GALAXY FORMATION AND EVOLUTION

HOUJUN MO
University of Massachusetts

FRANK VAN DEN BOSCH
Yale University

SIMON WHITE
Max Planch Institute for Astrophysics

CAMBRIDGE
UNIVERSITY PRESS

shows that the variance of the matter density field on scales of $8h^{-1}$Mpc is about 0.7–0.9 (e.g. Van Waerbeke et al., 2001), slightly lower than that of the distribution of bright galaxies.

Since the matter distribution around a given galaxy or cluster will cause a distortion of its background galaxies, weak lensing can also be used to probe the matter distributions around galaxies and clusters. In the case of clusters, one can often detect a sufficient number of background galaxies to reliably measure the shear induced by its gravitational potential. Weak lensing therefore offers a means of measuring the total gravitational mass of an individual (massive) cluster. In the case of individual galaxies, however, one typically has only a few background galaxies available. Consequently, the weak lensing signal is far too weak to detect around individual galaxies. However, by stacking the images of many foreground galaxies (for example, according to their luminosity), one obtains sufficient signal-to-noise to measure the shear, which reflects the *average* mass distribution around the stacked galaxies. This technique is called galaxy–galaxy lensing, and has been used to demonstrate that galaxies are surrounded by extended dark matter halos with masses 10–100 times more massive than the galaxies themselves (e.g. Mandelbaum et al., 2006b).

2.8 The Intergalactic Medium

The intergalactic medium (IGM) is the medium that permeates the space in between galaxies. In the framework laid out in Chapter 1, galaxies form by the gravitational aggregation of gas in a medium which was originally quite homogeneous. In this scenario, the study of the IGM is an inseparable part of galaxy formation, because it provides us with the properties of the gas from which galaxies form.

The properties of the IGM can be probed observationally by its emission and by its absorption of the light from background sources. If the medium is sufficiently dense and hot, it can be observed in X-ray emission, as is the case for the intracluster medium described in §2.5.1. However, in general the density of the IGM is too low to produce detectable emission, and its properties have to be determined from absorption studies.

2.8.1 The Gunn–Peterson Test

Much information about the IGM has been obtained through its absorption of light from distant quasars. Quasars are not only bright, so that they can be observed out to large distances, but also have well-behaved continua, against which absorption can be analyzed relatively easily. One of the most important tests of the presence of intergalactic neutral hydrogen was proposed by Gunn & Peterson (1965). The Gunn–Peterson test makes use of the fact that the Lyα absorption of neutral hydrogen at $\lambda_\alpha = 1216\,$Å has a very large cross-section. When the ultraviolet continuum of a distant quasar (assumed to have redshift $z_{\rm Q}$) is shifted to 1216 Å at some redshift $z < z_{\rm Q}$, the radiation would be absorbed at this redshift if there were even a small amount of neutral hydrogen. Thus, if the Universe were filled with a diffuse distribution of neutral hydrogen, photons bluer than Lyα would be significantly absorbed, causing a significant decrement of flux in the observed quasar spectrum at wavelengths shorter than $(1+z_{\rm Q})\lambda_\alpha$. Using the hydrogen Ly$\alpha$ cross-section and the definition of optical depth (see Chapter 16 for details), one obtains that the proper number density of HI atoms obeys

$$n_{\rm HI}(z) \sim 2.42\times10^{-11}\,\tau(z)hH(z)/H_0\,{\rm cm}^{-3}, \tag{2.47}$$

where $H(z)$ is Hubble's constant at redshift z, and $\tau(z)$ is the absorption optical depth out to z that can be determined from the flux decrements in quasar spectra. Observations show that the Lyα absorption optical depth is much smaller than unity out to $z \lesssim 6$. The implied density

of neutral hydrogen in the diffuse IGM is thus much lower than the mean gas density in the Universe (which is about $10^{-7}\,\mathrm{cm}^{-3}$). This suggests that the IGM must be highly ionized at redshifts $z \lesssim 6$.

As we will show in Chapter 3, the IGM is expected to be highly neutral after recombination, which occurs at a redshift $z \sim 1000$. Therefore, the fact that the IGM is highly ionized at $z \sim 6$ indicates that the Universe must have undergone some phase transition, from being largely neutral to being highly ionized, a process called re-ionization. It is generally believed that photo-ionization due to energetic photons (with energies above the Lyman limit) are responsible for the re-ionization. This requires the presence of effective emitters of UV photons at high redshifts. Possible candidates include quasars, star-forming galaxies and the first generation of stars. But to this date the actual ionizing sources have not yet been identified, nor is it clear at what redshift re-ionization occurred. The highest redshift quasars discovered to date, which are close to $z = 6.5$, show almost no detectable flux at wavelengths shorter than $(1+z)\lambda_\alpha$ (Fan et al., 2006). Although this seems to suggest that the mass density of neutral hydrogen increases rapidly at around this redshift, it is not straightforward to convert such flux decrements into an absorption optical depth or a neutral hydrogen fraction, mainly because any $\tau \gg 1$ can result in an almost complete absorption of the flux. Therefore it is currently still unclear whether the Universe became (re-)ionized at a redshift just above 6 or at a significantly higher redshift. At the time of writing, several facilities are being constructed that will attempt to detect 21cm line emission from neutral hydrogen at high redshifts. It is anticipated that these experiments will shed important light on the detailed re-ionization history of the Universe, as we discuss in some detail in §16.3.4.

2.8.2 Quasar Absorption Line Systems

Although the flux blueward of $(1+z_{\rm Q})\lambda_\alpha$ is not entirely absorbed, quasar spectra typically reveal a large number of absorption lines in this wavelength range (see Fig. 2.39). These absorption lines are believed to be produced by intergalactic clouds that happen to lie along the line-of-sight from the observer to the quasar, and can be used to probe the properties of the IGM. Quasar absorption line systems are grouped into several categories:

- Lyα forest: These are narrow lines produced by HI Lyα absorption. They are numerous and appear as a 'forest' of lines blueward of the Lyα emission line of a quasar.
- Lyman-limit systems (LLS): These are systems with HI column densities $N_{\rm HI} \gtrsim 10^{17}\,\mathrm{cm}^{-2}$, at which the absorbing clouds are optically thick to the Lyman-limit photons (912 Å). These systems appear as continuum breaks in quasar spectra at the redshifted wavelength $(1+z_{\rm a}) \times 912$ Å, where $z_{\rm a}$ is the redshift of the absorber.
- Damped Lyα systems (DLAs): These systems are produced by HI Lyα absorption of gas clouds with HI column densities, $N_{\rm HI} \gtrsim 2\times10^{20}\,\mathrm{cm}^{-2}$. Because the Ly$\alpha$ absorption optical depth at such column densities is so large, the quasar continuum photons are completely absorbed near the line center and the line profile is dominated by the damping wing due to the natural (Lorentz) broadening of the absorption line. DLAs with column densities in the range $10^{19}\,\mathrm{cm}^{-2} < N_{\rm HI} < 2\times10^{20}\,\mathrm{cm}^{-2}$ also exhibit damping wings, and are sometimes called sub-DLAs (Péroux et al., 2002). They differ from the largely neutral DLAs in that they are still significantly ionized.
- Metal absorption line systems: In addition to the hydrogen absorption line systems listed above, QSO spectra also frequently show absorption lines due to metals. The best-known examples are MgII systems and CIV systems, which are caused by the strong resonance-line doublets MgII$\lambda\lambda$2796, 2800 and CIV$\lambda\lambda$1548, 1550, respectively. Note that both doublets have rest-frame wavelengths longer than $\lambda_{\rm Ly\alpha} = 1216$Å. Consequently, they can appear on the

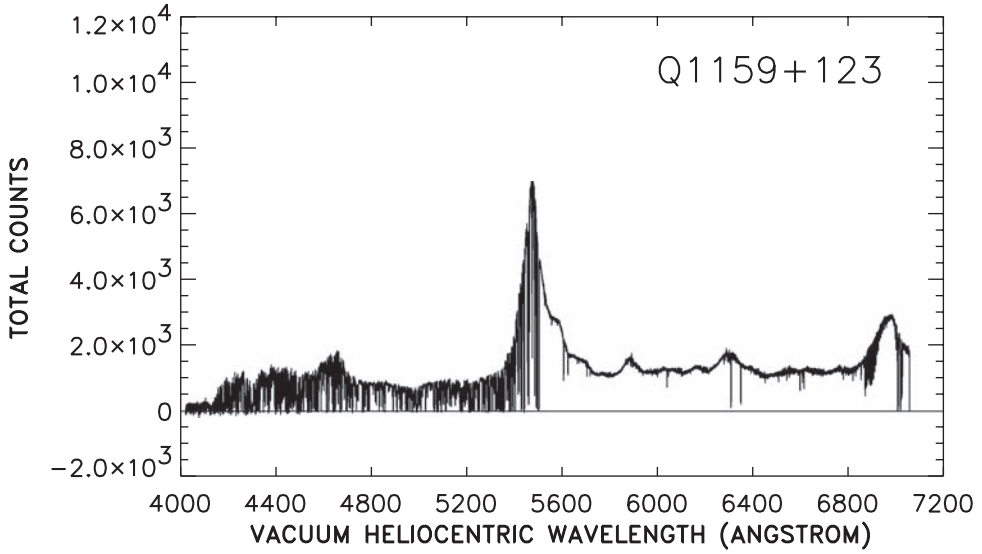

Fig. 2.39. The spectrum of a QSO that reveals a large number of absorption lines due to the IGM. The strongest peak at 5473 Å is the emission line due to Lyα at a rest-frame wavelength of 1216 Å. The numerous absorption lines at $\lambda < 5473$ Å make up the Lyα forest which is due to Lyα absorption of neutral hydrogen clouds between the QSO and the Earth. The break at 4150 Å is due to a Lyman-limit cloud which is optically thick at the hydrogen Lyman edge (rest-frame wavelength of 912 Å). The relatively sparse lines to the right of the Lyα emission line are due to absorption by metal atoms associated with the absorbing clouds. [Adapted from Songaila (1998) by permission of AAS]

red side of the Lyα emission line of the QSO, which makes them easily identifiable because of the absence of confusion from the Lyα forest.

Note that a single absorber may be detected as more than one absorption system. For example, an absorber at z_a may be detected as a HI Lyα line at $\lambda = (1+z_a)\times 1216$ Å, as a CIV system at $\lambda = (1+z_a)\times 1548$ Å, if it has a sufficiently large abundance of CIV ions, and as a Lyman-limit system at $\lambda = (1+z_a)\times 912$ Å, if its HI column density is larger than $\sim 10^{17}\,\mathrm{cm}^{-2}$.

In addition to the most common absorption systems listed above, other line systems are also frequently identified in quasar spectra. These include low ionization lines of heavy elements, such as CII, MgI, FeII, etc., and the more highly ionized lines, such as SiIV and NV. Highly ionized lines such as OVI and OVII are also detected in the UV and/or X-ray spectra of quasars. Since the ionization state of an absorbing cloud depends on its temperature, highly ionized lines, such as OVI and OVII, in general signify the existence of hot ($\sim 10^6$ K) gas, while low-ionization lines, such as HI, CII and MgII, are more likely associated with relatively cold ($\sim 10^4$ K) gas.

For a given quasar spectrum, absorption line systems are identified by decomposing the spectrum into individual lines with some assumed profiles (e.g. the Voigt profile, see §16.4.3). By modeling each system in detail, one can in principle obtain its column density, b parameter (defined as $b = \sqrt{2}\sigma$, where σ is the velocity dispersion of the absorbing gas), ionization state, and temperature. If both hydrogen and metal systems are detected, one may also estimate the metallicity of the absorbing gas. Table 2.8 lists the typical values of these quantities for the most commonly detected absorption systems mentioned above.

The evolution of the number of absorption systems is described by the number of systems per unit redshift, $\mathrm{d}\mathcal{N}/\mathrm{d}z$, as a function of z. This relation is usually fitted by a power law $\mathrm{d}\mathcal{N}/\mathrm{d}z \propto (1+z)^\gamma$, and the values of γ for different systems are listed in Table 2.9. The distribution of absorption line systems with respect to the HI column density is shown in Fig. 2.40. Over the

Table 2.8. Properties of common absorption lines in quasar spectra.

System	$\log(N_{\rm HI}/{\rm cm}^{-2})$	$b/({\rm km\,s}^{-1})$	$Z/Z_\odot$	$\log(N_{\rm HI}/N_{\rm H})$
Lyα forest	$12.5-17$	$15-40$	<0.01	<-3
Lyman limit	>17	~ 100	~ 0.1	>-2
sub-DLA	$19-20.3$	~ 100	~ 0.1	>-1
DLA	>20.3	~ 100	~ 0.1	~ 0
CIV	>15.5	~ 100	~ 0.1	>-3
MgII	>17	~ 100	~ 0.1	>-2

Table 2.9. Redshift evolution of quasar absorption line systems.

System	z range	γ	Reference
Lyα forest	$2.0-4.0$	~ 2.5	Kim et al. (1997)
Lyα forest	$0.0-1.5$	~ 0.15	Weymann et al. (1998)
Lyman limit	$0.3-4.1$	~ 1.5	Stengler-Larrea et al. (1995)
Damped Lyα	$0.1-4.7$	~ 1.3	Storrie-Lombardi et al. (1996a)
CIV	$1.3-3.4$	~ -1.2	Sargent et al. (1988)
MgII	$0.2-2.2$	~ 0.8	Steidel & Sargent (1992)

whole observed range, this distribution follows roughly a power law, $\mathrm{d}\mathcal{N}/\mathrm{d}N_{\rm HI} \propto N_{\rm HI}^{-\beta}$, with $\beta \sim 1.5$.

From the observed column density distribution, one can estimate the mean mass density of neutral hydrogen that is locked up in quasar absorption line systems:

$$\rho_{\rm HI}(z) = \left(\frac{\mathrm{d}l}{\mathrm{d}z}\right)^{-1} m_{\rm H} \int N_{\rm HI} \frac{\mathrm{d}^2\mathcal{N}}{\mathrm{d}N_{\rm HI}\,\mathrm{d}z}\,\mathrm{d}N_{\rm HI}, \tag{2.48}$$

where $\mathrm{d}l/\mathrm{d}z$ is the physical length per unit redshift at z (see §3.2.6). Given that $\mathrm{d}\mathcal{N}/\mathrm{d}N_{\rm HI}$ is a power law with index ~ -1.5, $\rho_{\rm HI}$ is dominated by systems with the highest $N_{\rm HI}$, i.e. by damped Lyα systems. Using the observed HI column density distribution, one infers that about 5% of the baryonic material in the Universe is in the form of HI gas at $z \sim 3$ (e.g. Storrie-Lombardi et al., 1996b). In order to estimate the total hydrogen mass density associated with quasar absorption line systems, however, one must know the neutral fraction, $N_{\rm HI}/N_{\rm H}$, as a function of $N_{\rm HI}$. This fraction depends on the ionization state of the IGM. Detailed modeling shows that the Lyα forest systems are highly ionized, and that the main contribution to the total (neutral plus ionized) gas density comes from absorption systems with $N_{\rm HI} \sim 10^{14}\,{\rm cm}^{-2}$. The total gas mass density at $z \sim 3$ thus inferred is comparable to the total baryon density in the Universe (e.g. Rauch et al., 1997; Weinberg et al., 1997b).

Quasar absorption line systems with the highest HI column densities are expected to be gas clouds in regions of high gas densities where galaxies and stars may form. It is therefore not surprising that these systems contain metals. Observations of damped Lyα systems show that they have typical metallicities about 1/10 of that of the Sun (e.g. Pettini et al., 1990; Kulkarni et al., 2005), lower than that of the ISM in the Milky Way. This suggests that these systems may be associated with the outer parts of galaxies, or with galaxies in which only a small fraction of the gas has formed stars. More surprising is the finding that most, if not all, of the Lyα forest lines also contain metals, although the metallicities are generally low, typically about $1/1000$ to $1/100$ of that of the Sun (e.g. Simcoe et al., 2004). There is some indication that the metallicity increases with HI column density, but the trend is not strong. Since star formation requires relatively high column densities of neutral hydrogen (see Chapter 9), the metals observed in absorption line

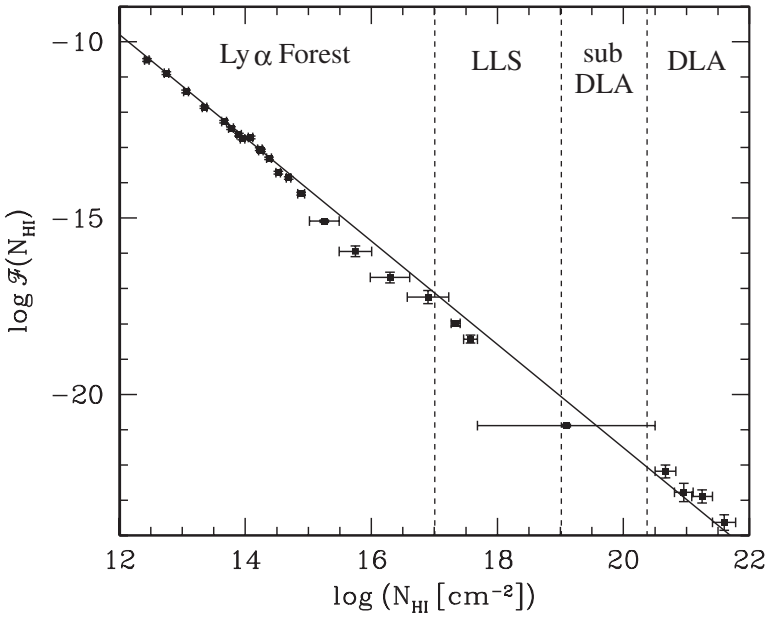

Fig. 2.40. The HI column density distribution of QSO absorption line systems. Here $\mathscr{F}(N_{\rm HI})$ is defined as the number of absorption lines per unit column density, per unit X (which is a quantity that is related to redshift according to Eq. [16.92]). The solid line corresponds to $\mathscr{F}(N_{\rm HI}) \propto N_{\rm HI}^{-1.46}$, which fits the data reasonably well over the full 10 orders of magnitude in column density. [Based on data published in Petitjean et al. (1993) and E. M. Hu et al. (1995)]

systems with low HI column densities most likely originate from, and have been expelled by, galaxies at relatively large distances.

2.9 The Cosmic Microwave Background

The cosmic microwave background (CMB) was discovered by Penzias and Wilson in 1965 when they were commissioning a sensitive receiver at centimeter wavelengths in Bell Telephone Laboratories. It was quickly found that this radiation background was highly isotropic on the sky and has a spectrum close to that of a blackbody with a temperature of about 3 K. The existence of such a radiation background was predicted by Gamow, based on his model of a Hot Big Bang cosmology (see §1.4.2), and it therefore did not take long before the cosmological significance of this discovery was realized (e.g. Dicke et al., 1965).

The observed properties of the CMB are most naturally explained in the standard model of cosmology. Since the early Universe was dense, hot and highly ionized, photons were absorbed and re-emitted many times by electrons and ions and so a blackbody spectrum could be established in the early Universe. As the Universe expanded and cooled and the density of ionized material dropped, photons were scattered less and less often and eventually could propagate freely to the observer from a last-scattering surface, inheriting the blackbody spectrum.

Because the CMB is so important for our understanding of the structure and evolution of the Universe, there have been many attempts in the 1970s and 1980s to obtain more accurate measurements of its spectrum. Since the atmospheric emission is quite close to the peak wavelength of a 3 K blackbody spectrum, most of these measurements were carried out using high-altitude balloon experiments (for a discussion of early CMB experiments, see Partridge, 1995).

A milestone in CMB experiments was the launch by NASA in November 1989 of the Cosmic Background Explorer (COBE), a satellite devoted to accurate measurements of the CMB over the entire sky. Observations with the Far InfraRed Absolute Spectrophotometer (FIRAS) on board COBE showed that the CMB has a spectrum that is perfectly consistent with a blackbody spectrum, to exquisite accuracy, with a temperature $T = 2.728 \pm 0.002\,\mathrm{K}$. As we will see in §3.5.4 the lack of any detected distortions from a pure blackbody spectrum puts strong constraints on any processes that may change the CMB spectrum after it was established in the early Universe.

Another important observational result from COBE is the detection, for the first time, of anisotropy in the CMB. Observations with the Differential Microwave Radiometers (DMR) on board COBE have shown that the CMB temperature distribution is highly isotropic over the sky, confirming earlier observational results, but also revealed small temperature fluctuations (see Fig. 2.41). The observed temperature map contains a component of anisotropy on very large angular scales, which is well described by a dipole distribution over the sky,

$$T(\alpha) = T_0 \left(1 + \frac{v}{c} \cos\alpha\right), \tag{2.49}$$

where α is the angle of the line-of-sight relative to a specific direction. This component can be explained as the Doppler effect caused by the motion of the Earth with a velocity $v = 369 \pm 3\,\mathrm{km\,s^{-1}}$ towards the direction $(l, b) = (264.31^\circ \pm 0.20^\circ, 48.05^\circ \pm 0.10^\circ)$ in Galactic coordinates

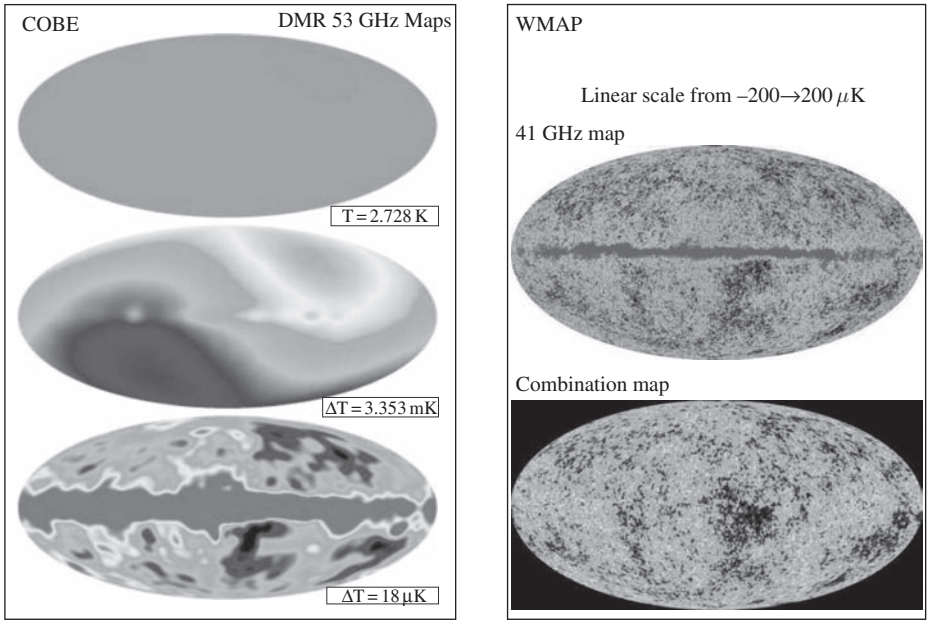

Fig. 2.41. Temperature maps of the CMB in galactic coordinates. The three panels on the left show the temperature maps obtained by the DMR on board the COBE satellite [Courtesy of NASA Goddard Space Flight Center]. The upper panel shows the near-uniformity of the CMB brightness; the middle panel is the map after subtraction of the mean brightness, showing the dipole component due to our motion with respect to the background; and the bottom panel shows the temperature fluctuations after subtraction of the dipole component. Emission from the Milky Way is evident in the bottom image. The two right panels show the temperature maps observed by WMAP from the first year of data [Courtesy of WMAP Science Team]; one is from the 41 GHz channel and the other is a linear combination of five channels. Note that the large-scale temperature fluctuations in the COBE map at the bottom are clearly seen in the WMAP maps, and that the WMAP angular resolution (about 0.5°) is much higher than that of COBE (about 7°).

(Lineweaver et al., 1996). Once this dipole component is subtracted, the map of the temperature fluctuations looks like that shown in the lower left panel of Fig. 2.41. In addition to emission from the Milky Way, it reveals fluctuations in the CMB temperature with an amplitude of the order of $\Delta T/T \sim 2 \times 10^{-5}$.

Since the angular resolution of the DMR is about 7°, COBE observations cannot reveal anisotropy in the CMB on smaller angular scales. Following the detection by COBE, there have been a large number of experiments to measure small-scale CMB anisotropies, and many important results have come out in recent years. These include the results from balloon-borne experiments such as Boomerang (de Bernardis et al., 2000) and Maxima (Hanany et al., 2000), from ground-based interferometers such as the Degree Angular Scale Interferometer (DASI; Halverson et al., 2002) and the Cosmic Background Imager (CBI; Mason et al., 2002), and from an all-sky satellite experiment called the Wilkinson Microwave Anisotropy Probe (WMAP; Bennett et al., 2003; Hinshaw et al., 2007). These experiments have provided us with extremely detailed and accurate maps of the anisotropies in the CMB, such as that obtained by WMAP shown in the right panels of Fig. 2.41.

In order to quantify the observed temperature fluctuations, a common practice is to expand the map in spherical harmonics,

$$\frac{\Delta T}{T}(\vartheta,\varphi) \equiv \frac{T(\vartheta,\varphi)-\overline{T}}{\overline{T}} = \sum_{\ell,m} a_{\ell m} Y_{\ell,m}(\vartheta,\varphi). \tag{2.50}$$

The angular power spectrum, defined as $C_\ell \equiv \langle |a_{\ell m}|^2 \rangle^{1/2}$ (where $\langle \ldots \rangle$ denotes averaging over m), can be used to represent the amplitudes of temperature fluctuations on different angular scales. Fig. 2.42 shows the temperature power spectrum obtained by the WMAP satellite. As one can

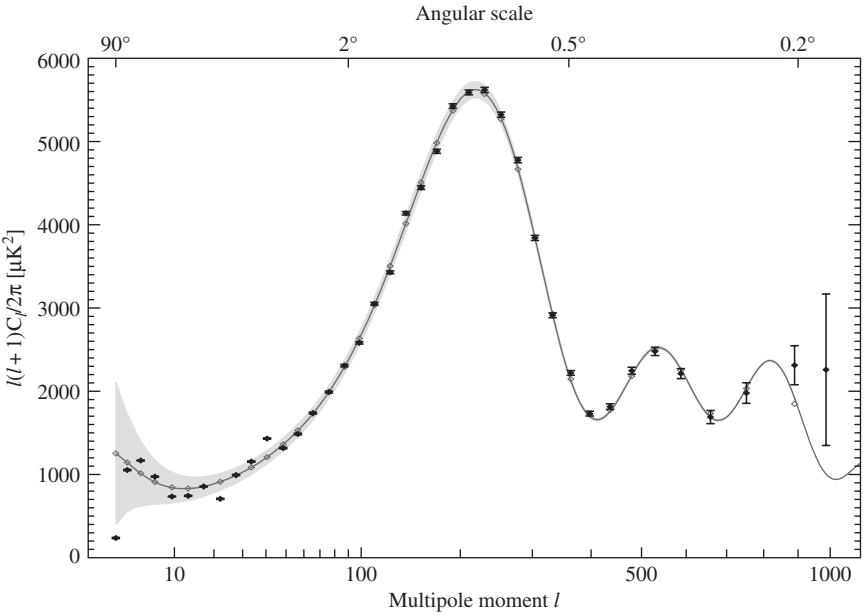

Fig. 2.42. The angular power spectrum, C_ℓ, of the CMB temperature fluctuations in the WMAP full-sky map. This shows the relative brightness of the 'spots' in the CMB temperature map vs. the size of the spots. The shape of this curve contains a wealth of information about the geometry and matter content of the Universe. The curve is the model prediction for the best-fit ΛCDM cosmology. [Adapted from Hinshaw et al. (2007) by permission of AAS]

see, the observed C_ℓ as a function of ℓ shows complex features. These observational results are extremely important for our understanding of the structure formation in the Universe. First of all, the observed high degree of isotropy in the CMB gives strong support for the assumption of the standard cosmology that the Universe is highly homogeneous and isotropic on large scales. Second, the small temperature fluctuations observed in the CMB are believed to be caused by the density perturbations at the time when the Universe became transparent to CMB photons. These same density perturbations are thought to be responsible for the formation of structures in the Universe. So the temperature fluctuations in the CMB may be used to infer the properties of the initial conditions for the formation of galaxies and other structures in the Universe. Furthermore, the observations of CMB temperature fluctuations can also be used to constrain cosmological parameters. As we will discuss in detail in Chapter 6, the peaks and valleys in the angular power spectrum are caused by acoustic waves present at the last scattering surface of the CMB photons. The heights (depths) and positions of these peaks (valleys) depend not only on the density of baryonic matter, but also on the total mean density of the Universe, Hubble's constant and other cosmological parameters. Modeling the angular power spectrum of the CMB temperature fluctuations can therefore provide constraints on all of these cosmological parameters.

2.10 The Homogeneous and Isotropic Universe

As we will see in Chapter 3, the standard cosmological model is based on the 'cosmological principle' according to which the Universe is homogeneous and isotropic on large scales. As we have seen, observations of the CMB and of the large-scale spatial distribution of galaxies offer strong support for this cosmological principle. Since according to Einstein's general relativity the space-time geometry of the Universe is determined by the matter distribution in the Universe, this large-scale distribution of matter has important implications for the large-scale geometry of space-time.

For a homogeneous and isotropic universe, its global properties (such as density and pressure) at any time must be the same as those in any small volume. This allows one to study the global properties of the Universe by examining the properties of a small volume within which Newtonian physics is valid. Consider a (small) spherical region of fixed mass M. Since the Universe is homogeneous and isotropic, the radius R of the sphere should satisfy the Newtonian equation[6]

$$\ddot{R} = -\frac{GM}{R^2}. \tag{2.51}$$

Note that, because of the homogeneity, there is no force due to pressure gradients and that only the mass within the sphere is relevant for the motion of R. This follows directly from Birkhoff's theorem, according to which the gravitational acceleration at any radius in a spherically symmetric system depends only on the mass within that radius. For a given M, the above equation can be integrated once to give

$$\frac{1}{2}\dot{R}^2 - \frac{GM}{R} = E, \tag{2.52}$$

[6] As we will see in Chapter 3, in general relativity it is the combination of energy density ρ and pressure P, $\rho + 3P/c^2$, instead of ρ, that acts as the source of gravitational acceleration. Therefore, Eq. (2.51) is not formally valid, even though Eq. (2.53), which derives from it, happens to be correct.

where E is a constant, equal to the specific energy of the spherical shell. For simplicity, we write $R = a(t)R_0$, where R_0 is independent of t. It then follows that

$$\frac{\dot{a}^2}{a^2} - \frac{8\pi G\overline{\rho}}{3} = -\frac{Kc^2}{a^2}, \tag{2.53}$$

where $\overline{\rho}$ is the mean density of the Universe and $K = -2E/(cR_0)^2$. Unless $E = 0$, which corresponds to $K = 0$, we can always choose the value of R_0 so that $|K| = 1$. So defined, K is called the curvature signature, and takes the value $+1$, 0, or -1. With this normalization, the equation for a is independent of M. As we will see in Chapter 3, Eq. (2.53) is identical to the Friedmann equation based on general relativity. For a universe dominated by a non-relativistic fluid, this is not surprising, as it follows directly from the assumption of homogeneity and isotropy. However, as we will see in Chapter 3, it turns out that Eq. (2.53) also holds even if relativistic matter and/or the energy density associated with the cosmological constant are included.

The quantity $a(t)$ introduced above is called the scale factor, and describes the change of the distance between any two points fixed in the cosmological background. If the distance between a pair of points is l_1 at time t_1, then their distance at some later time t_2 is related to l_1 through $l_2 = l_1 a(t_2)/a(t_1)$. It then follows that at any time t the velocity between any two (comoving) points can be written as

$$\dot{l} = [\dot{a}(t)/a(t)]l, \tag{2.54}$$

where l is the distance between the two points at time t. Thus, $\dot{a} > 0$ corresponds to an expanding universe, while $\dot{a} < 0$ corresponds to a shrinking universe; the universe is static only when $\dot{a} = 0$. The ratio $\dot{a}/a$ evaluated at the present time, t_0, is called the Hubble constant,

$$H_0 \equiv \dot{a}_0/a_0, \tag{2.55}$$

where $a_0 \equiv a(t_0)$, and the relation between velocity and distance, $\dot{l} = H_0 l$, is known as Hubble's expansion law. Another quantity that characterizes the expansion of the Universe is the deceleration parameter, defined as

$$q_0 \equiv -\frac{\ddot{a}_0 a_0}{\dot{a}_0^2}. \tag{2.56}$$

This quantity describes whether the expansion rate of the Universe is accelerating ($q_0 < 0$) or decelerating ($q_0 > 0$) at the present time.

Because of the expansion of the Universe, waves propagating in the Universe are stretched. Thus, photons with a wavelength λ emitted at an earlier time t will be observed at the present time t_0 with a wavelength $\lambda_{\rm obs} = \lambda a_0/a(t)$. Since $a_0 > a(t)$ in an expanding universe, $\lambda_{\rm obs} > \lambda$ and so the wavelength of the photons is redshifted. The amount of redshift z between time t and t_0 is given by

$$z \equiv \frac{\lambda_{\rm obs}}{\lambda} - 1 = \frac{a_0}{a(t)} - 1. \tag{2.57}$$

Note that $a(t)$ is a monotonically increasing function of t in an expanding universe, and so redshift is uniquely related to time through the above equation. If an object has redshift z, i.e. its observed spectrum is shifted to the red relative to its rest-frame (intrinsic) spectrum by $\Delta\lambda = \lambda_{\rm obs} - \lambda = z\lambda$, then the photons we observe today from the object were actually emitted at a time t that is related to its redshift z by Eq. (2.57). Because of the constancy of the speed of light, an object's redshift can also be used to infer its distance.

From Eq. (2.53) one can see that the value of K is determined by the mean density $\overline{\rho}_0$ at the present time t_0 and the value of Hubble's constant. Indeed, if we define a critical density

$$\rho_{\rm crit,0} \equiv \frac{3H_0^2}{8\pi G}, \tag{2.58}$$

and write the mean density in terms of the density parameter,

$$\Omega_0 \equiv \overline{\rho}_0/\rho_{\rm crit,0}, \tag{2.59}$$

then $K = H_0^2 a_0^2(\Omega_0 - 1)$. So $K = -1$, 0 and $+1$ corresponds to $\Omega_0 < 1$, $= 1$ and > 1, respectively. Before discussing the matter content of the Universe, it is illustrative to write the mean density as a sum of several possible components:

(i) non-relativistic matter whose (rest-mass) energy density changes as $\rho_{\rm m} \propto a^{-3}$;
(ii) relativistic matter (such as photons) whose energy density changes as $\rho_{\rm r} \propto a^{-4}$ (the number density changes as a^{-3} while energy is redshifted according to a^{-1});
(iii) vacuum energy, or the cosmological constant Λ, whose density $\rho_\Lambda = c^2\Lambda/8\pi G$ is a constant.

Thus,

$$\Omega_0 = \Omega_{\rm m,0} + \Omega_{\rm r,0} + \Omega_{\Lambda,0}, \tag{2.60}$$

and Eq. (2.53) can be written as

$$\left(\frac{\dot{a}}{a}\right)^2 = H_0^2 E^2(z), \tag{2.61}$$

where

$$E(z) = \left[\Omega_{\Lambda,0} + (1-\Omega_0)(1+z)^2 + \Omega_{\rm m,0}(1+z)^3 + \Omega_{\rm r,0}(1+z)^4\right]^{1/2} \tag{2.62}$$

with z related to $a(t)$ by Eq. (2.57). In order to solve for $a(t)$, we must know the value of H_0 and the energy (mass) content ($\Omega_{\rm m,0}$, $\Omega_{\rm r,0}$, $\Omega_{\Lambda,0}$) at the present time. The deceleration parameter defined in Eq. (2.56) is related to these parameters by

$$q_0 = \frac{\Omega_{\rm m,0}}{2} + \Omega_{\rm r,0} - \Omega_{\Lambda,0}. \tag{2.63}$$

A particularly simple case is the Einstein–de Sitter model in which $\Omega_{\rm m,0} = 1$, $\Omega_{\rm r,0} = \Omega_{\Lambda,0} = 0$ (and so $q_0 = 1/2$). It is then easy to show that $a(t) \propto t^{2/3}$. Another interesting case is a flat model in which $\Omega_{\rm m,0} + \Omega_{\Lambda,0} = 1$ and $\Omega_{\rm r,0} = 0$. In this case, $q_0 = 3\Omega_{\rm m,0}/2 - 1$, so that $q_0 < 0$ (i.e. the expansion is accelerating at the present time) if $\Omega_{\rm m,0} < 2/3$.

2.10.1 The Determination of Cosmological Parameters

As shown above, the geometry of the Universe in the standard model is specified by a set of cosmological parameters. The values of these cosmological parameters can therefore be estimated by measuring the geometrical properties of the Universe. The starting point is to find two observables that are related to each other only through the geometrical properties of the Universe. The most important example here is the redshift–distance relation. As we will see in Chapter 3, two types of distances can be defined through observational quantities. One is the luminosity distance, d_L, which relates the luminosity of an object, L, to its flux, f, according to $L = 4\pi d_L^2 f$. The other is the angular-diameter distance, d_A, which relates the physical size of an object, D, to its angular size, θ, via $D = d_A\theta$. In general, the redshift–distance relation can formally be written as

$$d(z) = \frac{cz}{H_0}\left[1 + \mathscr{F}_d(z; \Omega_{\rm m,0}, \Omega_{\Lambda,0}, \ldots)\right], \tag{2.64}$$

where d stands for either d_L or d_A, and by definition $\mathscr{F}_d \ll 1$ for $z \ll 1$. For redshifts much smaller than 1, the redshift–distance relation reduces to the Hubble expansion law $cz = H_0 d$, and so the Hubble constant H_0 can be obtained by measuring the redshift and distance of an object (ignoring, for the moment, that objects can have peculiar velocities). Redshifts are relatively easy to obtain from the spectra of objects, and in §2.1.3 we have seen how to measure the distances of a few classes of astronomical objects. The best estimate of the Hubble constant at the present comes from Cepheids observed by the HST, and the result is

$$H_0 = 100h\,\mathrm{km\,s^{-1}\,Mpc^{-1}}, \quad \text{with} \quad h = 0.72 \pm 0.08 \tag{2.65}$$

(Freedman et al., 2001).

In order to measure other cosmological parameters, one has to determine the nonlinear terms in the redshift–distance relation, which typically requires objects at $z \gtrsim 1$. For example, measuring the light curves of Type Ia supernovae out to $z \sim 1$ has yielded the following constraints:

$$0.8\Omega_{\mathrm{m},0} - 0.6\Omega_{\Lambda,0} \sim -0.2 \pm 0.1 \tag{2.66}$$

(e.g. Perlmutter et al., 1999). Using Eq. (2.63) and neglecting $\Omega_{\mathrm{r},0}$ because it is small, the above relation gives $q_0 \sim -0.33 - 0.83\Omega_{\mathrm{m},0}$. Since $\Omega_{\mathrm{m},0} > 0$, we have $q_0 < 0$, i.e. the expansion of the Universe is speeding up at the present time.

Important constraints on cosmological parameters can also be obtained from the angular spectrum of the CMB temperature fluctuations. As shown in Fig. 2.42, the observed angular spectrum C_ℓ contains peaks and valleys, which are believed to be produced by acoustic waves in the baryon–photon fluid at the time of photon–matter decoupling. As we will see in §6.7, the heights/depths and positions of these peaks/valleys depend not only on the density of baryonic matter in the Universe, but also on the total mean density, Hubble's constant and other cosmological parameters. In particular, the position of the first peak is sensitive to the total density parameter Ω_0 (or the curvature K). Based on the observational results shown in Fig. 2.42, one obtains

$$\begin{aligned} \Omega_0 = 1.02 \pm 0.02; \quad \Omega_{\mathrm{m},0}h^2 = 0.14 \pm 0.02; \\ h = 0.72 \pm 0.05; \quad \Omega_{\mathrm{b},0}h^2 = 0.024 \pm 0.001, \end{aligned} \tag{2.67}$$

where $\Omega_{\mathrm{m},0}$ and $\Omega_{\mathrm{b},0}$ are the density parameters of total matter and of baryonic matter, respectively (Spergel et al., 2007). Note that this implies that the Universe has an almost flat geometry, that matter accounts for only about a quarter of its total energy density, and that baryons account for only $\sim 17\%$ of the matter.

2.10.2 The Mass and Energy Content of the Universe

There is a fundamental difficulty in directly observing the mass (or energy) densities in different mass components: all that is gold does not glitter. There may well exist matter components with significant mass density which give off no detectable radiation. The only interaction which all components are guaranteed to exhibit is gravity, and thus gravitational effects must be studied if the census is to be complete. The global gravitational effect is the curvature of space-time which we discussed above. Independent information on the amount of gravitating mass can only be derived from the study of the inhomogeneities in the Universe, even though such studies may never lead to an unambiguous determination of the total matter content. After all, one can imagine adding a smooth and invisible component to any amount of inhomogeneously distributed mass, which would produce no detectable effect on the inhomogeneities.

The most intriguing result of such dynamical studies has been the demonstration that the total mass in large-scale structures greatly exceeds the amount of material from which emission can be

detected. This unidentified 'dark matter' (or 'invisible matter') is almost certainly the dominant contribution to the total mass density $\Omega_{\rm m,0}$. Its nature and origin remain one of the greatest mysteries of contemporary astronomy.

(a) Relativistic Components One of the best observed relativistic components of the Universe is the CMB radiation. From its blackbody spectrum and temperature, $T_{\rm CMB} = 2.73\,{\rm K}$, it is easy to estimate its energy density at the present time:

$$\rho_{\gamma,0} \approx 4.7 \times 10^{-34}\,{\rm g\,cm^{-3}}, \quad \text{or} \quad \Omega_{\gamma,0} = 2.5 \times 10^{-5} h^{-2}. \tag{2.68}$$

As we have seen in Fig. 2.2, the energy density of all other known photon backgrounds is much smaller. The only other relativistic component which is almost certainly present, although not yet directly detected, is a background of neutrinos. As we will see in Chapter 3, the energy density in this component can be calculated directly from the standard model, and it is expected to be 0.68 times that of the CMB radiation. Since the total energy density of the Universe at the present time is not much smaller than the critical density (see the last subsection), the contribution from these relativistic components can safely be ignored at low redshift.

(b) Baryonic Components Stars are made up of baryonic matter, and so a lower limit on the mass density of baryonic matter can be obtained by estimating the mass density of stars in galaxies. The mean luminosity density of stars in galaxies can be obtained from the galaxy luminosity function (see §2.4.1). In the B band, the best-fit Schechter function parameters are $\alpha \approx -1.2$, $\phi^* \approx 1.2 \times 10^{-2} h^3\,{\rm Mpc^{-3}}$ and $\mathscr{M}^* \approx -20.05 + 5\log h$ (corresponding to $L^* = 1.24 \times 10^{10} h^{-2}\,{\rm L_\odot}$), so that

$$\mathscr{L}_B \approx 2 \times 10^8 h\,{\rm L_\odot\,Mpc^{-3}}. \tag{2.69}$$

Dividing this into the critical density leads to a value for the mass per unit observed luminosity of galaxies required for the Universe to have the critical density. This critical mass-to-light ratio is

$$\left(\frac{M}{L}\right)_{B,{\rm crit}} = \frac{\rho_{\rm crit}}{\mathscr{L}_B} \approx 1500 h \left(\frac{\rm M_\odot}{\rm L_\odot}\right)_B. \tag{2.70}$$

Mass-to-light ratios for the visible parts of galaxies can be estimated by fitting their spectra with appropriate models of stellar populations. The resulting mass-to-light ratios tend to be in the range of 2 to $10({\rm M_\odot/L_\odot})$. Adopting $M/L = 5({\rm M_\odot/L_\odot})$ as a reasonable mean value, the global density contribution of stars is

$$\Omega_{\star,0} \sim 0.003 h^{-1}. \tag{2.71}$$

Thus, the visible parts of galaxies provide less than 1% of the critical density. In fact, combined with the WMAP constraints on $\Omega_{\rm b,0}$ and the Hubble constant, we find that stars only account for less than 10% of all baryons.

So where are the other 90% of the baryons? At low redshifts, the baryonic mass locked up in cold gas (either atomic or molecular), and detected via either emission or absorption, only accounts for a small fraction, $\Omega_{\rm cold} \sim 0.0005 h^{-1}$ (Fukugita et al., 1998). A larger contribution is due to the hot intracluster gas observed in rich galaxy clusters through their bremsstrahlung emission at X-ray wavelengths (§2.5.1). From the number density of X-ray clusters and their typical gas mass, one can estimate that the total amount of hot gas in clusters is about $(\Omega_{\rm HII})_{\rm cl} \sim 0.0016 h^{-3/2}$ (Fukugita et al., 1998). The total gas mass in groups of galaxies is uncertain. Based on X-ray data, Fukugita et al. obtained $(\Omega_{\rm HII})_{\rm group} \sim 0.003 h^{-3/2}$. However, the plasma in groups is expected to be colder than that in clusters, which makes it more difficult to detect in X-ray radiation. Therefore, the low X-ray emissivity from groups may also be due to low temperatures rather than due to small amounts of plasma. Indeed, if we assume that the gas/total mass

ratio in groups is comparable to that in clusters, then the total gas mass in groups could be larger by a factor of two to three. Even then, the total baryonic mass detected in stars, cold gas and hot gas only accounts for less than 50% of the total baryonic mass inferred from the CMB.

The situation is very different at higher redshifts. As discussed in §2.8, the average density of hydrogen inferred from quasar absorption systems at $z \sim 3$ is roughly equal to the total baryon density as inferred from the CMB data. Hence, although we seem to have detected the majority of all baryons at $z \sim 3$, at low redshifts roughly half of the expected baryonic mass is unaccounted for observationally. One possibility is that the gas has been heated to temperatures in the range 10^5–10^6 K at which it is very difficult to detect. Indeed, recent observations of OVI absorption line systems seem to support the idea that a significant fraction of the IGM at low redshift is part of such a warm-hot intergalactic medium (WHIM), whose origin may be associated with the formation of large-scale sheets and filaments in the matter distribution (see Chapter 16).

An alternative explanation for the 'missing baryons' is that a large fraction of the gas detected at $z \sim 3$ has turned into 'invisible' compact objects, such as brown dwarfs or black holes. The problem, though, is that most of these objects are stellar remnants, and their formation requires a star-formation rate between $z = 3$ and $z = 0$ that is significantly higher than normally assumed. Not only is this inconsistent with the observation of the global star-formation history of the Universe (see §2.6.8), but it would also result in an over-production of metals. This scenario thus seems unlikely. Nevertheless, some observational evidence, albeit controversial, does exist for the presence of a population of compact objects in the dark halo of our Milky Way. In 1986 Bohdan Paczyński proposed to test for the presence of massive compact halo objects (MACHOs) using gravitational lensing. Whenever a MACHO in our Milky Way halo moves across the line-of-sight to a background star (for example, a star in the LMC), it will magnify the flux of the background star, an effect called microlensing. Because of the relative motion of source, lens and observer, this magnification is time-dependent, giving rise to a characteristic light curve of the background source. In the early 1990s two collaborations (MACHO and EROS) started campaigns to monitor millions of stars in the LMC for a period of several years. This has resulted in the detection of about 20 events in total. The analysis by the MACHO collaboration suggests that about 20% of the mass of the halo of the Milky Way could consist of MACHOs with a characteristic mass of $\sim 0.5\,\mathrm{M}_\odot$ (Alcock et al., 2000). The nature of these objects, however, is still unclear. Furthermore, these results are inconsistent with those obtained by the EROS collaboration, which obtained an upper limit for the halo mass fraction in MACHOs of 8%, and rule out MACHOs in the mass range $0.6 \times 10^{-7}\,\mathrm{M}_\odot < M < 15\,\mathrm{M}_\odot$ as the primary occupants of the Milky Way halo (Tisserand et al., 2007).

(c) Non-Baryonic Dark Matter As is evident from the CMB constraints given by Eq. (2.67) on $\Omega_{\mathrm{m},0}$ and $\Omega_{\mathrm{b},0}$, baryons can only account for $\sim$ 15–20% of the total matter content in the Universe, and this is supported by a wide range of observations. As we will see in the following chapters, constraints from a number of other measurements, such as cosmic shear, the abundance of massive clusters, large-scale structure, and the peculiar velocity field of galaxies, all agree that $\Omega_{\mathrm{m},0}$ is of the order of 0.3. At the same time, the total baryonic matter density inferred from CMB observations is in excellent agreement with independent constraints from nucleosynthesis and the observed abundances of primordial elements. The inference is that the majority of the matter in the Universe (75–80%) must be in some non-baryonic form.

One of the most challenging tasks for modern cosmology is to determine the nature and origin of this dark matter component. Particle physics in principle allows for a variety of candidate particles, but without a direct detection it is and will be difficult to discriminate between

the various candidates. One thing that is clear from observations is that the distribution of dark matter is typically more extended than that of the luminous matter. As we have seen above, the mass-to-light ratios increase from $M/L \sim 30h(M/L)_\odot$ at a radius of about $30\,h^{-1}$kpc as inferred from the extended rotation curves of spiral galaxies, to $M/L \sim 100h(M/L)_\odot$ at the scale of a few hundred kpc, as inferred from the kinematics of galaxies in groups, to $M/L \sim 350h(M/L)_\odot$ in galaxy clusters, probing scales of the order of 1 Mpc. This latter value is comparable to that of the Universe as a whole, which follows from multiplying the critical mass-to-light ratio given by Eq. (2.70) with $\Omega_{\rm m,0}$, and suggests that the content of clusters, which are the largest virialized structures known, is representative of that of the entire Universe.

All these observations support the idea that galaxies reside in extended halos of dark matter. This in turn puts some constraints on the nature of the dark matter, namely that it has to be relatively cold (i.e. it needs to have initial peculiar velocities that are much smaller than the typical velocity dispersion within an individual galaxy). This coldness is required because otherwise the dark matter would not be able to cluster on galactic scales to form the dark halos around galaxies. Without a better understanding of the nature of the dark matter, we have to live with the vague term, cold dark matter (or CDM), when talking about the main mass component of the Universe.

(d) Dark Energy As we have seen above, the observed temperature fluctuations in the CMB show that the Universe is nearly flat, implying that the mean energy density of the Universe must be close to the critical density, $\rho_{\rm crit}$. However, studies of the kinematics of galaxies and of large-scale structure in the Universe give a mean mass density that is only about 1/4 to 1/3 of the critical density, in good agreement with the constraints on $\Omega_{\rm m,0}$ from the CMB itself. This suggests that the dominant component of the mass/energy content of the Universe must have a homogenous distribution so that it affects the geometry of the Universe but does not follow the structure in the baryonic and dark matter. An important clue about this dominant component is provided by the observed redshift–distance relation of high-redshift Type Ia supernovae. As shown in §2.10.1, this relation implies that the expansion of the Universe is speeding up at the present time. Since all matter, both baryonic and non-baryonic, decelerates the expansion of the Universe, the dominant component must be an energy component. It must also be extremely dark, because otherwise it would have been observed.

The nature of this dark energy component is a complete mystery at the present time. As far as its effect on the expansion of the Universe is concerned, it is similar to the cosmological constant introduced by Einstein in his theory of general relativity to achieve a stationary Universe (Einstein, 1917). The cosmological constant can be considered as an energy component whose density does not change with time. As the Universe expands, it appears as if more and more energy is created to fill the space. This strange property is due to its peculiar equation of state that relates its pressure, P, to its energy density, ρ. In general, we may write $P = w\rho c^2$, and so $w = 0$ for a pressureless fluid and $w = 1/3$ for a radiation field (see §3.1.5). For a dark energy component with constant energy density, $w = -1$, which means that the fluid actually gains internal energy as it expands, and acts as a gravitational source with a negative effective mass density ($\rho + 3P/c^2 = -2\rho < 0$), causing the expansion of the Universe to accelerate. In addition to the cosmological constant, dark energy may also be related to a scalar field (with $-1 < w < -1/3$). Such a form of dark energy is called quintessence, which differs from a cosmological constant in that it is dynamic, meaning that its density and equation of state can vary through both space and time. It has also been proposed that dark energy has an equation of state parameter $w < -1$, in which case it is called phantom energy. Clearly, a measurement of the value of w will allow us to discriminate between these different models. Currently, the value of w is constrained by a number of observations to be within a relatively

narrow range around -1 (e.g. Spergel et al., 2007), consistent with a cosmological constant, but also with both quintessence and phantom energy. The next generation of galaxy redshift surveys and Type Ia supernova searches aim to constrain the value of w to a few percent, in the hope of learning more about the nature of this mysterious and dominant energy component of our Universe.

3

Cosmological Background

Cosmology, the branch of science dealing with the origin, evolution and structure of the Universe on large scales, is closely related to the study of galaxy formation and evolution. Cosmology provides not only the space-time frame within which galaxy formation and evolution ought to be described, but also the initial conditions for the formation of galaxies. Modern cosmology is founded upon Einstein's theory of general relativity (GR), according to which the space-time structure of the Universe is determined by the matter distribution within it. This perspective on space-time is very different from that in classical physics, where space-time is considered eternal and absolute, independent of the existence of matter.

A complete description of GR is beyond the scope of this book. As a remedy, we provide a brief summary of the basics of GR in Appendix A and we refer the reader to the references cited there for details. It should be emphasized, however, that modern cosmology is a very simple application of GR, so simple that even a reader with little knowledge of GR can still learn it. This simplicity is owing to the simple form of the matter distribution in the Universe, which, as we have seen in the last chapter, is observed to be approximately homogeneous and isotropic on large scales. We do not yet have sufficient evidence to rule out inhomogeneity or anisotropy on very large scales, but the assumption of homogeneity and isotropy is no doubt a good basis for studying the observable Universe. If indeed the matter distribution in the Universe is completely homogeneous and isotropic, as is the ansatz on which modern cosmology is based,[1] GR would imply that space itself must also be homogeneous and isotropic. Such a space is the simplest among all possibilities. To see this more clearly, let us consider a two-dimensional space, i.e. a surface. We all know that the properties of a general two-dimensional surface can be very complicated. But if the surface is homogeneous and isotropic, we are immediately reminded of an infinite plane and a sphere. These two surfaces differ in their overall curvature. The plane is flat, while the sphere is said to have a positive curvature. In both cases the distance between any two infinitesimally close points on the surface can be written as

$$\mathrm{d}l^2 = a^2\left(\frac{\mathrm{d}r^2}{1-Kr^2} + r^2\,\mathrm{d}\vartheta^2\right), \tag{3.1}$$

where $K = 0$ for a plane and $K = 1$ for a sphere. In the case of a plane, (r, ϑ) are just the polar coordinates and a is a length scale (scale factor) relating the coordinate radius r to distance. To see that $K = 1$ corresponds to a sphere, we make the coordinate transformation $r = \sin\chi$. In terms of (χ, ϑ), the distance measure becomes $\mathrm{d}l^2 = a^2(\mathrm{d}\chi^2 + \sin^2\chi\,\mathrm{d}\vartheta^2)$, which is clearly that of a sphere in terms of the spherical coordinates, with a being the radius of the sphere. In this case, r is a spherical coordinate in the three-dimensional space in which the two-dimensional surface is embedded; r is *not* a distance measure on the surface, but

[1] Although on relatively small scales the present-day Universe deviates strongly from homogeneity and isotropy, we will see in Chapter 4 that these structures arise from small perturbations of an otherwise homogeneous and isotropic matter distribution.

rather a coordinate used to label positions on the surface. Actual distances have to be computed from the metric (3.1). Only in the case with $K = 0$ is r both coordinate and distance measure.

Mathematically it can be shown that there is another two-dimensional homogeneous and isotropic surface for which $K = -1$. Changing r to $\sinh\chi$, we can write the distance measure on such a surface as $\mathrm{d}l^2 = a^2(\mathrm{d}\chi^2 + \sinh^2\chi\,\mathrm{d}\vartheta^2)$, where the factor a is again a length scale relating coordinates to distance. This negatively curved, hyperbolic surface, which is locally similar to the surface near a saddle point, is not very familiar to us because it cannot be embedded in a three-dimensional Euclidean space. The existence of a low-dimensional 'space' which cannot be embedded in a space of higher dimensionality is, however, not as strange as it might seem; for example it is easy to envision that it is impossible to embed a hairspring (an intrinsically one-dimensional object) into a plane.

These examples show that the description of a homogeneous and isotropic two-dimensional surface is extremely simple. What we need to do is just to determine the value of K (1, 0, or -1), which specifies the global geometry of the surface, and the scale factor a, which relates coordinates to distances. In general, the scale factor a can change with time without violating the requirement of homogeneity and isotropy, corresponding to a surface that is uniformly expanding or contracting.

The above discussion can be extended to three-dimensional spaces. As we will see in §3.1, a homogeneous and isotropic space is also completely determined by the curvature signature K (again equal to 1 or 0 or -1), which determines the global geometry of the space, and the scale factor $a(t)$ as a function of time. Thus, as far as the space-time geometry is concerned, the task of modern cosmology is simply to determine the value of K and the functional form of $a(t)$ from the matter content of the Universe (see §3.2).

According to GR, the relationships among cosmological events are assumed to be governed by the physical laws that we are familiar with, while the effects of gravity are included in the properties of the space-time (i.e. in the transformations of reference frames). This equivalence principle (that a local gravitational field can be transformed away by choosing an appropriate frame of reference) allows one to derive physical equations in GR from their ordinary forms by general coordinate transformations (see Appendix A). Hence, once the value of K and the functional form of $a(t)$ are known, the relationships among cosmological events can be described in terms of physical laws. Similarly, if we believe that physical laws are applicable on cosmological scales, the predictions for these relationships will depend only on the space-time geometry, and so observations of such relationships can be used to test cosmological models.

One of the most important observations in cosmology is that the Universe is expanding [i.e. $a(t)$ increases with time], which implies that it must have been smaller in the past. Together with the observational fact that our Universe is filled with microwave photons, this time evolution of the scale factor determines the thermal history of the Universe. Because the Universe was denser in the past, it must also have been hotter. Since high density and temperature imply high probabilities for particles to collide with each other with high energy, the early Universe is an ideal place for the creation and transmutation of matter. As we will see in §3.3–§3.5, the applications of particle, nuclear and atomic physics to the thermal history of the early Universe lead to important predictions for the current matter content of the Universe. Although many of these predictions are still uncertain, they provide the basis for calculating relations between the dominant mass components of the Universe. Finally, in §3.6, we discuss some of the most fundamental problems of the standard model and show how the 'inflationary hypothesis' may help to solve them. Although this chapter gives a fairly detailed description of modern cosmology, readers interested in more details are referred to the textbooks by Kolb & Turner (1990), Peebles (1993), Peacock (1999), Coles & Lucchin (2002), Padmanabhan (2002), Börner (2003) and Weinberg (2008).

3.1 The Cosmological Principle and the Robertson–Walker Metric

3.1.1 The Cosmological Principle and its Consequences

The cosmological principle is the hypothesis that, on sufficiently large scales, the Universe can be considered spatially homogeneous and isotropic. While this may appear a reasonable extrapolation from current observations (see Chapter 2), it was originally proposed for quite different reasons. As stated by Milne (1935), this hypothesis follows from the belief that 'Not only the laws of Nature, but also the events occurring in Nature, the world itself, must appear the same to all observers.' In this sense, the cosmological principle can be thought of as a generalized Copernican principle: our location in the Universe should be typical, and should not be distinguished in any fundamental way from any other. The cosmological principle is, however, stronger than this simple statement implies, since it also eliminates the possibility of a self-similar, fractal structure on the largest scales. All points of such a structure are equivalent, but there are no scales on which it approaches homogeneity. Milne's statement is also incomplete, since it is possible to have a universe which appears the same from each point but is anisotropic, as in Gödel's model (Gödel, 1949).

An even stronger hypothesis is the perfect cosmological principle of Bondi & Gold (1948) and Hoyle (1948). This requires invariance not only under rotations and displacements in space, but also under displacements in time. The Universe looks the same in all directions, from all locations, and at all times. This hypothesis led to the steady state cosmology which requires a continuous creation of matter to keep the mean matter density constant with time. However, the discovery of the cosmic microwave background radiation, and in particular the demonstration that it has a perfect blackbody spectrum, has proven an unsurmountable problem for this cosmology. Additional evidence against the steady state cosmology comes from numerous detections of evolution in the galaxy population. We therefore will not discuss this theory further in this book.

What are the consequences of the cosmological principle for the geometric structure of the Universe? To answer this question, we put the cosmological principle in a slightly different form. The cosmological principle can also be stated as the existence of a *fundamental observer* at each location, to whom the Universe appears isotropic. The concept of a fundamental observer is required because two observers at the same point, but in relative motion, cannot both see the surrounding Universe as isotropic. The fundamental observer thus defines a cosmological 'rest frame' at each location in space. To better understand the meaning of a fundamental observer, let us define the fundamental observer, or the cosmological rest frame, in our neighborhood. As discussed in Chapter 2, galaxies in the Universe are strongly clustered on scales $\lesssim 10h^{-1}\mathrm{Mpc}$, and have random motions of the order of 100 to 1,000 $\mathrm{km\,s^{-1}}$ with respect to each other. It is thus unlikely that our own Galaxy defines a cosmological rest frame. On the other hand, we expect the mean motion of galaxies within a radius much larger than $10h^{-1}\mathrm{Mpc}$ around us to be small with respect to the cosmological rest frame. In particular the cosmic microwave background (CMB) should appear isotropic to such a frame. As shown in Chapter 2, the CMB map given by the COBE satellite appears very isotropic around us, when the dipole component is subtracted. The dipole in the CMB map is best explained by the motion of the Local Group of galaxies relative to the CMB with a velocity $(627 \pm 22)\,\mathrm{km\,s^{-1}}$ (Lineweaver et al., 1996). Thus, an observer in our neighborhood, traveling at the same speed relative to the Local Group but in the opposite direction, should be close to a fundamental observer. If the cosmological principle is correct, then the rest frame defined by the mean motion of galaxies within a large radius around us should converge to the one defined by the CMB. There are indeed indications of such convergence in present observational data (e.g. Schmoldt et al., 1999).

Since the Universe is isotropic to a fundamental observer, the velocity field in her neighborhood cannot have any preferred direction. The only allowed motion is therefore pure expansion (or pure contraction),

$$\delta \mathbf{v} = H\, \delta \mathbf{x}\,, \tag{3.2}$$

where $\delta\mathbf{x}$ and $\delta\mathbf{v}$ are the position and velocity of a particle relative to the fundamental observer, and H is a constant. Once some definition of distance is adopted, we can consider the set of all observers, O', which are equidistant from a given observer O at some given local time of O. Because of the isotropy, all the observers O' must measure the same local values of density, temperature, expansion rate, and other physical quantities. Furthermore, they must remain equidistant from O at any later time recorded by the clock of O. Thus they can in principle synchronize their clocks using a light signal from O, and once synchronized, the clocks must remain so. Since the original fundamental observer O is arbitrary, this argument shows that there exists a three-dimensional hypersurface in space-time, on which density, temperature, expansion rate, and all other locally defined properties are uniform and evolve according to a universally agreed time. Such a time is called the *cosmic time*. Since quantities such as the temperature of the CMB and the mean density of the Universe are monotonic functions of cosmic expansion, the value of these quantities can be used to label the cosmic time, as we will see below.

The isotropic and homogeneous three-dimensional hypersurfaces discussed above are maximally symmetric. As a result their metric can be written as

$$\mathrm{d}l^2 = a^2(t)\left[\frac{\mathrm{d}r^2}{1-Kr^2} + r^2(\mathrm{d}\vartheta^2 + \sin^2\vartheta\,\mathrm{d}\varphi^2)\right]. \tag{3.3}$$

A proof of this can be found in Weinberg (1972). In this formula $a(t)$ is a time-dependent scale factor which relates the coordinate labels (r,ϑ,φ) of the fundamental observers to true physical distances, and K is a constant which can take the values $+1$, 0, and -1. The radial coordinate r is dimensionless in Eq. (3.3). When physical distances are required, a length scale can be assigned to the scale factor.

To understand better the geometric meanings of $a(t)$ and K, consider an expanding or contracting three-sphere (the three-dimensional analog of the two-dimensional surface of an expanding or shrinking spherical balloon) whose radius is $R(t) = a(t)R_0$ at time t. The scale factor $a(t)$ therefore simply relates the radius of the three-sphere at time t to its *comoving* radius, R_0, whose value does not change as the sphere expands or contracts. (Thus the comoving radius is just the true radius measured in units of the scale factor.) In Cartesian coordinates (x,y,z,w), this three-surface is defined by

$$x^2 + y^2 + z^2 + w^2 = a^2(t)\,R_0^2\,. \tag{3.4}$$

With the change of coordinates from (x,y,z,w) to the polar coordinates (r,ϑ,φ):

$$\begin{cases} x = a(t)\, r \sin\vartheta\cos\varphi \\ y = a(t)\, r \sin\vartheta\sin\varphi \\ z = a(t)\, r\cos\vartheta \\ w = a(t)\left(R_0^2 - r^2\right)^{1/2}, \end{cases} \tag{3.5}$$

the line element in the four-dimensional Euclidean space is

$$\begin{aligned} \mathrm{d}l^2 &= \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2 + \mathrm{d}w^2 \\ &= a^2(t)\left[\frac{\mathrm{d}r^2}{1 - r^2/R_0^2} + r^2(\mathrm{d}\vartheta^2 + \sin^2\vartheta\,\mathrm{d}\varphi^2)\right]. \end{aligned} \tag{3.6}$$

The curvature scalar of such a three-sphere is

$$\mathscr{R} = \frac{6}{R_0^2 a^2(t)} \tag{3.7}$$

(see Appendix A). Comparing Eqs. (3.3) and (3.6) we immediately see that Eq. (3.3) with $K = +1$ is the metric of a three-sphere with comoving radius $R_0 = 1$, and with the true radius at time t given by the value of $a(t)$. This three-sphere has a finite volume $V = 2\pi^2 a^3(t)$, and the dimensionless radial coordinate $r \in [0, 1]$.

For $K = 0$, metric (3.3) is the same as that given by Eq. (3.6) with $R_0 \to \infty$, and so it describes a Euclidean flat space with infinite volume. In this case the scale factor $a(t)$ describes the change of the length scale due to the uniform expansion (or contraction) of the space.

Metric (3.3) with $K = -1$ can be obtained by the replacement $R_0 \to i$ in Eq. (3.6). The same replacement in Eq. (3.7) shows that such a metric describes a negatively curved three-surface with curvature radius set by $a(t)$. Such a three-surface cannot be embedded in a four-dimensional Euclidean space, but can be embedded in a four-dimensional Minkowski space with line element $\mathrm{d}l^2 = \mathrm{d}w^2 - \mathrm{d}x^2 - \mathrm{d}y^2 - \mathrm{d}z^2$. In this space, the negatively curved three-surface with curvature radius $a(t)$ can be written as $x^2 + y^2 + z^2 - w^2 = a^2(t)$. Thus, the metric (3.3) with $K = -1$ describes a hyperbolic three-surface, with unit comoving curvature radius, embedded in a four-dimensional Minkowski space. Such a three-surface has no boundaries and has infinite volume.

3.1.2 Robertson–Walker Metric

Since the isotropic and homogeneous three-dimensional surfaces described above are the space-like hypersurfaces corresponding to a constant cosmic time t, the four-metric of the space-time can be written as

$$\begin{aligned} \mathrm{d}s^2 &= c^2\mathrm{d}t^2 - \mathrm{d}l^2 \\ &= c^2\mathrm{d}t^2 - a^2(t)\left[\frac{\mathrm{d}r^2}{1-Kr^2} + r^2(\mathrm{d}\vartheta^2 + \sin^2\vartheta\,\mathrm{d}\varphi^2)\right], \end{aligned} \tag{3.8}$$

with c the speed of light. This is the Robertson–Walker metric. As in special relativity, the space-time interval, $\mathrm{d}s$, is real for two events with a time-like separation, is zero for two events on the same light path (null geodesic), and is imaginary for two events with a space-like separation. As before, the coordinates (r, ϑ, φ), which label fundamental observers, are called comoving coordinates, and the function $a(t)$ is the cosmic scale factor. If we define the proper time of an observer as the one recorded by the clock at rest with the observer, then the cosmic time t is the proper time of all fundamental observers. A proper distance l can be defined for any two fundamental observers at any given cosmic time t: $l = \int \mathrm{d}l$. Without losing generality we can assume one of the observers to be at the origin $r = 0$ and the other at $(r_1, \vartheta, \varphi)$. The proper distance can then be written as

$$l = a(t)\int_0^{r_1} \frac{\mathrm{d}r}{\sqrt{1-Kr^2}} = a(t)\chi(r_1), \tag{3.9}$$

where

$$\chi(r) = \begin{cases} \sin^{-1} r & (K = +1) \\ r & (K = 0) \\ \sinh^{-1} r & (K = -1). \end{cases} \tag{3.10}$$

The χ in the above equations is called the comoving distance between the two fundamental observers; it is the proper distance l measured in units of the scale factor. It is often useful to change the time variable from proper time t to a conformal time,

$$\tau(t) = \int_0^t \frac{c\,\mathrm{d}t'}{a(t')}\,. \tag{3.11}$$

In terms of χ and τ the Robertson–Walker metric can be written in another useful form:

$$\mathrm{d}s^2 = a^2(\tau)\left[\mathrm{d}\tau^2 - \mathrm{d}\chi^2 - f_K^2(\chi)(\mathrm{d}\vartheta^2 + \sin^2\vartheta\,\mathrm{d}\varphi^2)\right], \tag{3.12}$$

where

$$f_K(\chi) = r = \begin{cases} \sin\chi & (K=+1) \\ \chi & (K=0) \\ \sinh\chi & (K=-1). \end{cases} \tag{3.13}$$

This form of the metric is especially useful to gain insight into the causal properties of space-time.

It is instructive to look at the metric on a hypersurface with constant φ. In the $K=+1$ case the spatial part of the metric is $\mathrm{d}l^2 = a^2(\tau)(\mathrm{d}\chi^2 + \sin^2\chi\,\mathrm{d}\vartheta^2)$, which is just the metric of a two-dimensional sphere in terms of the 'polar angle' χ and the 'azimuthal angle' ϑ (see Fig. 3.1). We see that χ is the (comoving) geodesic distance, because it measures the length of the shortest path (arc) connecting two points on the hypersurface, while the radial coordinate r is *not* a distance measure on the surface. This conclusion is also true for the case of $K=-1$. Only for a flat space ($K=0$) where $r=\chi$, is the radial coordinate r also a geodesic distance.

The Hubble parameter, $H(t)$, at a cosmic time t is defined to be the rate of change of the proper distance l between any two fundamental observers at time t in units of l: $\mathrm{d}l/\mathrm{d}t \equiv H(t)l$. It then follows from Eq. (3.9) that

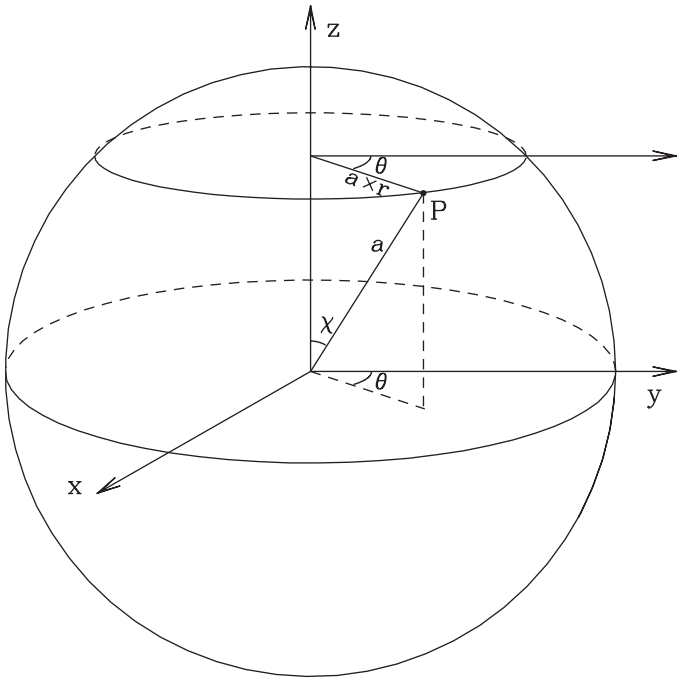

Fig. 3.1. The φ = constant section of a Robertson–Walker metric with $K=1$, showing the geometric meanings of various coordinates.

$$H(t) = \frac{\dot{a}(t)}{a(t)}\,, \tag{3.14}$$

where an over-dot denotes the derivative with respect to t. The Hubble parameter at the present time is called the Hubble constant, and is denoted by H_0. Quantities that depend on the value of H_0 are often expressed in terms of

$$h \equiv \frac{H_0}{100\,\mathrm{km\,s^{-1}\,Mpc^{-1}}}\,. \tag{3.15}$$

The time dependence of the scale factor $a(t)$ is determined by general relativity and the equation of state appropriate for the matter content of the Universe. This will be discussed in §3.2. However, some kinematic properties of an isotropic and homogeneous universe can already be inferred from the form of the metric [either Eq. (3.8) or Eq. (3.12)] without specifying the form of $a(t)$. Such discussion is useful, because it is based only on the cosmological principle, and is valid even if general relativity fails on cosmological scales or if our knowledge about the matter content of the Universe is incomplete. In the following four subsections, we examine these 'kinematic' properties of the Robertson–Walker metric.

3.1.3 Redshift

Almost all observations about astronomical objects are made through light signals. It is therefore important to understand how photons propagate in a homogeneous and isotropic universe. Without losing generality, consider a light signal propagating to the origin along a radial direction ($\mathrm{d}\vartheta = \mathrm{d}\varphi = 0$). Since photons travel along null geodesics on which $\mathrm{d}s = 0$, their trajectories can be written as

$$\mathrm{d}\tau = \mathrm{d}\chi \tag{3.16}$$

[see Eq. (3.12)]. Thus, if a wave crest is emitted at the time t_e from a fundamental observer $(r_\mathrm{e}, \vartheta_\mathrm{e}, \varphi_\mathrm{e})$, then the time t_0 when it reaches the origin is given by

$$\tau(t_0) - \tau(t_\mathrm{e}) = \chi(r_\mathrm{e}) - \chi(0) = \chi(r_\mathrm{e})\,. \tag{3.17}$$

Since the comoving distance $\chi(r_\mathrm{e})$ between the fundamental observer and the origin does not change with time, a successive wave crest emitted at a later time $t_\mathrm{e} + \delta t_\mathrm{e}$ reaches the origin at a time $t_0 + \delta t_0$ given by

$$\tau(t_0 + \delta t_0) - \tau(t_\mathrm{e} + \delta t_\mathrm{e}) = \chi(r_\mathrm{e})\,. \tag{3.18}$$

Combining Eqs. (3.17) and (3.18) gives

$$\tau(t_0 + \delta t_0) - \tau(t_0) = \tau(t_\mathrm{e} + \delta t_\mathrm{e}) - \tau(t_\mathrm{e})\,. \tag{3.19}$$

In real applications $\delta t_\mathrm{e} \ll t_\mathrm{e}$ and $\delta t_0 \ll t_0$, and so we can use the definition of τ to obtain

$$\frac{\delta t_0}{a(t_0)} = \frac{\delta t_\mathrm{e}}{a(t_\mathrm{e})}\,. \tag{3.20}$$

Thus the period of the wave, and hence its wavelength, increases (or its frequency decreases) in proportion to the scale factor:

$$\frac{\lambda_0}{\lambda_\mathrm{e}} = \frac{\nu_\mathrm{e}}{\nu_0} = \frac{\delta t_0}{\delta t_\mathrm{e}} = \frac{a(t_0)}{a(t_\mathrm{e})}\,. \tag{3.21}$$

Defining the relative change of wavelength by a redshift parameter, $z \equiv (\lambda_0 - \lambda_\mathrm{e})/\lambda_\mathrm{e}$, we have

$$1 + z \equiv \frac{\lambda_0}{\lambda_\mathrm{e}} = \frac{a(t_0)}{a(t_\mathrm{e})}\,. \tag{3.22}$$

If the light wave is emitted from the transitions of a given kind of atoms between two energy levels E_1 and E_2, and if these atoms are at rest with respect to the fundamental observer $(r_e, \vartheta_e, \varphi_e)$ at time t_e, then $\nu_e = |E_1 - E_2|/h_P$ (where h_P is Planck's constant). Eq. (3.21) then describes the relation between the observed wavelength and the rest-frame wavelength which can be determined by the observer in his local laboratory. In an expanding universe $a(t_0) > a(t_e)$ so that $z > 0$ and spectral features are shifted redwards (redshift). On the other hand, in a contracting universe $a(t_0) < a(t_e)$, so that $z < 0$ and spectral features are shifted bluewards (blueshift). As we have seen in Chapter 2, distant galaxies in the Universe are all observed to show redshifted spectra, indicating that the Universe is expanding.

3.1.4 Peculiar Velocities

As we will see later in Chapter 4, small perturbations in the background energy density distribution cause the growth of structures, which in turn induce velocities that deviate from pure expansion. These velocities with respect to the cosmological rest frame of fundamental observers are called peculiar velocities.

The proper velocity of a particle with respect to a fundamental observer at the origin is defined as $v = \mathrm{d}l/\mathrm{d}t$, with $l(t)$ the proper distance between the particle and observer. Using Eq. (3.9) we can write this as

$$v(t) = \dot{a}(t)\chi(t) + a(t)\dot{\chi}(t) = v_{\mathrm{exp}} + v_{\mathrm{pec}}\,, \tag{3.23}$$

where $v_{\mathrm{exp}} = H(t)l(t)$ is the velocity component due to the universal expansion, and v_{pec} is the peculiar velocity.

Let $\mathscr{O}_1$ be a fundamental observer at the same location as a particle $\mathscr{P}$ which has a peculiar velocity v_{pec} with respect to $\mathscr{O}_1$. Since locally the geometry at $\mathscr{O}_1$ is that of a Minkowski space, $\mathscr{O}_1$ will observe the light from $\mathscr{P}$ with a Doppler redshift

$$1 + z_{\mathrm{pec}} = \sqrt{\frac{1 + v_{\mathrm{pec}}/c}{1 - v_{\mathrm{pec}}/c}}\,. \tag{3.24}$$

But what is the redshift of $\mathscr{P}$ observed by a fundamental observer $\mathscr{O}_2$, located at a proper distance δl_{12} from $\mathscr{O}_1$? For simplicity we assume that the peculiar velocity of $\mathscr{P}$ is along the geodesic connecting $\mathscr{O}_1$ and $\mathscr{O}_2$. Using the definition of redshift in Eq. (3.22) we can write for the observed redshift

$$1 + z_{\mathrm{obs}} = \frac{\lambda_2}{\lambda_P} = \frac{\lambda_1}{\lambda_P}\,\frac{\lambda_2}{\lambda_1}\,, \tag{3.25}$$

where λ_P is the wavelength emitted by $\mathscr{P}$, and λ_1 and λ_2 are the wavelengths observed by $\mathscr{O}_1$ and $\mathscr{O}_2$, respectively. The physical correspondence of the second equality is a simple relay station at $\mathscr{O}_1$ that passes the information from $\mathscr{P}$ on to $\mathscr{O}_2$. The first factor on the right-hand side of Eq. (3.25) is simply the Doppler redshift of Eq. (3.24), while the second factor corresponds to the cosmological redshift z_{cos} of $\mathscr{O}_1$, and thus also of $\mathscr{P}$. Therefore

$$1 + z_{\mathrm{obs}} = (1 + z_{\mathrm{pec}})(1 + z_{\mathrm{cos}})\,, \tag{3.26}$$

which shows that the observed redshift of any object consists of a contribution due to the universal expansion and one due to its peculiar velocity along the line-of-sight. In the non-relativistic case we can approximate Eq. (3.24) with $z_{\mathrm{pec}} = v_{\mathrm{pec}}/c$, so that Eq. (3.26) reduces to

$$z_{\mathrm{obs}} = z_{\mathrm{cos}} + \frac{v_{\mathrm{pec}}}{c}(1 + z_{\mathrm{cos}})\,. \tag{3.27}$$

Thus, for a cluster of galaxies at redshift z, the (peculiar) velocity dispersion of galaxies, σ_v, is related to the observed dispersion in redshifts, σ_z, as

$$\sigma_v = \sigma_z \frac{c}{1+z}\,. \tag{3.28}$$

Next let us consider the motion of a non-relativistic particle $\mathscr{P}$ in a homogeneous and isotropic universe. Consider once again the fundamental observers $\mathscr{O}_1$ and $\mathscr{O}_2$, and let $\mathscr{P}$ pass $\mathscr{O}_1$ at time t_1 with a peculiar velocity v_1 in the direction of $\mathscr{O}_2$. If $\mathscr{P}$ moves freely to $\mathscr{O}_2$, what is the peculiar velocity of $\mathscr{P}$ when it passes $\mathscr{O}_2$ at time t_2? To answer this question, focus first on the velocity of $\mathscr{P}$ at $t = t_2$ with respect to $\mathscr{O}_1$. This velocity consists of two components: a peculiar velocity v_2 as well as a velocity $v_{\rm exp} = H(t_2)\delta l_{12}$ due to the universal expansion. Since $\mathscr{P}$ has not been accelerated with respect to $\mathscr{O}_1$, the sum of these two velocities has to be equal to v_1, such that from the perspective of $\mathscr{O}_1$ the line-of-sight velocity of $\mathscr{P}$ has not changed. Therefore

$$\delta v \equiv v_2 - v_1 = -\frac{\dot{a}(t_2)}{a(t_2)}\delta l_{12}\,. \tag{3.29}$$

Using Taylor expansion we can write, to first order in $\delta t = t_2 - t_1$, the proper distance between $\mathscr{O}_1$ and $\mathscr{O}_2$ as $\delta l_{12} = v_1 \delta t$. Substitution in Eq. (3.29) and integration then yields

$$v_2 = v_1 \frac{a(t_1)}{a(t_2)}\,. \tag{3.30}$$

Therefore, the peculiar velocity of a free, non-relativistic particle decreases as the inverse of the scale factor:

$$v_{\rm pec}(t) \propto a^{-1}(t)\,. \tag{3.31}$$

Since the momentum p of a non-relativistic particle is proportional to its peculiar velocity, Eq. (3.31) also implies that $p(t) \propto a^{-1}(t)$. Note that for a photon with zero rest mass $pc = E = h_{\rm P}\nu$. As is evident from Eq. (3.21) $\nu \propto a^{-1}$, so that the decay law $p \propto a^{-1}$ holds for photons as well as for massive particles.

3.1.5 Thermodynamics and the Equation of State

The homogeneous and isotropic properties of the expanding Universe also allow an analysis of its thermodynamic properties. Let us consider a uniform, perfect gas contained in a (small) comoving volume $V \propto a^3(t)$ which expands with the Universe. Since the Universe is homogeneous and isotropic, there should not be any net heat flow across the boundaries of V. This implies that we can consider V as an adiabatic system, and since V can be chosen arbitrarily small, no GR is required to describe its thermodynamic properties.

According to the first law of thermodynamics, the increase in internal energy, $\mathrm{d}U$, is equal to the heat, $\mathrm{d}Q$, transferred into the system plus the work, $\mathrm{d}W$, done *on* the system: $\mathrm{d}U = \mathrm{d}Q + \mathrm{d}W$. The second law of thermodynamics is related to the entropy S, and states that $\mathrm{d}S = \mathrm{d}Q/T$, with T the temperature. For our adiabatically expanding volume V we therefore have

$$\mathrm{d}U + P\,\mathrm{d}V = 0\,; \quad \mathrm{d}S = 0\,, \tag{3.32}$$

with P the pressure. This shows that the entropy per unit comoving volume is conserved, and that the expansion of the Universe causes a decrease or increase of its internal energy depending on whether $P > 0$ or $P < 0$.

In order to be able to apply the first law to both relativistic and non-relativistic fluids, we write the internal energy, U, in terms of the energy density ρc^2. In principle there may be many different sources contributing to the energy density of the Universe: matter (both non-relativistic and relativistic), radiation, vacuum energy, scalar fields, etc. As we shall see later in this chapter,

the Universe transited from a radiation dominated phase early on to a matter dominated phase at later stages. In addition, the Universe may have become dominated by vacuum energy in the recent past. In what follows we therefore focus on these three energy components only, so that the total energy density may be written as

$$\rho c^2 = \rho_{\rm m} c^2 + \rho_{\rm m}\varepsilon + \frac{4\sigma_{\rm SB}}{c} T^4 + \rho_{\rm vac} c^2 \,. \tag{3.33}$$

Here $\rho_{\rm m}$ is the matter density, and ε the internal energy per unity mass ($\varepsilon = \frac{3}{2} k_{\rm B} T/m$ for a monatomic ideal gas, with $k_{\rm B}$ Boltzmann's constant). The first two terms of Eq. (3.33) therefore express the energy density due to non-relativistic matter, split in a contribution of rest-mass energy and internal energy. The third term indicates the energy density of the radiation, with $\sigma_{\rm SB}$ the Stefan–Boltzmann constant.[2] Finally, $\rho_{\rm vac} c^2$ is the energy density of the vacuum.

In terms of the energy density, the first law of thermodynamics for our adiabatically expanding volume can now be written as

$$V\,{\rm d}\rho + (\rho + P/c^2){\rm d}V = 0 \,. \tag{3.34}$$

Using that $V \propto a^3$, and differentiating with respect to a we obtain

$$\frac{{\rm d}\rho}{{\rm d}a} + 3\left(\frac{\rho + P/c^2}{a}\right) = 0 \,. \tag{3.35}$$

For a given equation of state, $P(\rho)$, this equation gives the density and pressure as functions of a. It is common practice to introduce the equation of state parameter w and to write

$$P = w\rho c^2 \,. \tag{3.36}$$

If w is time-independent, then substitution of Eq. (3.36) into Eq. (3.35) gives

$$\rho \propto a^{-3(1+w)} \,. \tag{3.37}$$

To describe the evolution of ρ, P, and T during the matter dominated phase, we approximate the Universe as an ideal gas, for which $PV = N k_{\rm B} T$, with N the number of atoms of the gas. For a monatomic gas consisting of particles of mass m we have $\rho_{\rm m} = mN/V$, so that

$$P_{\rm m} = \frac{k_{\rm B} T}{mc^2} \rho_{\rm m} c^2 \,. \tag{3.38}$$

Note that since $\rho_{\rm m} \neq \rho$, this does *not* imply that $w = k_{\rm B} T/mc^2$. To determine the true equation of state parameter, it is useful to write the equation of state as function of the adiabatic index γ (for a monatomic gas $\gamma = 5/3$):

$$P_{\rm m} = (\gamma - 1)(\rho - \rho_{\rm m})c^2 \,. \tag{3.39}$$

Note that Eq. (3.39) makes it explicit that, in the non-relativistic limit, the rest-mass energy does not contribute to the pressure of the gas. Combining Eqs. (3.38)–(3.39) we can write the pressure in the form of Eq. (3.36) with

$$w = w(T) = \frac{k_{\rm B} T}{mc^2} \left(1 + \frac{1}{\gamma - 1} \frac{k_{\rm B} T}{mc^2}\right)^{-1} \,. \tag{3.40}$$

Since $k_{\rm B} T \ll mc^2$ we immediately see that $w(T) \ll 1$. A non-relativistic gas is thus well approximated by a fluid of zero pressure ($w = 0$), often referred to as a dust fluid. Since $\rho \propto a^{-3(1+w)}$ a

[2] Since the Universe is homogeneous and isotropic, the radiation fluid is in thermal equilibrium, and its energy density follows from integrating the Planck function corresponding to a blackbody of temperature T.

Table 3.1. Thermodynamics of a homogeneous and isotropic universe.

Dominant component	w	ρ	P	T
matter	0	a^{-3}	a^{-5}	a^{-2}
radiation	$1/3$	a^{-4}	a^{-4}	a^{-1}
vacuum energy	-1	a^{0}	a^{0}	

dust fluid has $\rho_{\rm m} \propto a^{-3}$, as expected. To obtain the relation between T and a we use kinetic theory which relates the gas temperature to the peculiar motions of the gas particles: $k_{\rm B}T_{\rm m} \propto m\langle v^2\rangle$. Since $v \propto a^{-1}$ [see Eq. (3.31)], we have that $T_{\rm m} \propto a^{-2}$. Finally, using Eq. (3.38) we find that $P_{\rm m} \propto a^{-5}$. This rapid decrease of pressure with the scale factor indicates that the Universe quickly approaches a dust fluid once it becomes matter dominated.

At early times the Universe is radiation dominated. To investigate how ρ, P and T scale with a during this period, we approximate the fluid as an ultra-relativistic radiation fluid for which $w = 1/3$. This implies that $\rho_{\rm r} \propto a^{-4}$, which is consistent with the fact that the number density of photons scales as a^{-3}, while the energy per photon, $E = h_{\rm P}\nu$, scales as a^{-1} [see Eq. (3.21)]. From the equation of state we obtain that $P_{\rm r} \propto a^{-4}$, while the scaling relation for the temperature, $T_{\rm r} \propto a^{-1}$, follows from the fact that for radiation $\rho \propto T^4$ [see Eq. (3.33)]. As a result, a blackbody radiation field remains blackbody with a temperature decreasing as a^{-1}. This is an important result which explains how the cosmic microwave background radiation maintains its blackbody form as the Universe expands.

Finally, if the energy density is dominated by vacuum energy, it only depends on the energy difference between the true and false vacua and so is independent of a. It then follows from Eq. (3.35) that

$$P_{\rm vac} = -\rho_{\rm vac}c^2\,, \tag{3.41}$$

i.e. $w = -1$. This equation of state can be understood as follows: in order to keep a constant energy density $\rho_{\rm vac}$ as the Universe expands, the pressure $P_{\rm vac}$ must be negative so that the $P{\rm d}V$ work in Eq. (3.32) is a positive contribution to the total internal energy in a given comoving volume as it expands.

Although the above relations are derived from the application of thermodynamics to a small volume in the Universe, they are applicable to the Universe as a whole, because the Universe is assumed to be homogeneous and isotropic. These relations are important, because they allow us to obtain the mean density, temperature and pressure of the Universe at any redshift from their values at the present time. Table 3.1 summarizes how energy density, pressure, and temperature evolve with the scale factor a for different dominating components of the energy density. Before we continue, it is important to emphasize that these scaling relations only hold while the equation of state remains constant. In the early Universe, however, the adiabatic cooling due to the expansion of the Universe may cause various particle species to change from relativistic to non-relativistic. During these transitions, the *true* scaling relations follow from an application of the entropy conservation law (see § 3.3).

3.1.6 Angular-Diameter and Luminosity Distances

The comoving distance χ and the proper distance $a(t)\chi$ from a source are not directly observable, because the light from a distant source observed at the present time was emitted at an earlier time. In this subsection we consider two other distances that can be measured directly from astronomical observations. Consider an object of size D and intrinsic luminosity L at some distance d. The observable properties of such an object are the angular size ϑ subtended by the object,

and the flux F. These allow us to define the angular-diameter distance, $d_{\rm A}$, and the luminosity distance, $d_{\rm L}$, according to

$$\vartheta = \frac{D}{d_{\rm A}}\,, \tag{3.42}$$

and

$$F = \frac{L}{4\pi d_{\rm L}^2}\,. \tag{3.43}$$

In a static space, $d_{\rm A} = d_{\rm L} = d$, consistent with our everyday experience. However, when cosmic distances are concerned in an expanding Universe, $d_{\rm A}$, $d_{\rm L}$, and d may all have different values, as we will see in the following.

To obtain an expression for $d_{\rm A}$ in a Robertson–Walker metric, we recall that the proper size D can be considered as the proper distance between two light signals, sent from two points with the same radial coordinate $r_{\rm e}$ at a given cosmic time $t_{\rm e}$, and reaching the origin at the time t_0. Thus, the value of D is just the integral of dl in Eq. (3.8) over the transverse direction:

$$D = a_{\rm e} r_{\rm e} \int {\rm d}\vartheta = \frac{a_0 r_{\rm e}}{1+z}\vartheta\,, \tag{3.44}$$

where $a_0 = a(t_0)$ and $a_{\rm e} = a(t_{\rm e})$. It then follows from Eq. (3.42) that

$$d_{\rm A} = \frac{a_0 r_{\rm e}}{1+z_{\rm e}} = a_{\rm e} r_{\rm e}\,. \tag{3.45}$$

To get an expression for $d_{\rm L}$, we consider a proper area, $\mathscr{A}$, which is at the origin (the position of the observer) and subtends a solid angle, ω, at the object. By definition of the angular-diameter distance $d_{\rm A}$, such a solid angle at the origin corresponds to a proper area $\omega d_{\rm A}^2$ at the position of the object. If the universe were static, this area would, by symmetry arguments, be equal to $\mathscr{A}$. Because of expansion, however, the proper area at the origin subtended by a fixed solid angle at a given object is stretched by a factor in proportion to the square of the scale factor, and so

$$\mathscr{A} = \omega d_{\rm A}^2 (a_0/a_{\rm e})^2 = (a_0 r_{\rm e})^2 \omega\,. \tag{3.46}$$

Without losing generality, we can assume that the object emits monochromatic radiation with rest-frame frequency $\nu_{\rm e}$. The number of photons emitted from the object into the solid angle ω within a time interval $\delta t_{\rm e}$ is $L\delta t_{\rm e}\omega/(4\pi h_{\rm P}\nu_{\rm e})$. If the same number of photons pass through the area $\mathscr{A}$ in a time interval δt_0, we have

$$\frac{L\,\delta t_{\rm e}\,\omega}{4\pi h_{\rm P}\nu_{\rm e}} = \frac{F\,\delta t_0\,\mathscr{A}}{h_{\rm P}\nu_0}\,, \tag{3.47}$$

where ν_0 is the observed frequency of the photons at the origin. It then follows from Eqs. (3.21) and (3.46) that

$$F = \frac{\omega}{4\pi}\frac{L}{\mathscr{A}}\left[\frac{a_{\rm e}}{a_0}\right]^2 = \frac{L}{4\pi[a_0 r_{\rm e}(1+z)]^2}\,. \tag{3.48}$$

The luminosity distance defined in Eq. (3.43) can thus be written as

$$d_{\rm L} = a_0 r_{\rm e}(1+z)\,. \tag{3.49}$$

Since we observe the object using photons, the quantity $a_0 r_{\rm e}$ in the expressions of $d_{\rm L}$ and $d_{\rm A}$ is related to the redshift z by Eqs. (3.17) and (3.22).[3] This relation can be obtained once the dynamical equations have been solved to specify $a(t)$. Although we will address the dynamical

[3] In the case of a flat universe ($K = 0$), $a_0 r_{\rm e}$ is equal to the proper distance between object and observer at the time of observation.

behavior of $a(t)$ in detail in §3.2, we can make a simple approximation by using the first few terms of its Taylor expansion:

$$a(t) = a_0 \left[1 + H_0(t - t_0) - \frac{1}{2} q_0 H_0^2 (t - t_0)^2 + \dots \right], \tag{3.50}$$

where

$$q_0 \equiv -\frac{\ddot{a}_0 a_0}{\dot{a}_0^2} \tag{3.51}$$

is known as the deceleration parameter. Using Eq. (3.17), the power series can be manipulated to give

$$a_0 r_e \approx \frac{c}{H_0} \left[z - \frac{1}{2} z^2 (1 + q_0) + \dots \right]. \tag{3.52}$$

Inserting this into Eqs. (3.44) and (3.48), we can obtain D as a function of ϑ and z, and L as a function of F and z, respectively. Thus, for given values of H_0 and q_0, the proper size D (or the intrinsic luminosity L) of an object can be obtained by measuring its redshift z and its angular size ϑ (or its flux F). Similarly, if the proper sizes (or intrinsic luminosities) of a set of objects are known, one can estimate the values of H_0 and q_0 by measuring ϑ (or F) as a function of redshift. Although this way of using Eq. (3.52) to interpret observational data is common practice, it is valid only for $z \ll 1$. It is therefore preferable to use the exact equations for $a_0 r$ as function of z (derived in the next section) rather than this small-z approximation. Nevertheless, the present values of H_0 and q_0 are often used to characterize cosmological models.

Finally, Eqs. (3.44) and (3.48) can be combined to give the apparent surface brightness of an object,

$$S \equiv \frac{F}{\frac{1}{4}\pi\vartheta^2} = \frac{L}{\pi^2 D^2} (1 + z)^{-4}. \tag{3.53}$$

Unlike d_A and d_L, the apparent surface brightness S is independent of the relationship between $a_0 r_e$ and z_e, and so is independent of the dynamical evolution of $a(t)$. This arises because Eq. (3.53) depends only on the local thermodynamics of the radiation field, and follows, in fact, directly from $S \propto T^4$. For given L and D, the apparent surface brightness decreases with redshift as $(1 + z)^{-4}$, which is usually referred to as cosmological surface brightness dimming.

3.2 Relativistic Cosmology

In general relativity, the geometric properties of space-time are determined by the distribution of matter/energy. The standard model of cosmology arises from the application of general relativity to the very special class of matter/energy distributions implied by the cosmological principle, i.e. homogeneous and isotropic distributions. As we have seen above, the geometric properties of a homogeneous and isotropic universe are described by the Robertson–Walker metric which, in turn, is specified by the scale factor $a(t)$ and the curvature signature K. The task of this section is to obtain an expression for $a(t)$ and the value of K for any given homogeneous and isotropic matter/energy content.

3.2.1 Friedmann Equation

In the standard model of cosmology, the geometry of space-time is determined by the matter/energy content of the Universe through the Einstein field equation (see Appendix A):

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R - g_{\mu\nu}\Lambda = \frac{8\pi G}{c^4} T_{\mu\nu} \,. \tag{3.54}$$

Here $R_{\mu\nu}$ is the Ricci tensor, describing the local curvature of space-time, R is the curvature scalar, $g_{\mu\nu}$ is the metric, $T^{\mu\nu}$ is the energy–momentum tensor of the matter content of the Universe, and Λ is the cosmological constant, which was introduced by Einstein to obtain a static universe. Contracting Eq. (3.54) with $g^{\mu\nu}$ yields the trace of the field equation,

$$R + 4\Lambda = -\frac{8\pi G}{c^4} T \,, \tag{3.55}$$

where $T = T^{\lambda}{}_{\lambda}$. This allows the field equation to be written in the form

$$R_{\mu\nu} + g_{\mu\nu}\Lambda = \frac{8\pi G}{c^4} \left(T_{\mu\nu} - \frac{1}{2} g_{\mu\nu} T \right) . \tag{3.56}$$

For a uniform ideal fluid,

$$T^{\mu\nu} = (\rho + P/c^2) U^{\mu} U^{\nu} - g^{\mu\nu} P \,, \tag{3.57}$$

with ρc^2 the energy density, P the pressure, and $U^{\mu} = c\mathrm{d}x^{\mu}/\mathrm{d}s$ the four velocity of the fluid. In a homogeneous and isotropic universe, the density and pressure depend only on the cosmic time, and the four-velocity is $U^{\mu} = (c, 0, 0, 0)$ (i.e. no peculiar motion is allowed). This implies that $T^{\mu}{}_{\nu} = \mathrm{diag}(\rho c^2, -P, -P, -P)$ and $T = \rho c^2 - 3P$.

For a homogeneous and isotropic universe, $g_{\mu\nu}$ is given by the Robertson–Walker metric, which allows the Ricci tensor $R_{\mu\nu}$ and curvature scalar R to be expressed in terms of the scale factor $a(t)$ and the curvature signature K (see Appendix A). Inserting the results into Eq. (3.56), and using the energy–momentum tensor of a perfect fluid given in Eq. (3.57), one obtains

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3} \left(\rho + 3\frac{P}{c^2} \right) + \frac{\Lambda c^2}{3} \tag{3.58}$$

for the time-time component, and

$$\frac{\ddot{a}}{a} + 2\frac{\dot{a}^2}{a^2} + 2\frac{Kc^2}{a^2} = 4\pi G \left(\rho - \frac{P}{c^2} \right) + \Lambda c^2 \tag{3.59}$$

for the space-space components. It then follows from substituting Eq. (3.58) into Eq. (3.59) that

$$\left(\frac{\dot{a}}{a} \right)^2 = \frac{8\pi G}{3} \rho - \frac{Kc^2}{a^2} + \frac{\Lambda c^2}{3} \,. \tag{3.60}$$

As one sees from Eqs. (3.58)–(3.60), the cosmological constant can be considered as an energy component with 'mass' density $\rho_{\Lambda} = \Lambda c^2 / 8\pi G$ and pressure $P_{\Lambda} = -\rho_{\Lambda} c^2$. Indeed, the term of Einstein's cosmological constant in Eq. (3.54) can be included as an energy–momentum tensor, $T_{\mu\nu} = (c^4 \Lambda / 8\pi G) g_{\mu\nu}$, on the right-hand side of the field equation.

Eq. (3.60) is the Friedmann equation, and a cosmology that obeys it is called a Friedmann–Robertson–Walker (FRW) cosmology. Together with Eq. (3.35), an equation of state, and an initial condition, it determines the time dependence of a, ρ, P, and other properties of the Universe.

It is interesting to note that one can derive the Friedmann equation (without the cosmological constant term) for a matter dominated universe purely from Newtonian gravity (see §2.10). This follows from the assumption that the Universe is homogeneous and isotropic so that the global

properties of the Universe can be represented by those in a small region where Newtonian physics applies. The Newtonian derivation, however, does not contain the pressure term, $3P/c^2$, in the equation for the acceleration, which can be considered a relativistic correction. As is evident from Eq. (3.58), in general relativity this pressure term acts as a source of gravity.

The density which appears in Eq. (3.60) can be made up of various components. At the moment we distinguish a non-relativistic matter component, a radiation component, and a possible vacuum energy (cosmological constant) component. We denote their energy densities (written in terms of mass densities) at the present time t_0 by $\rho_{\mathrm{m},0}$, $\rho_{\mathrm{r},0}$ and $\rho_{\Lambda,0}$, respectively. As the Universe expands, these quantities scale with a in different ways, as described in §3.1.5. We can then write the Friedmann equation as

$$\left(\frac{\dot{a}}{a}\right)^2 = H^2(t) = \frac{8\pi G}{3}\left[\rho_{\mathrm{m},0}\left(\frac{a_0}{a}\right)^3 + \rho_{\mathrm{r},0}\left(\frac{a_0}{a}\right)^4 + \rho_{\Lambda,0}\right] - \frac{Kc^2}{a^2}\,, \tag{3.61}$$

where $a_0 = a(t_0)$.[4] Using the fact that the Universe is in its expanding phase at the present time (i.e. $H_0 = \dot{a}_0/a_0 > 0$), we can examine the behavior of $a(t)$ in various cases, even without solving the Friedmann equation explicitly.

If $\Lambda \geq 0$ and if $K = 0$ or $K = -1$, the right-hand side of Eq. (3.61) is always larger than zero, and $a(t)$ always increases with t. If $K = +1$ and $\Lambda = 0$, the right-hand side of Eq. (3.61) becomes zero in the future as the scale factor increases until the curvature term, K/a^2, is as large as the sum of the matter and radiation terms. Thereafter $a(t)$ decreases with t, and the Universe contracts until $a = 0$. If $K = +1$ and $\Lambda > 0$, the situation is similar to that with $K = +1$ and $\Lambda = 0$, provided that the Λ term in Eq. (3.61) is smaller than the matter plus radiation terms at the present time. If the Λ term is sufficiently large at the present time, there may have been a minimum value of a at some previous epoch. This corresponds to a time when the right-hand side of Eq. (3.61) is equal to zero, and an initially contracting universe 'bounced' on its vacuum energy density and started to re-expand. As one can see from Eq. (3.61), this re-expansion will continue forever. For positive Λ a static (but unstable) solution is also possible – Einstein's original static model – as are solutions which asymptotically approach this model in the infinite future or infinite past. Finally, if $\Lambda < 0$, the expansion will eventually halt and be followed by recollapse, giving a history qualitatively similar to that of a $K = +1$, $\Lambda = 0$ universe.

3.2.2 The Densities at the Present Time

To solve Eq. (3.61), we need to know K and the various densities at the present time, $\rho_{\mathrm{m},0}$, $\rho_{\mathrm{r},0}$ and $\rho_{\Lambda,0}$. Here we summarize constraints on these quantities based on observational and theoretical considerations.

The total rest mass density of non-relativistic matter in the Universe is conventionally expressed as

$$\rho_{\mathrm{m},0} = \Omega_{\mathrm{m},0}\rho_{\mathrm{crit},0} \approx 1.88 \times 10^{-29}\Omega_{\mathrm{m},0}h^2\,\mathrm{g\,cm^{-3}}\,, \tag{3.62}$$

where, for reasons that will soon become clear, the density

$$\rho_{\mathrm{crit}}(t) \equiv \frac{3H^2(t)}{8\pi G} \tag{3.63}$$

is known as the *critical* density at time t. The subscript '0' denotes the values at the present time. The dimensionless quantity, $\Omega_{\mathrm{m},0}$, is the present cosmic density parameter for non-relativistic

[4] Note, however, that Eq. (3.61) only applies if there is no transformation from one density component to another. If such transformation occurs, the time dependence of the equation of state must be taken into account.

matter, and h is defined in Eq. (3.15). As discussed in §2.10, current observational constraints suggest

$$\Omega_{\mathrm{m},0} = 0.27 \pm 0.05; \quad h = 0.72 \pm 0.05\,. \tag{3.64}$$

The current density in the relativistic component appears to be dominated by the cosmic microwave background which is, to high accuracy, a blackbody at temperature $T_\gamma = 2.73\,\mathrm{K}$. Thus, using $\rho_\gamma = 4\sigma_{\mathrm{SB}} T^4/c^3$ with σ_{SB} the Stefan–Boltzmann constant, we have

$$\rho_{\gamma,0} \approx 4.7 \times 10^{-34}\,\mathrm{g\,cm^{-3}} \quad \text{or} \quad \Omega_{\gamma,0} \equiv \rho_{\gamma,0}/\rho_{\mathrm{crit},0} \approx 2.5 \times 10^{-5} h^{-2}\,. \tag{3.65}$$

In addition, if the three species of neutrinos and their antiparticles are all massless (or relativistic at the present time), they will have a temperature $T_\nu = (4/11)^{1/3} T_\gamma$ (see §3.3). Because each neutrino has only one spin state (while a photon has two) and because neutrinos are fermions (and so for a given temperature the statistical weight of each degree of freedom is only $7/8$ of that for photons; see §3.3 for details), the energy density in neutrinos at the present time is $3 \times (7/8) \times (4/11)^{4/3}$ times that of the CMB photons. This brings the total energy density in the relativistic component to

$$\rho_{\mathrm{r},0} \approx 7.8 \times 10^{-34}\,\mathrm{g\,cm^{-3}} \quad \text{or} \quad \Omega_{\mathrm{r},0} \approx 4.2 \times 10^{-5} h^{-2}\,. \tag{3.66}$$

Combining Eqs. (3.62) and (3.66) shows that the ratio of the energy densities in the non-relativistic and relativistic components varies with redshift as

$$\frac{\rho_{\mathrm{m}}}{\rho_{\mathrm{r}}} \approx 2.4 \times 10^4 \Omega_{\mathrm{m},0} h^2 (1+z)^{-1}\,, \tag{3.67}$$

where we have used that $\rho_{\mathrm{m}} \propto a^{-3}$ and $\rho_{\mathrm{r}} \propto a^{-4}$ (see Table 3.1). Thus, provided the Universe did not bounce in the recent past due to a large cosmological constant, it has been matter dominated and effectively pressure-free since the epoch of matter/radiation equality defined by $\rho_{\mathrm{r}} = \rho_{\mathrm{m}}$, i.e. since the redshift given by

$$1 + z_{\mathrm{eq}} \approx 2.4 \times 10^4 \Omega_{\mathrm{m},0} h^2\,. \tag{3.68}$$

To constrain the present day energy density provided by the cosmological constant, we use the Friedmann equation (3.61), which we rewrite as

$$\frac{8\pi G}{3}\rho_{\Lambda,0} = H_0^2[1 - \Omega_{\mathrm{m},0} - \Omega_{\mathrm{r},0}] + \frac{Kc^2}{a_0^2}\,. \tag{3.69}$$

As discussed in §2.9, observations of the microwave background show that our Universe is almost flat and that the current density in non-relativistic matter is significant [see Eq. (3.64)]. This excludes the possibility of a bounce in the recent past due to a large cosmological constant. Such an expansion history is also excluded by the observation of objects out to redshifts beyond 6, so we will not consider such cosmological models any further. Setting $K = 0$ in Eq. (3.69) we obtain

$$\rho_{\Lambda,0} = \rho_{\mathrm{crit},0}(1 - \Omega_{\mathrm{m},0} - \Omega_{\mathrm{r},0}) \quad \text{i.e.} \quad \Omega_{\Lambda,0} = 1 - \Omega_{\mathrm{m},0} - \Omega_{\mathrm{r},0}\,. \tag{3.70}$$

Data from WMAP combined with other observations give $\Omega_{\Lambda,0} \sim 0.75 \pm 0.02$ (Spergel et al., 2007).

3.2.3 Explicit Solutions of the Friedmann Equation

(a) The Evolution of Cosmological Quantities Taking $t = t_0$, the Friedmann equation can be rewritten as

$$\Omega_{K,0} \equiv -\frac{Kc^2}{H_0^2 a_0^2} = 1 - \Omega_0\,, \tag{3.71}$$

where

$$\Omega_0 = \Omega_{\rm m,0} + \Omega_{\Lambda,0} + \Omega_{\rm r,0} \tag{3.72}$$

is the total density parameter at the present time. As is immediately evident from Eq. (3.71), the curvature of space-time depends on the matter density of the Universe. In particular, Ω_0 is less than 1 for a negatively curved, open universe, is equal to 1 for a flat universe, and is bigger than 1 for a positively curved, closed universe. The terminology 'open' and 'closed' only has a logical meaning for a $\Lambda = 0$ universe; open (and flat) universes expand forever, while closed universes recollapse in the future. For non-zero Λ, however, open and flat universes can recollapse and closed universes can expand forever, depending on the values of the various density parameters (see discussion at the end of §3.2.1). Since Ω_0 is just the total energy density of the Universe in units of $\rho_{\rm crit,0}$, it follows that $\rho_{\rm crit,0}$ defines a critical density for closure. Note that Eq. (3.71) defines the scale factor a_0 at the present time:

$$a_0 = \frac{c}{H_0}\sqrt{\frac{K}{\Omega_0 - 1}}\,, \tag{3.73}$$

which goes to infinity as Ω_0 approaches 1 from either side. This follows from our definition of the coordinate r in Eqs. (3.3) and (3.8). Since a_0 is only a scale factor, its value does not have physical meaning and so can be set to any positive value. A choice for the value of a_0 corresponds to a choice in the definition of the coordinate r. In fact, physical distances are all related to a_0 through the combination $a_0 r$, which is well behaved near $\Omega_0 = 1$ and independent of the choice of a_0. It is common practice to adopt $a_0 = 1$.

Substituting Eq. (3.71) into Eq. (3.61) gives

$$H(z) \equiv \left(\frac{\dot{a}}{a}\right)(z) = H_0 E(z)\,, \tag{3.74}$$

where

$$E(z) = \left[\Omega_{\Lambda,0} + (1-\Omega_0)(1+z)^2 + \Omega_{\rm m,0}(1+z)^3 + \Omega_{\rm r,0}(1+z)^4\right]^{1/2}. \tag{3.75}$$

Defining the cosmic density parameters at cosmic time t as

$$\Omega(t) \equiv \frac{\rho(t)}{\rho_{\rm crit}(t)}\,, \tag{3.76}$$

we have

$$\Omega_\Lambda(z) = \frac{\Omega_{\Lambda,0}}{E^2(z)}\,; \quad \Omega_{\rm m}(z) = \frac{\Omega_{\rm m,0}(1+z)^3}{E^2(z)}\,; \quad \Omega_{\rm r}(z) = \frac{\Omega_{\rm r,0}(1+z)^4}{E^2(z)}\,. \tag{3.77}$$

Thus, once H, Ω_Λ, $\Omega_{\rm m}$ and $\Omega_{\rm r}$ are known at the present time, Eqs. (3.74)–(3.77) can be used to obtain their values at any given redshift. It is also clear from Eqs. (3.61) and (3.71) that the geometry of a FRW universe is completely determined by the values of H_0, $\Omega_{\Lambda,0}$, $\Omega_{\rm m,0}$ and $\Omega_{\rm r,0}$. Since $\Omega_{\rm r,0} \ll \Omega_{\rm m,0}$ (see §3.2.2), the deceleration parameter, q_0, defined in Eq. (3.51) can be written as

$$q_0 = \Omega_{\rm m,0}/2 - \Omega_{\Lambda,0}\,, \tag{3.78}$$

where we have used Eq. (3.58) with $P = 0$, as appropriate for a matter dominated universe.

Finally, using Eq. (3.71) and the definition of $E(z)$, we can write down the redshift evolution of the total density parameter

$$\Omega(z) - 1 = (\Omega_0 - 1)\frac{(1+z)^2}{E^2(z)}\,. \tag{3.79}$$

As long as $\Omega_{\rm m,0}$ or $\Omega_{\rm r,0}$ are non-zero, $\Omega(z)$ always approaches unity at high redshifts, independent of the present day values of H_0, $\Omega_{\Lambda,0}$, $\Omega_{\rm m,0}$ and $\Omega_{\rm r,0}$. Therefore, every FRW universe with non-zero matter or radiation content must have started out with a total density parameter very close to unity. As we will see in §3.6, this results in the so-called flatness problem.

(b) Radiation Dominated Epoch In the absence of a contracting phase in the past, the right-hand side of Eq. (3.61) is dominated by the radiation term at $z \gg z_{\rm eq}$. In this case, integration of Eq. (3.61) yields

$$\frac{a}{a_0} = \left(\frac{32\pi G\rho_{\rm r,0}}{3}\right)^{1/4} t^{1/2}. \tag{3.80}$$

Using that $\rho_{\rm r} \propto a^{-4}$, $\rho_{\rm m} \propto a^{-3}$ and $T_{\rm r} \propto a^{-1}$ (see Table 3.1), this gives the following rough scalings for the early Universe:

$$\frac{T}{10^{10}\,{\rm K}} \sim \frac{k_{\rm B}T}{1\,{\rm MeV}} \sim \left[\frac{\rho}{10^7\,{\rm g\,cm^{-3}}}\right]^{1/4} \sim \left[\frac{\rho_{\rm m}}{1\,{\rm g\,cm^{-3}}}\right]^{1/3} \sim \frac{1+z}{10^{10}} \sim \left[\frac{t}{1\,{\rm s}}\right]^{-1/2}. \tag{3.81}$$

These relations are approximately correct for $0 < t < 10^{10}\,{\rm s}$, or $z > 10^5$. The arbitrarily high temperatures and densities which are achieved at sufficiently early times have given this standard cosmological model its generic name, the Hot Big Bang.

(c) Matter Dominated Epoch and $\Omega_{\Lambda,0} = 0$ At redshift $z \ll z_{\rm eq}$, the radiation content of the Universe has little effect on its global dynamics, and assuming $\Lambda = 0$, Eq. (3.61) reduces to

$$\left(\frac{\dot a}{a}\right)^2 = H_0^2\left[\Omega_{\rm m,0}\left(\frac{a_0}{a}\right)^3 - \frac{Kc^2}{H_0^2a_0^2}\left(\frac{a_0}{a}\right)^2\right]. \tag{3.82}$$

For $K = 0$ the solution is particularly simple:

$$\frac{a}{a_0} = \left(\frac{3}{2}H_0t\right)^{2/3}. \tag{3.83}$$

This is the solution for an Einstein–de Sitter (EdS) universe. For $K = -1$, the solution can be expressed in parametric form:

$$\frac{a}{a_0} = \frac{1}{2}\frac{\Omega_{\rm m,0}}{(1-\Omega_{\rm m,0})}(\cosh\vartheta - 1); \quad H_0t = \frac{1}{2}\frac{\Omega_{\rm m,0}}{(1-\Omega_{\rm m,0})^{3/2}}(\sinh\vartheta - \vartheta), \tag{3.84}$$

where ϑ goes from 0 to ∞. At early epochs, $a \propto t^{2/3}$, which follows directly from the fact that the curvature term in Eq. (3.82) can be neglected when a is sufficiently small. At later epochs when $\vartheta \gg 1$ and $\sinh\vartheta = \cosh\vartheta$ so that $a \propto t$, the universe enters a phase of free expansion, unretarded by gravity.

The corresponding parametric solution for a $K = +1$ universe is

$$\frac{a}{a_0} = \frac{1}{2}\frac{\Omega_{\rm m,0}}{(\Omega_{\rm m,0}-1)}(1-\cos\vartheta); \quad H_0t = \frac{1}{2}\frac{\Omega_{\rm m,0}}{(\Omega_{\rm m,0}-1)^{3/2}}(\vartheta - \sin\vartheta), \tag{3.85}$$

where $0 \le \vartheta \le 2\pi$. Such models reach a maximum size, $a_{\rm max}$, at a time, $t_{\rm max}$, given by

$$\frac{a_{\rm max}}{a_0} = \frac{\Omega_{\rm m,0}}{\Omega_{\rm m,0}-1}; \quad H_0t_{\rm max} = \frac{\pi}{2}\frac{\Omega_{\rm m,0}}{(\Omega_{\rm m,0}-1)^{3/2}}. \tag{3.86}$$

This maximum expansion is followed by recollapse to a singularity. At early epochs, $a \propto t^{2/3}$, for the same reason as that for the $K = -1$ case.

Note that H_0t_0 depends only on $\Omega_{\rm m,0}$ in these models. Since the normalization time, t_0, can be chosen arbitrarily, it is easy to see that $H(t)t$ depends only on $\Omega_{\rm m}(t)$.

(d) Flat ($\Omega_{\mathrm{m},0}+\Omega_{\Lambda,0}=1$) Models at $z \ll z_{\mathrm{eq}}$ In this case Eq. (3.61) can be written as

$$\left(\frac{\dot{a}}{a}\right)^2 = H_0^2\left[\Omega_{\mathrm{m},0}\left(\frac{a_0}{a}\right)^3 + \Omega_{\Lambda,0}\right]. \tag{3.87}$$

When the matter term is negligible, the model is called a de Sitter universe for which the solution of Eq. (3.87) is particularly simple:

$$\frac{a}{a_0} = \exp\left[H_0(t-t_0)\right], \tag{3.88}$$

and the universe expands exponentially without an initial singularity. For $0 < \Omega_{\mathrm{m},0} < 1$, using the fact that $H_0 \equiv \dot{a}/a > 0$, Eq. (3.87) can be easily solved to give

$$\frac{a}{a_0} = \left(\frac{\Omega_{\mathrm{m},0}}{\Omega_{\Lambda,0}}\right)^{1/3}\left[\sinh\left(\frac{3}{2}\Omega_{\Lambda,0}^{1/2}H_0 t\right)\right]^{2/3}. \tag{3.89}$$

At early epochs, $a \propto t^{2/3}$ as in an Einstein–de Sitter universe; when t is large, $a \propto \exp(\Omega_{\Lambda,0}^{1/2}H_0 t)$ so that the universe approximates the de Sitter model.

(e) Open and Closed Models with $\Omega_{\Lambda,0} \neq 0$ at $z \ll z_{\mathrm{eq}}$ The Friedmann equation in this case is

$$\left(\frac{\dot{a}}{a}\right)^2 = H_0^2\left[\Omega_{\mathrm{m},0}\left(\frac{a_0}{a}\right)^3 - \frac{Kc^2}{H_0^2 a_0^2}\left(\frac{a_0}{a}\right)^2 + \Omega_{\Lambda,0}\right]. \tag{3.90}$$

This equation can be cast into a dimensionless form:

$$\frac{1}{2}\left(\frac{\mathrm{d}x}{\mathrm{d}\eta}\right)^2 = \frac{1}{x} - \kappa + \lambda x^2, \tag{3.91}$$

where $x = a/a_0$, $\eta = \sqrt{\Omega_{\mathrm{m},0}/2}\,H_0 t$, $\lambda = \Omega_{\Lambda,0}/\Omega_{\mathrm{m},0}$, and $\kappa = Kc^2/(H_0^2 a_0^2 \Omega_{\mathrm{m},0})$. The evolution of x can thus be discussed in terms of the Newtonian motion of a particle with total energy $\varepsilon = -\kappa$ in a potential $\phi(x) = -1/x - \lambda x^2$.

When $\lambda < 0$, the potential $\phi(x)$ monotonically increases from 0 to ∞ so that x is confined, and all solutions evolve from an initial singularity into a final singularity.

When $\lambda > 0$, the potential $\phi(x)$ is always negative, and x can go to infinity if $\varepsilon > 0$ or $K = -1$. Hence an open universe with $\Omega_{\Lambda,0} > 0$ expands from an initial singularity forever. If $\lambda > 0$ and $K = +1$, the potential $\phi(x)$ has a maximum, $\phi_{\max} = -(27\lambda/4)^{1/3}$, at $x = x_{\max} = 1/(2\lambda)^{1/3}$. In this case, if the total energy $\varepsilon > \phi_{\max}$, i.e.

$$\lambda > \lambda_c \equiv \frac{4}{27}\left[\frac{c^2}{H_0^2 a_0^2 \Omega_{\mathrm{m},0}}\right]^3, \tag{3.92}$$

the universe still expands forever, starting from an initial singularity. If, however, $\varepsilon < \phi_{\max}$ or $\lambda < \lambda_c$, then there is the possibility that the universe contracts from large radii to a minimum radius, $a_{\min}$, given by $\phi(a_{\min}/a_0) = \varepsilon$, and expands thereafter to infinity. This happens if the universe starts with a radius $a > a_0 x_{\max}$. If the universe starts with an initial singularity, then it will evolve into a final singularity, giving a situation similar to that of a closed universe without cosmological constant.

If $\lambda > 0$ and $K = +1$, a special situation occurs when $\varepsilon = \phi_{\max}$ or $\lambda = \lambda_c$. In this case, there is a static solution with a constant radius $a_{\mathrm{E}} = a_0/(2\lambda_c)^{1/3}$. Such a model is called the 'Einstein universe'. If the universe expands from an initial singularity, or contracts from a large radius, it will coast asymptotically towards the radius a_{E}. If the universe expands from an initial radius larger than a_{E}, it will do so forever.

3.2.4 Horizons

A light ray emitted by an event (r_e, t_e) reaches an observer at the origin at time, t_0, given by

$$\chi(r_e) = \int_{t_e}^{t_0} \frac{c\,dt}{a(t)} = \int_{a_e}^{a_0} \frac{da}{a} \left[\frac{8\pi G\rho(a)a^2}{3c^2} - K \right]^{-1/2}, \tag{3.93}$$

with $\rho = \rho_m + \rho_r + \rho_\Lambda$. The second equality follows from substituting $dt = da/\dot{a}$ and using the Friedmann equation (3.60). If the t (or a) integral converges, as $t_e \to 0$, to a value $\chi_h = \chi(r_h)$, then there may exist particles (fundamental observers) for which $\chi(r) > \chi_h$ and from which no communication can have reached the origin by time t_0. Such particles (or values of r) are said to lie beyond the particle horizon of the origin at time t_0. From the form of the last integral in Eq. (3.93) it is clear that convergence requires $\rho a^2 \to \infty$ as $a \to 0$. Thus particle horizons exist in a universe which is matter or radiation dominated at the earliest times, but do not exist in a universe which was initially dominated by vacuum energy density. As t_0 increases, χ_h becomes bigger, all particle horizons expand, and signals can be received from more and more distant particles.

If the t integral in Eq. (3.93) converges as $t_0 \to \infty$ (or as t_0 approaches the recollapse time for a universe with a finite lifetime), there may exist events which the observer at the origin will never see, and which therefore can never influence him/her by any physical means. Such events are said to lie beyond the event horizon of this observer. Event horizons exist in closed models and in models that are vacuum dominated at late times, but do not exist in flat or open universes with zero cosmological constant. In the latter case, therefore, any event will eventually be able to influence every fundamental observer in the Universe.

The existence of particle horizons in the Big Bang model has important implications, because it means that many parts of the presently observable Universe may not have been in causal contact at early times. This gives rise to certain difficulties, as we will see in §3.6.

3.2.5 The Age of the Universe

In currently viable models the Universe has been expanding since the Big Bang, so that $\dot{a} > 0$ holds over its entire history. The age of the Universe at redshift z can then be obtained from Eqs. (3.22) and (3.74):

$$t(z) \equiv \int_0^{a(z)} \frac{da}{\dot{a}} = \frac{1}{H_0} \int_z^\infty \frac{dz}{(1+z)E(z)}, \tag{3.94}$$

where $E(z)$ is given by Eq. (3.75). With this, the lookback time at redshift z, defined as $t_0 - t(z)$, can also be obtained. For a given set of cosmological parameters, $t(z)$ can be calculated easily from Eq. (3.94) by numerical integration. In some special cases, the integration can even be carried out analytically.

In the radiation dominated epoch (i.e. at $z \gg z_{eq}$), the solution of the Friedmann equation is given by Eq. (3.80), and the age of the Universe is

$$t(z) \approx \left(\frac{1+z}{10^{10}} \right)^{-2} \mathrm{s}. \tag{3.95}$$

In the matter dominated epoch ($z \ll z_{eq}$), we can neglect the radiation term in $E(z)$. It can then be shown that for an EdS universe (i.e. for $\Omega_{m,0} = 1$ and $\Omega_{\Lambda,0} = 0$),

$$t(z) = \frac{1}{H_0} \frac{2}{3} (1+z)^{-3/2} \approx \frac{2}{3} (1+z)^{-3/2} \times 10^{10} h^{-1}\,\mathrm{yr}. \tag{3.96}$$

For an open universe with $\Omega_{\Lambda,0} = 0$ and $\Omega_0 = \Omega_{\mathrm{m},0} < 1$,

$$t(z) = \frac{1}{H_0} \frac{\Omega_0}{2(1-\Omega_0)^{3/2}} \left[\frac{2\sqrt{(1-\Omega_0)(\Omega_0 z+1)}}{\Omega_0(1+z)} - \cosh^{-1}\left(\frac{\Omega_0 z - \Omega_0 + 2}{\Omega_0 z + \Omega_0} \right) \right]. \tag{3.97}$$

For a closed universe with $\Omega_{\Lambda,0} = 0$ and $\Omega_0 = \Omega_{\mathrm{m},0} > 1$,

$$t(z) = \frac{1}{H_0} \frac{\Omega_0}{2(\Omega_0-1)^{3/2}} \left[-\frac{2\sqrt{(\Omega_0-1)(\Omega_0 z+1)}}{\Omega_0(1+z)} + \cos^{-1}\left(\frac{\Omega_0 z - \Omega_0 + 2}{\Omega_0 z + \Omega_0} \right) \right]. \tag{3.98}$$

Finally, for a flat universe with $\Omega_{\mathrm{m},0} + \Omega_{\Lambda,0} = 1$,

$$t(z) = \frac{1}{H_0} \frac{2}{3\sqrt{\Omega_{\Lambda,0}}} \ln \left[\frac{\sqrt{\Omega_{\Lambda,0}(1+z)^{-3}} + \sqrt{\Omega_{\Lambda,0}(1+z)^{-3} + \Omega_{\mathrm{m},0}}}{\sqrt{\Omega_{\mathrm{m},0}}} \right]. \tag{3.99}$$

In all these cases, the behavior at $z \gg 1$ is

$$t(z) \approx \frac{2}{3H_0} \Omega_{\mathrm{m},0}^{-1/2} (1+z)^{-3/2}. \tag{3.100}$$

Fig. 3.2 shows the product of the Hubble parameter, h, defined by Eq. (3.15), and the lookback time, $t_0 - t(z)$, as a function of $(1+z)$ for models with $\Omega_{\Lambda,0} = 0$, and for flat models with a

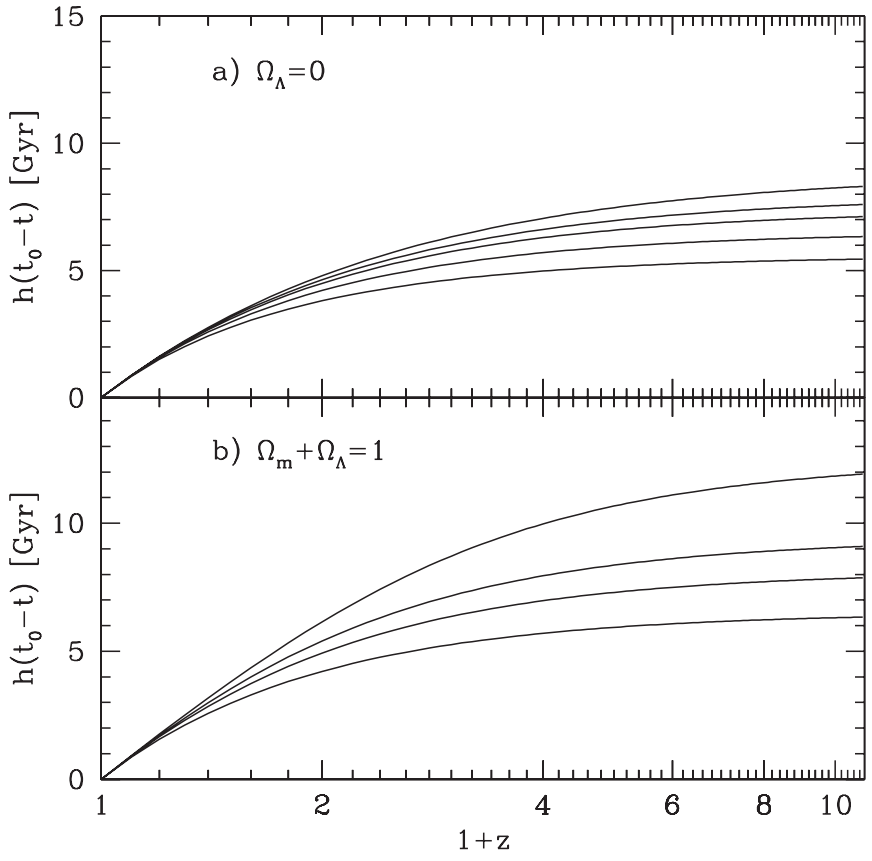

Fig. 3.2. The lookback time as a function of redshift for (a) models with $\Omega_{\Lambda,0} = 0$ and $\Omega_{\mathrm{m},0} = 0.1$, 0.3, 0.5, 1, 2 (from top down); and (b) flat models ($\Omega_{\mathrm{m},0} + \Omega_{\Lambda,0} = 1$) with $\Omega_{\mathrm{m},0} = 0.1$, 0.3, 0.5, 1 (from top down).

cosmological constant ($\Omega_{\rm m,0}+\Omega_{\Lambda,0}=1$). It is clear that for given h and $\Omega_{\rm m,0}$, the age of the Universe is larger in models with a cosmological constant. By definition, the age of the Universe at the present time should be larger than that of the oldest objects it contains. The oldest objects whose ages can be determined reliably are a class of star clusters called globular clusters, which have ages ranging up to 13 Gyr (e.g. Carretta et al., 2000). This requires $h \lesssim 0.5$ for an EdS universe, and $h \lesssim 0.7$ for a flat universe with $\Omega_{\rm m,0}=0.3$ and $\Omega_{\Lambda,0}=0.7$.

3.2.6 Cosmological Distances and Volumes

As defined in §3.1.6, the luminosity distance, $d_{\rm L}$, and the angular-diameter distance, $d_{\rm A}$, are related to the redshift, z, and the comoving coordinate, r, by

$$d_{\rm L}=\left(\frac{L}{4\pi F}\right)^{1/2}=a_0 r(1+z)\,; \qquad d_{\rm A}=\frac{D}{\vartheta}=\frac{a_0 r}{1+z}\,. \tag{3.101}$$

In order to write $d_{\rm L}$ and $d_{\rm A}$ in terms of observable quantities, we need to express the unobservable coordinate r as a function of z. To do this, recall that $r(t)$ is the comoving coordinate of a light signal (an event) that originates at cosmic time t and reaches us at the origin at the present time t_0. It then follows from Eq. (3.17) that the comoving distance corresponding to r is

$$\chi(r)=\tau(t_0)-\tau(t)=c\int_{a(t)}^{a_0}\frac{{\rm d}a}{a\dot{a}}\,, \tag{3.102}$$

where we have used the definition of the conformal time in Eq. (3.11) and the fact that ${\rm d}t={\rm d}a/\dot{a}$. Using Eq. (3.74) and the fact that $a(z)=a_0/(1+z)$ this can be rewritten as

$$\chi(r)=\frac{c}{H_0 a_0}\int_0^z\frac{{\rm d}z}{E(z)}\,, \tag{3.103}$$

where $E(z)$ is given by Eq. (3.75). Using Eqs. (3.10) and (3.13), this gives

$$r=f_K\left[\frac{c}{H_0 a_0}\int_0^z\frac{{\rm d}z}{E(z)}\right]\,. \tag{3.104}$$

Note that r is the angular-diameter distance in comoving units. In general Eq. (3.103) can be integrated numerically for a given set of cosmological parameters. When $z\ll z_{\rm eq}$ and $\Omega_{\Lambda,0}=0$, a closed expression can be derived for all three values of K,

$$a_0 r=\frac{2c}{H_0}\,\frac{\Omega_0 z+(2-\Omega_0)\left[1-(\Omega_0 z+1)^{1/2}\right]}{\Omega_0^2(1+z)}\,, \tag{3.105}$$

which is known as Mattig's formula (Mattig, 1958). For a flat ($\Omega_{\rm m,0}+\Omega_{\Lambda,0}=1$) universe $r=\chi$, so that for $z\ll z_{\rm eq}$

$$a_0 r=\frac{c}{H_0}\int_0^z\frac{{\rm d}z}{\left[\Omega_{\Lambda,0}+\Omega_{\rm m,0}(1+z)^3\right]^{1/2}}\,. \tag{3.106}$$

Luminosity (or angular-diameter) distances can be measured directly for objects of known intrinsic luminosity (or proper size). Such objects are known as 'standard candles' (or 'standard rulers'). Since the relation of redshift to these distances depends on cosmological parameters, in particular on H_0, $\Omega_{\rm m,0}$ and $\Omega_{\Lambda,0}$, measuring the redshift of properly calibrated standard candles (or standard rulers) can provide estimates of these parameters.

One of the most reliable and historically most important standard candles is a class of pulsating stars known as Cepheids, for which the pulsation period is tightly correlated with their mean intrinsic luminosity (see §2.1.3). Using the HST, Cepheids have been measured out to distances of about 10 Mpc. At such distances, the $d_{\rm L}$-z relation is still linear, $d_{\rm L}\approx cz/H_0$, so interesting

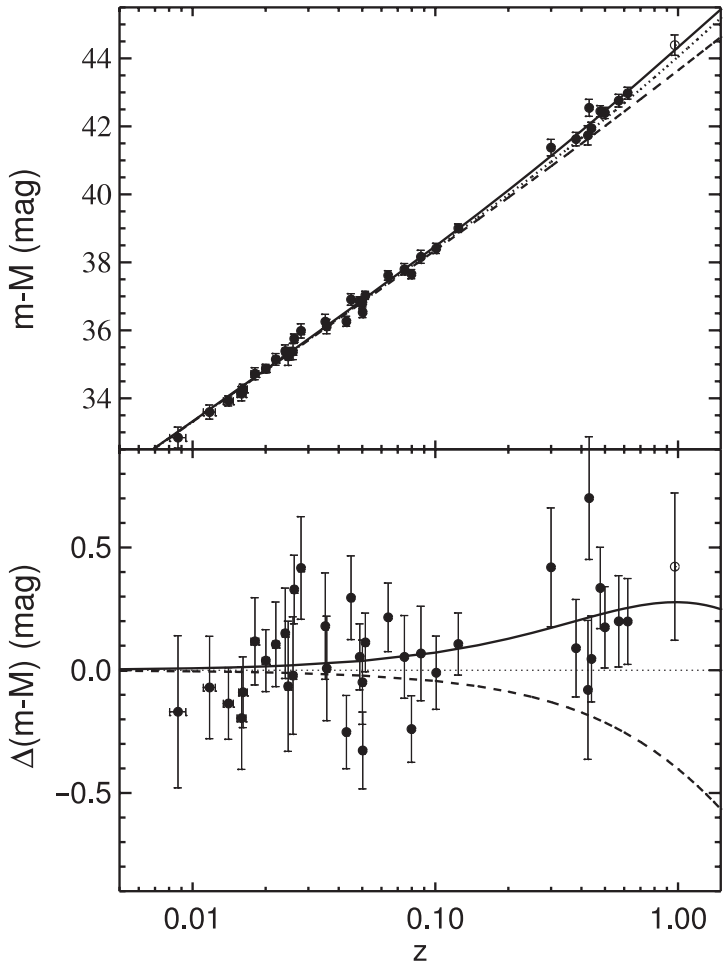

Fig. 3.3. The upper panel shows distance modulus, $(m - \mathscr{M}) = 5\log(d_{\rm L}/10\,{\rm pc})$, against redshift for Type Ia supernovae for which the light curve shape has been used to estimate their absolute magnitudes (data points). The predicted relations for three cosmological models are indicated by dashed ($\Omega_{\rm m,0} = 1$, $\Omega_{\Lambda,0} = 0$), dotted ($\Omega_{\rm m,0} = 0.2$, $\Omega_{\Lambda,0} = 0$) and solid ($\Omega_{\rm m,0} = 0.28$, $\Omega_{\Lambda,0} = 0.72$) curves. The lower panel shows the difference between the distance modulus and the prediction for the ($\Omega_{\rm m,0} = 0.2$, $\Omega_{\Lambda,0} = 0$) model. [Adapted from Riess et al. (1998) by permission of AAS]

constraints can be obtained only for the Hubble constant. The current best estimate is $H_0 = (72 \pm 8)\,{\rm km\,s^{-1}\,Mpc^{-1}}$ (e.g. Freedman et al., 2001).

In order to measure other cosmological parameters we must go to sufficiently large distances so that nonlinear terms in the distance–redshift relation are important, i.e. to $z \gtrsim 1$. In Chapter 2 we have seen that Type Ia supernovae can be used as standard candles and that they have now been observed out to $z \sim 1$. In Fig. 3.3 the observed luminosity distance–redshift relation for Type Ia supernovae is compared with the predictions of a number of cosmological models. Detailed analyses of these data give the following constraint:

$$0.8\Omega_{\rm m,0} - 0.6\Omega_{\Lambda,0} \simeq -0.2 \pm 0.1 \tag{3.107}$$

(Perlmutter et al., 1999).

The proper-distance element at time t is $\mathrm{d}l = a(t)\,\mathrm{d}\chi$. Using Eq. (3.103) we have

$$\frac{\mathrm{d}l}{\mathrm{d}z} = \frac{c}{H_0}\frac{1}{(1+z)}\frac{1}{E(z)}\,. \tag{3.108}$$

This gives the proper distance per unit redshift at redshift z. Suppose that there is a population of objects with proper number density $n(z) = n_0(z)(1+z)^3$ (so that n_0 is a constant if the number of the objects is conserved) and with average proper cross-section $\sigma(z)$. The number of intersections between such objects and a sightline in a unit redshift interval around z is

$$\frac{\mathrm{d}\mathscr{N}}{\mathrm{d}z} = n_0(z)(1+z)^3\sigma(z)\frac{\mathrm{d}l}{\mathrm{d}z} = n_0(z)\sigma(z)\frac{c}{H_0}\frac{(1+z)^2}{E(z)}\,. \tag{3.109}$$

The 'optical depth' for the intersection of objects up to redshift z is therefore

$$\tau(z) = \int_0^z \mathrm{d}\mathscr{N}(z) = \frac{c}{H_0}\int_0^z n_0(z)\sigma(z)\frac{(1+z)^2}{E(z)}\,\mathrm{d}z\,. \tag{3.110}$$

These quantities are relevant for the discussion of QSO absorption line systems (see Chapter 16). In this case $n_0(z)$ is the comoving number density, $\sigma(z)$ is the average absorption cross-section of absorbers, and $\mathrm{d}\mathscr{N}/\mathrm{d}z$ is just the expected number of absorption systems per unit redshift. Another application of Eqs. (3.109) and (3.110) concerns the interpretation of the observed number of gravitational lensing events caused by foreground objects. In this case, $n_0(z)$ is the comoving number density of lenses, and $\sigma(z)$ is the average lensing cross-section. A third application is to the scattering of the microwave background by ionized intergalactic gas. Here, $\sigma(z)$ is the Thomson cross-section and $n_0(z)$ is the comoving number density of free electrons.

Consider next the proper volume element at a redshift z. The proper-length element in the radial direction is again $a(t)\,\mathrm{d}\chi$, and the proper distance subtended by an angle element $\mathrm{d}\vartheta$ is $a(t)r\,\mathrm{d}\vartheta$. The proper-volume element at redshift z corresponding to a solid angle $\mathrm{d}\omega = \mathrm{d}\vartheta^2$ and a depth $\mathrm{d}z$ is thus

$$\mathrm{d}^2V_\mathrm{p} = a^3(t)r^2\,\mathrm{d}\chi\,\mathrm{d}\omega = \frac{c}{H_0}\frac{\mathrm{d}z}{(1+z)E(z)}\frac{[a_0r(z)]^2\,\mathrm{d}\omega}{(1+z)^2}\,, \tag{3.111}$$

where $r(z)$ is related to z by Eq. (3.104). Using Eq. (3.111), the total, proper volume out to redshift z is

$$\begin{aligned} V_\mathrm{p}(z) &= 4\pi a^3(t)\int_0^{r(z)}\frac{r'^2\,\mathrm{d}r'}{\sqrt{1-Kr'^2}} \\ &= \begin{cases} 2\pi a^3(t)\left(\sin^{-1}r - r\sqrt{1-r^2}\right) & (K=+1) \\ \frac{4\pi}{3}a^3(t)r^3 & (K=0) \\ 2\pi a^3(t)\left(r\sqrt{1+r^2} - \sinh^{-1}r\right) & (K=-1). \end{cases} \end{aligned} \tag{3.112}$$

We can also use Eq. (3.111) to compute the total number of objects per unit volume. Assuming the proper number density of objects at redshift z to be $n(z) = n_0(z)(1+z)^3$, the predicted count of objects per unit redshift and per unit solid angle is

$$\frac{\mathrm{d}^2N}{\mathrm{d}z\,\mathrm{d}\omega} = n(z)\frac{\mathrm{d}^2V_\mathrm{p}}{\mathrm{d}z\,\mathrm{d}\omega} = n_0(z)\frac{c}{H_0}\frac{[a_0r(z)]^2}{E(z)}\,. \tag{3.113}$$

Thus, if the z dependence of n_0 is known, one can use Eq. (3.113) to put constraints on cosmological parameters by simply counting objects (e.g. galaxies) as a function of z (see Loh & Spillar (1986) for a discussion).

Another important quantity in cosmology is the comoving distance between any two observed objects in the Robertson–Walker metric. Suppose that these two objects (labeled $\mathscr{O}_1$ and $\mathscr{O}_2$) are

at redshifts z_1 and z_2, and are separated by an angle α on the sky. Their comoving distances from an observer at the origin are given by Eq. (3.102), and we denote them by χ_1 and χ_2, respectively. As shown in §3.1.2, for $K = +1$ the comoving distance, χ_{12}, between $\mathcal{O}_1$ and $\mathcal{O}_2$ is equal to the distance on the unit sphere between two points with polar angles χ_1 and χ_2 and with azimuthal angles differing by α. Thus

$$\cos\chi_{12} = \cos\chi_1\cos\chi_2 + \sin\chi_1\sin\chi_2\cos\alpha\,. \tag{3.114}$$

The corresponding equation for $K = -1$ is

$$\cosh\chi_{12} = \cosh\chi_1\cosh\chi_2 - \sinh\chi_1\sinh\chi_2\cos\alpha\,, \tag{3.115}$$

and for $K = 0$ is

$$\chi_{12}^2 = \chi_1^2 + \chi_2^2 - 2\chi_1\chi_2\cos\alpha\,. \tag{3.116}$$

Finally, consider the case in which α is zero (or very small). In this case, the angular-diameter distance from $\mathcal{O}_1$ to $\mathcal{O}_2$ can be written as

$$d_{\mathrm{A},12} = \frac{a_0 r_{12}}{1+z_2}\,, \tag{3.117}$$

where

$$r_{12} \equiv f_K(\chi_{12}) = f_K(\chi_2-\chi_1) = r(z_2)\sqrt{1-Kr^2(z_1)} - r(z_1)\sqrt{1-Kr^2(z_2)}\,. \tag{3.118}$$

For the $\Omega_{\Lambda,0} = 0$ case this gives

$$d_{\mathrm{A},12} = \frac{2c}{H_0}\frac{\sqrt{1+\Omega_0 z_1}(2-\Omega_0+\Omega_0 z_2) - \sqrt{1+\Omega_0 z_2}(2-\Omega_0+\Omega_0 z_1)}{\Omega_0^2(1+z_2)^2(1+z_1)} \tag{3.119}$$

(Refsdal, 1966). Note that $|d_{\mathrm{A},12}| \neq |d_{\mathrm{A},21}|$, and that, as required, Eq. (3.119) reduces to Mattig's formula for $z_1 = 0$. Eq. (3.119) plays an important role in gravitational lensing, where z_1 and z_2 are the redshifts of the lens and the source, respectively (see §6.6).

3.3 The Production and Survival of Particles

An important feature of the standard cosmology is that the temperature of the Universe was arbitrarily high at the beginning of the Big Bang [see Eq. (3.81)] and has decreased continuously as the Universe expanded to its present state. As we have seen in §3.1.5, the thermal history of the Universe follows from a simple application of thermodynamics to a small patch of the homogeneous and isotropic Universe. In this section we show that this thermal history, together with particle, nuclear and atomic physics, allows a detailed prediction of the matter content of the Universe at each epoch. The reason for this is simple: when the temperature of the Universe was higher than the rest mass of a kind of charged particles, the photon energy is high enough to create these particles and their antiparticles. This, in turn, could give rise to other kinds of particles. For example, when the temperature of the Universe was higher than the rest mass of an electron, i.e. $k_{\mathrm{B}}T > m_{\mathrm{e}}c^2 \approx 0.511\,\mathrm{MeV}$ (corresponding to $T \sim 5.8\times10^9\,\mathrm{K}$), electrons and positrons could be generated via pair production, $\gamma+\gamma \leftrightarrow \mathrm{e}+\bar{\mathrm{e}}$, and electronic neutrinos could be produced via neutral current reactions, such as $\mathrm{e}+\bar{\mathrm{e}} \leftrightarrow \nu_{\mathrm{e}}+\overline{\nu}_{\mathrm{e}}$. When the density of the Universe was sufficiently high, the creation and annihilation of $(\mathrm{e},\bar{\mathrm{e}})$ pairs, and the Compton scattering between $(\mathrm{e},\bar{\mathrm{e}})$ and photons, could establish a thermal equilibrium among these particles, while the neutrinos established such an equilibrium via their neutral current coupling to the electrons. Consequently, the Universe was filled with a hot plasma that included γ, e, $\bar{\mathrm{e}}$, ν_{e} and $\overline{\nu}_{\mathrm{e}}$, all in thermal equilibrium at the same temperature.

In order to maintain thermodynamic equilibrium the frequency of interactions among the various particle species involved needs to be sufficiently high. The interaction rate is $\Gamma \equiv n\langle v\sigma\rangle$, where n is the number density of particles, v is their relative velocity, and σ is the interaction cross-section (which usually depends on v). As the Universe expands and the temperature drops, this rate in general decreases. When it becomes smaller than the expansion rate of the Universe, given by the Hubble parameter, $H(t)$, the particles 'decouple' from the photon fluid, and, as long as the particles are stable, their comoving number density 'freezes-out' at its current value. Except for possible particle species created out of thermal equilibrium (e.g. axions), and for particles that have been created more recently in high-energy processes, all elementary particles in the present-day Universe are thermal relics that have decoupled from the photon fluid at some time in the past.

In what follows we first present a brief outline of the chronology of the early Universe, and then discuss the production and survival of particles during a number of important epochs. Since the early Universe was dominated by relativistic particles, Eq. (3.81) can be used to relate temperature T (or energy $k_{\rm B}T$) to the cosmic time. As this section is concerned with high energy physics, we will use the natural unit system in which the speed of light c, Boltzmann's constant $k_{\rm B}$, and Planck's constant $\hbar = h_{\rm P}/2\pi$ are all set to 1. In cgs units $[c] = {\rm cm\,s^{-1}}$, $[\hbar] = {\rm g\,cm^2\,s^{-1}}$, and $[k_{\rm B}] = {\rm g\,cm^2\,s^{-2}\,K^{-1}}$. Therefore, making these constants dimensionless implies that

$$[\text{energy}] = [\text{mass}] = [\text{temperature}] = [\text{time}]^{-1} = [\text{length}]^{-1}\,, \tag{3.120}$$

and all physical quantities can be expressed in one unit, usually mass or energy. However, they can also be expressed in one of the other units using the following conversion factors: $1\,{\rm MeV} = 1.602\times10^{-6}\,{\rm erg} = 1.161\times10^{10}\,{\rm K} = 1.783\times10^{-27}\,{\rm g} = 5.068\times10^{10}\,{\rm cm^{-1}} = 1.519\times10^{21}\,{\rm s^{-1}}$. Whenever needed, the 'missing' powers of c, $k_{\rm B}$, and $\hbar$ in equations can be reinserted straightforwardly from a simple dimensional analysis.

3.3.1 The Chronology of the Hot Big Bang

Since our understanding of particle physics is only robust below energies of $\sim 1\,{\rm GeV}$ ($\sim 10^{13}\,{\rm K}$), the physics of the very early Universe ($t \lesssim 10^{-6}\,{\rm s}$) is still very uncertain. In popular, although speculative, extensions of the standard model for particle physics, this era is characterized by a number of symmetry-breaking phase transitions. Particle physicists have developed a number of models which suggest the existence of many exotic particles as a result of these symmetry breakings, and it is a popular idea that the elusive dark matter consists of one or more of such particle species. However, it should be kept in mind that the theories predicting the existence of these exotic particles are not well established and that there is not yet any convincing, direct experimental evidence for their existence.

For the purpose of the discussion here, the two most important events that (probably) took place during this early period after the Big Bang are inflation and baryogenesis. Inflation is a period of exponential expansion that resulted from a phase transition associated with some unknown scalar field. Inflation is invoked to solve several important problems for the standard Hot Big Bang cosmology, and is described in detail in §3.6. Baryogenesis is a mechanism that is needed to explain the observed asymmetry between baryons and antibaryons: one does not observe a significant abundance of antibaryons. If they were there, their continuous annihilation with baryons would produce a much greater gamma-ray background than observed, unless they are spatially segregated from the baryons, which is extremely contrived. Apparently, the Universe has a non-zero baryon number. If baryon number is conserved, this asymmetry between baryons and antibaryons must have originated at very early times through a process called baryogenesis. The details of this process are still poorly understood, and will not be discussed in this book (see Kolb & Turner, 1990).

In what follows we give a brief overview of some of the most important events that took place in the early Universe after it had cooled down to a temperature of $\sim 10^{13}\,\mathrm{K}$. At this point in time, the temperature of the Universe was still higher than the binding energy of hadrons (baryons and mesons). Quarks were not yet bound into hadronic states. Instead, the matter in the Universe was in a form referred to as quark soup, which consists of quarks, leptons and photons.

- At $T \sim 3\times 10^{12}\,\mathrm{K}$ ($t \sim 10^{-5}\,\mathrm{s}$), corresponding to an energy of 200–300 MeV, the quark–hadron phase transition occurs, confining quarks into hadrons. If the phase transition was strongly first order, it may have induced significant inhomogeneities in the baryon-to-photon ratio, and affected the later formation of elements, a topic discussed further in §3.4. Once the transition was complete, the Universe was filled with a hot plasma consisting of three types of (relativistic) pions (π^+, π^-, π^0), (non-relativistic) nucleons (protons, p, and neutrons, n), charged leptons (e, $\bar{\mathrm{e}}$, μ, $\overline{\mu}$; the τ and $\overline{\tau}$ have already annihilated), their associated neutrinos (ν_e, $\overline{\nu}_\mathrm{e}$, ν_μ, $\overline{\nu}_\mu$), and photons, all in thermal equilibrium. In addition, the Universe comprises several decoupled species, such as the tau-neutrinos (ν_τ and $\overline{\nu}_\tau$) – their coupling has to be through their reactions with τ and $\overline{\tau}$, and possible exotic particles that make up the (non-baryonic) dark matter.
- At $T \sim 10^{12}\,\mathrm{K}$ ($t \sim 10^{-4}\,\mathrm{s}$) the (π^+, π^-) pairs annihilate while the neutral pions π^0 decay into photons. From this point on the nucleons (and a small abundance of their antiparticles which escaped annihilation) are the only hadronic species left. At around the same time, the muons start to annihilate, and their number density becomes negligibly small as T drops to about $10^{11}\,\mathrm{K}$. At this time, ν_μ and $\overline{\nu}_\mu$ also decouple from the hot plasma, and expand freely with the Universe.
- When T drops below $10^{11}\,\mathrm{K}$, the number of neutrons becomes smaller than that of protons by a factor of about $\exp(-\Delta m/T)$, where $\Delta m \approx 1.3\,\mathrm{MeV}$ is the mass difference between a neutron and a proton. This asymmetry in the numbers of n and p continues to grow until the reaction rate between neutrons and protons becomes negligible.
- At $T \sim 5\times 10^{9}\,\mathrm{K}$ ($t \sim 4\,\mathrm{s}$), the annihilations of $(\mathrm{e},\bar{\mathrm{e}})$ pairs begins. As the number density of $(\mathrm{e},\bar{\mathrm{e}})$ pairs drops, ν_e and $\overline{\nu}_\mathrm{e}$ decouple from the hot plasma. Since the $(\mathrm{e},\bar{\mathrm{e}})$ annihilations heat the photons but not the decoupled neutrinos, the neutrinos expand freely with a temperature that is lower than that of the photons. Because of the reduction in the number of $(\mathrm{e},\bar{\mathrm{e}})$ pairs and the cooling of ν_e and $\overline{\nu}_\mathrm{e}$, reactions such as $\mathrm{n}+\nu_\mathrm{e} \leftrightarrow \mathrm{p}+\mathrm{e}$ and $\mathrm{n}+\bar{\mathrm{e}} \leftrightarrow \mathrm{p}+\overline{\nu}_\mathrm{e}$ are no longer effective. Consequently, the n/p ratio freezes out at a value of about $\exp(-\Delta m/T) \sim 1/10$. Note that this ratio does not change much due to beta decay of the neutrons, because the half-time of the decay (about 10 minutes) is much longer than the age of the Universe at this time.
- At $T \sim 10^{9}\,\mathrm{K}$ ($t \sim$ few minutes), nucleosynthesis starts, synthesizing protons and neutrons to produce D, He and a few other elements. Since the temperature is still too high for the formation of neutral atoms, all these elements are highly ionized. Consequently, the Universe is now filled with freely expanding neutrinos (and possibly exotic particles) and a plasma of electrons and highly ionized atoms (mainly protons and He^{++}). However, as the temperature continues to decrease, electrons start to combine with the ions to produce neutral atoms.
- At $T \sim 4000\,\mathrm{K}$ ($t \sim 2\times 10^{5}\,\mathrm{yr}$) roughly 50% of the baryonic matter is in the form of neutral atoms. This point in time is often called the time of recombination. Because of the resulting drop in the number density of free electrons, the Universe suddenly becomes transparent to photons. These photons are observed today as the cosmic microwave background. From this point on, photons, neutrinos, H, He and other atoms all expand freely with the Universe. At around the same time, the energy density in relativistic particles has become smaller than that in the rest mass of non-relativistic matter, and the Universe enters the matter dominated epoch.

Once the processes involved are known from particle, nuclear and atomic physics, it is in principle straightforward to calculate the matter content at different epochs summarized above. A detailed treatment of such calculations is beyond the scope of this book, and can be found in Börner (2003) and Kolb & Turner (1990), for example. In what follows, we present a brief discussion about the basic principles involved and their applications to some important examples.

3.3.2 Particles in Thermal Equilibrium

As discussed above, at any given epoch, some particles are in thermal equilibrium with the hot plasma, some in free expansion with the Universe, and others are in transition between the two states. The number density n, energy density ρ, and pressure P of a given particle species can be written in terms of its distribution function $f(\mathbf{x},\mathbf{p},t)$. Since the Universe is homogeneous and isotropic $f(\mathbf{x},\mathbf{p},t) = f(p,t)$, with $p = |\mathbf{p}|$, so that

$$n(t) = 4\pi \int f(p,t)\, p^2\, \mathrm{d}p\,, \tag{3.121}$$

$$\rho(t) = 4\pi \int E(p) f(p,t)\, p^2\, \mathrm{d}p\,, \tag{3.122}$$

$$P(t) = 4\pi \int \frac{p^2}{3E(p)} f(p,t)\, p^2\, \mathrm{d}p\,, \tag{3.123}$$

where the energy E is related to the momentum p as $E(p) = (p^2 + m^2)^{1/2}$. Eq. (3.123) follows from kinetic theory, according to which the pressure is related to momentum and velocity as $P = \frac{1}{3} n \langle p v \rangle$. Using the components of the four-momentum, we have $v = pc^2/E$, so that $P = n\langle p^2 c^2/3E\rangle$.

For a particle species in thermal equilibrium

$$f(\mathbf{p},t)\mathrm{d}^3\mathbf{p} = \frac{g}{(2\pi)^3}\left\{\exp\left[\frac{E(p)-\mu}{T(t)}\right] \pm 1\right\}^{-1} \mathrm{d}^3\mathbf{p}\,, \tag{3.124}$$

where μ is the chemical potential of the species, and $T(t)$ is its temperature at time t. The signature, $\pm$, takes the positive sign for Fermi–Dirac species and the negative sign for Bose–Einstein species. The factor $1/(2\pi)^3$ is due to Heisenberg's uncertainty principle, which states that no particle can be localized in a phase-space volume smaller than the fundamental element $(2\pi\hbar)^3$ (recall that we use $\hbar = c = k_\mathrm{B} = 1$), and g is the spin-degeneracy factor (neutrinos have $g = 1$, photons and charged leptons have $g = 2$, and quarks have $g = 6$).

Substituting Eq. (3.124) in Eqs. (3.121)–(3.123) yields

$$n_\mathrm{eq} = \frac{g}{2\pi^2} \int_m^\infty \frac{(E^2 - m^2)^{1/2} E\, \mathrm{d}E}{\exp[(E-\mu)/T] \pm 1}\,; \tag{3.125}$$

$$\rho_\mathrm{eq} = \frac{g}{2\pi^2} \int_m^\infty \frac{(E^2 - m^2)^{1/2} E^2\, \mathrm{d}E}{\exp[(E-\mu)/T] \pm 1}\,; \tag{3.126}$$

$$P_\mathrm{eq} = \frac{g}{6\pi^2} \int_m^\infty \frac{(E^2 - m^2)^{3/2}\, \mathrm{d}E}{\exp[(E-\mu)/T] \pm 1}\,. \tag{3.127}$$

Let us consider two special cases. In the non-relativistic limit, i.e. when $T \ll m$, the number density, the energy density and pressure are the same for both Bose–Einstein and Fermi–Dirac species, and can be written in the following analytic forms:

$$n_{\rm eq} = g\left(\frac{mT}{2\pi}\right)^{3/2} e^{(\mu-m)/T}\,, \tag{3.128}$$

$$\rho_{\rm eq} = nm\,, \quad P_{\rm eq} = nT\,. \tag{3.129}$$

For a relativistic ($T \gg m$ and $E = p$), non-degenerate ($\mu \ll T$) gas, the corresponding analytical expressions are

$$n_{\rm eq} = \begin{cases} \left[\zeta(3)/\pi^2\right] gT^3 & \text{(Bose–Einstein)} \\ (3/4)\left[\zeta(3)/\pi^2\right] gT^3 & \text{(Fermi–Dirac),} \end{cases} \tag{3.130}$$

$$\rho_{\rm eq} = \begin{cases} \left(\pi^2/30\right) gT^4 & \text{(Bose–Einstein)} \\ (7/8)\left(\pi^2/30\right) gT^4 & \text{(Fermi–Dirac).} \end{cases} \tag{3.131}$$

$$P_{\rm eq} = \rho_{\rm eq}/3\,, \tag{3.132}$$

where $\zeta(3) \approx 1.2021...$ is the Riemann zeta function of 3.

In general, in order to use Eqs. (3.125)–(3.127) to calculate the density and pressure, one needs to know the chemical potential μ. The principle for determining the chemical potential of a species is that chemical potential is an additive quantity which is conserved during a 'chemical' reaction (e.g. Landau & Lifshitz, 1959). Thus, if species 'i' takes part in a reaction like $i + j \leftrightarrow k + l$, then $\mu_i + \mu_j = \mu_k + \mu_l$. The values of the chemical potentials therefore depend on the various conservation laws under which the various reactions take place. For example, since the number of photons is not a conserved quantity for a thermodynamic system, the chemical potential of photons must be zero. This is consistent with the fact that photons at thermal equilibrium have the Planck distribution. It then follows that the chemical potential for a particle is the negative of that for its antiparticle (because particle–antiparticle pairs can be annihilated to photons). Put differently, the difference in the number density of particles and antiparticles depends only on the chemical potential. Similar to electric charge, particle reactions are thought to generally conserve baryon number (which explains the long lifetime of the proton, of $> 10^{34}$ years) and lepton number. Since the number densities of baryons and leptons are found to be (or, in the case of leptons, believed to be) much smaller than the number density of photons, the chemical potential of all species may be set to zero to good approximation in computing the mean energy density and pressure in the early Universe.

There is one caveat, however. Since the chemical potential of a particle is the negative of that of its antiparticle, it follows from Eq. (3.121) that, for fermions, their difference in number densities is given by

$$n - \bar{n} = \frac{gT^3}{6\pi^2}\left[\pi^2\left(\frac{\mu}{T}\right) + \left(\frac{\mu}{T}\right)^3\right]. \tag{3.133}$$

When the Universe cools to temperatures below the rest mass of the particles, all particle–antiparticle pairs will be annihilated[5] leaving only this small excess, which is zero when $\mu = 0$. Therefore, the fact that we do have non-zero baryon and lepton number densities in the Universe today implies that μ cannot have been strictly zero at all times. In the early Universe, some physics must have occurred that did not conserve baryon number or lepton number, and that resulted in the present-day number densities of protons and electrons. The actual physics of this

[5] In principle, because of the expansion of the Universe, tiny fractions of particles and antiparticles may survive, but their number densities are negligibly small.

baryon- and lepton-genesis are poorly understood, and will not be discussed further in this book. Detailed descriptions can be found in Kolb & Turner (1990).

With all chemical potentials set to zero, it is evident from Eqs. (3.125) and (3.130) that the number density of non-relativistic particles is suppressed exponentially with respect to that of relativistic species. This reflects the coupling to the photon fluid. When $T \gg m$ the photons have sufficient energy to create a thermal background number density of particle–antiparticle pairs. However, when $T \ll m$ only an exponential tail of the photon distribution function has sufficient energy for pair creation, causing a similar suppression of their number density. Consequently, particles in thermal equilibrium with the photon gas can only contribute significantly to the energy density and pressure when they are relativistic. Thus, to good accuracy, we can write the total energy density, number density and pressure of the Universe, in the radiation dominated era, as

$$\rho(T) = \frac{\pi^2}{30} g_* T^4, \quad n(T) = \frac{\zeta(3)}{\pi^2} g_{*,n} T^3, \quad P(T) = \rho(T)/3, \tag{3.134}$$

with

$$g_* = \sum_{i\in\text{Boson}} g_i \left(\frac{T_i}{T}\right)^4 + \frac{7}{8} \sum_{i\in\text{Fermion}} g_i \left(\frac{T_i}{T}\right)^4, \tag{3.135}$$

$$g_{*,n} = \sum_{i\in\text{Boson}} g_i \left(\frac{T_i}{T}\right)^3 + \frac{3}{4} \sum_{i\in\text{Fermion}} g_i \left(\frac{T_i}{T}\right)^3. \tag{3.136}$$

Note that we have included the possibility that the temperature of a species T_i may be different from that of the radiation background T. The values of g_* and $g_{*,n}$ at a given time can be calculated once the existing relativistic species are identified. For example, at $T \ll 1\,\text{MeV}$, the only relativistic species are photons at temperature T and three species of neutrinos and their antiparticles (all assumed to be massless) at temperature $T_\nu = (4/11)^{1/3} T$ (as we will see in §3.3.3). Therefore $g_* = g_\gamma + (7/8)(3 \times 2 \times g_\nu)(T_\nu/T)^4 \approx 3.36$. At higher T (earlier times) more species are relativistic, so that the degeneracy factors are larger. Fig. 3.4 shows g_* as a function of T obtained from the standard model of particle physics. It increases from 3.36 at the present-day temperature of 2.73 K to 106.75 at $T \gtrsim 300\,\text{GeV}$.

3.3.3 Entropy

An important thermodynamic quantity for describing the early Universe is the entropy $S = S(V,T)$. If we continue to ignore the chemical potential, the second law of thermodynamics, as applied to a comoving volume $V \propto a^3(t)$, states that

$$\mathrm{d}S(V,T) = \frac{1}{T}\left\{\mathrm{d}\left[\rho(T)V\right] + P(T)\,\mathrm{d}V\right\}, \tag{3.137}$$

where ρ is the equilibrium energy density of the gas.

Alternatively, we can write the differential of S in terms of its general form

$$\mathrm{d}S(V,T) = \frac{\partial S}{\partial V}\mathrm{d}V + \frac{\partial S}{\partial T}\mathrm{d}T. \tag{3.138}$$

Using Eq. (3.137) to identify the two partial derivatives, the integrability condition,

$$\frac{\partial^2 S}{\partial T \partial V} = \frac{\partial^2 S}{\partial V \partial T}, \tag{3.139}$$

yields

$$\frac{\mathrm{d}P}{\mathrm{d}T} = \frac{\rho(T) + P(T)}{T}. \tag{3.140}$$

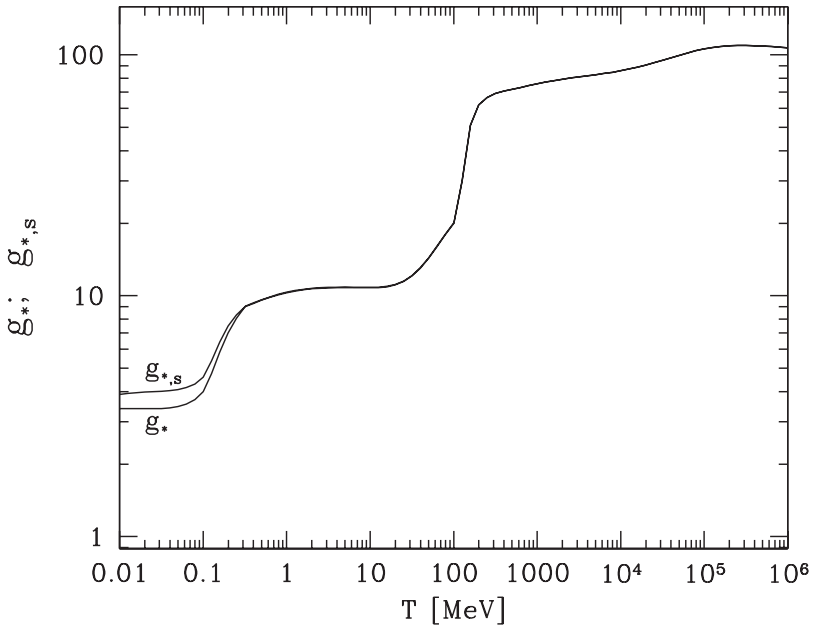

Fig. 3.4. The statistical weights g_* and $g_{*,s}$ as functions of temperature, T, in the standard $SU(3)\times SU(2)\times U(1)$ model of particle physics.

Inserting this in Eq. (3.137) we obtain

$$\mathrm{d}S(V,T) = \frac{1}{T}\mathrm{d}\{[\rho(T)+P(T)]V\} - \frac{V}{T^2}[\rho(T)+P(T)]\mathrm{d}T\,, \tag{3.141}$$

which may be integrated to show that, up to an additive constant, the entropy density, $s(T) \equiv S(V,T)/V$, is given by

$$s(T) = \frac{\rho(T)+P(T)}{T}\,. \tag{3.142}$$

It is easy to show with the use of Eqs. (3.35) and (3.140) that

$$\frac{\mathrm{d}S}{\mathrm{d}a} \propto \frac{\mathrm{d}(sa^3)}{\mathrm{d}a} = 0\,, \tag{3.143}$$

which is the 'entropy conservation law', owing to the adiabaticity of the universal expansion (see §3.1.5).

Using Eqs. (3.131) and (3.132) the entropy density for non-degenerate, relativistic particles in thermal equilibrium is

$$s_{\mathrm{eq}}(T) = \frac{2\pi^2}{45}gT^3\,. \tag{3.144}$$

The entropy density of non-relativistic particles in thermal equilibrium with the photon fluid can be expressed in terms of the entropy density of photons, $s_\gamma(T)$, as

$$\frac{s_{\mathrm{eq}}(T)}{s_\gamma(T)} = \frac{3}{4}\frac{\rho(T)}{\rho_\gamma(T)}\left(1+\frac{P}{\rho}\right)\,. \tag{3.145}$$

Since $\rho \ll \rho_\gamma$,[6] the contribution of non-relativistic particles to the total entropy density is negligible. To good accuracy, therefore, the total entropy density of the Universe is obtained by summing over all relativistic species:

$$s(T) = \frac{2\pi^2}{45} g_{*,s} T^3 \tag{3.146}$$

with

$$g_{*,s} = \sum_{i \in \text{Boson}} g_i \left(\frac{T_i}{T}\right)^3 + \frac{7}{8} \sum_{i \in \text{Fermion}} g_i \left(\frac{T_i}{T}\right)^3 . \tag{3.147}$$

Combining Eq. (3.146) with the entropy conservation law we see that $g_{*,s} T^3 a^3$ is a conserved quantity, so that

$$g_{*,s}^{1/3}(T) T \propto a^{-1} . \tag{3.148}$$

Therefore, as long as $g_{*,s}$ remains constant, $T \propto a^{-1}$, consistent with the thermodynamic derivation in §3.1.5. However, as the Universe cools, every now and then particle species become non-relativistic and stop contributing (significantly) to the entropy density of the Universe. Their entropy is transferred to the remaining relativistic particle species, causing T to decrease somewhat slower. An interesting application of this is the decoupling of light neutrinos. Although neutrinos do not couple directly to the photons, they can maintain thermal equilibrium via weak reactions such as $\mathrm{e} + \overline{\mathrm{e}} \leftrightarrow \nu_\mathrm{e} + \overline{\nu}_\mathrm{e}$, etc. At a freeze-out temperature of $T_\mathrm{f} \sim 1\,\mathrm{MeV}$ the interaction rate for these reactions drops below the expansion rate of the Universe, and the neutrinos decouple from the photon fluid. From this point on, their temperature will decrease strictly as $T_\nu \propto a^{-1}$, while the photon temperature, T_γ, obeys Eq. (3.148). Since the neutrinos are relativistic both before and after decoupling, their freeze-out leaves $g_{*,s}$ invariant. Consequently, despite being decoupled, the temperature of the neutrinos remains exactly the same as that of the photons. This changes a little time later, when the temperature has dropped to $T \sim 0.51\,\mathrm{MeV}$ and electrons start to annihilate and freeze-out from the photon fluid. The entropy released in this process is given to the photons, but not to the decoupled neutrinos (who conserve their entropy density separately). Consequently, after electron annihilation, $T_\gamma > T_\nu$. Their ratio follows from the entropy conservation law, according to

$$\frac{T_{\gamma,\text{after}}}{T_{\nu,\text{after}}} = \frac{T_{\gamma,\text{after}}}{T_{\gamma,\text{before}}} = \left[\frac{g_{*,s}(T_\text{before})}{g_{*,s}(T_\text{after})}\right]^{1/3} , \tag{3.149}$$

where we have used that $T_{\nu,\text{after}} = T_{\nu,\text{before}} = T_{\gamma,\text{before}}$. Before electron annihilation, the relativistic species in the Universe are photons, electrons, positrons, and three flavors of neutrinos with their antiparticles, all at the same temperature. Therefore, $g_{*,s}(T_\text{before}) = g_\gamma + (7/8)(g_\mathrm{e} + g_{\overline{\mathrm{e}}} + 3g_\nu + 3g_{\overline{\nu}}) = 2 + (7/8)(2+2+3+3) = 43/4$. After electron annihilation, $g_{*,s}(T_\text{after}) = g_\gamma + (7/8)(3g_\nu + 3g_{\overline{\nu}})(T_{\nu,\text{after}}/T_{\gamma,\text{after}})^3$. Substitution of these degeneracy parameters into Eq. (3.149) yields

$$T_{\nu,\text{after}} = \left(\frac{4}{11}\right)^{1/3} T_{\gamma,\text{after}} . \tag{3.150}$$

It is thus expected that the present-day Universe contains a relic neutrino background with a temperature of $T_{\nu,0} \simeq 0.71 \times 2.73\,\mathrm{K} = 1.95\,\mathrm{K}$. This difference in the temperature of the two relativistic species (neutrinos and photons) is also apparent from Fig. 3.4. At $T \gtrsim 0.5\,\mathrm{MeV}$, $g_{*,s}(T)$ is identical to g_*, indicating that all relativistic particle species have a common temperature. At

[6] The rest-mass density of particles should not be included as part of the equilibrium energy density of the gas, because there is no creation or annihilation of particles.

lower temperatures, however, electron annihilation has increased T_γ with respect to T_ν, causing an offset of $g_{*,s}$ with respect to g_*.

3.3.4 Distribution Functions of Decoupled Particle Species

In §3.3.2 we discussed the distribution functions of particles in thermal equilibrium. We now turn our attention to species that have dropped out of thermal equilibrium, and have decoupled from the hot plasma. If particle species i decoupled at a time t_f, where the subscript 'f' stands for 'freeze-out', its temperature is approximately equal to the photon temperature at that time, i.e. $T \approx T_\mathrm{f} \equiv T_\gamma(t_\mathrm{f})$. After decoupling, the mean interaction rate of the particle drops below the expansion rate, and the particle basically moves on a geodesic. As we have seen in §3.1.4, the momentum of the particle then scales as $p \propto a^{-1}$, which is valid for both relativistic and non-relativistic species. Since the *relative* momenta are conserved, the actual distribution function at $t > t_\mathrm{f}$ can be written as

$$f(\mathbf{p},t) = f\left(\mathbf{p}\frac{a(t)}{a(t_\mathrm{f})}, t_\mathrm{f}\right) . \tag{3.151}$$

In other words, the form of the distribution function is 'frozen-in' the moment the particles decouple from the hot plasma.

If a species is still relativistic after decoupling, we have $E = p$, so that

$$f(\mathbf{p},t)\mathrm{d}^3\mathbf{p} = \frac{g}{(2\pi)^3}\left\{\exp\left[\frac{p\,a(t)}{T_\mathrm{f}a(t_\mathrm{f})}\right] \pm 1\right\}^{-1} \mathrm{d}^3\mathbf{p} . \tag{3.152}$$

Thus, the distribution function of a decoupled, relativistic species is self-similar to that of a relativistic species in thermal equilibrium, but with a temperature

$$T = T_\mathrm{f}\frac{a(t_\mathrm{f})}{a(t)} . \tag{3.153}$$

Note that this differs from the temperature scaling of species still in thermal equilibrium, which is instead given by Eq. (3.148). As we discussed in §3.3.3 this explains why the present-day temperature of the neutrino background is lower than that of the CMB.

If the species is already non-relativistic when it decouples, its energy is given by $E = m + p^2/2m$. Since for non-relativistic species we can ignore the ± 1 term, the distribution function is given by

$$f(\mathbf{p},t)\mathrm{d}^3\mathbf{p} = \frac{g}{(2\pi)^3}\exp\left[-\frac{m}{T_\mathrm{f}}\right]\exp\left[-\frac{p^2}{2mT}\right]\mathrm{d}^3\mathbf{p} \tag{3.154}$$

with

$$T = T_\mathrm{f}\left[\frac{a(t_\mathrm{f})}{a(t)}\right]^2 . \tag{3.155}$$

Note that Eq. (3.154) is a Maxwell–Boltzmann distribution, and that the temperature scales as expected from kinetic theory (see §3.1.5).

As is immediately evident from substituting Eq. (3.151) in Eq. (3.121), the number density of decoupled particles (both relativistic and non-relativistic) is given by

$$n(t) = \left[\frac{a(t_\mathrm{f})}{a(t)}\right]^3 n_\mathrm{eq}(t_\mathrm{f}) , \tag{3.156}$$

so that $n \propto a^{-3}$, as expected. For relativistic species, we can contrast this number density against that of the photons:

$$\frac{n(t)}{n_\gamma(t)} = \frac{g_{\rm eff}}{2}\left(\frac{T_{\rm f}}{T}\right)^3\left[\frac{a(t_{\rm f})}{a(t)}\right]^3 = \frac{g_{\rm eff}}{2}\frac{g_{*,s}(T)}{g_{*,s}(T_{\rm f})} \tag{3.157}$$

with $g_{\rm eff} = g$ for bosons and $g_{\rm eff} = (3/4)g$ for fermions, where we have used that the photon temperature, T, scales as in Eq. (3.148). This illustrates that the number density of any relic background of relativistic particles is comparable to the number density of photons. Note that Eq. (3.156) remains valid even if the particles become non-relativistic some time after decoupling.

3.3.5 The Freeze-Out of Stable Particles

Having discussed the distribution functions of particles before and after decoupling, we now turn to discuss the actual process by which a species decouples ('freezes out') from the hot plasma. We first consider cases where the particles involved are stable (i.e. their half-time of decay is much longer than the age of the Universe), and derive their relic abundances. We distinguish between 'hot' relics, which correspond to species that decouple in the relativistic regime, and 'cold' relics, whose decoupling takes place when the particles have already become non-relativistic.

The evolution of the particle number density is governed by the Boltzmann equation, which, for a given species 'i', can be written as

$$\frac{{\rm d}f_i}{{\rm d}t} = C_i[f]\,, \tag{3.158}$$

where $C_i[f]$ (called the collisional term) describes the change of the distribution function of species 'i' due to the interactions with other species. Since the Universe is homogeneous and isotropic, f_i depends only on the cosmic time, t, and the value of the momentum, $p \propto a^{-1}(t)$. It then follows from Eq. (3.158) that

$$\frac{\partial f_i}{\partial t} - H(t)p\frac{\partial f_i}{\partial p} = C_i[f]\,, \tag{3.159}$$

where $H = \dot{a}/a$ is the Hubble parameter. Integrating both sides of Eq. (3.159) over momentum space, and using the definition of n_i, we obtain

$$\frac{{\rm d}n_i}{{\rm d}t} + 3H(t)\,n_i = \int C_i[f]\,{\rm d}^3\mathbf{p}\,. \tag{3.160}$$

Here the second term on the left-hand side (often called the Hubble drag term) describes the dilution of the number density due to the expansion of the Universe, while the right-hand side describes the change in number density due to interactions. Note that in the limit $C_i[f] \to 0$ the number density scales as $n_i \propto a^{-3}$, as expected.

In general, the collisional term $C_i[f]$ depends on f_i and on the distribution functions of all other species that interact with 'i'. If the cross-sections of all these interactions are known (from relevant physics), we can obtain the functional form of $C_i[f]$. Species that do not have any channel to interact with 'i' collisionally can still affect the distribution function of 'i' via their contributions to the general expansion of the Universe. Thus, the evolution of the matter content of the Universe is described by a coupled set of Boltzmann equations for all important species in the Universe, which can in principle be solved once the initial conditions are given.

For illustration, consider a case in which species 'i' takes part *only* in the following two-body interactions:

$$i + j \leftrightarrow a + b\,. \tag{3.161}$$

If the production and destruction rates of 'i' due to this reaction are $\alpha(T)$ and $\beta(T)$, respectively, then Eq. (3.160) can be written as

$$\frac{\mathrm{d}n_i}{\mathrm{d}t} + 3H(t)\,n_i = \alpha(T) n_a n_b - \beta(T) n_i n_j\,. \tag{3.162}$$

The meaning of this equation is clear: particles of species 'i' are destroyed due to their reactions with species 'j', and are created due to the reactions between species 'a' and 'b'. A similar equation can be written for 'j'. Subtracting these two equations gives $(n_i - n_j)a^3 = \text{constant}$. Now suppose that '$a$' and '$b$' are in thermal equilibrium with the general hot plasma, so that their distribution functions are given by Eq. (3.124) with $T_a = T_b = T$, while 'i' and 'j' are coupled to the hot plasma through their reactions with 'a' and 'b'. We define an equilibrium density for 'i', $n_{i,\mathrm{eq}}$, and an equilibrium density for 'j', $n_{j,\mathrm{eq}}$, so that

$$\beta(T) n_{i,\mathrm{eq}} n_{j,\mathrm{eq}} = \alpha(T) n_a n_b\,. \tag{3.163}$$

Thus defined, $n_{i,\mathrm{eq}}$ and $n_{j,\mathrm{eq}}$ are just the number densities of 'i' and 'j' under the assumption that they are in thermal equilibrium with the hot plasma. Consider the case in which 'j' and 'b' are the antiparticles of 'i' and 'a', respectively. As long as the chemical potential of 'i' is small, the number densities of 'i' and 'j' will be virtually identical [see Eq. (3.133)]. In what follows we therefore set $n_i = n_j$, but note that the discussion is easily extended to cases where $n_i \neq n_j$ by using that $(n_i - n_j)a^3 = \text{constant}$. With these definitions, we can write the rate equation (3.162) as

$$\frac{\mathrm{d}n_i}{\mathrm{d}t} + 3H(t)\,n_i = \beta(T)(n_{i,\mathrm{eq}}^2 - n_i^2)\,. \tag{3.164}$$

Since the entropy density s is proportional to a^{-3} (see §3.3.3), it is convenient to define both n_i and $n_{i,\mathrm{eq}}$ in units of s:

$$Y_i \equiv \frac{n_i}{s}, \qquad Y_{i,\mathrm{eq}} \equiv \frac{n_{i,\mathrm{eq}}}{s}\,. \tag{3.165}$$

Using $\mathrm{d}s/\mathrm{d}t = -3Hs$, Eq. (3.164) becomes

$$\frac{\mathrm{d}Y_i}{\mathrm{d}t} = \beta(T) s(T) (Y_{i,\mathrm{eq}}^2 - Y_i^2)\,. \tag{3.166}$$

If we now introduce the dimensionless variable, $x \equiv m_i/T$, and use the fact that, in the radiation dominated era, $t \propto a^2 \propto T^{-2}$ (or $t = t_{\mathrm{m}} x^2$, where t_{m} is the cosmic time when $x = 1$), the rate equation can be written in the following form:

$$\frac{x}{Y_{i,\mathrm{eq}}} \frac{\mathrm{d}Y_i}{\mathrm{d}x} = -\frac{\Gamma(x)}{H(x)} \left[\left(\frac{Y_i}{Y_{i,\mathrm{eq}}} \right)^2 - 1 \right], \tag{3.167}$$

where $\Gamma(x) \equiv n_{i,\mathrm{eq}}(x)\beta(x)$ and $H = (2t)^{-1} = (2t_{\mathrm{m}}x^2)^{-1}$ (which follows from $a \propto t^{1/2}$).

Given a particle species' rest mass m_i and its interaction cross-section $\beta(T) = \langle \sigma v \rangle(T)$, thermally averaged over all reactions in which 'i' partakes, the rate equation (3.167) can be solved for $Y_i(x)$ numerically. The initial conditions follow from the fact that for $x \ll 1$ the solution is given by $Y_i = Y_{i,\mathrm{eq}}$. Fig. 3.5 shows the solutions of Y_i thus obtained for different values of β [here assumed to be constant, $\beta(T) = \beta_0$]. A larger interaction cross-section (larger β_0) implies that the species can maintain thermal equilibrium for a longer time. As long as β_0 is such that decoupling occurs in the relativistic regime ($x \ll 1$), the final freeze-out abundance will be comparable to that of the photons [see Eq. (3.157)], and depend very little on the exact value of β_0. For sufficiently large β_0, the particles remain in thermal equilibrium well into the non-relativistic regime ($x \gg 1$), causing an exponential suppression of their final freeze-out abundance. In this regime the relic abundances are extremely sensitive to β, and thus to the exact epoch of decoupling.

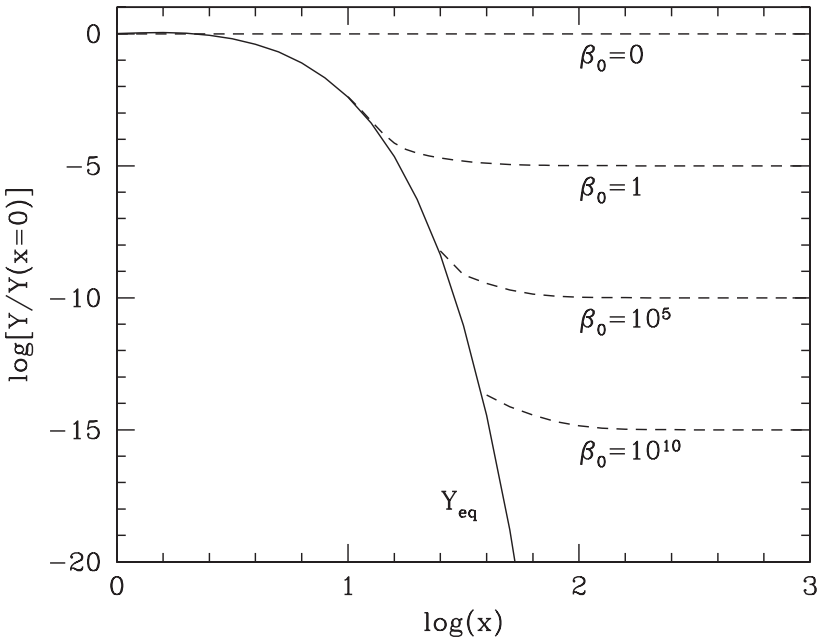

Fig. 3.5. The solution of Eq. (3.167) assuming a constant annihilation cross-section; $\beta = \beta_0$ (dashed curves). The solid curve shows the equilibrium abundance.

In what follows we present a simple, but relatively accurate, estimate of the relic abundances of various particle species. Rather than solving Eq. (3.167), which needs to be done numerically or by other approximate methods, we make the assumption that freeze-out occurs at a temperature $T_{\rm f}$, corresponding to $x_{\rm f}$, when $\Gamma/H = 1$, and that the relic abundance is simply given by $Y_i(x \to \infty) = Y_{i,{\rm eq}}(x_{\rm f})$. Using Eqs. (3.128) and (3.130) for $n_{i,{\rm eq}}$, and Eq. (3.146) for s, we have

$$Y_{i,{\rm eq}}(x) = \begin{cases} (45\zeta(3)/2\pi^4)[g_{i,{\rm eff}}/g_{*,s}(x)] & (x \ll 1) \\ (90/(2\pi)^{7/2})[g_i/g_{*,s}(x)]x^{3/2}e^{-x} & (x \gg 1), \end{cases} \tag{3.168}$$

where $g_{i,{\rm eff}} = g_i$ for bosons and $g_{i,{\rm eff}} = (3/4)g_i$ for fermions. The freeze-out temperature follows from $\Gamma(x_{\rm f}) = n_{i,{\rm eq}}(x_{\rm f})\beta(x_{\rm f}) = H(x_{\rm f})$. From Eq. (3.61) we have that in the radiation dominated era $H^2(t) = (8\pi G/3)\rho_{\rm r}(t)$. Substitution of Eq. (3.134) then gives

$$H(x) = \left(\frac{m_i m_{\rm Pl}}{x}\right)^2 \sqrt{\frac{4\pi^3 g_*(x)}{45}}\,, \tag{3.169}$$

where $m_{\rm Pl} = G^{-1/2}$ is the Planck mass in the natural units used here. Our definition of freeze-out then yields

$$\begin{aligned} x_{\rm f} &= \sqrt{\frac{45}{\pi^7}}\frac{\zeta(3)}{2}\frac{g_{i,{\rm eff}}}{\sqrt{g_{*,s}(x_{\rm f})}}\, m_{\rm Pl}\, m_i\, \beta(x_{\rm f}) \quad (x_{\rm f} \ll 1)\,; \\ x_{\rm f}^{-1/2} e^{x_{\rm f}} &= \sqrt{\frac{45}{32\pi^6}}\frac{g_i}{\sqrt{g_{*,s}(x_{\rm f})}}\, m_{\rm Pl}\, m_i\, \beta(x_{\rm f}) \qquad (x_{\rm f} \gg 1)\,. \end{aligned} \tag{3.170}$$

Note that since $x_{\rm f}$ appears on both sides of these equations, they typically need to be solved numerically.

Let us first consider the case of hot relics that have remained relativistic to the present day, i.e. their rest mass $m_i \ll T_0 = 2.4 \times 10^{-4}\,{\rm eV}$. Its energy density follows from Eq. (3.131), which can

be written in terms of the photon energy density, as is done in Eq. (3.157) for the number density. Expressing this energy density in terms of the critical density for closure, we obtain

$$\Omega_{i,0}h^2 = \frac{g_{i,\mathrm{eff}}}{2}\left[\frac{g_{*,s}(x)}{g_{*,s}(x_\mathrm{f})}\right]^{4/3}\Omega_{\gamma,0}h^2\,. \tag{3.171}$$

Since $g_{*,s}(x) \leq g_{*,s}(x_\mathrm{f})$, and since $\Omega_{\gamma,0}h^2 = 2.5\times10^{-5}$ [see Eq. (3.65)], we immediately see that a relic particle that is still relativistic today (e.g. zero mass neutrinos) contributes negligibly to the total energy density of the Universe at the present time.

Next we consider the case of weakly interacting massive particles, usually called WIMPs. Examples of WIMPs are massive neutrinos and stable, light supersymmetric particles. Note that WIMPs can be either hot or cold, depending on whether $x_\mathrm{f} \ll 1$ or $x_\mathrm{f} \gg 1$. The present-day mass density of massive relics is $\rho_{i,0} = m_i Y_{i,\mathrm{eq}}(x_\mathrm{f})s_0$, with s_0 the present-day value of the entropy density. After electron annihilation, $g_{*,s} = 2 + (7/8)\times3\times2\times1\times(4/11) = 3.91$. Substituting this in Eq. (3.146) and using $T_0 = 2.73\,\mathrm{K}$ gives $s_0 = 2{,}906\,\mathrm{cm}^{-3}$. For hot relics, we then obtain

$$\Omega_{i,0}h^2 \approx 7.64\times10^{-2}\left[\frac{g_{i,\mathrm{eff}}}{g_{*,s}(x_\mathrm{f})}\right]\left(\frac{m_i}{\mathrm{eV}}\right). \tag{3.172}$$

This abundance depends only very weakly on the exact moment of freeze-out, x_f, reflecting the fact that $Y_i(x)$ is virtually constant for $x \ll 1$. Since $\Omega_0 h^2 \lesssim 1$, we obtain a cosmological bound to the mass of hot relics,

$$m_i \lesssim 13.1\,\mathrm{eV}\left[\frac{g_{*,s}(x_\mathrm{f})}{g_{i,\mathrm{eff}}}\right]. \tag{3.173}$$

For massive neutrinos, $g_{*,s}(x_\mathrm{f}) = 43/4$ and $g_{i,\mathrm{eff}} = 6/4$ (assuming $g_i = 2$ to account for antiparticles), the limit is $m_i \lesssim 93.8\,\mathrm{eV}$.

Finally we examine cold WIMPs, which are considered to be candidates for the cold dark matter. Solving Eq. (3.170) for e^{-x_f}, and substituting the result in Eq. (3.168) gives

$$Y_{i,\mathrm{eq}}(x) = \sqrt{\frac{45}{\pi}}\frac{x_\mathrm{f}}{\sqrt{g_{*,s}(x_\mathrm{f})}}[m_\mathrm{pl}\,m_i\,\beta(x_\mathrm{f})]^{-1}\,. \tag{3.174}$$

Using the present-day entropy density s_0 we obtain a density parameter for cold relics:

$$\Omega_{i,0}h^2 \approx 0.86\frac{x_\mathrm{f}}{\sqrt{g_{*,s}(x_\mathrm{f})}}\left[\frac{\beta(x_\mathrm{f})}{10^{10}\,\mathrm{GeV}^{-2}}\right]^{-1}. \tag{3.175}$$

Contrary to the case of hot relics, $\Omega_{i,0}h^2$ now depends strongly on the interaction cross-section, owing to the exponential decrease of $Y_{i,\mathrm{eq}}(x)$ in the non-relativistic regime. As an example, consider a (hypothetical) stable neutrino species with $m_i \gg 1\,\mathrm{MeV}$ but less than $m_Z \sim 100\,\mathrm{GeV}$ (the mass of the Z boson). Because of its large mass, $x_\mathrm{f} \gg 1$ and its relic abundance follows from Eq. (3.175). For neutrinos, the annihilation rate can be approximately written as

$$\beta(x) \approx \frac{c_2}{2\pi}G_\mathrm{F}^2 m_i^2 x^{-b}\,, \tag{3.176}$$

with G_F the Fermi coupling constant, and c_2 a constant depending on the type of neutrinos ('Dirac' or 'Majorana'). The value of b is determined by the details of the annihilation processes involved, but is typically of the order unity. Substituting Eq. (3.176) in Eq. (3.175) yields

$$\Omega_{i,0}h^2 \approx \frac{3.95}{c_2}\frac{x_\mathrm{f}^{b+1}}{\sqrt{g_{*,s}(x_\mathrm{f})}}\left[\frac{m_i}{\mathrm{GeV}}\right]^{-2}. \tag{3.177}$$

For Dirac-type neutrinos $c_2 \sim 5$ and $b = 0$ (Kolb & Turner, 1990). Taking $g_i = 2$ (to also account for the antiparticles) and $g_{*,s} \sim 60$ at around the time of freeze-out, and solving Eq. (3.170) for x_f gives

$$x_\mathrm{f} \approx 17.8 + 3\ln(m_i/\,\mathrm{GeV})\,, \tag{3.178}$$

so that

$$\Omega_{i,0}h^2 \approx 1.82\left(\frac{m_i}{\mathrm{GeV}}\right)^{-2}\left[1 + 0.17\ln\left(\frac{m_i}{\mathrm{GeV}}\right)\right]. \tag{3.179}$$

The cosmological bound, $\Omega_0 h^2 \lesssim 1$, to the mass of massive neutrinos is thus

$$m_i \gtrsim 1.4\,\mathrm{GeV}\,. \tag{3.180}$$

Note that $\Omega_{i,0}$ *decreases* with increasing particle mass. This reflects the fact that the annihilation cross-section in Eq. (3.176) increases as m_i^2, so that more massive species can stay in thermal equilibrium longer, resulting in a lower freeze-out abundance. The cross-section will not continue to grow as m_i^2 indefinitely, however. For particles with $m_i \gg m_Z \simeq 100\,\mathrm{GeV}$ the cross-section actually *decreases* with particle mass as m_i^{-2}. Using the same argument as above and inserting the appropriate numbers, we find

$$\Omega_{i,0}h^2 \approx \left(\frac{m_i}{3\,\mathrm{TeV}}\right)^2. \tag{3.181}$$

Therefore, the cosmological bound to the mass of such species is

$$m_i \lesssim 3\,\mathrm{TeV}\,. \tag{3.182}$$

Fig. 3.6 summarizes the relation between the WIMP mass (assumed to interact as a Dirac-type neutrino) and its relic contribution to the cosmological density parameter. At $m_\mathrm{wimp} \lesssim$ MeV the WIMPs produce 'hot' relics for which $\Omega_\mathrm{wimp}h^2 \propto m_\mathrm{wimp}$.[7] At particle masses above $\sim 1\,\mathrm{MeV}$, decoupling occurs in the non-relativistic regime, resulting in 'cold' relics for which $\Omega_\mathrm{wimp}h^2 \propto m_\mathrm{wimp}^{-2}$. Finally, for particle masses above that of the Z boson ($m_\mathrm{wimp} \gtrsim 100\,\mathrm{GeV}$) the scaling changes to $\Omega_\mathrm{wimp}h^2 \propto m_\mathrm{wimp}^2$. Combining these results with observational constraints on the cosmological density parameter ($0.1 \lesssim \Omega_0 h^2 \lesssim 1.0$), we find that there are only three narrow mass ranges of WIMPs allowed, at $\sim 30\,\mathrm{eV}$, $\sim 2\,\mathrm{GeV}$ and $\sim 2\,\mathrm{TeV}$ (see Fig. 3.6). Note, however, that these constraints are only valid under the assumption that the WIMPs have the same interaction cross-sections as neutrinos. Since the nature of the dark matter particles is still unknown, there are large uncertainties regarding the possible interaction cross-sections. Consequently, the observational constraints on $\Omega_0 h^2$ currently only constrain the combination of interaction cross-section and WIMP mass, and large ranges of WIMP masses are still allowed.

3.3.6 Decaying Particles

So far we have discussed the freeze-out of stable particles (those with a lifetime much larger than the age of the Universe) and their cosmological consequences. For unstable particles, the situation is different. In particular, if massive particles decay into photons and other relativistic particles, they will release energy into the Universe, and depending on how effectively this energy is thermalized, the decay may produce a radiation background, increasing the entropy of the Universe. Consider a heavy particle, 'h', with mass m_h and with a mean lifetime τ_h, which decays

[7] When their mass is this low, one normally would not speak of WIMPs, but of weakly interacting particles instead. For brevity, we also refer to these particles as WIMPs in Fig. 3.6.

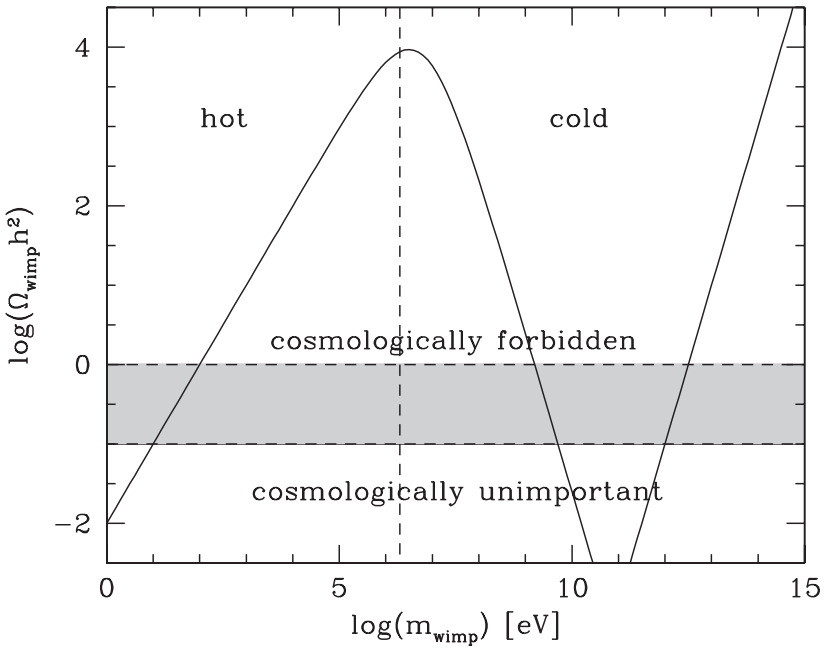

Fig. 3.6. Cosmological constraints on the mass of weakly interactive dark matter particles under the assumption that they interact as a Dirac-type neutrino. The solid curve shows the predicted cosmological density parameter of the WIMPs as a function of WIMP mass, while the shaded area roughly brackets the observed range of the cosmological density parameter. The mass ranges in which the particles make up 'hot' and 'cold' dark matter are indicated.

into light particles while it is non-relativistic. The number of decay events per proper volume at any time t is $n_h(t)/\tau_h$, with $n_h(t)$ given by

$$\frac{\mathrm{d}n_h}{\mathrm{d}t} + 3H(t)n_h = \alpha(T)n_a n_b - \beta(T)n_h n_j - n_h/\tau_h\,, \tag{3.183}$$

where, as an example, we assume that 'h' takes part in the reaction $h + j \leftrightarrow a + b$ in addition to the decay. Without implicitly solving Eq. (3.183), we can directly infer the evolution of $n_h(t)$ at two extremes. At early time when the reaction rate ($\sim \beta n_j$) is higher than both the decay rate ($1/\tau_h$) and the expansion rate (H), the species 'h' has the equilibrium abundance, and basically behaves as stable particles. At later times, when the right-hand side of Eq. (3.183) is dominated by decay, it is easy to show that

$$n_h(t) = n_h(t_\mathrm{D})\left[\frac{a(t)}{a(t_\mathrm{D})}\right]^{-3} \exp\left(-t/\tau_h\right), \tag{3.184}$$

where t_D is the time when the decay becomes more important than other reactions. If the rest mass of the decaying particles is thermalized, then the entropy density per unit comoving volume (see §3.1.5) increases with time as

$$\mathrm{d}S = -\frac{\mathrm{d}\left(n_h m_h a^3\right)}{T} = \frac{\rho_h a^3}{T}\frac{\mathrm{d}t}{\tau_h}\,, \tag{3.185}$$

where $\rho_h \equiv m_h n_h$. Using Eqs. (3.134) and (3.146), we have

$$\frac{\mathrm{d}S}{S} = \frac{3}{4}\frac{g_*}{g_{*,s}}\frac{\rho_h}{\rho_\mathrm{r}}\frac{\mathrm{d}t}{\tau_h}\,, \tag{3.186}$$

where $\rho_{\rm r}$ is the total energy density in relativistic particles. The entropy of the Universe can therefore be increased significantly if $\rho_h(\tau_h) \gtrsim \rho_{\rm r}(\tau_h)$, i.e. if the Universe is dominated by species 'h' at the time of decay. Since $\rho_h \propto a^{-3}$ while $\rho_{\rm r} \propto a^{-4}$, we can define a time of equality for species 'h' by $\rho_{\rm r}(t_{{\rm eq},h}) = \rho_h(t_{{\rm eq},h})$, and express the relative increase in $\rho_{\rm r}$ due to the decay of 'h' in terms of the ratio of $t_{{\rm eq},h}$ to the decay time τ_h:

$$\frac{\Delta\rho_{\rm r}}{\rho_{\rm r}} = \frac{\rho_h(\tau_h)}{\rho_{\rm r}(\tau_h)} = \frac{a(\tau_h)}{a(t_{{\rm eq},h})} = \left(\frac{\tau_h}{t_{{\rm eq},h}}\right)^{2/3} . \tag{3.187}$$

Therefore, any species with a decay time $t_{{\rm eq},h} \lesssim \tau_h \lesssim t_0$ can have caused a significant increase of $\rho_{\rm r}$. Such an increase can have profound impacts on the evolution of the Universe. If it occurs before radiation–matter equality it may cause a delay in the time $t_{\rm eq}$ when the Universe eventually becomes dominated by matter. Since perturbations cannot grow before the Universe becomes matter dominated, as we will see in the next chapter, such a particle decay can have a significant impact on the development of large-scale structure. An increase in $\rho_{\rm r}$ also causes the Universe to expand faster in the period $\tau_h \lesssim t \lesssim t_{\rm eq}$, affecting the production of other particle species during that era. For example, as we will see in the next section, the abundance of helium can be significantly affected if the decay occurs before primordial nucleosynthesis.

If the decay product contains photons, there are additional stringent limits on the mass and lifetime of the decaying particle. If the lifetime were comparable to the present age of the Universe, we would observe a strong radiation background in X-ray and gamma-ray produced by the decay. The lack of such background requires that either $\tau_h \gg t_0$ (i.e. the particle is almost stable) or that the decay occurs at a time when the Universe is still opaque to high-energy photons (so that they can be down-graded by scattering with matter). Another stringent constraint comes from the fact that the observed CMB has a blackbody spectrum to a very high degree of accuracy. This requires the decay occur at a time when high-energy photons can be effectively thermalized (see §3.5).

3.4 Primordial Nucleosynthesis

We all know that the Universe contains not only hydrogen (whose nuclei are single protons) but also heavier elements like helium, lithium, etc. An important question is therefore how these heavier elements were synthesized. Since nuclear reactions are known to be taking place in stars – for example, the luminosity of the Sun is powered mainly by the burning of hydrogen into helium – one possibility is that all heavier elements are synthesized in stars. However, the observed mass fraction of helium is roughly a constant everywhere in the Universe, suggesting that most of the helium is in fact primordial. In this section we examine how nucleosynthesis proceeds in the early Universe.

3.4.1 Initial Conditions

All nuclei are built up of protons and neutrons. Before we explore the nuclear reactions that synthesize deuterium, helium, lithium, etc., we therefore examine the abundances of their building blocks. Protons and neutrons have a very comparable rest mass of $\sim 940\,{\rm MeV}$, which implies that they become non-relativistic at very early times ($t \simeq 10^{-6}\,{\rm s}$, $T \simeq 10^{13}\,{\rm K}$). Down to a temperature of $\sim 0.8\,{\rm MeV}$ they maintain thermal equilibrium through weak interactions like

$$\mathrm{p} + \mathrm{e} \leftrightarrow \mathrm{n} + \nu_{\rm e}\,, \qquad \mathrm{n} + \overline{\mathrm{e}} \leftrightarrow \mathrm{p} + \overline{\nu}_{\rm e}\,. \tag{3.188}$$

In thermal equilibrium, their number densities follow from Eq. (3.128):

$$n_{\rm n,p} = 2\left(\frac{m_{\rm n,p}T}{2\pi}\right)^{3/2}\exp\left[-\frac{m_{\rm n,p}-\mu_{\rm n,p}}{T}\right], \tag{3.189}$$

where we have used that both protons and neutrons have two helicity states ($g_{\rm n}=g_{\rm p}=2$). Writing the mass difference as $Q\equiv m_{\rm n}-m_{\rm p}=1.294\,{\rm MeV}$, and using that $m_{\rm n}/m_{\rm p}\simeq 1$, we obtain the ratio between the number densities of protons and neutrons in thermal equilibrium:

$$\frac{n_{\rm n}}{n_{\rm p}} = \exp\left(-\frac{Q}{T}+\frac{\mu_{\rm n}-\mu_{\rm p}}{T}\right)\approx\exp\left(-\frac{Q}{T}\right), \tag{3.190}$$

where $\mu_{\rm n}-\mu_{\rm p}=\mu_{\rm e}-\mu_\nu\approx 0$ (see §3.3). When $T\gg 10^{10}\,{\rm K}$ the reactions (3.188) go equally fast in both directions and there are as many protons as neutrons. When the temperature decreases towards $\sim 1\,{\rm MeV}$, however, the number density of neutrons starts to drop with respect to that of protons, because neutron is slightly more massive. If thermal equilibrium were to be maintained, the ratio would continue to decrease to very small values. However, as we have seen in §3.3.3, at about the same temperature of $\sim 1\,{\rm MeV}$, neutrinos start to decouple. Therefore, the rate of the weak reactions (3.188) is no longer fast enough to establish thermal equilibrium against the expansion rate of the Universe, and the ratio $n_{\rm n}/n_{\rm p}$ will eventually 'freeze out' at a value of $\sim\exp(-1.294/0.8)\sim 0.2$. However, neutrons are unstable to beta decay,

$$\mathrm{n}\rightarrow\mathrm{p}+\mathrm{e}+\overline{\nu}_{\rm e}, \tag{3.191}$$

so that even after freeze-out the neutron-to-proton ratio continuous to decrease. If we define the neutron abundance as

$$X_{\rm n}\equiv\frac{n_{\rm n}}{n_{\rm n}+n_{\rm p}}, \tag{3.192}$$

then it evolves due to the neutron decay as

$$X_{\rm n}\propto\exp\left[-\frac{t}{\tau_{\rm n}}\right], \tag{3.193}$$

where $\tau_{\rm n}=(887\pm 2)\,{\rm s}$ is the mean lifetime of neutrons. The main reason that the present-day Universe contains a large abundance of neutrons is that, shortly before the Universe reaches an age $t=\tau_{\rm n}$, most neutrons have already ended up in helium nuclei (which stabilizes them against beta decay due to Pauli's exclusion principle) through the process of nucleosynthesis to be described below.

3.4.2 Nuclear Reactions

Nuclei can form in abundant amounts as soon as the temperature of the Universe has cooled down to temperatures corresponding to their binding energy, and the number densities of protons and neutrons are sufficiently high. For a (non-relativistic) species with mass number A and charge number Z [such a species will be called A(Z), and contains Z protons and $A-Z$ neutrons], the equilibrium number density can be obtained from Eq. (3.128):

$$n_{\rm A} = g_{\rm A}\left(\frac{m_{\rm A}T}{2\pi}\right)^{3/2}\exp\left(-\frac{m_{\rm A}-\mu_{\rm A}}{T}\right). \tag{3.194}$$

The chemical potential $\mu_{\rm A}$ is related to those of protons and neutrons as

$$\mu_{\rm A} = Z\mu_{\rm p}+(A-Z)\mu_{\rm n}, \tag{3.195}$$

which allows us to rewrite Eq. (3.194) as

$$n_{\rm A} = g_{\rm A}\left(\frac{m_{\rm A}T}{2\pi}\right)^{3/2}\exp\left(-\frac{m}{T}\right)\left[\exp\left(\frac{\mu_{\rm p}}{T}\right)\right]^{Z}\left[\exp\left(\frac{\mu_{\rm n}}{T}\right)\right]^{(A-Z)}. \tag{3.196}$$

Writing $\exp(\mu_{\rm p}/T)$ and $\exp(\mu_{\rm n}/T)$ in terms of the proton and neutron mass densities given by Eq. (3.189), respectively, and defining the nucleon mass $m_{\rm N} \equiv m_{\rm A}/A \approx m_{\rm n} \approx m_{\rm p}$, we obtain

$$n_{\rm A} = \frac{g_{\rm A}A^{3/2}}{2^A}n_{\rm p}^{Z}n_{\rm n}^{A-Z}\left(\frac{m_{\rm N}T}{2\pi}\right)^{3(1-A)/2}\exp\left(\frac{B_{\rm A}}{T}\right), \tag{3.197}$$

where

$$B_{\rm A} \equiv Zm_{\rm p} + (A-Z)m_{\rm n} - m_{\rm A} \tag{3.198}$$

is the binding energy of the species A(Z). Next we define the 'mass fraction' or 'abundance' of nucleus A as

$$X_{\rm A} \equiv \frac{An_{\rm A}}{n_{\rm b}}. \tag{3.199}$$

Here $n_{\rm b} \equiv n_{\rm n} + n_{\rm p} + \sum_i A_i n_{{\rm A},i}$ is the number density of baryons in the Universe, with the summation over all nuclear species so that $\sum_i X_{{\rm A},i} = 1$. Substituting Eq. (3.197) in Eq. (3.199) we obtain

$$X_{\rm A} = \frac{g_{\rm A}}{2}A^{5/2}\left[\frac{4\zeta(3)}{\sqrt{2\pi}}\right]^{A-1}X_{\rm p}^{Z}X_{\rm n}^{A-Z}\eta^{A-1}\left(\frac{m_{\rm N}}{T}\right)^{3(1-A)/2}\exp\left(\frac{B_{\rm A}}{T}\right), \tag{3.200}$$

where $\eta \equiv n_{\rm b}/n_\gamma$ is the present-day baryon-to-photon ratio. Since $n_\gamma = [2\zeta(3)/\pi^2]T^3$ [see Eq. (3.130)], and $T_0 = 2.73\,{\rm K}$, we have

$$\eta \equiv n_{\rm b}/n_\gamma \approx 2.72\times10^{-8}\Omega_{\rm b,0}h^2, \tag{3.201}$$

where $\Omega_{\rm b,0}$ is the present-day baryon density in terms of the critical density for closure. Eq. (3.200) reveals that species A(Z), with $A > 1$, can only be produced in appreciable amounts once the temperature has dropped to a value $T_{\rm A}$ given by

$$T_{\rm A} \sim \frac{|B_{\rm A}|}{(A-1)\left[|\ln\eta| + \frac{3}{2}\ln(m_{\rm N}/T)\right]}. \tag{3.202}$$

The binding energies of the lightest nuclei, such as deuterium and helium, are all of the order of a few MeV, corresponding to a temperatures of a few $\times10^{10}\,{\rm K}$. However, because of the small number of baryons per photon ($10^{-10} \lesssim \eta \lesssim 10^{-9}$), or, in other words, the high entropy per baryon, their synthesis has to wait until the Universe has cooled down to temperatures of the order of $(1\rightarrow3)\times10^9\,{\rm K}$.

At such low temperatures, however, the number densities of protons and neutrons are already much too low to form heavy elements by direct many-body reactions, such as $2{\rm n}+2{\rm p}\rightarrow{}^4{\rm He}$. Therefore, nucleosynthesis must proceed through a chain of two-body reactions. The dominant reactions in this chain are:

$$\mathrm{p}\,(\mathrm{n},\gamma)\,\mathrm{D} \tag{3.203}$$

$$\mathrm{D}\,(\mathrm{n},\gamma)\,{}^3\mathrm{H},\quad \mathrm{D}\,(\mathrm{D},\mathrm{p})\,{}^3\mathrm{H} \tag{3.204}$$

$$\mathrm{D}\,(\mathrm{p},\gamma)\,{}^3\mathrm{He},\quad \mathrm{D}\,(\mathrm{D},\mathrm{n})\,{}^3\mathrm{He},\quad {}^3\mathrm{H}\,(\mathrm{p},\mathrm{n})\,{}^3\mathrm{He},\quad {}^3\mathrm{H}\,(\,,\mathrm{e}\,\overline{\nu}_{\rm e})\,{}^3\mathrm{He} \tag{3.205}$$

$${}^3\mathrm{H}\,(\mathrm{p},\gamma)\,{}^4\mathrm{He},\quad {}^3\mathrm{H}\,(\mathrm{D},\mathrm{n})\,{}^4\mathrm{He},\quad {}^3\mathrm{He}\,(\mathrm{n},\gamma)\,{}^4\mathrm{He},\quad {}^3\mathrm{He}\,(\mathrm{D},\mathrm{p})\,{}^4\mathrm{He}, \tag{3.206}$$

$$2\,{}^3\mathrm{He}\,(\,,2\mathrm{p})\,{}^4\mathrm{He},\quad {}^7\mathrm{Li}\,(\mathrm{p},\,)\,2\,{}^4\mathrm{He} \tag{3.207}$$

$${}^4\mathrm{He}\,({}^3\mathrm{H},\gamma)\,{}^7\mathrm{Li},\quad {}^4\mathrm{He}\,({}^3\mathrm{He},\gamma)\,{}^7\mathrm{Be},\quad {}^7\mathrm{Be}\,(\mathrm{e},\nu_{\rm e})\,{}^7\mathrm{Li}. \tag{3.208}$$

Here the notation $X\,(a,b)\,Y$ indicates a reaction of the form $X + a \rightarrow Y + b$. Since the cross-sections for almost all these reactions are accurately known, the reaction network can be integrated numerically to compute the final abundances of all elements.

Note that the reaction network does not produce any elements heavier than lithium. This is a consequence of the fact that there are no stable nuclei with atomic weight 5 or 8. Since direct many-body reactions at earlier epoch are very inefficient in producing heavy elements, we can conclude that elements heavier than lithium are not produced by primordial nucleosynthesis. Indeed, as we will see in Chapter 10, heavy elements can be synthesized in stars where the density of helium is so high that a short-lived ^{8}Be, formed through ^{4}He–^{4}He collisions, can quickly capture another ^{4}He to form a stable carbon nucleus (^{12}C), thus allowing further nuclear reactions to proceed.

Inspection of the reaction network of primordial nucleosynthesis given above reveals that it can only proceed if the first step, the production of deuterium, is sufficiently efficient. Since deuterium has the lowest binding energy of all nuclei in the network, its production serves as a 'bottleneck' to get nucleosynthesis started. The production of deuterium through $\mathrm{p(n,\gamma)D}$ has a rate per free neutron given by

$$\begin{aligned}\Gamma &= (4.55\times10^{-20}\,\mathrm{cm^3\,s^{-1}})n_\mathrm{p} \\ &\approx 2.9\times10^4 X_\mathrm{p}\Omega_{\mathrm{b},0}h^2\left(\frac{T}{10^{10}\,\mathrm{K}}\right)^3\,\mathrm{s^{-1}}\,,\end{aligned} \tag{3.209}$$

which is much larger than the expansion rate $H \sim (T/10^{10}\,\mathrm{K})^2\,\mathrm{s^{-1}}$. Therefore, for temperatures $T \gtrsim 5\times10^8\,\mathrm{K}$, deuterium nuclei are always produced with the equilibrium abundance:

$$X_\mathrm{D} \approx 16.4\eta X_\mathrm{n}X_\mathrm{p}\eta\left(\frac{m_\mathrm{N}}{T}\right)^{-3/2}\exp\left(\frac{2.22\,\mathrm{MeV}}{T}\right). \tag{3.210}$$

From this we see that large amounts of deuterium are only produced once the temperature drops to $T_\mathrm{D} \sim 10^9\,\mathrm{K}$ [see also Eq. (3.202)]. This occurs when the Universe is about 100 seconds old, and signals the onset of primordial nucleosynthesis. The subsequent reaction chain proceeds very quickly, because at $T \simeq T_\mathrm{D}$ all nuclei heavier than deuterium can possess high equilibrium abundances. However, nuclei heavier than helium are still rare because of the instability of nuclei with $A = 5$ and $A = 8$, and because the temperature is already too low to effectively overcome the large Coulomb barrier in reactions like $^4\mathrm{He}(^3\mathrm{H},\gamma)^7\mathrm{Li}$ and $^4\mathrm{He}(^3\mathrm{He},\gamma)^7\mathrm{Be}$. As a result, almost all free neutrons existing at the onset of nucleosynthesis will be bound into ^{4}He, the most tightly bound species with $A < 5$. The mass fraction of ^{4}He can therefore be approximately written as

$$Y \equiv X_{^4\mathrm{He}} \approx \frac{4(n_\mathrm{n}/2)}{n_\mathrm{n}+n_\mathrm{p}} = \frac{2(n_\mathrm{n}/n_\mathrm{p})_\mathrm{D}}{1+(n_\mathrm{n}/n_\mathrm{p})_\mathrm{D}}\,, \tag{3.211}$$

where $(n_\mathrm{n}/n_\mathrm{p})_\mathrm{D}$ is the neutron-to-proton ratio at $T = T_\mathrm{D}$.

3.4.3 Model Predictions

Once the relevant reactions are specified and their cross-sections are given, the nucleosynthesis reaction network can be integrated forwards from the initial conditions at early times to make detailed predictions for the abundances of all species. This was first done with a complete network by Wagoner et al. (1967), and subsequent work using updated cross-sections and modernized computer codes (e.g. Wagoner, 1973; Walker et al., 1991; Cyburt et al., 2008) has modified their conclusions rather little. Detailed calculations show that the bulk of nucleosynthesis occurs at $t \approx 300\,\mathrm{s}$ ($T \approx 0.8\times10^9\,\mathrm{K} = 0.07\,\mathrm{MeV}$), in agreement with the qualitative

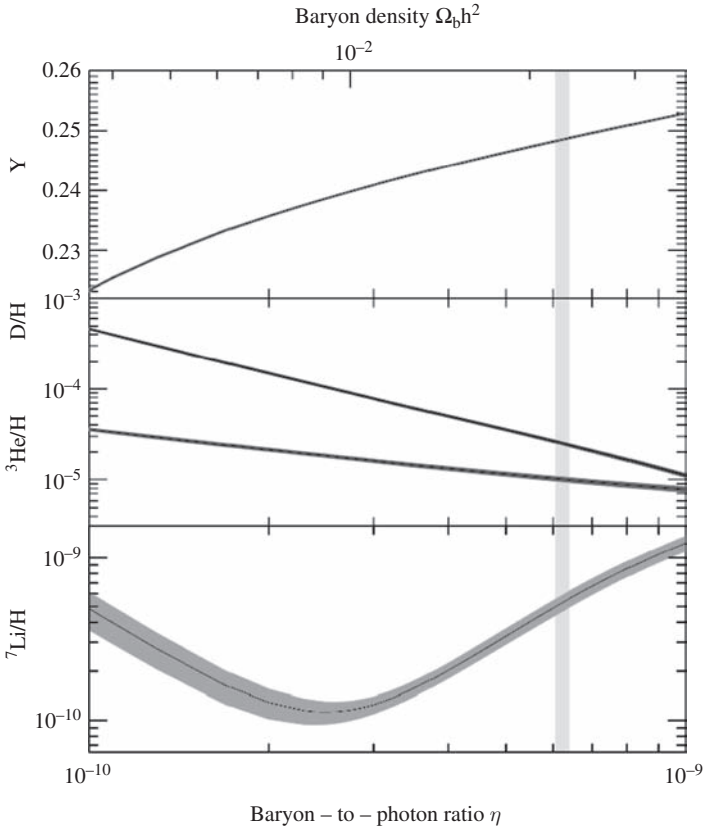

Fig. 3.7. Primordial abundances of light elements as a function of the baryon-to-photon ratio, η. The line thicknesses in each panel reflect the remaining theoretical uncertainties, while the vertical shaded band shows the range of η consistent with the WMAP measurements of fluctuations in the microwave background. [Courtesy of R. Cyburt; see Cyburt et al. (2008)]

arguments given above. At this point in time, the neutron-to-proton ratio n_n/n_p is about $1/7$. Using Eq. (3.211) this implies a final abundance of primordial ^{4}He of

$$Y_p \equiv X_{^4\mathrm{He}} \approx 1/4\,. \tag{3.212}$$

Observations of the mass fraction of helium everywhere and always give values of about 24%, which would be very difficult to understand if such an abundance were not primordial. This prediction (3.212) is therefore considered a great success of the standard Big Bang model.

The primordial abundances predicted by an updated version of the code of Wagoner et al. (1967) are shown in Fig. 3.7. Note that the abundances of deuterium and ^{3}He are about three orders of magnitude below that of ^{4}He, while that of ^{7}Li is nine orders of magnitude smaller; all other nuclei are expected to be much less abundant. The predicted abundances of light elements depend on three parameters: the baryon-to-photon ratio, η, the mean lifetime of the neutron, τ_n, and $g_*(T \sim 10^{10}\,\mathrm{K})$, which measures the number of degrees of freedom of effectively massless particles at the relevant temperature $T \sim 10^{10}\,\mathrm{K}$ [see Eq. (3.135)]. Given the discussion earlier in this section, we can understand the sensitivity of the abundances to all these parameters.

As η increases ($|\ln\eta|$ decreases), nucleosynthesis of D, ^{3}He and ^{3}H starts slightly earlier [see Eq. (3.202)]. As a result, the synthesis of ^{4}He commences at an earlier epoch when the depletion of neutrons by beta decay is less significant, and so more neutrons are bound into

^{4}He. This explains why the ^{4}He abundance increases with η. Since the age of the Universe at temperature T_D is smaller than τ_n, the neutron-to-proton ratio decreases only slowly at the time when nucleosynthesis begins. Therefore, the η-dependence of the ^{4}He abundance is weak. However, since the burning rates of D and ^{3}He are proportional to their equilibrium abundances, which increase with η as $X_A \propto \eta^{A-1}$ [see Eq. (3.200)], a larger baryon-to-photon ratio results in a smaller abundance of these two nuclei. The more complex behavior of the ^{7}Li abundance is a result of competition between the formation and destruction reactions in the network. Direct formation dominates at small η, and formation via ^{7}Be dominates at large η.

The neutron mean lifetime affects the predicted helium abundance by influencing the number density of neutrons at the onset of nucleosynthesis. An increase of τ_n leads to an increase in the number of neutrons and so to an increase in Y_p. In the relevant range of η, $\Delta Y_p \sim 2 \times 10^{-4} (\Delta\tau_n / 1\,\mathrm{s})$, implying that the uncertainty in Y_p arising from that of τ_n is quite small.

Finally, substituting Eq. (3.134) in Eq. (3.80), and using that $\rho_r \propto a^{-4}$, one finds

$$t = \left(\frac{45}{16\pi^3 G g_*} \right)^{1/2} T^{-2}. \tag{3.213}$$

Since $H = (2t)^{-1}$, the expansion rate $H \propto \sqrt{g_*}\, T^2$. Consequently, an increase of g_* leads to a faster expansion rate for given T. This raises the temperature at which the reaction rates equal the expansion rate, thus increasing the neutron-to-photon ratio at 'freeze-out'. Consequently, the predicted ^{4}He abundance increases with increasing g_*. In the relevant range of η, $\Delta Y_p \sim 0.01 \Delta g_*$. Therefore, the abundance of primordial helium provides a stringent constraint on the number of relativistic species at $T \gtrsim 10^9$ K. The standard model of primordial nucleosynthesis assumes these species to be photons and three species of massless neutrinos.

3.4.4 Observational Results

The predictions of primordial nucleosynthesis are of vital importance in the standard cosmology, and therefore much effort has been devoted to the observational determination of the primordial abundances of the light elements. Such determination can be used not only to constrain the number of relativistic species at the time of nucleosynthesis, but also to constrain η and so the number density of baryons in the Universe through Eq. (3.201). Unfortunately, precise determination of the primordial abundances is far from trivial. They usually rely on the emission or absorption of gas clouds due to the ions of the element in consideration. Turning this into an abundance often requires careful modeling of the properties of the observed cloud. An even greater problem comes from the fact that the material we observe today may have been processed through stars, so that (often uncertain) corrections have to be applied in order to derive a 'primordial' abundance. In the following we give a brief summary of the present observational situation.

- **Helium-4:** Because the abundance of helium is large, it is relatively easy to determine. Most measurements are made from HII clouds where the gas is highly ionized, and the abundance of both helium and hydrogen can be inferred from the strengths of their recombination lines. Since ^{4}He is also synthesized in stars, some of the observed ^{4}He may not be primordial. In order to reduce this contamination, it is desirable to use metal-poor clouds, as stars which produce the ^{4}He contamination also produce metals. Observations have been made for clouds with different metalicities, and an extrapolation to zero metalicity gives $Y_p = 0.24 \pm 0.01$ (e.g. Fields & Olive, 1998). From Fig. 3.7 we see that this observational result requires $\eta = (1.2 \rightarrow 8) \times 10^{-10}$. Since the predicted Y_p depends only weakly on η, extremely precise measurements are needed to give a more stringent constraint.

- **Deuterium:** Because of its strong dependence on η, the measurement of the primordial deuterium abundance is crucial in determining $\Omega_{b,0}$. Accurate determinations of the deuterium abundance have been obtained from UV absorption measurements in the local interstellar medium (ISM). The deuterium-to-hydrogen ratio (in mass) is found to be $[\mathrm{D/H}]_{\mathrm{ISM}} \approx 1.6\times10^{-5}$ (e.g. Linsky et al., 1995). Since deuterium is weakly bound, it is easy to destroy but hard to produce in stars. Therefore, this observed ISM value represents a lower limit on the primordial abundance. An alternative estimate of the deuterium abundance can be obtained from the absorption strength in Lyman-α clouds along the line-of-sight to quasars at high redshift. Since these high-redshift clouds are metal poor and perhaps not yet severely contaminated by stars, the deuterium abundance thus derived may actually be close to the primordial one. The observational data are still relatively sparse. The values of [D/H] obtained originally ranged from $\sim 2.4\times10^{-5}$ (Tytler et al., 1996) to $\sim 2\times10^{-4}$ (Webb et al., 1997) but now seem to have settled at $2.82\pm0.53\times10^{-5}$ (Pettini et al., 2008). This agrees well with the value of η inferred from WMAP data on microwave background fluctuations.
- **Helium-3:** The abundance of ^{3}He has been measured both in the solar system (using meteorites and the solar wind) and in HII regions (based on the strength of the $^3\mathrm{He}^+$ hyperfine line, the equivalent of the 21 cm hyperfine line of neutral hydrogen). The abundance inferred from HII regions is $[^3\mathrm{He/H}] = (1.3 \rightarrow 3.0)\times10^{-5}$ (e.g. Gloeckler & Geiss, 1996). A similar abundance, $[^3\mathrm{He/H}] = (1.4\pm0.4)\times10^{-5}$, is obtained from the oldest meteorites, the carbonaceous chondrites. Since these meteorites are believed to have formed at about the same time as the solar system, the observed abundance may be representative of pre-solar material. The abundance of ^{3}He in the solar wind has been determined by analyzing gas-rich meteorites and lunar soil. Because D is burned to ^{3}He during the Sun's approach towards the main sequence, the observed ^{3}He in the solar system may be a good measure of the pre-solar sum $(\mathrm{D}+{}^3\mathrm{He})$. All the measurements are consistent with $[(\mathrm{D}+{}^3\mathrm{He})/\mathrm{H}] \approx (4.1\pm0.6)\times10^{-5}$. Although ^{3}He can be reduced by stellar burning, it is much more difficult to destroy than deuterium and the reduction factor is no more than a factor of 2. The measurements in the solar system therefore give an upper limit on the primordial abundance of $[(\mathrm{D}+{}^3\mathrm{He})/\mathrm{H}]_{\mathrm{p}} \lesssim 10^{-4}$, corresponding to a lower limit of $\eta \gtrsim 3\times10^{-10}$.
- **Lithium-7:** Estimates of the ^{7}Li abundance come from stellar atmospheres. Since ^{7}Li is quite fragile, it can be depleted by circulation through the centers of stars. The observational estimates therefore vary from one stellar population to another. Since mass circulation (convection) does not go as deep in metal-poor stars as in metal-rich ones, it is desirable to use metal-poor stars where the depletion of ^{7}Li from the atmosphere is expected to be smaller. There have been attempts to observe ^{7}Li lines in the atmospheres of old stars with very low metalicity (e.g. Spite & Spite, 1982), from which the primordial ^{7}Li abundance was originally inferred to be $[^7\mathrm{Li/H}]_{\mathrm{p}} \approx (1.1\pm0.4)\times10^{-10}$. More recent attempts paying close attention to systematics give values in the range 1.0 to 1.5×10^{-10} (Asplund et al., 2006). From Fig. 3.7 we see that this abundance is inconsistent by a factor of about 4 with the value $[^7\mathrm{Li/H}]_{\mathrm{p}} = 5.24\pm0.7\times10^{-10}$ inferred from the five-year WMAP data on fluctuations in the microwave background.

At the present time, Big Bang nucleosynthesis is essentially a parameter-free theory. Improvements in experimental determinations of the neutron lifetime have shrunk the uncertainties so that they are no longer significant for this problem; the standard model of particle physics is now sufficiently constrained by accelerator experiments that the number of light particle species present at nucleosynthesis cannot differ significantly from the standard value; and WMAP measurements of the power spectrum of the cosmic microwave background lead to a photon-to-baryon ratio estimate, $\eta = 6.23\pm0.17\times10^{-10}$ (see §2.10.1). With these parameters the theory gives quite precise predictions for all the light element abundances. These agree with observational estimates of the

observed abundance of ^{4}He and D, where the first is only weakly constraining because of its logarithmic dependence on parameters, but the second can be considered a major success. The situation with ^{3}He is too complex for a meaningful comparison to be possible, and the results for ^{7}Li appear to disagree with observation. While this discrepancy may still reflect observational difficulties in inferring the primordial abundance of ^{7}Li, it may also be an indicator of unexpected physics in the early Universe. Notice that independent of inferences from microwave background observations, the baryon density required for successful primordial nucleosynthesis is much too small to be consistent with the large amounts of dark matter required to bind groups and clusters of galaxies, thus providing an independent argument in favor of non-baryonic dark matter (see §2.5).

3.5 Recombination and Decoupling

Immediately after primordial nucleosynthesis (when $T \sim 0.1\,\mathrm{MeV} \sim 10^9\,\mathrm{K}$) the Universe consists mainly of the following particles: hydrogen nuclei (i.e. protons), ^{4}He nuclei, electrons, photons, and decoupled neutrinos. Since the temperature is already lower than $m_\mathrm{e} = 0.51\,\mathrm{MeV}$, baryons and electrons can all be considered non-relativistic. All the particles (except the decoupled neutrinos) interact through electromagnetic processes, such as free–free interactions among charged particles, Compton scattering between charged particles and photons, and the recombinations[8] of ions with electrons to form atoms. In this section, we examine these processes in connection to several important cosmological events at $T < 10^9\,\mathrm{K}$.

3.5.1 Recombination

As soon as the temperature of the Universe drops below $\sim 13.6\,\mathrm{eV}$, electrons and protons start to combine to form hydrogen atoms. Here we examine how this 'recombination' process proceeds. In addition, we compute the fractions of electrons and protons that remain unbound after recombination, namely the 'freeze-out' abundances of free electrons and protons. For simplicity, we ignore all elements heavier than hydrogen.

Let us start from an early enough time when recombination and ionization can maintain equilibrium among the reacting particles. The number densities of electrons, protons, and hydrogen atoms are then all given by Eq. (3.128) with $i = \mathrm{e}$, p or H. As we will see below, the temperatures of all three species are identical to that of the photons, so that $T_i = T$. Since the chemical potentials are related by $\mu_\mathrm{H} = \mu_\mathrm{p} + \mu_\mathrm{e}$, we can write the equilibrium density of H as the Saha equation:

$$n_{\mathrm{H,eq}} = \left(\frac{g_\mathrm{H}}{g_\mathrm{p} g_\mathrm{e}}\right) n_{\mathrm{p,eq}} n_{\mathrm{e,eq}} \left(\frac{m_\mathrm{e} T}{2\pi}\right)^{-3/2} \exp\left(\frac{B_\mathrm{H}}{T}\right), \tag{3.214}$$

where $B_\mathrm{H} = m_\mathrm{p} + m_\mathrm{e} - m_\mathrm{H} = 13.6\,\mathrm{eV}$ is the binding energy of a hydrogen atom, and we have used $(m_\mathrm{H}/m_\mathrm{p})^{3/2} \approx 1$ in the prefactor. Expressing the particle number densities in terms of the baryon number density, $n_\mathrm{b} = n_\mathrm{p} + n_\mathrm{H}$, and the ionization fraction, $X_\mathrm{e} \equiv n_\mathrm{e}/n_\mathrm{b} = n_\mathrm{p}/n_\mathrm{b}$, then yields

$$\frac{1 - X_{\mathrm{e,eq}}}{X_{\mathrm{e,eq}}^2} = \sqrt{\frac{32}{\pi}}\, \zeta(3)\, \eta \left(\frac{m_\mathrm{e}}{T}\right)^{-3/2} \exp\left(\frac{B_\mathrm{H}}{T}\right), \tag{3.215}$$

[8] Note that the term 'recombination' is somewhat unfortunate, as this will be the *first* time in the history of the Universe that the electrons combine with nuclei to form atoms.

where we have used that $g_e = g_p = 2$, $g_H = 4$, and $n_b = \eta n_\gamma$. This is the Saha equation for the ionization fraction in thermal equilibrium, which holds as long as the reaction rate $p + e \leftrightarrow H$ is larger than the expansion rate.

Assuming for the moment that thermal equilibrium holds, we can use Eq. (3.215) to compute the temperature, T_{rec}, and redshift, z_{rec}, of recombination. For example, if we define recombination as the epoch at which $X_e = 0.1$, we obtain that

$$\theta_{rec}^{3/2} \exp\left(13.6/\theta_{rec}\right) = 3.2 \times 10^{17} \left(\Omega_{b,0} h^2\right)^{-1} , \tag{3.216}$$

where

$$\theta \equiv (T/1\,\mathrm{eV}) \approx (1+z)/4250 . \tag{3.217}$$

Taking logarithms and iterating once we get an approximate solution for θ_{rec}:

$$\theta_{rec}^{-1} \approx 3.084 - 0.0735 \ln\left(\Omega_{b,0} h^2\right) , \tag{3.218}$$

which corresponds to a redshift given by

$$(1 + z_{rec}) \approx 1367 \left[1 - 0.024 \ln\left(\Omega_{b,0} h^2\right)\right]^{-1} . \tag{3.219}$$

Assuming $\Omega_{b,0} h^2 = 0.02$, we get $T_{rec} \approx 0.3\,\mathrm{eV}$ and $z_{rec} \approx 1{,}300$. Note that $T_{rec} \ll B_H$, which is a reflection of the high entropy per baryon (i.e. the small value of η); since there are many times more photons than baryons, there can still be sufficient photons with $h_P \nu > 13.6\,\mathrm{eV}$ in the Wien tail of the blackbody spectrum to keep the majority of the hydrogen atoms ionized, even when the temperature has dropped below the ionization value.

As the Universe expands and the number densities of electrons and protons decrease, the rate at which recombination and ionization can proceed may become smaller than the expansion rate. The assumption of equilibrium will then no longer be valid. In order to examine in detail how recombination proceeds, we need to understand the main reactions involved. In a normal cloud of ionized hydrogen (HII cloud), recombination occurs mainly via two processes: (i) direct recombination to the ground state, and (ii) the capture of an electron to an excited state which then cascades to the ground level. In the first case, a Lyman continuum photon (with energy larger than 13.6 eV) is produced, while in the second case one of the recombination photons must have an energy higher than or equal to that of Lyα. If the cloud is optically thin, all recombination photons can escape and do not contribute to further ionization. In the case of cosmological recombination, however, recombination photons will be absorbed again because they cannot escape from the Universe. In fact, the direct capture of electrons to the ground state does not contribute to the net recombination, because the resulting photon is energetic enough to ionize another hydrogen atom from its ground state. The normal cascade process is also ineffective, because the Lyman series photons produced can excite hydrogen atoms from their ground states, so that multiple absorptions lead to re-ionization. Therefore, recombination in the early Universe must have proceeded by different means.

There are two main channels by which cosmological recombination can proceed. One is the two-photon decay from the metastable $2S$ level to the ground state ($1S$). In this process two photons must be emitted in order to conserve both energy and angular momentum, and it is possible that the energies of the emitted photons fall below the ionization threshold. This process is forbidden to first order and so it has a slow rate: $\Gamma_{2\gamma} \approx 8.23\,\mathrm{s}^{-1}$. The second process is the elimination of Lyα photons by cosmological redshift. Once redshifted to a lower energy, the Lyα photons produced in the cascade will no longer be able to excite hydrogen atoms from their ground state. The details of these recombination processes have been worked out by several authors (Peebles, 1968; Zel'dovich et al., 1968; Peebles, 1993). They show that, of the two processes discussed,

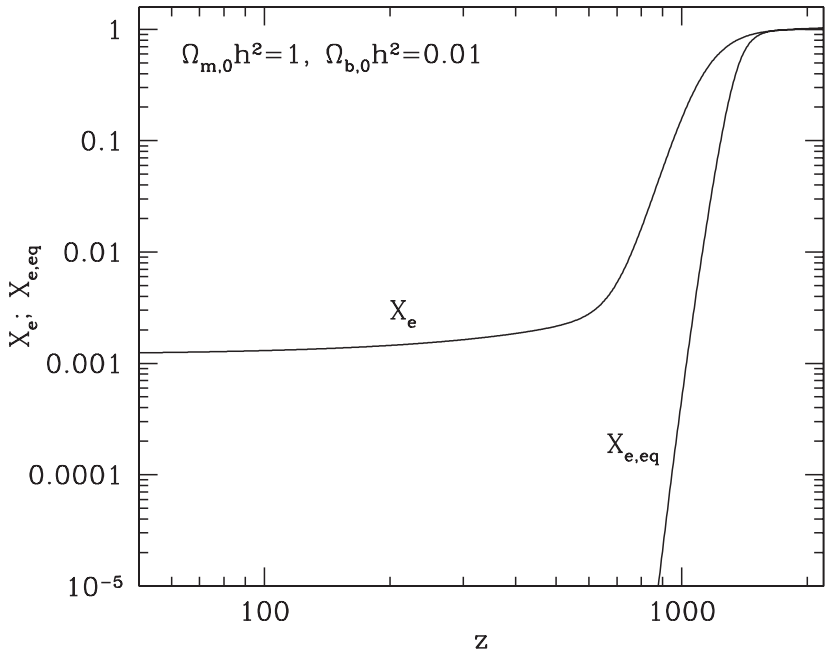

Fig. 3.8. The ionization fraction as a function of redshift, z. The curve marked $X_{\rm e,eq}$ shows the redshift evolution of the equilibrium ionization fraction, while the one marked $X_{\rm e}$ shows the actual ionization fraction for a cosmology with $\Omega_{\rm m,0}h^2 = 1$ and $\Omega_{\rm b,0}h^2 = 0.01$.

the two-photon emission dominates, and that the ionization fraction drops from approximately unity at $z \gtrsim 2000$, to a 'freeze-out' value of

$$X_{\rm e} \approx 1.2 \times 10^{-5} \left(\frac{\sqrt{\Omega_{\rm m,0}}}{\Omega_{\rm b,0} h} \right) \tag{3.220}$$

at $z \lesssim 200$. An example of the evolution of $X_{\rm e}$ with redshift is shown in Fig. 3.8.

3.5.2 Decoupling and the Origin of the CMB

Charged particles and photons interact with each other via Thomson scattering. The rate at which a photon collides with an electron is $\Gamma_{\rm T} = n_{\rm e}\sigma_{\rm T}c$, where

$$\sigma_{\rm T} = \frac{8\pi}{3} \left(\frac{q_{\rm e}^2}{m_{\rm e}c^2} \right)^2 \approx 6.65 \times 10^{-25}\,{\rm cm}^2 \tag{3.221}$$

is the Thomson cross-section, with $q_{\rm e}$ the charge of an electron. In what follows we only consider this scattering between electrons and photons, since the interaction rate with ions is much lower. Substituting the electron number density with $n_{\rm e} = X_{\rm e}\eta n_\gamma$, and using the Saha equation to compute $X_{\rm e}$ in the limit $X_{\rm e} \ll 1$, we obtain

$$\Gamma_{\rm T} = 1.01 \left(\Omega_{\rm b,0}h^2\right)^{1/2} \theta^{9/4} \exp\left(-6.8/\theta\right) {\rm s}^{-1}. \tag{3.222}$$

In order to estimate at what redshift the photons decouple from the matter, we compare this interaction rate with the expansion rate. At $z \gg 1$, we can use Eq. (3.74) to write

$$H(T) = \begin{cases} 3.8 \times 10^{-13} \theta^2 \, \mathrm{s}^{-1} & (\text{for } z > z_{\rm eq}) \\ 9.0 \times 10^{-13} \left(\Omega_{\rm m,0} h^2\right)^{1/2} \theta^{3/2} \, \mathrm{s}^{-1} & (\text{for } z < z_{\rm eq}), \end{cases} \tag{3.223}$$

where $z_{\rm eq}$ is the redshift at which the Universe becomes matter dominated, and we have used $g_* = 3.36$ in calculating the energy density of relativistic species. Equating Eqs. (3.222) and (3.223) with the assumption that decoupling occurs as $z < z_{\rm eq}$, we obtain the decoupling temperature:

$$\theta_{\rm dec}^{-1} \approx 3.927 + 0.074 \ln \left(\Omega_{\rm b,0}/\Omega_{\rm m,0}\right). \tag{3.224}$$

Taking $\Omega_{\rm b,0}/\Omega_{\rm m,0} = 0.1$ we get

$$T_{\rm dec} \approx 0.26 \, \mathrm{eV}; \qquad (1 + z_{\rm dec}) \approx 1{,}100. \tag{3.225}$$

As expected, the decoupling of matter and radiation occurs shortly after the number density of free electrons has suddenly decreased due to recombination.

A somewhat more accurate derivation of the redshift of decoupling can be obtained by defining an optical depth of Thomson scattering from an observer at $z = 0$ to a surface at a redshift z:

$$\tau(z) = \int_0^z n_{\rm e} \sigma_{\rm T} \frac{{\rm d}t}{{\rm d}z} \, {\rm d}z. \tag{3.226}$$

Using the solution of $X_{\rm e}(z)$ shown in Fig. 3.8, rather than the equilibrium ionization fraction used in the previous estimate, one finds to good approximation

$$\tau(z) = 0.37 \left(z/1{,}000\right)^{14.25}. \tag{3.227}$$

The probability that a photon was last scattered in the redshift interval $z \pm {\rm d}z/2$ can be approximated as

$$P(z) = e^{-\tau} \frac{{\rm d}\tau}{{\rm d}z} \approx 5.26 \times 10^{-3} \left(\frac{z}{1{,}000}\right)^{13.25} \exp\left[-0.37 \left(\frac{z}{1{,}000}\right)^{14.25}\right]. \tag{3.228}$$

This distribution peaks sharply at $z \approx 1{,}067$ and has a width $\Delta z \approx 80$ (e.g. Jones & Wyse, 1985). This represents the last scattering surface of photons, which is the surface probed by the cosmic microwave background (CMB) radiation. Similar to the photosphere of the Sun, it acts as a kind of photon barrier. No information carried by photons originating from $z \gtrsim 1100$ can reach the Earth, as the photons involved will be scattered many times.

As discussed in §2.9, one of the most important properties of the observed CMB is that its spectrum is very close to that of a blackbody. This implies that the emission must have originated when the Universe was highly opaque. In the standard cosmology, such an epoch is expected because photons and other particles are tightly coupled at $z > 10^6$. However, the CMB photons have been scattered many times by electrons and ions between their redshift of origin and the last scattering surface. An important question therefore is whether the background radiation can retain a blackbody spectrum during this process. The answer is yes and the reason is, as we show below, that the high entropy content of the Universe can keep the gas particles at the same temperature as that of the photons. In this case, there is no net energy transfer between the photons and electrons, ensuring that the radiation field remains blackbody. Furthermore, although the finite thickness of the last scattering surface ($\Delta z \approx 80$) implies a spread in photon temperatures at their last scattering event, this does not lead to observable distortions in the CMB temperature spectrum. The reason is that the higher initial temperature of a photon that decoupled somewhat earlier is exactly compensated by the larger redshift it experiences before reaching the observer. Thus, the blackbody nature of the CMB is naturally explained in the standard cosmology. In

what follows we examine more closely the temperature evolution of matter and radiation from the epoch of electron–positron annihilation to that of decoupling.

3.5.3 Compton Scattering

By far the most dominant electromagnetic interaction during the era of decoupling is the Coulomb interaction, which is sufficiently strong to maintain thermal equilibrium among various matter components. In the absence of any interactions between matter and radiation, the temperature of the former would decrease as $T_{\rm m} \propto a^{-2}$, while the photon temperature $T_\gamma \propto a^{-1}$ (see Table 3.1). However, as we have seen above, photons and electrons interact with each other via Compton scattering. As long as $T_{\rm e} > T_\gamma$ there will be a net energy transfer from the electrons to the photons, and vice versa. The mean free path is $l_\gamma = 1/(n_{\rm e}\sigma_{\rm T})$ for the photons, and $l_{\rm e} = 1/(n_\gamma \sigma_{\rm T})$ for electrons. Their ratio can be expressed in terms of the ionization fraction $X_{\rm e} = n_{\rm e}/n_{\rm b}$ as

$$\frac{l_{\rm e}}{l_\gamma} = X_{\rm e}\eta = 2.72\times 10^{-8}(X_{\rm e}\Omega_{\rm b,0}h^2)\,. \tag{3.229}$$

Since $X_{\rm e} \leq 1$, we have $l_{\rm e} \ll l_\gamma$. This shows that it is much easier for photons to change the energy distribution of the electrons than the other way around. This is, once again, a consequence of the high entropy per baryon, or, put differently, of the fact that the heat capacity of the radiation is many orders of magnitude larger than that of the electrons. Therefore, as long as the Compton interaction rate is sufficiently large compared to the expansion rate, the matter temperature will follow that of the photons.

To compute the redshift at which the matter temperature will finally decouple from that of the radiation, we proceed as follows. The average energy transfer per Compton collision is

$$\Delta E = \frac{4}{3}\left(\frac{v}{c}\right)^2 h_{\rm P}\bar{\nu} = 4\left(\frac{k_{\rm B}T_{\rm e}}{m_{\rm e}c^2}\right)\frac{\varepsilon_\gamma}{n_\gamma}\,, \tag{3.230}$$

where we have used that the average electron energy is $\frac{1}{2}m_{\rm e}v^2 = \frac{3}{2}k_{\rm B}T_{\rm e}$ and that the mean energy of the photons is $h_{\rm P}\bar{\nu} = \varepsilon_\gamma/n_\gamma$ with ε_γ the photon energy density (see §B1.3.6). The rate at which the energy density of the matter, $\varepsilon_{\rm m}$, changes due to Compton interactions with the radiation field is

$$\frac{{\rm d}\varepsilon_{\rm m}}{{\rm d}t} = n_{\rm e}n_\gamma\sigma_{\rm T}c\Delta E = 4n_{\rm e}\sigma_{\rm T}\varepsilon_\gamma\left(\frac{k_{\rm B}T_{\rm e}}{m_{\rm e}c}\right)\,. \tag{3.231}$$

This allows us to define the Compton rate at which electrons can adjust their energy density to that of the photons as

$$\Gamma_{\gamma\rightarrow{\rm e}} \equiv \frac{1}{\varepsilon_{\rm m}}\frac{{\rm d}\varepsilon_{\rm m}}{{\rm d}t} = 8.9\times 10^{-6}\left(\frac{X_{\rm e}}{X_{\rm e}+1}\right)\theta^4\,{\rm s}^{-1}\,, \tag{3.232}$$

where we have used that $\varepsilon_{\rm m} = \frac{3}{2}nk_{\rm B}T_{\rm e}$, with $n = n_{\rm e}+n_{\rm b}$, and $\varepsilon_\gamma = (4\sigma_{\rm SB}/c)T_\gamma^4$. Comparing this to the expansion rate given by Eq. (3.223), we find that decoupling of matter from radiation occurs at a redshift

$$1+z = 6.8\left(\frac{X_{\rm e}}{X_{\rm e}+1}\right)^{-2/5}\left(\Omega_{\rm m,0}h^2\right)^{1/5}\,. \tag{3.233}$$

As we have seen above, before the onset of the first ionizing sources, the residual ionization fraction at $z \lesssim 200$ is $X_{\rm e} \sim 10^{-5}\Omega_{\rm m,0}^{1/2}/(\Omega_{\rm b,0}h)$ (see Fig. 3.8). Substituting this in Eq. (3.233), and adopting $\Omega_{\rm b,0}h^2 = 0.02$, yields a redshift, $z \simeq 150$, at which the temperatures of matter and radiation decouple. This is a much lower redshift than the redshift of decoupling defined by an

optical depth of unity for Compton scattering . This reflects the small values of η and X_e, which ensure that there are about $3 \times 10^{12} (\Omega_{m,0} h^2)^{1/2}$ photons for every free electron. The electron temperature can remain coupled to that of the photons even if only a tiny fraction of photons are scattered by the electrons.

3.5.4 Energy Thermalization

In addition to $\Gamma_{\gamma \to e}$ defined above, we can also define the rate at which Compton scattering can adjust the photon energy density to that of the electrons:

$$\Gamma_{e \to \gamma} \equiv \frac{1}{\varepsilon_\gamma} \frac{d\varepsilon_\gamma}{dt} = 1.3 \times 10^{-13} \left(X_e \Omega_{b,0} h^2 \right) \theta^4 \, s^{-1} , \tag{3.234}$$

where we have used that in thermal equilibrium $|d\varepsilon_\gamma / dt| = |d\varepsilon_m / dt|$. This is equal to the expansion rate in Eq. (3.223) at a redshift

$$1 + z = 7.2 \times 10^3 \left(X_e \Omega_{b,0} h^2 \right)^{-1/2} . \tag{3.235}$$

At $z \gtrsim 2{,}000$, $X_e = 1$ to good approximation. Using $\Omega_{b,0} h^2 = 0.02$ we thus find that Compton scattering can significantly modify the energy distribution of the photon fluid at $z \gtrsim 5 \times 10^4$. Since Compton scattering ($e + \gamma \to e + \gamma$) does not change the number of photons, this process alone cannot lead to a Planck distribution. However, since the photon fluid starts out in thermal equilibrium with the matter, it will remain properly thermalized (i.e. the Compton scattering does not lead to any net energy transfer between matter and radiation). On the other hand, one might envision scenarios in which physical processes (e.g. turbulence, black hole evaporation, decay of heavy unstable leptons) heat the electrons to a temperature above that of the photons. If this occurs at $z \gtrsim 5 \times 10^4$, Compton scattering is sufficiently efficient that photons experience multiple scattering events, which bring them into thermal equilibrium with the electrons. Since there is no change in the photon number, such scattering results in a modification of the photon energy distribution from a Planck distribution to a Bose–Einstein distribution with a negative chemical potential ($\mu < 0$). Such a distortion is usually referred to as a μ-distortion. In the absence of any photon-producing processes, an increase of the electron temperature therefore leads to a Comptonization of the CMB, which is observable as a μ-distortion of its spectrum.

Two examples of photon producing processes, which may thermalize the injected energy and bring the photon energy distribution back to that of a blackbody, are bremsstrahlung (also called free–free emission) and the double-photon Compton process ($e + \gamma \to e + 2\gamma$). In a medium with relatively high photon density, such as that in the radiation dominated era, double Compton emission is the dominant photon producing process and its rate is higher than the expansion rate of the Universe at

$$z \gtrsim 2.0 \times 10^6 \left(\frac{\Omega_{b,0} h^2}{0.02} \right)^{-2/5} \left(1 - \frac{Y_p}{2} \right)^{-2/5} , \tag{3.236}$$

where Y_p is the helium abundance in mass (e.g. Danese & de Zotti, 1982). Thus, any energy input into the radiation field at $z > 2 \times 10^6$ can be effectively thermalized into a blackbody distribution. If the energy ejection occurs at $z < 5 \times 10^4$ the Compton rate is insufficient to establish a new thermal equilibrium. Therefore, only energy injection in the redshift range $5 \times 10^4 \lesssim z \lesssim 2 \times 10^6$ can lead to a μ-distortion in the CMB. Detailed observations with the COBE satellite have established that the CMB has a blackbody spectrum to very high accuracy; the corresponding limit on the chemical potential is $|\mu| \leq 9 \times 10^{-5}$ (Fixsen et al., 1996). Apparently, there have not been any major energy ejections into the baryonic gas in the above mentioned redshift interval.

Because *multiple* Compton scattering becomes rare at $z \lesssim 5 \times 10^4$, any energy input into the electron distribution no longer drives the photon field towards a Bose–Einstein distribution to produce a μ-distortion. However, *single* Compton scattering of low-energy photons in the Rayleigh–Jeans tail of the CMB can still cause those photons to gain energy. Although this does not bring the photons in thermal equilibrium with the electrons, it does result in a distortion of the photon energy distribution. This kind of distortion is called *y*-distortion, because it is proportional to the Compton *y*-parameter defined in §B1.3.6. Such distortions can be produced by the hot intracluster medium, which is called the Sunyaev–Zel'dovich (SZ) effect, and is discussed in detail in §6.7.4.

3.6 Inflation

So far we have seen that the standard relativistic cosmology provides a very successful framework for interpreting observations. There are, however, a number of problems that cannot be solved within the standard framework. Here we summarize some of these problems and show how an 'inflationary hypothesis' can help to solve them.

3.6.1 The Problems of the Standard Model

(a) The Horizon Problem As shown in §3.2.4, the comoving radius of the particle horizon for a fundamental observer, $\mathscr{O}$, at the origin at cosmic time t is

$$\chi_{\rm h} = \int_0^t \frac{c{\rm d}t'}{a(t')}\,. \tag{3.237}$$

For a universe which did not have a contracting phase in its history, radiation was the dominant component of the cosmic energy density at $z > z_{\rm eq}$, and the scale factor $a(t) \propto t^{1/2}$. In this case $\chi_{\rm h}$ has a finite value, so that there must be fundamental observers (denoted by $\mathscr{O}'$) whose comoving distances to $\mathscr{O}$ are larger than $\chi_{\rm h}$. No physical processes at any $\mathscr{O}'$ could have influenced $\mathscr{O}$ by time t. To get a rough idea of the size of the particle horizon at the time of decoupling, assume for simplicity an Einstein–de Sitter universe ($\Omega_{\rm m,0} = 1$), and ignore for the moment that the Universe was radiation dominated at $z > z_{\rm eq}$. It then follows from Eqs. (3.74)–(3.75) that

$$\chi_{\rm h}(z) = 6{,}000\,h^{-1}{\rm Mpc}\,(1+z)^{-1/2}\,. \tag{3.238}$$

At the time of decoupling, which occurs at a redshift of $\sim 1{,}100$ (see §3.5), the comoving radius of the particle horizon is $\sim 180\,h^{-1}$Mpc. The comoving distance from us to the last scattering surface is $\sim 5{,}820\,h^{-1}$Mpc, so that the particle horizon at decoupling subtends an angle of about 1.8 degrees on the sky. This implies that many regions that we observe on the CMB sky have not been in causal contact. Yet, as discussed in §2.9, once measurements are corrected to the frame of the fundamental observer at the position of the Sun, the temperature of the CMB radiation is the same in all directions to an accuracy of better than one part in 10^5. The problem is how all these causally disconnected regions can have extremely similar temperatures. This problem is known as the horizon problem of the standard model.

(b) The Flatness Problem This problem concerns the processes which determine the density, age, and size of the Universe at the present time. In the standard model, these properties are assumed to 'arise' as initial conditions at the Planck time, when the Universe emerged from the quantum gravity epoch. The problem arises if $\Omega = \Omega_{\rm m} + \Omega_\Lambda + \Omega_{\rm r}$ differs mildly from unity at the present time, because such a universe requires extreme 'fine-tuning' of Ω at the Planck time. A simple way to illustrate the situation is to focus on the quantity, $\Omega^{-1} - 1$, which measures the

fractional deviation of the total density from the critical density. Using the Friedmann equation, we can write that

$$\Omega(a)^{-1} - 1 = -\frac{3Kc^2}{8\pi G\rho(a)a^2}\,, \tag{3.239}$$

which is proportional to a^2 at $z > z_{\rm eq}$ and to a at $z < z_{\rm eq}$. Therefore, in the standard model,

$$\frac{\Omega_{\rm Pl}^{-1} - 1}{\Omega_0^{-1} - 1} \sim \frac{T_0}{T_{\rm eq}}\left(\frac{T_{\rm eq}}{T_{\rm Pl}}\right)^2 \sim 10^{-60}\,, \tag{3.240}$$

where subscripts 'eq' and 'Pl' denote the values at the time of radiation/matter equality, $t_{\rm eq} \sim 10^4\,{\rm yr}$ (corresponding to a temperature $T_{\rm eq} \sim 10^4\,{\rm K}$), and the Planck time, $t_{\rm Pl} \equiv (\hbar G/c^5)^{1/2} \sim 10^{-43}\,{\rm s}$ (corresponding to a temperature $T_{\rm Pl} \sim 10^{32}\,{\rm K}$), respectively. This demonstrates that $\Omega_{\rm Pl}$ is about 60 orders of magnitude closer to unity than Ω_0. For example, if $\Omega = 0.1$ today, it must have been $1 - 10^{-59}$ at the Planck time, which clearly constitutes a fine-tuning problem. A 'trivial' way out of this problem is to postulate that Ω_0 is exactly equal to unity, in which case it has been exactly unity throughout the history of the Universe. However, this cannot be considered a proper solution unless it has a proper physical explanation. This problem is known as the flatness problem.

(c) Monopole Problem In the early stages of the Hot Big Bang, particle energies are well above the threshold at which grand unification (GUT) is expected to occur ($T_{\rm GUT} \sim 10^{14}$–$10^{15}\,{\rm GeV}$). As the temperature drops through this threshold, a phase transition associated with spontaneous symmetry breaking (SSB) can occur. One speaks of SSB when the fundamental equations of a system possesses a symmetry which the ground state does not have. For example, one may have a situation in which the Lagrangian density is invariant under a gauge transformation, while the vacuum state, the state of the least energy, does not possess this symmetry. SSB plays a crucial role in quantum field theory, where it provides a mechanism for assigning masses to the gauge bosons without destroying the gauge invariance. As we will see below, SSB also plays an important role in inflation.

Depending on the properties of the symmetry breaking, the phase transition can produce topological defects, such as magnetic monopoles, strings, domain walls or textures (see Vilenkin & Shellard (1994) for a detailed description). In the case of the GUT phase transition, one expects the formation of magnetic monopoles with a density of about one per horizon volume at that epoch. The mass of each monopole is expected to be of the order of the energy scale in consideration, i.e. $m \sim T_{\rm GUT}$. This predicts a present-day energy density in magnetic monopoles of

$$\rho_{\rm mono,0} \sim \frac{T_{\rm GUT}}{t_{\rm GUT}^3}\left(\frac{T_0}{T_{\rm GUT}}\right)^3 \sim \left(\frac{T_{\rm GUT}}{10^{11}\,{\rm GeV}}\right)^4 \rho_{\gamma,0}\,, \tag{3.241}$$

where T_0 and $\rho_{\gamma,0}$ are the temperature and energy density of the cosmic microwave background at the present time, and we have used Eq. (3.81) to relate $t_{\rm GUT}$ to $T_{\rm GUT}$. With $\Omega_{\gamma,0} \approx 2.5 \times 10^{-5} h^{-2}$ and $T_{\rm GUT} \sim 10^{15}\,{\rm GeV}$, we see that monopoles are expected to completely dominate the present matter density with $\Omega_0 \sim 5 \times 10^{11}$, in fatal conflict with observations. Since monopoles are expected to arise in almost any GUT, there is a monopole problem in the standard cosmology.

(d) Structure Formation Problem This problem concerns the origin of the large-scale structure in the Universe. The observed structures such as the clusters of galaxies have an amplitude which may be characterized by their dimensionless binding energy per unit mass, $\mathscr{E}/c^2 \sim 10^{-5}$. Such structures are coherent over a mass of about $10^{15}\,{\rm M}_\odot$ (corresponding to $\sim 10\,{\rm Mpc}$ in comoving size) and are presumed to have grown via gravitational instability from small initial perturbations. Since both the mass and binding energy of a perturbation are approximately conserved during gravitational evolution, the perturbation must have been generated while its entire

mass was within the particle horizon (i.e. when $\chi_{\rm h} > 10\,{\rm Mpc}$), in order to explain its coherence. This requires that the perturbations associated with present-day clusters be generated at $z \lesssim 10^6$. Since the standard scenario of structure formation via gravitational instability does not include any processes which could produce the binding energy of clusters at such low redshift, the origin of large, coherent density perturbations constitutes another problem for the standard cosmology.

It should be pointed out, however, that this particular problem is not fully generic for the standard cosmology. In particular, the problem may be avoided if we abandon the assumption that structures form via gravitational instability. For example, density perturbations with large amplitudes may be generated in the early Universe within patches of the horizon size at the time of generation. If these perturbations collapse and form objects which can eject energy to large distances, structures of much larger scales may form out of the perturbations created by these ejecta. Such non-gravitational models for the formation of large-scale structure have, for example, been considered by Ostriker & Cowie (1981). However, as we will see later in the book, the large-scale structure observed in the Universe is best explained by gravitational instability, implying that the structure formation problem must be considered seriously.

(e) Initial Condition Problem It should be pointed out that the problems mentioned above do not falsify the standard cosmology in any way. All of these problems can be incorporated into the standard cosmology as initial conditions, even though the standard cosmology does not explain them. In this sense, standard cosmology only provides a consistent theory to explain the state of the observable Universe with some assumed initial conditions, but does not explain their origin.

For many years it was believed that the initial conditions for standard cosmology would arise from quantum cosmology (a quantum treatment of space-time) at very early times when the Universe was so small that classical cosmology is no longer valid. Unfortunately, such theory is still highly incomplete and no reliable predictions can be presented. However, the situation changed dramatically in the early 1980s when it was realized that a new concept, called inflation, can solve all the aforementioned problems within the classical theory of space-time. Inflation basically provides an explanation for the initial conditions, and it operates at an energy scale that is much lower than the Planck scale, so that gravity can be treated classically. In what follows we present a brief overview of cosmological inflation, and illustrate how it solves the problems mentioned here. A more detailed treatment of this topic can be found in Kolb & Turner (1990) and Liddle & Lyth (2000).

3.6.2 The Concept of Inflation

As discussed above, the horizon problem arises because the comoving radius of the particle horizon of a fundamental observer (at time t),

$$\chi_{\rm h} = \int_0^t \frac{{\rm d}ct'}{a(t')} = \int_0^a \frac{{\rm d}a'}{a'} \left[\frac{8\pi G\rho(a')a'^2}{3c^2} - K\right]^{-1/2}, \tag{3.242}$$

is finite in the standard model, where $\rho(a) \propto a^{-4}$ as $a \to 0$. To get rid of this problem, $\chi_{\rm h}$ must diverge, making the radius of the particle horizon infinite. From Eq. (3.242) one sees that this requires $\rho(a) \propto a^{-\beta}$ with $\beta < 2$ as $a \to 0$. Inserting this a-dependence of ρ into the first law of thermodynamics, Eq. (3.35), one obtains

$$\rho + 3P/c^2 < 0\,, \tag{3.243}$$

which, in Eq. (3.58), gives $\ddot{a} > 0$. Such a phase of accelerated expansion is called inflation, and arises when the Universe is dominated by an energy component whose equation of state satisfies Eq. (3.243).

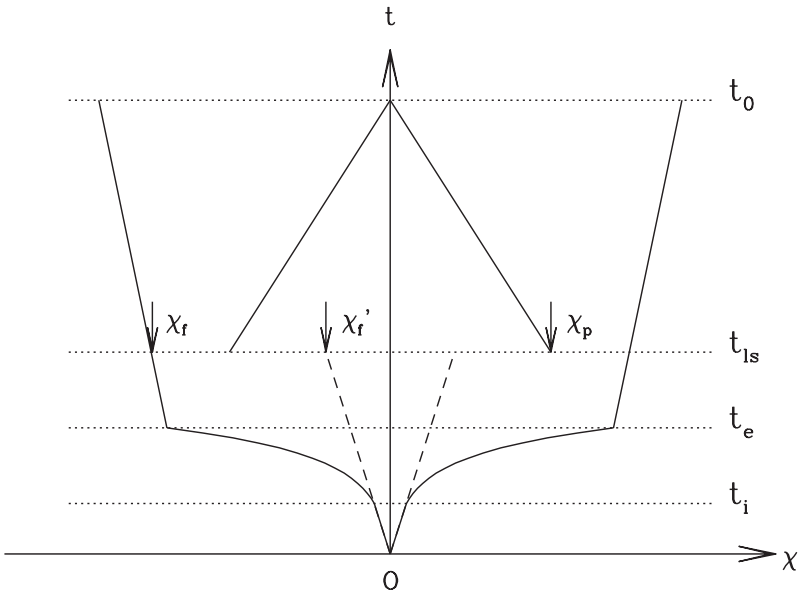

Fig. 3.9. A sketch of the light-cone structure in an inflationary universe. The cosmic time flows from bottom up (with the Big Bang labeled by O) and the horizontal axis marks the comoving radius, χ, of the light cone. In the absence of inflation, the forward light cone (the dashed lines) would be smaller than our past light cone, $\chi_{\rm p}$, at the last scattering surface (corresponding to $t = t_{\rm ls}$), resulting in the causality problem discussed in §3.6.1. With a period of inflation (from $t_{\rm i}$ to $t_{\rm e}$), however, the forward light cone can be (much) larger than the past light cone at $t_{\rm ls}$ (i.e. $\chi_{\rm f} > \chi_{\rm p}$).

An example of such an energy component is vacuum energy, whose equation of state is $P = -\rho_{\rm vac}c^2$. In this case, the solution of the Friedmann equation corresponds to an exponentially expanding universe,

$$a \propto e^{Ht} \quad \text{where} \quad H = \sqrt{8\pi G\rho_{\rm vac}/3} \tag{3.244}$$

(see §3.2.3). Fig. 3.9 illustrates how such an inflationary period can solve the horizon problem. Suppose that inflation begins at some very early time $t_{\rm i}$ and ends at some later time $t_{\rm e}$. The period of inflation is therefore $\Delta t = t_{\rm e} - t_{\rm i}$. During inflation, the forward light cone expands exponentially, whereas the past light cone of an observer at the present time t_0 is not affected by the exponential expansion for $t > t_{\rm e}$. Therefore, if Δt is sufficiently long, the size of the forward light cone on the last scattering surface of the CMB photons, $\chi_{\rm f}(t_{\rm ls})$, can be larger than the size of the past light cone, $\chi_{\rm p}(t_{\rm ls})$. Since $t_{\rm ls} \ll t_0$, the size of the past light cone is $\chi_{\rm p}(t_{\rm ls}) = \int_{t_{\rm ls}}^{t_0} {\rm d}t/a(t) \approx 3t_0$. The size of the forward light cone at $t_{\rm ls}$ is $\chi_{\rm f}(t_{\rm ls}) \sim \int_{t_{\rm i}}^{t_{\rm ls}} {\rm d}t/a(t) = (1/H)[e^{H\Delta t} - 1]a^{-1}(t_{\rm e})$. The condition that $\chi_{\rm f}(t_{\rm ls}) > \chi_{\rm p}(t_{\rm ls})$ therefore requires

$$e^{H\Delta t} > 3Ha(t_{\rm e})t_0 \sim a(t_{\rm e})\frac{t_0}{t_{\rm e}} \sim \frac{1}{\sqrt{1+z_{\rm eq}}}\frac{T_{\rm e}}{T_0} \sim 10^{25}, \tag{3.245}$$

where the final value is for $T_{\rm e} \sim 10^{14}\,{\rm GeV}$ (roughly the GUT energy scale) and $T_0 \sim 10^{-13}\,{\rm GeV}$ (the temperature of the CMB). Thus, in order to solve the horizon problem, an inflationary period of

$$\Delta t \gtrsim 60H^{-1} \tag{3.246}$$

is required, corresponding to 60 e-foldings in the scale factor.

Inflation can also solve the flatness problem. To see this we use Eq. (3.79) to write

$$\frac{\Omega^{-1}(t_{\rm e})-1}{\Omega^{-1}(t_{\rm i})-1}=\left[\frac{a(t_{\rm i})}{a(t_{\rm e})}\right]^2\lesssim 10^{-52}\,, \tag{3.247}$$

where we have inserted the number of e-foldings implied by Eq. (3.245). Therefore, even when $\Omega(t_{\rm i})$ deviates substantially from unity, at the end of inflation $\Omega(t_{\rm e})\simeq 1$ to very high accuracy. If we assume that inflation ends at about the GUT time (i.e. $T_{\rm e}\sim 10^{15}\,{\rm GeV}$), then the present-day value of Ω is related to that at the beginning of inflation according to

$$\frac{\Omega^{-1}(t_0)-1}{\Omega^{-1}(t_{\rm i})-1}=\frac{\Omega^{-1}(t_0)-1}{\Omega^{-1}(t_{\rm e})-1}\times\frac{\Omega^{-1}(t_{\rm e})-1}{\Omega^{-1}(t_{\rm i})-1}\lesssim 10^{-52}\left(\frac{T_{\rm eq}}{T_0}\right)\left(\frac{T_{\rm e}}{T_{\rm eq}}\right)^2\sim 1\,. \tag{3.248}$$

Thus the same number of e-foldings needed to solve the horizon problem also solves the flatness problem. Since the value of Ω at the present time depends very sensitively on the number of e-foldings, unless it is exactly unity, extreme fine-tuning is required to give $\Omega\neq 1$. In this sense, inflation predicts that the Universe is spatially flat, with $\Omega_{\rm m,0}+\Omega_{\Lambda,0}=1$. This can be understood as follows. Because of inflation the curvature radius (measured in physical scale) increases exponentially, and the observed piece of space in the past light cone looks essentially flat after inflation even if it had a large curvature before.

If monopoles are produced before inflation, their number density will be diluted exponentially during inflation. At the end of inflation, the number density would be reduced by a factor $\sim\left(e^{H\Delta t}\right)^3\sim 10^{78}$, making the contribution of monopoles to the cosmic density completely negligible. Thus, inflation also solves the monopole problem discussed in §3.6.1.

Finally, inflation also provides a mechanism to explain why structures like clusters can form in a causal way. Because of inflation, small-scale structures present before and during inflation can be blown up exponentially. Thus, the different parts of a perturbation responsible for a cluster and a larger-scale structure, although not in causal contact after inflation, were actually in causal contact before or during inflation. It is therefore possible to have causality if the perturbations responsible for the formation of clusters were generated before or during inflation. In fact, inflation not only allows such perturbations to exist, but also provides a mechanism to generate them, as we will discuss in detail in §4.5.1.

3.6.3 Realization of Inflation

The above discussion shows that inflation can solve many problems (or, more appropriately, puzzles) regarding the initial conditions of the Big Bang cosmology, as long as it operates for a sufficiently long time. All that is needed is a dominant energy component with an equation of state obeying Eq. (3.243). As already mentioned, the cosmological constant is an example of such a component. However, it cannot serve to describe inflation for the simple reason that it will never stop. By definition, Λ is a constant, and once it dominates the energy density of the Universe it will continue to do so eternally. A successful inflation model, however, needs to stop after some time, and it needs to end in a particular way. After all, at the end of inflation the matter and radiation density of the Universe will be virtually zero, so will be its temperature: the Universe is basically a vacuum. We thus need a mechanism, called reheating, which at the end of inflation creates matter and radiation. In other words, at the end of inflation the Universe needs to undergo a cosmological phase transition. The temperature at the end of this phase transition has to be sufficiently low so that during the subsequent evolution no new phase transition can create large quantities of magnetic monopoles. Otherwise we are back to where we started. In addition, the temperature needs to be sufficiently high so that the process of baryogenesis can still operate.

It was Guth (1981) who first realized that all these requirements can be realized in a natural way with scalar fields. These quantum fields, which describe scalar (spin-0) particles, play an

important role in quantum field theory. Their dominant role is to cause spontaneous symmetry breaking (SSB) via the Higgs mechanism. These so-called Higgs fields have a non-zero vacuum expectation value. As a result, the interactions of the fermion and boson fields with the Higgs field give a finite potential energy to the fermions and bosons, which is expressed as an effective mass. Before the symmetry is broken, the Higgs field has a zero expectation value, and the fermions and bosons are massless. In what follows we show that under certain conditions, a similar scalar field can also cause inflation. The key point here is that the zero-point energy (vacuum energy) of such fields can mimic a cosmological constant. A scalar field that causes inflation is generally called an inflaton.

The Lagrangian density of a scalar field $\varphi(\mathbf{x},t)$ is

$$\mathscr{L} = \frac{1}{2}\partial_\mu\varphi\partial^\mu\varphi - V(\varphi)\,, \tag{3.249}$$

where $V(\varphi)$ is the potential of the field. Different inflationary models, i.e. different inflatons, correspond to different choices for $V(\varphi)$. The energy–momentum tensor of the inflaton is

$$T^{\mu\nu} = \partial^\mu\varphi\partial^\nu\varphi - g^{\mu\nu}\mathscr{L}\,. \tag{3.250}$$

If the inhomogeneity in φ is small, this $T^{\mu\nu}$ has the form of a perfect fluid, Eq. (3.57), with energy density and pressure given by

$$\rho = \frac{\dot\varphi^2}{2} + \frac{(\nabla\varphi)^2}{2a^2} + V(\varphi) \quad \text{and} \quad P = \frac{\dot\varphi^2}{2} - \frac{(\nabla\varphi)^2}{6a^2} - V(\varphi)\,, \tag{3.251}$$

where $\dot\varphi \equiv (\partial\varphi/\partial t)$, and ∇ is the derivative with respect to the comoving coordinates $\mathbf{x}$. We therefore have

$$\rho + 3P = 2\left[\dot\varphi^2 - V(\varphi)\right]\,, \tag{3.252}$$

and the condition for inflation becomes

$$\dot\varphi^2 \ll V(\varphi)\,, \tag{3.253}$$

which is called the slow-roll approximation. Note that in this case $\rho = V(\varphi)$, and for inflation to happen $V(\varphi)$ thus needs to be sufficiently large to dominate the total energy density. As soon as inflation operates it drives the curvature to zero so that the Friedmann equation (3.60) becomes

$$H = \sqrt{8\pi G V(\varphi)/3} = \frac{1}{m_{\rm Pl}}\sqrt{\frac{8\pi V}{3}}\,, \tag{3.254}$$

where $m_{\rm Pl} \equiv (\hbar c/G)^{1/2}$ is the Planck mass and we have used $\hbar = c = 1$.

Since the scale factor a increases exponentially during inflation, the spatial derivative term $\nabla\varphi/a$ in ρ and P rapidly becomes negligible, provided V is large enough for inflation to start in the first place. Therefore any spatial inhomogeneities can be neglected and, under the slow-roll approximation, one obtains $P = -\rho$, an equation of state similar to that for the cosmological constant.

To translate the slow-roll requirement into a requirement for the shape of the potential, $V(\varphi)$, which is the 'free parameter' in the construction of inflation models, we need to look at the dynamics of a scalar field. The classical equation of motion is obtained by writing down the action

$$S = \int \mathscr{L}\sqrt{-g}\,\mathrm{d}^4x\,, \tag{3.255}$$

where g is the determinant of the metric tensor $g_{\mu\nu}$. The Euler–Lagrange equation, which follows from the least-action principle, $\delta S = 0$, then yields

$$\ddot\varphi + 3H\dot\varphi + \mathrm{d}V/\mathrm{d}\varphi = 0\,, \tag{3.256}$$

where we have ignored any spatial inhomogeneity ($\nabla\varphi = 0$). Equivalently, Eq. (3.256) follows from conservation of the energy–momentum tensor (${T^{\mu\nu}}_{;\nu} = 0$), or from substituting Eq. (3.251) in the continuity equation for a FRW cosmology, $\dot{\rho} = -3H(\rho + P)$ [see Eq. (3.35)]. Note that Eq. (3.256) is similar to the equation of motion of a ball moving under the influence of a potential V in the presence of friction (the Hubble drag) proportional to $3H$. Using that $\dot{\varphi} \sim \varphi/t$, so that $\ddot{\varphi} \sim \varphi/t^2 \sim \dot{\varphi}^2/\varphi$, the slow-roll approximation implies that $\ddot{\varphi} \ll V(\varphi)/\varphi \sim \mathrm{d}V/\mathrm{d}\varphi$. Therefore, the first term in Eq. (3.256) is negligible, and

$$3H\dot{\varphi} + \mathrm{d}V/\mathrm{d}\varphi = 0\,. \tag{3.257}$$

This equation expresses that the acceleration due to the gradient in the potential is balanced by the Hubble drag due to the expansion. This together with Eq. (3.253) leads to the following slow-roll condition:

$$\varepsilon \equiv \frac{m_{\rm Pl}^2}{16\pi}\left(\frac{\mathrm{d}V/\mathrm{d}\varphi}{V}\right)^2 = \frac{m_{\rm Pl}^2}{16\pi}\frac{(3H\dot{\varphi})^2}{V^2} \ll m_{\rm Pl}^2\frac{(3H)^2}{V} \sim 1\,, \tag{3.258}$$

where we have used the Friedmann equation (3.254). Similarly, since $(\mathrm{d}^2V/\mathrm{d}\varphi^2)/V \sim (\mathrm{d}V/\mathrm{d}\varphi)/(\varphi V) \sim (\mathrm{d}V/\mathrm{d}\varphi)^2/V^2$, we have

$$\eta \equiv \frac{m_{\rm Pl}^2}{8\pi}\frac{1}{V}\frac{\mathrm{d}^2V}{\mathrm{d}\varphi^2} \ll 1\,. \tag{3.259}$$

Conditions (3.258) and (3.259) indicate the intuitively obvious, namely that for the slow-roll condition to be satisfied the potential needs to be very flat. Any scalar field that obeys these two constraints will cause an inflationary phase, whose duration increases with the flatness of $V(\varphi)$.

As emphasized above, inflation is only successful if it can also stop and reheat the Universe. Below we illustrate how this comes about with scalar fields using three specific examples. In each of these the end of inflation and the reheating mechanism are somewhat different.

3.6.4 Models of Inflation

(a) Old Inflation The 'old inflation' model, proposed by Guth (1981), is based on a scalar field which initially gets trapped in a false vacuum at $\varphi = 0$ and which at some point undergoes spontaneous symmetry breaking to its true vacuum state via a first order phase transition. The prototype of such a potential has the form

$$V(\varphi) = \frac{1}{4}\varphi^4 - \frac{1}{3}(\alpha+\beta)|\varphi|^3 + \frac{1}{2}\alpha\beta\varphi^2 + V_0\,, \tag{3.260}$$

where $V_0 = \alpha^3(\alpha - 2\beta)/12 > 0$ and $\alpha > 2\beta > 0$. The field is assumed to be in thermal equilibrium with a radiation field at temperature T, and so the effective potential of the field is

$$V_{\rm eff}(\varphi) = V(\varphi) + \frac{1}{2}\tilde{\lambda}T^2\varphi^2 \tag{3.261}$$

according to finite-temperature field theory (e.g. Brandenberger, 1995), where $\tilde{\lambda}$ is a coupling constant. Fig. 3.10 a shows $V_{\rm eff}(\varphi)$ at different temperatures. When the temperature is high, the effective potential has a single minimum at $\varphi = 0$. As the temperature decreases, two other minima develop. This occurs at a critical temperature $T_{\rm c} = (\alpha - \beta)/(2\tilde{\lambda}^{1/2})$. For $T \ll T_{\rm c}$, the three minima are at $\varphi = 0, \pm\alpha$, and the values of the potential at these points are $V_{\rm eff}(0) = V_0$ and $V_{\rm eff}(\pm\alpha) = 0$. Thus, the vacua at $\varphi = \pm\alpha$ represent two true vacua of the field, while the one at $\varphi = 0$ is called a false vacuum state. When $T \gg T_{\rm c}$ the expectation value of the inflaton is $\varphi = 0$. At this stage no inflation occurs simply because the energy density of the radiation still exceeds that of the inflaton. When the temperature drops below $T_{\rm c}$, the field gets trapped in the

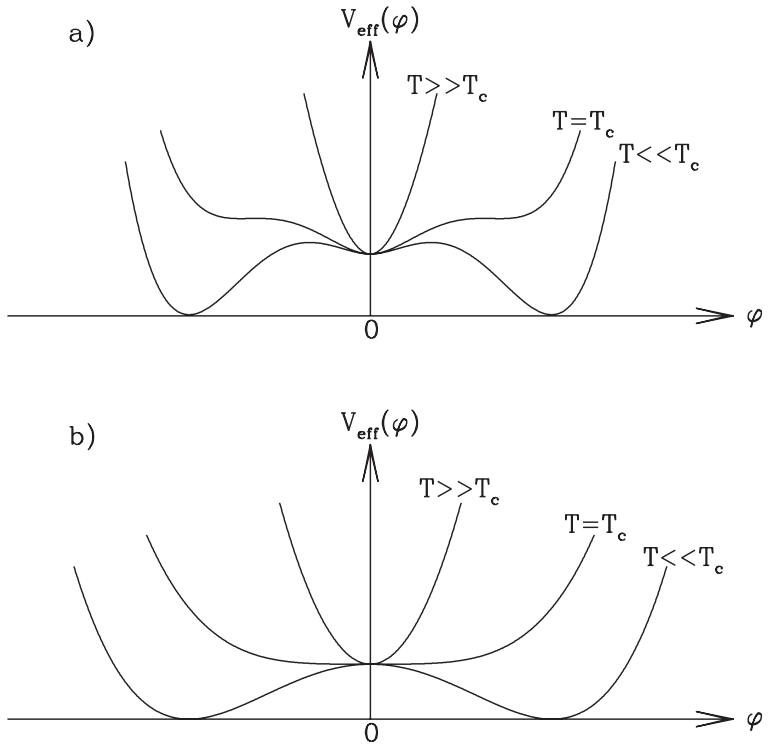

Fig. 3.10. Two examples of effective scalar potentials, at three different temperatures, that can lead to inflation. In example (a) φ experiences a first-order phase transition, characteristic of the old inflation models, while in (b) the phase transition is of second order.

false vacuum at $\varphi = 0$, and the system is said to undergo supercooling. At this point, the slow-roll condition is satisfied, and $\rho \sim V(\varphi = 0)$ is dominated by the energy density of the inflaton. Consequently, $\varphi(\mathbf{x})$ acts like a cosmological constant, the Universe enters a de Sitter phase with a Friedmann equation of the form (3.254), and the Universe expands exponentially. The epoch of inflation only ends when thermal fluctuations or quantum tunneling moves φ across the barrier so that it can proceed towards its true vacuum. This transition is a spontaneous symmetry breaking, and since the field value changes discontinuously, it is of first order. During the transition the energy $V(\varphi = 0)$ associated with the inflaton field, the so-called latent heat, is rapidly liberated and can be used for reheating. If the system stays in the false vacuum sufficiently long, the Universe can be inflated by a sufficiently large number of e-foldings. It thus appears that this model fulfills all requirements.

However, it was realized that this model has a 'graceful exit' problem (Guth, 1981; Guth & Weinberg, 1981). Because the transition is of first order, it proceeds through the nucleation of bubbles of the true vacuum in a surrounding sea of false vacuum. These bubbles must grow in a causal way, and so their sizes at the end of inflation cannot be larger than the horizon size at that time, which is much smaller than our past light cone. In addition, the latent heat needed for reheating is stored in the kinetic energy of the nucleated bubbles, and reheating only occurs when this kinetic energy is thermalized via bubble collisions. Thus, unless bubbles can collide and homogenize in the Hubble radius, the model will predict too large inhomogeneities to match the observed isotropy of the CMB and too little reheating. However, since the space between the bubbles is filled with exponentially expanding false vacuum, while the volume of a bubble expands only with a low power of time, percolation and homogenization of bubbles can never

happen. Instead, inflation continues indefinitely, and the bubbles of true vacuum have only a small volume filling factor at any time. The volume filling factor can be increased by increasing the nucleation rate of true-vacuum bubbles, but this would require a high tunneling rate, making the inflation period too short.

(b) New Inflation Because of the 'graceful exit' problem, a modified scenario has been proposed by Linde (1982) and Albrecht & Steinhardt (1982). The prototype potential in this scenario has the form

$$V(\varphi) = \frac{1}{4}\lambda\left(\varphi^2 - \sigma^2\right)^2, \tag{3.262}$$

and the effective potential $V_{\rm eff}(\varphi)$ is plotted in Fig. 3.10 b for different temperatures. At high temperature, the effective potential has a single minimum at $\varphi = 0$, but when the temperature drops below a critical value, $T_{\rm c} = \sigma(\lambda/\tilde{\lambda})^{1/2}$, the minimum at $\varphi = 0$ disappears (and becomes a local maxima) while two new minima develop. As in old inflation, the scalar field is confined to the neighborhood of $\varphi(\mathbf{x}) = 0$ by the thermal force at $T \gg T_{\rm c}$, when the Universe is dominated by radiation. As the temperature drops to $T \sim T_{\rm c}$ (when the vacuum energy of φ begins to dominate over radiation), the field configuration at $\varphi(\mathbf{x}) = 0$ becomes unstable and it evolves towards $\varphi = \pm\sigma$ as the temperature decreases. The change from $\varphi = 0$ to $\varphi = \pm\sigma$ is smooth everywhere, and so the spontaneous symmetry breaking occurs via a second-order phase transition. As long as the evolution obeys the slow-roll requirements derived above, inflation will occur. When φ approaches σ (or $-\sigma$), the field rolls rapidly towards the minimum (because of the large potential gradient). Since this violates the slow-roll requirement, it signals the end of inflation. The inflaton φ subsequently oscillates around the minimum with a frequency ω given by $\omega^2 = (\mathrm{d}^2V/\mathrm{d}\varphi^2)_\sigma = \lambda\sigma^2$. If the field is coupled to the radiation field, these oscillations will be damped by the decay of φ into photons and other particles, and the Universe is reheated to a temperature $T \sim \omega \sim T_{\rm i}$, with $T_{\rm i}$ the temperature at the onset of inflation. The Universe then enters the radiation dominated era of the ordinary FRW cosmology.

The spatial fluctuations in $\varphi(\mathbf{x})$ are expected to be correlated over some microphysical scale and, as a result, the field is homogeneous within domains with sizes typically of the correlation length. Since the correlated domains are established before the onset of inflation, any domain boundaries are inflated outside the present Hubble radius and the inflation in a domain stops when $|\varphi| \sim \sigma$. Since our Universe is thus contained within a single domain, there is no 'graceful exit' problem in this model. Hence the new inflation model is an improvement over the old one. Unfortunately it also has problems. In order to obtain inflation, we must have $\mathrm{d}^2V/\mathrm{d}\varphi^2 \ll V/m_{\rm Pl}^2$ [see Eq. (3.259)] which, for $V = (\varphi^2 - \sigma^2)^2/4$, requires $\sigma \gg m_{\rm Pl}$. This is obviously an unnatural condition, since $m_{\rm Pl}$ is the highest energy scale expected in particle physics. There is also a more general problem. In order to ensure a large-enough number of e-foldings, the initial value of φ must satisfy $|\varphi_{\rm i}| \ll \sigma$. However, since the thermal fluctuations of φ at the initial time (when $T = T_{\rm i}$) are expected to be of the order $\lambda^{-1/4}T_{\rm i} \sim [V(0)/\lambda]^{1/4} \sim \sigma$, fine-tuning is needed to get the required initial condition, $|\varphi_{\rm i}| \ll \sigma$.

(c) Chaotic Inflation Chaotic inflation was proposed to give a more natural explanation for the initial conditions leading to inflation (Linde, 1986). Unlike in the old and new inflation models, no phase transition is involved here so that no initial thermal bath is required. In this model, one starts with a simple potential, e.g. $V(\varphi) = m\varphi^2/2$, and inflation simply arises because of the slow motion of φ from some initial value, $\varphi_{\rm i}$, towards the potential minimum. At any given point $\mathbf{x}$, the initial field configuration is assumed to be set by some chaotic processes. The values of φ are expected to be the same within regions with a size set by the correlation length. Inflation only occurs in those regions where the conditions needed for inflation are attained; the other regions simply never inflate. Chaotic inflation therefore predicts that the Universe is locally homogeneous, but globally inhomogeneous. In a region where inflation persists for a sufficiently long

period, the boundary of this region can be blown out of the current particle horizon, leaving a universe in which the initial inhomogeneities generated by the chaotic processes have no observable consequences. In this scenario our Universe is assumed to have emerged from one of such regions.

In order to solve the horizon and flatness problems, the number of e-foldings must be $N \gtrsim 60$. Using the slow-roll approximation, we can write the number of e-foldings between t_i and t_e (the times when inflation starts and terminates) as

$$N = \int_{t_\mathrm{i}}^{t_\mathrm{e}} H \, \mathrm{d}t \sim -\frac{1}{m_\mathrm{Pl}^2} \int_{\varphi_\mathrm{i}}^{\varphi_\mathrm{e}} \frac{V}{|\mathrm{d}V/\mathrm{d}\varphi|} \, \mathrm{d}\varphi \,. \tag{3.263}$$

For a smooth potential such as $V(\varphi) = m^2\varphi^2/2$, $|\mathrm{d}V/\mathrm{d}\varphi| \sim V/\varphi$ and so $N \sim (\varphi_\mathrm{i}/m_\mathrm{Pl})^2$ (assuming that $\varphi_\mathrm{e} \ll \varphi_\mathrm{i}$). It then follows that $\varphi_\mathrm{i} \gg m_\mathrm{Pl}$ is needed to have successful inflation. If inflation starts near the Planck time, the fluctuations in V are about m_Pl^4, and for the potential in consideration $m \ll m_\mathrm{Pl}$ is required. It is unclear if such a small mass scale can be achieved in a Planck time, because the most natural mass scale at this time is m_Pl. Indeed, if inflation happened at the Planck time, it may not be really possible to construct a realistic inflation model without a proper understanding of quantum gravity. In this sense, our initial hope that inflation models would solve some of the problems in the standard model within the classical space-time framework is not realized.

The schemes and problems discussed above are typical of many other inflation models suggested. At the present time, it is fair to say that, although the concept of inflation can help to solve several outstanding problems in standard cosmology, a truly successful model is still lacking.

Appendix A
Basics of General Relativity

General relativity (hereafter GR) is the subject dealing with the structure of space-time and with how to describe physical laws in any given space-time. The perspective of space-time in GR is very different from that in Newtonian physics. In Newtonian physics, space is considered to be flat, infinite and eternal, time is considered to flow uniformly, and physical processes are considered to act in this external space-time frame. In the framework of GR, however, space-time is a four-dimensional manifold which may be curved and the properties of space-time itself are determined by dynamical processes.

This appendix provides a brief summary of the aspects of GR that are used in this book. More details can be found in the excellent textbooks by Weinberg (1972), Misner et al. (1973), Rindler (1977), and Carroll (2004).

A1.1 Space-time Geometry

In order to gain some insight in how to describe space-time as a four-dimensional manifold (hypersurface), consider a two-dimensional analog. To describe a two-dimensional surface, we can construct a coordinate system and label each point on the surface by its coordinates. The geometrical properties of the surface can be obtained by considering the distance between each pair of infinitesimally close points on the surface in terms of the differences in coordinates. In general, the square of this distance can be written as

$$\mathrm{d}l^2 = \sum_{i,j=1}^{2} g_{ij}(\mathbf{x})\,\mathrm{d}x^i\,\mathrm{d}x^j, \tag{A1.1}$$

where $\mathbf{x} = (x^1, x^2)$ are the coordinates and $g_{ij}(\mathbf{x})$ is the metric which gives the distance in terms of the difference in coordinates. As an example, if we use Cartesian coordinates (x, y), then $g_{ij} = \delta_{ij}$ is the metric for a plane, because $\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2$. Similarly,

$$\mathrm{d}s^2 = \frac{R^2 - y^2}{R^2 - x^2 - y^2}\,\mathrm{d}x^2 + \frac{R^2 - x^2}{R^2 - x^2 - y^2}\,\mathrm{d}y^2 + \frac{2xy}{R^2 - x^2 - y^2}\,\mathrm{d}x\,\mathrm{d}y \tag{A1.2}$$

gives the metric of a sphere with radius R. This is evident by using the spherical coordinates:

$$\mathrm{d}l^2 = R^2(\mathrm{d}\vartheta^2 + \sin^2\vartheta\,\mathrm{d}\varphi^2), \tag{A1.3}$$

where (ϑ, φ) is related to (x, y) by $x = R\sin\vartheta\cos\varphi$, $y = R\sin\vartheta\sin\varphi$. This shows that the metric not only depends on the properties of the surface, but also on the choice of the coordinate system. In general, one chooses a coordinate system which simplifies the problem at hand.

The geometrical properties of the space-time can be described in a similar manner. Each point on the four-dimensional space-time hypersurface is an event, represented by a time coordinate

and three spatial coordinates. The 'distance' between any two points (events) on this hypersurface is the interval ds. For a flat space-time this interval has the same form as in special relativity:

$$\mathrm{d}s^2 = c^2\mathrm{d}t^2 - \mathrm{d}x^2 - \mathrm{d}y^2 - \mathrm{d}z^2 = \eta_{\mu\nu}\,\mathrm{d}x^\mu\,\mathrm{d}x^\nu = c^2\mathrm{d}t^2 - \delta_{ij}\,\mathrm{d}x^i\mathrm{d}x^j, \tag{A1.4}$$

where x, y, z are the Cartesian coordinates, $(x^0, x^1, x^2, x^3) = (ct, x, y, z)$, δ_{ij} is the Kronecker delta function, and

$$\eta_{\mu\nu} = \mathrm{diag}(1, -1, -1, -1) \tag{A1.5}$$

is the Minkowski metric. In Eq. (A1.4) and in the following, a pair of repeated upper and lower indices implies summation over their range, Greek indices run from 0 to 3 while Latin indices run from 1 to 3. For a general space-time, the interval can be written as

$$\mathrm{d}s^2 = g_{\mu\nu}\,\mathrm{d}x^\mu\,\mathrm{d}x^\nu, \tag{A1.6}$$

where x^μ ($\mu = 0, 1, 2, 3$) are general space-time coordinates and the metric $g_{\mu\nu}$ gives the interval in terms of the difference in space-time coordinates. Note that $g_{\mu\nu} = g_{\nu\mu}$.

Since ds is invariant under coordinate transformation, the metric must transform as

$$g_{\mu\nu}(x) \to g'_{\mu\nu}(x') = \frac{\partial x^\alpha}{\partial x'^\mu}\frac{\partial x^\beta}{\partial x'^\nu} g_{\alpha\beta}(x), \tag{A1.7}$$

under a general coordinate transformation $x \to x'$. The inverse four-metric $g^{\mu\nu}$ is the inverse of $g_{\mu\nu}$:

$$g_{\mu\alpha}g^{\alpha\nu} = \delta_\mu{}^\nu, \tag{A1.8}$$

and so transforms as

$$g^{\mu\nu}(x) \to g'^{\mu\nu}(x') = \frac{\partial x'^\mu}{\partial x^\alpha}\frac{\partial x'^\nu}{\partial x^\beta} g^{\alpha\beta}(x). \tag{A1.9}$$

From the space-time metric, one can derive other useful geometric quantities. The affine connection, $\Gamma^\mu{}_{\alpha\beta}$, which connects vectors in nearby tangent spaces, is defined as

$$\Gamma^\mu{}_{\alpha\beta} = \frac{1}{2}g^{\mu\sigma}\left(\partial_\beta g_{\sigma\alpha} + \partial_\alpha g_{\sigma\beta} - \partial_\sigma g_{\alpha\beta}\right), \tag{A1.10}$$

where $\partial_\mu \equiv \partial/\partial x^\mu$. The Riemann–Christoffel curvature tensor, $R^\mu{}_{\nu\alpha\beta}$, which describes the curvature of the space-time manifold, is defined as

$$R^\mu{}_{\nu\alpha\beta} = \partial_\alpha\Gamma^\mu{}_{\nu\beta} - \partial_\beta\Gamma^\mu{}_{\nu\alpha} + \Gamma^\mu{}_{\sigma\alpha}\Gamma^\sigma{}_{\nu\beta} - \Gamma^\mu{}_{\sigma\beta}\Gamma^\sigma{}_{\nu\alpha}. \tag{A1.11}$$

The Ricci tensor and the curvature scalar are defined as

$$R_{\mu\nu} \equiv R^\sigma{}_{\mu\sigma\nu} \quad \text{and} \quad R \equiv g^{\mu\nu}R_{\mu\nu}, \tag{A1.12}$$

respectively.

For the Robertson–Walker metric,

$$\mathrm{d}s^2 = c^2\mathrm{d}t^2 - a^2(t)\left[\frac{\mathrm{d}r^2}{1-Kr^2} + r^2(\mathrm{d}\vartheta^2 + \sin^2\vartheta\,\mathrm{d}\varphi^2)\right], \tag{A1.13}$$

the non-zero components of the affine connection are

$$\begin{aligned}
&\Gamma^0_{11} = c^{-1}a\dot{a}/(1-Kr^2); && \Gamma^0_{22} = c^{-1}a\dot{a}r^2; && \Gamma^0_{33} = c^{-1}a\dot{a}r^2\sin^2\vartheta;\\
&\Gamma^1_{01} = \Gamma^2_{02} = \Gamma^3_{03} = \dot{a}/ca; && \Gamma^2_{12} = \Gamma^3_{13} = 1/r; &&\\
&\Gamma^1_{11} = Kr/(1-Kr^2); && \Gamma^1_{22} = -r(1-Kr^2); && \Gamma^1_{33} = -r(1-Kr^2)\sin^2\vartheta;\\
&\Gamma^2_{33} = -\sin\vartheta\cos\vartheta; && \Gamma^3_{23} = \cot\vartheta, &&
\end{aligned} \tag{A1.14}$$

where $(x^0,x^1,x^2,x^3) = (ct,r,\vartheta,\varphi)$ and $\dot{a} = \mathrm{d}a/\mathrm{d}t$. The non-zero components of the Ricci tensor are

$$R_{00} = -\frac{3}{c^2}\frac{\ddot{a}}{a}, \quad R_{ij} = -\frac{1}{c^2}\left[\frac{\ddot{a}}{a} + 2\frac{\dot{a}^2}{a^2} + \frac{2c^2K}{a^2}\right] g_{ij}, \tag{A1.15}$$

and the curvature scalar is

$$R = -\frac{6}{c^2}\left[\frac{\ddot{a}}{a} + \frac{\dot{a}^2}{a^2} + \frac{Kc^2}{a^2}\right]. \tag{A1.16}$$

For small perturbations of Minkowski space-time, the perturbed metric can in general be written as

$$\mathrm{d}s^2 = c^2(1+2\Psi/c^2)\,\mathrm{d}t^2 - 2cw_i\,\mathrm{d}t\,\mathrm{d}x^i - \left[(1-2\Phi/c^2)\delta_{ij} + H_{ij}\right]\mathrm{d}x^i\,\mathrm{d}x^j, \tag{A1.17}$$

where the perturbation quantities $|\Psi|/c^2$, $|\Phi|/c^2$, $|w_i|$, and $|H_{ij}|$ are all $\ll 1$. To the first order of the perturbation quantities, the non-zero components of the affine connection are:

$$\begin{aligned}
&\Gamma^0_{00} = \partial_0\Psi; && \Gamma^0_{i0} = \partial_i\Psi;\\
&\Gamma^i_{00} = \partial_i\Psi + \partial_0 w_i; && \Gamma^i_{j0} = \tfrac{1}{2}(\partial_j w_i - \partial_i w_j) + \tfrac{1}{2}\partial_0 h_{ij};\\
&\Gamma^0_{jk} = -\tfrac{1}{2}(\partial_j w_k + \partial_k w_j) + \tfrac{1}{2}\partial_0 h_{jk}; && \Gamma^i_{jk} = \tfrac{1}{2}(\partial_j h_{ki} + \partial_k h_{ji}) - \tfrac{1}{2}\partial_i h_{jk},
\end{aligned} \tag{A1.18}$$

where $h_{ij} = H_{ij} - 2\Phi\delta_{ij}$. The components of the Ricci tensor are

$$\begin{aligned}
&R_{00} = \delta^{ij}\partial_i\partial_j\Psi + 3\partial_0^2\Phi + \partial_0\partial_j w^j;\\
&R_{0j} = -\tfrac{1}{2}\delta^{kl}\partial_k\partial_l w_j + \tfrac{1}{2}\partial_j\partial_k w^k + 2\partial_0\partial_j\Phi + \tfrac{1}{2}\partial_0\partial_k H_j{}^k;\\
&R_{ij} = \partial_i\partial_j(\Phi-\Psi) - \tfrac{1}{2}\partial_0(\partial_i w_j + \partial_j w_i) + \delta_{ij}\eta^{\mu\nu}\partial_\mu\partial_\nu\Phi - \tfrac{1}{2}\eta^{\mu\nu}\partial_\mu\partial_\nu H_{ij} + \tfrac{1}{2}\partial_k(\partial_i H_j{}^k + \partial_j H_i{}^k).
\end{aligned} \tag{A1.19}$$

These results can also be used in dealing with small metric perturbations of a flat, expanding universe. Here the perturbed metric can be written as

$$\mathrm{d}s^2 = c^2(1+2\Psi/c^2)\,\mathrm{d}t^2 - 2caw_i\,\mathrm{d}t\,\mathrm{d}x^i - \left[(1-2\Phi/c^2)\delta_{ij} + H_{ij}\right]a^2\,\mathrm{d}x^i\,\mathrm{d}x^j, \tag{A1.20}$$

where $a(t)$ is the scale factor. It is evident that if we use a new set of space-time coordinates, (ct,x'^1,x'^2,x'^3) where $\mathrm{d}x'^i = a\mathrm{d}x^i$, the affine connection and Ricci tensor corresponding to metric (A1.20) will have the same forms as given by Eqs. (A1.18) and (A1.19), except that all spatial derivatives are with respect to x'^i. The quantities in terms of the comoving coordinates, x^i, can then be obtained by using general coordinate transformations (see below).

A1.2 The Equivalence Principle

According to the principle of general relativity, all reference frames are equivalent, and a physical law should have the same form under general coordinate transformation (the general covariance). Because of this, physical fields defined on a space-time must transform according to a set of rules under the general coordinate transformation. Depending on whether it is a scalar, S, a vector, $\mathbf{V}$, or a tensor, $\mathbf{T}$, a physical field transforms as

$$S'(x') = S(x); \;\; V'^{\mu}(x') = \frac{\partial x'^{\mu}}{\partial x^{\alpha}}V^{\alpha}(x); \;\; T'^{\mu\nu}(x') = \frac{\partial x'^{\mu}}{\partial x^{\alpha}}\frac{\partial x'^{\nu}}{\partial x^{\beta}}T^{\alpha\beta}(x), \tag{A1.21}$$

under the general coordinate transformation $x^{\mu} \to x'^{\mu}$. From a given vector or a given tensor, we can define another vector or another tensor as

$$V_{\mu} = g_{\mu\nu}V^{\nu}; \quad T_{\mu\nu} = g_{\mu\alpha}g_{\nu\beta}T^{\alpha\beta}. \tag{A1.22}$$

In general, new tensors can be obtained by using $g_{\mu\nu}$ to lower indices and by using $g^{\mu\nu}$ to raise indices. As in special relativity, V_α and V^α are called the covariant and contravariant components of $\mathbf{V}$. Generally, a lower index is called the covariant index while an upper index is called a contravariant index. It is easy to prove that under general coordinate transformation, V_μ and $T_{\mu\nu}$ transform as

$$V'_\mu(x') = \frac{\partial x^\alpha}{\partial x'^\mu} V_\alpha(x); \quad T'_{\mu\nu}(x') = \frac{\partial x^\alpha}{\partial x'^\mu}\frac{\partial x^\beta}{\partial x'^\nu} T_{\alpha\beta}(x). \tag{A1.23}$$

Thus, both $V^\mu V_\mu$ and $T^{\mu\nu}T_{\mu\nu}$ are invariant under general coordinate transformations.

From the perspective of GR, gravitation is manifested as curved space, and so the space-time must be (locally) Minkowskian in a frame which is in free fall in a gravitational field. An important aspect of GR is embodied in the equivalence principle that can be stated as follows: In a reference frame which is in free fall in a gravitational field, all physical laws have their special relativistic form, except the gravitational force which disappears. Together with the principle of general relativity, the equivalence principle enables us to find physical equations valid for any general reference frame: what we need to do is just to write the usual special relativistic equations in covariant forms.

Since physical equations generally involve derivatives with respect to the space-time coordinates, we need to find the covariant forms of the derivatives of physical fields. The covariant derivative with respect to a space-time coordinate x^μ is usually denoted by a subscript '$;\mu$'. For a scalar field it is defined as

$$S_{;\mu} \equiv \partial_\mu S, \tag{A1.24}$$

while for vector fields it is defined as

$$V^\alpha{}_{;\beta} = \partial_\beta V^\alpha + \Gamma^\alpha{}_{\mu\beta} V^\mu, \quad V_{\alpha;\beta} = \partial_\beta V_\alpha - \Gamma^\mu{}_{\alpha\beta} V_\mu. \tag{A1.25}$$

It is easy to show that, under general coordinate transformation, $S_{;\mu}$ transforms as a vector, while $V^\alpha{}_{;\beta}$ and $V_{\alpha;\beta}$ transform as tensors. In general, to obtain the covariant derivative of the tensor $T^{\cdots}_{\cdots}$ with respective to x^α, we add to the ordinary derivative $\partial_\alpha T^{\cdots}_{\cdots}$ a term $-\Gamma^\mu{}_{\beta\alpha} T^{\cdots}_{\cdot\mu\cdot}$ for each covariant index β ($T^{\cdots}_{\cdot\beta\cdot}$), and a term $\Gamma^\beta{}_{\mu\alpha} T^{\cdot\mu\cdot}_{\cdots}$ for each contravariant index β ($T^{\cdot\beta\cdot}_{\cdots}$). Note that the affine connection itself is *not* a tensor.

Using the definition of the metric, one can show that

$$\sqrt{-g}\,\mathrm{d}^4x = \sqrt{-g'}\,\mathrm{d}^4x', \tag{A1.26}$$

where g is the determinant of $g_{\mu\nu}$. This means that $\sqrt{-g}\,\mathrm{d}^4x$ is an invariant volume element. Thus, if $S(x)$ is a scalar field, then $\int S(x)\sqrt{-g}\,\mathrm{d}^4x$ is independent of the choice of coordinates.

A1.3 Geodesic Equations

As an application of the equivalence principle, consider the motion of a free particle with non-zero mass in a gravitational field. In the reference frame comoving with the particle (where, according to the principle of equivalence, the space-time must be locally Minkowskian with metric $\eta_{\mu\nu}$), the motion of the particle is given by

$$\frac{\mathrm{d}^2\xi^\mu}{\mathrm{d}s^2} = 0, \tag{A1.27}$$

where ds/c is the proper time interval measured in the free-fall frame, and ξ^μ is the space-time coordinates of the particle. For a general reference frame with coordinates x^μ related to ξ^ν by $x^\mu(\xi)$, the metric is related to $\eta_{\mu\nu}$ by

$$g_{\mu\nu} = \eta_{\alpha\beta}\frac{\partial\xi^\alpha}{\partial x^\mu}\frac{\partial\xi^\beta}{\partial x^\nu}. \tag{A1.28}$$

In the x-frame, the equation of motion (A1.27) becomes

$$\frac{d^2x^\mu}{ds^2} = -\Gamma^\mu{}_{\alpha\beta}\frac{dx^\alpha}{ds}\frac{dx^\beta}{ds}, \tag{A1.29}$$

where

$$\Gamma^\mu{}_{\alpha\beta} = \frac{\partial x^\mu}{\partial\xi^\nu}\frac{\partial^2\xi^\nu}{\partial x^\alpha\partial x^\beta}. \tag{A1.30}$$

One can prove that $\Gamma^\mu{}_{\alpha\beta}$ is just the affine connection defined by Eq. (A1.10). This can be done by using the relation

$$\partial_\lambda g_{\mu\nu} = \Gamma^\alpha{}_{\lambda\mu}g_{\alpha\nu} + \Gamma^\alpha{}_{\lambda\nu}g_{\alpha\mu}, \tag{A1.31}$$

and the results of cyclically permuting the three indices. Thus, in the x-frame there is a force exerting on the free-fall particle. This is gravity. But in the perspective of GR it is because the particle is moving in a curved space (non-zero affine connection). Free particles move along geodesics, and so Eq. (A1.29) is also the geodesic equation. If we define the four-momentum as

$$p^\mu = mU^\mu, \qquad U^\mu = c\frac{dx^\mu}{ds}, \tag{A1.32}$$

where m is the rest mass of the particle, Eq. (A1.29) can be written in the form

$$\frac{p^0}{c}\frac{dp^\mu}{dt} = -\Gamma^\mu{}_{\alpha\beta}p^\alpha p^\beta, \tag{A1.33}$$

where $p^0 = mcdx^0/ds = mc^2dt/ds$. Another useful form of Eq. (A1.29) is

$$\frac{p^0}{c}\frac{dp_\mu}{dt} = \frac{1}{2}(\partial_\mu g_{\alpha\beta})p^\alpha p^\beta, \tag{A1.34}$$

where $p_\mu = g_{\mu\nu}p^\nu$. Note that

$$g_{\mu\nu}p^\mu p^\nu = m^2c^2. \tag{A1.35}$$

In time-orthogonal coordinates, where $g_{00} = 1$ and $g_{0i} = 0$, we have

$$(p^0)^2 + g_{ij}p^ip^j = m^2c^2. \tag{A1.36}$$

If we define the magnitude of the three-momentum as $p^2 = -g_{ij}p^ip^j$, then

$$(p^0)^2 = p^2 + m^2c^2, \tag{A1.37}$$

and cp^0 can be considered the total energy of the particle in the time-orthogonal frame.

For massless particles $ds \to 0$ and so Eq. (A1.29) is invalid. However, both Eqs. (A1.33) and (A1.34) are well defined for massless particles, provided that p^μ is properly defined. One possibility is to define the four-momentum as

$$p^\mu = \frac{p^0}{c}\frac{dx^\mu}{dt}, \tag{A1.38}$$

which is the same as Eq. (A1.32) for massive particles if we choose $p^0 = mc^2\,dt/ds$. For massless particles, $g_{\mu\nu}p^\mu p^\nu = 0$. It can then be shown that in time-orthogonal coordinates, cp^0 is the

energy of the particle. To cast the equation of motion for massless particles in the form of the geodesic equation (A1.29), we introduce an affine parameter λ by the equation

$$p^0 \equiv \frac{\mathrm{d}x^0}{\mathrm{d}\lambda}. \tag{A1.39}$$

Eq. (A1.33) can then be written in the form

$$\frac{\mathrm{d}^2x^\mu}{\mathrm{d}\lambda^2} = -\Gamma^\mu{}_{\alpha\beta}\frac{\mathrm{d}x^\alpha}{\mathrm{d}\lambda}\frac{\mathrm{d}x^\beta}{\mathrm{d}\lambda}. \tag{A1.40}$$

A1.4 Energy–Momentum Tensor

If a charge Q is invariant under Lorentz transformation, the equation of charge conservation can be written in the form

$$\frac{\partial(nQ)}{\partial t} + \nabla\cdot\mathbf{j} = 0, \tag{A1.41}$$

where $\mathbf{j} = nQ\mathbf{v}$ is the current density. In covariant form this is

$$J^\mu{}_{;\mu} = 0, \tag{A1.42}$$

where J^μ is the four-current density vector. One might consider applying this to the mass to obtain a covariant form for the continuity equation. However, mass is not invariant under Lorentz transformation; it depends on momentum because of its connection to energy. Thus, a covariant continuity equation must involve both energy and momentum. The conserved quantity we are seeking is expected to have 16 components: the energy and energy current in three directions, plus momenta in three directions and their currents (each momentum has three components). Thus the quantity must be a 4×4 tensor which we call the energy–momentum tensor and denote by $T^{\mu\nu}$. The conservation of energy–momentum can then be written in the covariant form

$$T^{\mu\nu}{}_{;\mu} = 0. \tag{A1.43}$$

In many cosmological applications, the material content can be approximated by a fluid. To obtain the corresponding energy–momentum tensor, we again use the equivalence principle. A fluid is characterized by the density, $\rho(\mathbf{x})$, and pressure, $P(\mathbf{x})$, both measured by an observer *comoving* with the fluid at the point $\mathbf{x}$, and the velocity of the fluid element relative to some reference frame. Note that ρ and P defined in this way are invariant under general coordinate transformation. In the rest frame of a fluid element, the energy–momentum tensor is

$$T^{\mu\nu} = \mathrm{diag}(\rho c^2, P, P, P) = (\rho + P/c^2)U^\mu U^\nu - P\eta^{\mu\nu}, \tag{A1.44}$$

where $U^\mu = (c,0,0,0)$ is the four-velocity of the fluid element in the comoving frame. We can make a Lorentz transformation to get the energy–momentum tensor in a reference frame which is in free fall in the gravitational field at the point of the fluid element:

$$T^{\mu\nu} = (\rho + P/c^2)U^\mu U^\nu - P\eta^{\mu\nu}, \tag{A1.45}$$

where U^μ is the four-velocity of the fluid element in the free-fall reference frame. Thus, using the principle of equivalence, the energy–momentum tensor in a general coordinate system is

$$T^{\mu\nu} = (\rho + P/c^2)U^\mu U^\nu - Pg^{\mu\nu}, \tag{A1.46}$$

where $U^\mu = c\mathrm{d}x^\mu/\mathrm{d}s$.

A1.5 Newtonian Limit

One interesting question is what form the space-time metric takes in the Newtonian limit of gravity. Such a metric tells us how Newtonian gravity (the gravitational potential) is interpreted in terms of geometric quantities, thereby providing a hint how to construct the field equation in GR by generalizing the Newtonian field equation (Poisson's equation). To start with, consider a reference frame O' which is in free fall in a Newtonian gravitational potential Φ which is zero at some large distance. In this frame, the metric has the Minkowski form:

$$\mathrm{d}s^2 = c^2\mathrm{d}t'^2 - \mathrm{d}x'^2. \tag{A1.47}$$

Now consider another reference frame O relative to which O' has the free-fall velocity given by $v^2 = -2\Phi$ (assumed to be in the x-direction). According to Lorentz transformation, we have

$$\mathrm{d}t' = (1+2\Phi/c^2)^{1/2}\,\mathrm{d}t; \qquad \mathrm{d}x' = (1-2\Phi/c^2)^{1/2}\,\mathrm{d}x. \tag{A1.48}$$

Thus, the metric in terms of the coordinates in the O system can be written as

$$\mathrm{d}s^2 = c^2\left(1+2\Phi/c^2\right)\mathrm{d}t^2 - \left(1-2\Phi/c^2\right)\left(\mathrm{d}x^2+\mathrm{d}y^2+\mathrm{d}z^2\right). \tag{A1.49}$$

This is the metric in the Newtonian limit.

A1.6 Einstein's Field Equation

In the Newtonian limit, the 0–0 component of the energy–momentum tensor has the form $T_{00} = \rho c^2$, and $g_{00} = (1+2\Phi/c^2)$. The Poisson equation for gravity therefore takes the form

$$\nabla^2 g_{00} = 8\pi G T_{00}/c^4. \tag{A1.50}$$

This is a relation between the energy–momentum tensor and the derivatives of the metric. In general, the field equation must be a covariant extension of the above relation. We therefore expect the right-hand side of the above equation to be replaced by $8\pi G T_{\mu\nu}/c^4$, and the left-hand side to be replaced by a 4×4 tensor constructed from the metric and its derivatives. Einstein proposed a tensor (the Einstein tensor) of the form

$$G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R, \tag{A1.51}$$

and so the Einstein field equation takes the form

$$G_{\mu\nu} = \frac{8\pi G}{c^4}T_{\mu\nu}. \tag{A1.52}$$

Using that $g_{\mu\nu}g^{\mu\nu} = 4$, we see that the trace of the field equation is $R = -8\pi G T/c^4$, where $T = T_\mu{}^\mu$. The field equation can then be written in the form

$$R_{\mu\nu} = \frac{8\pi G}{c^4}\left(T_{\mu\nu} - \frac{1}{2}g_{\mu\nu}T\right). \tag{A1.53}$$

It can be shown that, in the Newtonian limit (A1.49), this equation reduces to the Poisson equation.

Einstein also realized that he could add to $G_{\mu\nu}$ a term $-\Lambda g_{\mu\nu}$ and write the field equation as

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R - \Lambda g_{\mu\nu} = \frac{8\pi G}{c^4}T_{\mu\nu}, \tag{A1.54}$$

where Λ is a constant called the cosmological constant. Using the expression of $T_{\mu\nu}$ for an ideal fluid [see Eq. (A1.46)], we see that the Λ term can be included in the energy–momentum tensor as an ideal fluid with $\rho = -P/c^2 = c^2\Lambda/8\pi G$.