{2\zeta(3)}{\pi^2} \mathcal{T}^3 \approx 411 \,\mathrm{cm}^{-3} \,, \tag{1.67}$$

where the numerical evaluation corresponds to today's value. This can be contrasted with the number density of matter particles, which, as we calculated in (1.39), is a billion times lower. Thus one may think of the universe as being mostly filled with CMB radiation on average.

The cosmic background radiation indeed exists and, along with big bang nucleosynthesis, it constitutes the best evidence that we have for the hot big bang model. However, just as for nucleosynthesis, it is important to realize that it is not evidence for the big bang itself. Rather, it provides clear evidence that the universe was once very dense and more than a thousand times hotter than presently.

https://doi.org/10.1017/9781009516815.002 Published online by Cambridge University Press

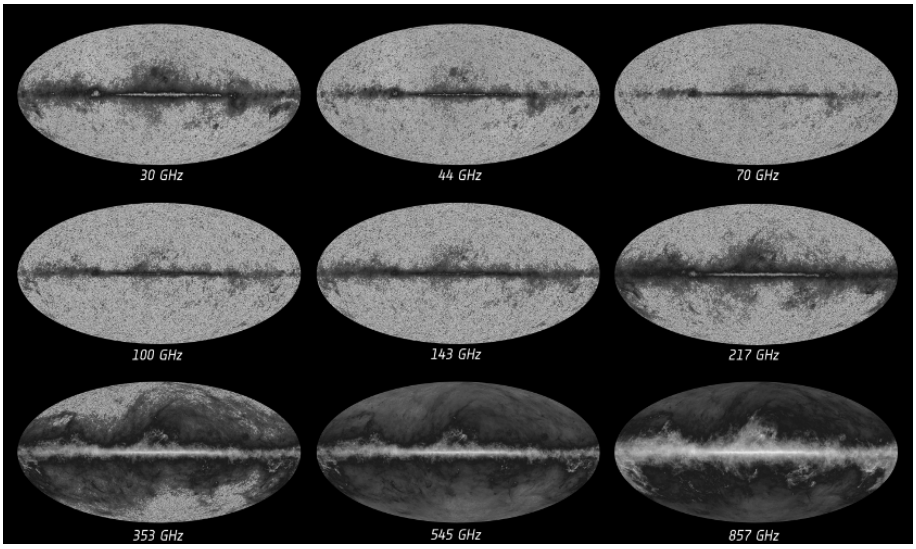


Figure 1.4 All-sky maps of the cosmic microwave background radiation, as observed by the Planck satellite in different frequency bands. The Milky Way galaxy is seen as the large horizontal structure in each picture. Foregrounds such as the Milky Way must be subtracted in order to be left purely with the primordial signals. Figure reproduced with permission from Planck Collaboration; Aghanim et al. (2020a). *Credit: ESA/Planck Collaboration.*

The CMB can be measured via radio telescopes, with the most precise measurements having been carried out by specifically dedicated satellite missions, such as the Wilkinson Microwave Anisotropy Probe (WMAP) and Planck; see Fig. 1.4. As the figure shows, foregrounds must be carefully removed before the primordial data can be analyzed. Once that is done, one finds that with high precision the temperature is the same in all directions in the sky. Thus the CMB provides the best evidence that we have for the isotropy of the universe, though only in a special frame.

The microwave background radiation itself defines a preferred frame in the universe, which can be considered to be the (comoving) "rest frame." To someone traveling at a considerable fraction of the speed of light with respect to that frame, the universe would not look isotropic at all. And in fact measurements of the CMB show a dipole with a maximal intensity of $\Delta\mathcal{T} \approx 3.3\,\mathrm{mK}$, which can be interpreted as indicating that we are moving with respect to the rest frame of the universe with a speed of about $370\,\mathrm{km\,s^{-1}}$, roughly in the direction of the constellation Leo (this dipole can be thought of as a Doppler effect due to our motion relative to the background radiation). The motion of the Local Group of galaxies is even

https://doi.org/10.1017/9781009516815.002 Published online by Cambridge University Press

faster (about $620\,\mathrm{km\,s^{-1}}$), because of the motion of the solar system within the Milky Way and its motion, in turn, within the Local Group. When the dipole is subtracted from the data, the universe appears isotropic to one part in 10^4 in all directions. Beyond the dipole, there is a wealth of information in higher multipoles, offering us the best clues we have as to what may have preceded the hot big bang phase. This will be discussed in detail in Chapter 3.

1.3 The big bang and assorted puzzles

The hot big bang model we have just described is in very precise agreement with cosmological observations. Amongst others, it explains the observed redshifts of galaxies, the abundances of primordial elements, as well as the existence of the CMB, and it provides a consistent picture of the ages and development of astrophysical objects.

The big bang model also resolves one apparent paradox about our universe that was known for quite some time, namely the observation that the sky is dark at night (this is sometimes called *Olbers's paradox*). If stars were visible out to infinite distances, then the whole sky ought to be filled with starlight, and even when the Sun is not visible the sky should be bright. However, the big bang effectively puts a limit to how far we can see, and thus also puts a limit on the number of stars that may fill up the night sky. The big bang has left the sky filled with the cosmic background radiation discussed in Section 1.2.1, but, as we saw, it is made up of microwave radiation invisible to the naked eye. Thus, the sky indeed appears dark at night.

Nevertheless, some of the features of and assumptions inherent in the big bang model lead to the conclusion that it cannot possibly be the final word on the origin of the universe. We will go over the main open issues now.

1.3.1 Singularity puzzle

The most obvious shortcoming of the big bang model is the big bang itself. When the scale factor goes to zero, curvature invariants (whose numerical value is independent of the choice of coordinates) like the Ricci scalar (1.31), or the Riemann tensor squared, generally become infinite. For example, we have

$$R^{\mu}{}_{\nu\rho\sigma}R_{\mu}{}^{\nu\rho\sigma} = 12\frac{\ddot{a}^2}{a^2} + 12\frac{\dot{a}^4}{a^4} + 24K\frac{\dot{a}^2}{a^4} + 12\frac{K^2}{a^4} + \text{anisotropy terms}\,, \quad (1.68)$$

where we explicitly wrote out the terms involving solely the scale factor a. Clearly, during a radiation-dominated phase with $a(t) \propto a^{1/2}$, or for other

https://doi.org/10.1017/9781009516815.002 Published online by Cambridge University Press

power-law solutions, this quantity blows up as $a \to 0$. But this is true far more generally: The Penrose–Hawking singularity theorems (which would take us too far off course to review in detail) demonstrate that anisotropies cannot compensate for the blowup of the volume terms, but rather reinforce the singularity; the theorems imply that a curvature singularity occurs under very generic circumstances.

A singularity means trouble. It implies for instance that the matter side of the Einstein equations also blows up or becomes otherwise pathological. Infinite quantities are neither measurable nor, consequently, physical. The model simply breaks down at the big bang. In this vein, it also seems premature to assume, as has often been done in the past, that the big bang marked the origin of space and time. A more sensible conclusion is to say that, in order to explain the origin of the universe, we need something that goes beyond the classical big bang model. There is a widespread hope that quantum gravity will be able to resolve this singularity and replace it with something well defined. In later chapters, we will discuss possibilities along these lines.

It is not only the big bang itself that presents a puzzle, but there are several dynamical features of the hot big bang model that require additional explanation and justification. The two most important ones are the flatness and horizon puzzles, which we will turn to next.

1.3.2 Flatness puzzle

If we consider the Friedmann equation again, for instance in the form of Eq. (1.45), we can see that the term representing the homogeneous curvature of the universe contributes a fractional amount $|k|/(aH)^2$ to the total energy content of the universe. The dependence on the scale factor and Hubble rate implies that, as the universe expands during the radiation and matter-dominated epochs, this fractional contribution becomes larger and larger, in proportion to t and $t^{2/3}$, respectively. From Eq. (1.50) we know that present-day observations imply an upper bound

$$\frac{|K|}{(aH)^2}\,|_{t_0} \lesssim 10^{-3}\,. \tag{1.69}$$

But this means then that at early times, the average curvature of the universe must have been extremely tiny.

More quantitatively, at radiation–matter equality, we must have had

$$\frac{|K|}{(aH)^2}\,|_{t_{\rm eq}} \lesssim 10^{-3}\frac{(aH)_0^2}{(aH)_{\rm eq}^2} = 10^{-6}\,, \tag{1.70}$$

https://doi.org/10.1017/9781009516815.002 Published online by Cambridge University Press

while during the radiation phase one obtains

$$\frac{|K|}{(aH)^2} \lesssim 10^{-6}\frac{a^2}{a_{\rm eq}^2} = 10^{-6}\frac{t}{t_{\rm eq}} \qquad (t < t_{\rm eq})\,. \tag{1.71}$$

It seems reasonable to extrapolate back at least to energies of 1 TeV, since these are the energies that can be probed by current particle physics experiments and are well understood. Then one has $|K|/(aH)^2\,|_{t_{\rm TeV}} \lesssim 10^{-31}$. This is an incredibly small quantity. How could the spatial part of the universe have been so precisely flat at very early times?

1.3.3 Horizon puzzle

Let us start by drawing a (highly simplified) conformal map of the universe. This is useful not only to appreciate the observational situation we find ourselves in, but also to understand causal structures in the universe. Recall that one may write the FLRW line element, now specializing to $K = 0$ as suggested by observations, in conformal time τ as

$$\mathrm{d}s^2 = a^2(\tau)[-\mathrm{d}\tau^2 + \mathrm{d}r^2 + r^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)]\,. \tag{1.72}$$

We can gain knowledge about causal relationships in spacetime by following the paths of lights rays (null geodesics). We can simplify the analysis by orienting the coordinates such that the light rays propagate at constant θ, ϕ values. Then the equation for null geodesics becomes remarkably simple,

$$\mathrm{d}s^2 = 0 \quad \rightarrow \quad \mathrm{d}\tau = \pm\mathrm{d}r\,, \tag{1.73}$$

where the sign depends on whether the light ray is outgoing or incoming. Thus if we observe an object at comoving distance r, the light was emitted at conformal time $\tau(t) = \tau(t_0) - r$.

If we now specialize to null geodesics originating at the big bang, we can evaluate the conformal time using its definition in terms of physical time t and using the Friedmann equation (1.45),

$$\begin{aligned}\tau(t) &= \int_0^t \frac{\mathrm{d}t}{a} \\ &= \int_0^a \frac{\mathrm{d}\ln(a)}{aH} \qquad (1.74)\\ &= \frac{1}{H_0}\int_0^a \frac{\mathrm{d}a}{\sqrt{\Omega_r + \Omega_m a + \Omega_\Lambda a^4}}\,. \qquad (1.75)\end{aligned}$$

If we evaluate the integral up to the present time, $a = a_0 = 1$, we find $\eta_{\rm BB} = 3.196/H_0$. This implies a current distance $a_0 r = r = \eta_{\rm BB}$ of about

https://doi.org/10.1017/9781009516815.002 Published online by Cambridge University Press

46.2 billion light-years to the furthest region from which a particle could have reached us by today. In other words, this is the current particle horizon, see Fig. 1.5 for an illustration.

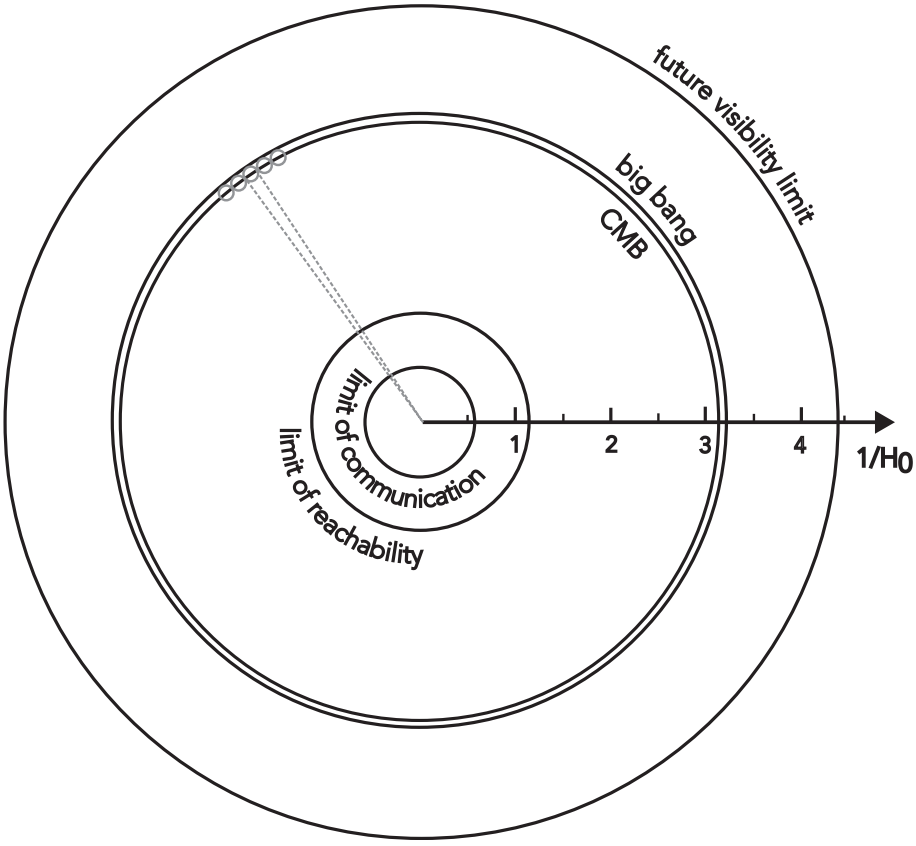


Figure 1.5 A conformal map of the observable universe. We are located in the center of the picture. Distances are given in units of the current Hubble radius $1/H_0 \approx 14.5$ billion light-years. The map shows the location of events in comoving coordinates. We can currently see out to the CMB, at comoving distance $3.13/H_0$ or redshift $z = 1089$. The small circles indicate the size of spheres of causality at last scattering. The future visibility limit (at $4.34/H_0$) is the furthest we will be able to see (in the infinite future) if the dark energy remains constant at the present value. The limit of reachability is the furthest we can send a message to, while the limit of communication is the current maximal distance to which we can send a message and hope to receive a response in the future – it is astonishingly close, at comoving distance $0.57/H_0$ or redshift $z = 0.69$.

The CMB was emitted at scale factor $a_{\rm CMB} = 1/1090$ and consequently $\tau_{\rm CMB} = 3.132/H_0$ or about 45.3 billion light-years. This is the current dis-

https://doi.org/10.1017/9781009516815.002 Published online by Cambridge University Press

tance to the regions from which the CMB (the part that we see today) was emitted. When comparing with the age of the universe, one should note that this is larger than 13.8 billion light-years due to the expansion of the universe. At the time that the CMB was emitted, this region was smaller by a factor 1 090, i.e., it was a sphere with only a 41.5 million light-year radius.

The comoving distance between the big bang and the CMB was thus $\tau_{\rm BB} - \tau_{\rm CMB} = 0.063/H_0$. This is the distance that light rays could cover by the time of last scattering. In other words, this is the maximal region that could have been in causal contact up to that time. Corresponding regions are drawn as small gray circles in Fig. 1.5. On the current sky, they subtend an angle of $(\tau_{\rm BB} - \tau_{\rm CMB})/\tau_{\rm CMB}$ radians or about 1.15 degrees. This observation should give one pause: We observe the same temperature to very good accuracy in all directions in the sky, yet the radiation is coming from regions that could not have been in causal contact with each other since the big bang. How then can they all be describing the same blackbody spectrum at the exact same temperature?

This surprising feature is known as the *horizon puzzle*. Note that it gets worse when we extrapolate further back from the CMB. If we assume that the universe was at thermal equilibrium at earlier moments then, as should be clear by inspecting Fig. 1.5, the corresponding regions of causality were even smaller, and the puzzle sharper. One may trace the origin of the problem back to the behavior of the comoving horizon $1/(aH)$, which can be seen to determine the conformal time τ via Eq. (1.74). Just as for the flatness puzzle discussed above, the problem is that during the radiation-dominated phase (and also during matter domination) the comoving horizon $1/(aH)$ steadily grows. In this way more and more regions, which up to that moment had not been in causal contact, are suddenly able to communicate. The puzzle is that, when this happens, everyone seems to know already exactly what the others have been saying all along.

Before moving on to further puzzles, let us use the opportunity to point out a few additional features of interest on the conformal map in Fig. 1.5. For instance, we can let the scale factor range from zero at the big bang all the way to infinity in the infinite future. This still results in a finite integral (1.75), yielding $\tau_\infty = 4.34/H_0$. This is the largest distance that a particle can travel from the big bang onward. Thus it is the future visibility limit, beyond which we will never be able to observe anything, if the dark energy remains constant at the present value. In the figure, this is the circle to which the big bang circle will recede in the infinite future.

If we evaluate the integral (1.75) from today ($a = 1$) up to the infinite future ($a = \infty$), we obtain a comoving distance of $1.14/H_0$. This is the fur-

https://doi.org/10.1017/9781009516815.002 Published online by Cambridge University Press

thest we can send a signal to, and it corresponds to the limit of reachability shown in the figure. If we would also like to receive a response, then that comoving distance gets halved to $0.57/H_0$, corresponding to galaxies that currently reside at a redshift of only $z = 0.69$. Anything further away than this redshift is beyond our limit of communication. This perhaps surprisingly small region constitutes the part of the universe that we can hope one day to interact with.

1.3.4 Brief list of additional puzzles

There are further aspects of the hot big bang model that seem contrived or, better said, that will require justification. We will simply list a few of these open issues.

- The CMB is isotropic to a high degree, but there are temperature fluctuations of order $\Delta\mathcal{T}/\mathcal{T} \sim \mathcal{O}(10^{-4})$, which are in fact responsible for all the gravitationally bound and collapsed structures we see in the universe today. How did these primordial temperature perturbations arise, and what determined their properties? This is the *anisotropy puzzle*.
- At high energies the gauge couplings of the standard model of particle physics approach each other (and when supersymmetry is included they appear to evolve to a common value at energies of about 10^{16} GeV). This suggests *grand unification*, namely that at high energies a larger symmetry group united the $SU(3)$, $SU(2)$, and $U(1)$ symmetries of the standard model. But if this is indeed the case, then topological defects would have formed when the symmetry was broken as the universe cooled. None such topological defects, which would manifest themselves as monopoles or cosmic strings, for example, have been observed. Why not? This the *topological defects puzzle*.
- There must have existed a slight excess of matter over antimatter in the early universe. This is inferred from the fact that our observable universe appears to be made purely out of matter, and not antimatter (otherwise we would observe the light emitted when matter and antimatter annihilate each other). To prevent a total annihilation of matter and antimatter in the early universe, there must have been a slight overabundance of matter. Various theoretical models exist that can potentially explain this so-called "baryon asymmetry." However, it remains too early to know which (if any) occurred in our universe. This is the *antimatter puzzle*.
- The second law of thermodynamics says that the entropy always grows. But then this means that in the early universe, the entropy must have

https://doi.org/10.1017/9781009516815.002 Published online by Cambridge University Press

been comparatively low. In other words, the early universe must have been in a special, unlikely state. This is the *entropy puzzle*. But then what determined this state? It seems difficult for a classical theory to achieve this; rather, we may hope that a proper quantum treatment of the early universe may result in certain states being strongly favored over others.

- The hot big bang model assumes that already very early on, space and time can be treated purely classically. However, we tend to believe that nature's most fundamental laws are quantum laws. So how is it that space and time came to behave so classically? In the late universe, decoherence can be invoked to explain why macroscopic objects, and also space and time, appear classical to us. This is because the interactions of large numbers of elementary particles lead to a strong suppression of quantum interference effects. But if one thinks about the possible beginning stages of the universe, then it is not clear that decoherence can be applied, as decoherence is a process that occurs over time, and time itself may not have been classical at the creation of the universe, given that the very production of spacetime and of matter would have had to be a quantum process. This is the *classicality puzzle*.

We should point out that none of these puzzles are strict inconsistencies of the hot big bang model. Rather, they can all be seen as specifying special initial conditions for the universe to start out from. It would certainly be much more satisfying if one could explain these features by physical arguments, perhaps either being due to necessity or being caused by suitable dynamical processes occurring prior to the hot big bang. We will start exploring the most prominent suggestions for such "pre-histories" in the next chapter.

In closing, let us mention one final oddity: In the hot big bang model, dark energy plays essentially no role (and the history of the universe up to now would have been quite similar without dark energy), yet it has come to dominate the present-day energy budget of the universe. Why then is dark energy there? One feels that there must be a deeper reason.

Exercises

1.1 Estimate the density inside our Galaxy. At what redshift did the universe have the equivalent average density? For consistency of the big bang model, check that this was before the first galaxies formed.

1.2 Roughly how far away into deep space do you have to go before the ex-

https://doi.org/10.1017/9781009516815.002 Published online by Cambridge University Press

pansion of space becomes significant? *Hint:* It may be useful to consider the Schwarzschild–de Sitter solution.

1.3 Calculate the current distance to galaxies at redshifts $z_g = 0.1, 1, 10$ and the cosmic time at which we are seeing them.

1.4 Give an order-of-magnitude estimate as to how many galaxies recede away from us each year such that they can no longer be communicated with.

1.5 Assume that the observable universe is the full extent of the universe, and assume radiation domination throughout its history. Then compare the age and size of the universe at temperature 1 TeV and at 1 Planck time.

https://doi.org/10.1017/9781009516815.002 Published online by Cambridge University Press