

PREPARED FOR SUBMISSION TO JHEP

An Introduction to Big Bang Cosmology¹

C.P. Burgess

*Department of Physics & Astronomy, McMaster University
and Perimeter Institute for Theoretical Physics*

¹ ©C Burgess, Physics 789 McMaster University 2005 & PSI Explorations in Particle Theory 2016

Contents

| | | |
|----------|---|-----------|
| 1 | Kinematics of an Expanding Universe | 1 |
| 1.1 | The FRW Metric | 2 |
| 1.1.1 | Flat Spatial Curvature | 3 |
| 1.1.2 | Positive Spatial Curvature | 3 |
| 1.1.3 | Negative Spatial Curvature | 4 |
| 1.2 | Particle Motion | 4 |
| 1.2.1 | Hubble Flow and Peculiar Motion | 5 |
| 1.2.2 | Light Rays and Redshift | 6 |
| 1.3 | Distance vs Redshift | 8 |
| 1.3.1 | Proper Distance | 8 |
| 1.3.2 | Luminosity Distance | 9 |
| 1.3.3 | Angular-Diameter Distance | 10 |
| 1.4 | Examples | 11 |
| 1.4.1 | The Recent Universe | 12 |
| 1.4.2 | Power-Law Expansion | 13 |
| 1.4.3 | Exponential Expansion | 15 |
| 2 | Dynamics of an Expanding Universe | 15 |
| 2.1 | Relating Cosmic Expansion to Matter Content | 16 |
| 2.1.1 | Homogeneous and Isotropic Stress-Energy | 16 |
| 2.1.2 | Einstein's Equations | 17 |
| 2.1.3 | Cosmic Acceleration and Matter | 17 |
| 2.2 | Equations of State | 18 |
| 2.2.1 | Empty Space | 20 |
| 2.2.2 | Radiation | 20 |
| 2.2.3 | Non-relativistic Matter | 20 |
| 2.2.4 | The Vacuum | 22 |
| 2.3 | Multi-Component Fluids | 23 |
| 2.3.1 | The Present-Day Energy Content | 23 |
| 2.3.2 | Earlier Epochs | 26 |

| | | |
|----------|--|-----------|
| 3 | Thermal Evolution of the Universe | 29 |
| 3.1 | Big Bang Cosmology | 29 |
| 3.1.1 | The Known Particle Content | 29 |
| 3.2 | Temperature Evolution - Thermodynamics | 31 |
| 3.2.1 | Relativistic Particles | 31 |
| 3.2.2 | Nonrelativistic Particles | 31 |
| 3.2.3 | Multi-Component Fluids | 32 |
| 3.3 | Temperature Evolution - Statistical Mechanics | 33 |
| 3.3.1 | Equilibrium Distributions | 34 |
| 3.3.2 | Statistical Mechanics in Special Relativity | 36 |
| 3.3.3 | Statistical Mechanics in an Expanding Universe | 40 |
| 3.4 | Equilibrium and Decoupling | 41 |
| 3.4.1 | Scattering Rate vs Expansion Rate | 43 |
| 3.4.2 | Energy Dependence of Interactions | 45 |
| 3.4.3 | Some Decoupling Examples | 47 |
| 4 | Cosmic Relics | 52 |
| 4.1 | A Thermal History of the Universe | 52 |
| 4.2 | Relict Neutrinos | 54 |
| 4.3 | Nucleosynthesis | 55 |
| 4.4 | The Cosmic Microwave Background | 58 |
| 4.4.1 | Recombination | 59 |
| 4.4.2 | Photon Decoupling | 60 |
| 4.4.3 | Last Scattering | 60 |
| 4.5 | WIMP Dark Matter | 64 |
| 4.6 | Baryogenesis | 66 |
| 5 | An early accelerated epoch | 68 |
| 5.1 | Peculiar initial conditions | 68 |
| 5.2 | Acceleration to the rescue | 72 |
| 5.3 | Inflation or a bounce? | 76 |
| 5.4 | Simple inflationary models | 79 |
| 5.4.1 | Higgs field as inflaton | 79 |
| 5.4.2 | New field as inflaton | 80 |
| 5.4.3 | Slow-Roll Inflation | 81 |
| 5.4.4 | Some illustrative examples | 83 |

| | | |
|----------|--|------------|
| 5.5 | Flies in the ointment | 87 |
| 6 | Density Perturbations | 89 |
| 6.1 | Nonrelativistic Density Perturbations | 89 |
| 6.1.1 | Perturbations About a Static Background | 89 |
| 6.1.2 | Perturbations About an Expanding Background | 92 |
| 6.1.3 | Multi-Component Fluids | 96 |
| 6.1.4 | The Power Spectrum | 98 |
| 6.1.5 | Late-time structure growth | 102 |
| 6.1.6 | Hot Dark Matter | 104 |
| 6.2 | Primordial fluctuations from inflation | 105 |
| 6.2.1 | Linear evolution of metric-inflaton fluctuations | 106 |
| 6.2.2 | Slow-roll evolution of scalar perturbations | 108 |
| 6.2.3 | Post-Inflationary evolution | 108 |
| 6.2.4 | Quantum origin of fluctuations | 109 |
| 6.2.5 | Predictions for the scalar power spectrum | 111 |
| 6.2.6 | Tensor fluctuations | 113 |
| A | General Relativity | 115 |
| A.1 | Metrics | 115 |
| A.2 | Particle Motion | 116 |
| A.3 | Einstein's Field Equations | 117 |

1 Kinematics of an Expanding Universe

These notes are meant to provide a brief overview of the Big Bang theory of cosmology. The emphasis is on the theoretical ideas, since the rest of the class has been devoted to its observational foundations.

We start with a section describing the geometry of spacetime on which all of the subsequent sections rely. The key underlying assumption in this section is that the universe is homogeneous and isotropic when seen on the largest distance scales. Until relatively recently this assertion about the homogeneity and isotropy of the universe was an assumption, often called the *Cosmological Principle*. More recently it has become possible to put this assertion on an observational footing, based on large-scale surveys of the distribution of matter and radiation within the observed universe.

1.1 The FRW Metric

The most general 4D geometry which is consistent with isotropy and homogeneity of its spatial slices is described by the Friedmann-Robertson-Walker (FRW) metric:

$$\begin{aligned} ds^2 &= -dt^2 + a^2(t) \left[\frac{dr^2}{1 - \kappa r^2/R_0^2} + r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2 \right] \\ &= -dt^2 + a^2(t) \left[d\ell^2 + r^2(\ell) d\theta^2 + r^2(\ell) \sin^2 \theta d\phi^2 \right], \end{aligned} \quad (1.1)$$

where R_0 is a constant and κ can take one of the following three values: $\kappa = 1, 0, -1$. The coordinate ℓ is related to r by $d\ell = dr/(1 - \kappa r^2/R_0^2)^{1/2}$, and so

$$r(\ell) = \begin{cases} R_0 \sin(\ell/R_0) & \text{if } \kappa = +1 \\ \ell & \text{if } \kappa = 0 \\ R_0 \sinh(\ell/R_0) & \text{if } \kappa = -1 \end{cases}. \quad (1.2)$$

Notice that the metric, eq. (1.1), is invariant under the following re-scaling of parameters: $a \rightarrow \lambda a$, $R_0 \rightarrow \lambda R_0$, provided we also re-scale the coordinate $\ell \rightarrow \lambda \ell$. This freedom is often used to choose convenient units, such as by choosing λ to ensure $R_0 = 1$ (if $\kappa \neq 0$), or perhaps to set $a(t_0) = 1$ for some t_0 .

The coordinates used all have the following simple physical interpretations.

- t represents the proper time along the time-like trajectories along which ℓ, θ and ϕ are fixed. The range over which t may run is defined by the region over which the function $a(t)$ is neither zero nor infinite.
- ℓ is simply related to the proper distance measured along the radial directions along which t, θ and ϕ are fixed, since this proper distance is given by

$$D(\ell, t) = \ell a(t). \quad (1.3)$$

If $\kappa = 0, -1$ then ℓ takes values in the range $0 < \ell < \infty$, but if $\kappa = +1$ then ℓ is restricted to run over $0 < \ell < \pi R_0$ because $r(\ell)$ vanishes at $\ell = \pi R_0$.

- $0 < \theta < \pi$ and $0 < \phi < 2\pi$ represent the usual angular coordinates of spherical polar coordinates. (Spherical coordinates furnish a convenient description of our view of the universe, with the origin of coordinates representing our vantage point.) The geometry is invariant under the $SO(3)$ rotations of the 2-dimensional spherical surfaces at fixed ℓ and t which these coordinates parameterize.

- $r(\ell)$ is simply related to the arc-length measured along these spherical surfaces of fixed ℓ and t in the sense that a small angular displacement, $d\theta$, is subtended by a proper arc-length

$$ds = a(t) r(\ell) d\theta, \quad (1.4)$$

at a proper distance ℓ . It follows that the sphere having proper radius $\ell a(t)$ has a proper circumference of $\mathcal{C} = 2\pi r(\ell) a(t)$ and its proper area is $\mathcal{A} = 4\pi r^2(\ell) a^2(t)$.

The quantities κ and R_0 characterize the curvature of the spatial slices at fixed t , in the following way.

1.1.1 Flat Spatial Curvature

If $\kappa = 0$ then $r(\ell) = \ell$ and the spatial part of the FRW metric reduces (apart from the overall factor, $a^2(t)$) to the metric of flat 3-dimensional space, written in spherical polar coordinates:

$$ds_3^2 = dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2), \quad (1.5)$$

as may be seen by performing the standard coordinate transformation

$$x = r \sin\theta \cos\phi, \quad y = r \sin\theta \sin\phi, \quad z = r \cos\theta \quad (1.6)$$

in the metric of eq. (??). In this case the parameter R_0 does not appear in the metric.

1.1.2 Positive Spatial Curvature

When $\kappa = 1$ we have $r(\ell) = R_0 \sin(\ell/R_0)$ and the metric for t fixed describes the geometry of a 3-dimensional sphere whose radius of curvature is R_0 . For instance, in this case the circumference of a circle of proper radius $a(t)\ell$ is

$$\mathcal{C} = 2\pi a(t) R_0 \sin\left(\frac{\ell}{R_0}\right), \quad (1.7)$$

which is strictly smaller than the corresponding flat result: $\mathcal{C} < 2\pi a(t)\ell$.

Furthermore, for fixed t , \mathcal{C} is a monotonically increasing function of ℓ until $\ell = \pi R_0/2$, but beyond this point \mathcal{C} decreases until it vanishes at $\ell = \pi R_0$. The maximum coordinate circumference obtained in this way is $\mathcal{C}_{\max} = 2\pi a(t) R_0$.

Notice also that the flat $\kappa = 0$ case is retrieved in the limit of infinite curvature radius: $R_0 \rightarrow \infty$.

1.1.3 Negative Spatial Curvature

When $\kappa = -1$ we have $r(\ell) = R_0 \sinh(\ell/R_0)$, which makes the metric for constant t describe the geometry of a 3-dimensional surface of negative constant curvature. (The surface of a saddle is close to being a 2-dimensional surface having constant negative curvature.) The radius of curvature of this space is R_0 . In this case the circumference of a circle of proper radius $a(t)\ell$ grows monotonically with ℓ ,

$$\mathcal{C} = 2\pi a(t) R_0 \sinh\left(\frac{\ell}{R_0}\right), \quad (1.8)$$

and is always larger than the corresponding flat-space result: $\mathcal{C} > 2\pi a(t)\ell$.

Again the flat $\kappa = 0$ case is retrieved in the limit of infinite curvature radius: $R_0 \rightarrow \infty$.

1.2 Particle Motion

For the purposes of cosmology galaxies are particles, and so their trajectories in this spacetime are given, as usual, by solutions to the geodesic equation, eq. (A.2)

$$\frac{d^2 x^\mu}{ds^2} + \Gamma_{\nu\lambda}^\mu[x(s)] \left(\frac{dx^\nu}{ds}\right) \left(\frac{dx^\lambda}{ds}\right) = 0, \quad (1.9)$$

with the Christoffel symbols, $\Gamma_{\nu\lambda}^\mu$, given by eq. (A.3).

For the FRW metric the only nonzero Christoffel symbols turn out to be given by

$$\begin{aligned} \Gamma_{\ell\ell}^t &= a\dot{a}, & \Gamma_{\theta\theta}^t &= a\dot{a}r^2, & \Gamma_{\phi\phi}^t &= a\dot{a}r^2 \sin^2\theta, \\ \Gamma_{t\ell}^\ell &= \Gamma_{\ell t}^\ell = \Gamma_{t\theta}^\theta = \Gamma_{\theta t}^\theta = \Gamma_{t\phi}^\phi = \Gamma_{\phi t}^\phi = \frac{\dot{a}}{a}, & & & & & (1.10) \\ \Gamma_{\theta\theta}^\ell &= -rr', & \Gamma_{\phi\phi}^\ell &= -rr' \sin^2\theta, & \Gamma_{\ell\theta}^\theta &= \Gamma_{\theta\ell}^\theta = \Gamma_{\ell\phi}^\phi = \Gamma_{\phi\ell}^\phi = \frac{r'}{r}, \\ \Gamma_{\phi\phi}^\theta &= -\sin\theta \cos\theta, & \Gamma_{\theta\phi}^\phi &= \cot\theta, & & & \end{aligned}$$

where the dots denote differentiation with respect to t and the primes represent derivatives with respect to ℓ . Several features of the geodesics may be seen by using these expressions for $\Gamma_{\nu\lambda}^\mu$ in the geodesic equations.

- *Radial Motion:* If $d\theta/ds = d\phi/ds = 0$ at one point, then these quantities remain zero along the entire geodesic. This shows that an initially radial motion continues

in the radial direction for all times. Radial free fall is described by the equations of motion

$$\frac{d^2t}{ds^2} + a \dot{a} \left(\frac{d\ell}{ds} \right)^2 = 0 \quad \text{and} \quad \frac{d^2\ell}{ds^2} + 2 \frac{\dot{a}}{a} \left(\frac{d\ell}{ds} \right) \left(\frac{dt}{ds} \right) = 0, \quad (1.11)$$

which together also imply the constancy of $(dt/ds)^2 - a^2(d\ell/ds)^2$, as expected on general grounds.

- *Inertial Motion:* If a galaxy is initially at rest — and so $d\ell/ds = d\theta/ds = d\phi/ds = 0$ — then it remains at rest, at fixed coordinate position, for all t .

1.2.1 Hubble Flow and Peculiar Motion

Consider now a particle moving more slowly than light, but for which some force keeps it from moving along a geodesic. This might happen for a galaxy, for instance, if some local density enhancement attracts it. In particular, consider for simplicity a galaxy having coordinates $(t, \ell = \ell(t), \theta = \theta_0, \phi = \phi_0)$, which moves on a purely radial trajectory. The proper distance to this galaxy from, say, the origin is given by $D(\ell, t) = \ell(t)a(t)$, and so its proper velocity relative to an observer at the origin is

$$V_p = \frac{dD}{dt} = \ell \frac{da}{dt} + a \frac{d\ell}{dt} = H D + a \frac{d\ell}{dt}, \quad (1.12)$$

where

$$H \equiv \frac{1}{a} \left(\frac{da}{dt} \right). \quad (1.13)$$

The first term of eq. (1.12) describes the galaxy's apparent motion due to the overall universal expansion, and expresses the *Hubble Law*: in the absence of other motions at any given instant all galaxies recede with a proper speed which is proportional to their proper distance. By contrast, the second term describes *peculiar* velocity,

$$V_{\text{pec}} = a \frac{d\ell}{dt}, \quad (1.14)$$

which expresses any non-free-fall motion which is not due to the overall FRW metric.

Measurements of H at the present epoch, $H_0 = H(t = t_0)$, give $H_0 = 70 \pm 10$ km/sec/Mpc, which for a galaxy 1,000 Mpc distant (using present-day proper distance) would represent an apparent Hubble velocity of $V_H = 70,000$ km/sec, or $v_H/c \sim 0.2$.

If the proper time of an observer riding in this galaxy, τ , is used as the parameter along its trajectory, then (see appendix)

$$g_{\mu\nu} \left(\frac{dx^\nu}{d\tau} \right) \left(\frac{dx^\nu}{d\tau} \right) = -1. \quad (1.15)$$

This expression allows the time dilation of observers in the galaxy to be related to the motion just described. Specialized to the radial motion $\ell = \ell(t)$ this last equation reads

$$\left(\frac{dt}{d\tau}\right)^2 - a^2 \left(\frac{d\ell}{d\tau}\right)^2 = \left(\frac{dt}{d\tau}\right)^2 [1 - V_{\text{pec}}^2] = 1, \quad (1.16)$$

and so the local time dilation is

$$\frac{dt}{d\tau} = \gamma_{\text{pec}} = \frac{1}{\sqrt{1 - V_{\text{pec}}^2}}. \quad (1.17)$$

We see that there is no time dilation in the absence of peculiar motion, so t describes the proper time of all observers who sit at fixed coordinate positions. In the presence of proper motion a time dilation arises, given by the usual special relativistic expression in terms of the peculiar velocity, V_{pec} .

1.2.2 Light Rays and Redshift

The trajectories of particles (like photons) moving at the speed of light similarly satisfy

$$g_{\mu\nu} \left(\frac{dx^\mu}{ds}\right) \left(\frac{dx^\nu}{ds}\right) = 0, \quad (1.18)$$

which for radial motion specializes to

$$\frac{dt}{ds} = \pm a \frac{d\ell}{ds}. \quad (1.19)$$

Consider now a photon which is sent to us (at the origin) along a radial trajectory from a galaxy which is situated at fixed coordinate position $\ell = L$. If we suppose the photon to arrive at our position at $t = 0$ then we may compute its departure time at the emitting galaxy, $t = -T$. Explicitly, the *look-back time*, T , is given by eq. (1.19) to be

$$L = \int_0^T \frac{dt}{a(t)}. \quad (1.20)$$

Imagine now repeating this calculation for a sequence of photons (or for a train of wave crests) which are emitted from the galaxy and are received here. Suppose two consecutive photons are emitted at events which are labelled by the coordinate positions $(-T, L, \theta_0, \phi_0)$ and $(-T + \delta T, L + \delta L, \theta_0, \phi_0)$, with the first of these received at the origin at time $t = 0$ and the second arriving at $(\delta t, \delta \ell, \theta_0, \phi_0)$. The redshift of such a wave train may be found by computing how δt depends on δT , the scale factor, $a(t)$, and the peculiar motions of the emitter and observer.

We know that the trajectories of both photons satisfy eq. (1.19), and so we know

$$L = \int_0^T \frac{dt}{a(t)} \quad \text{and} \quad (L + \delta L) - \delta \ell = \int_{-\delta t}^{T-\delta T} \frac{dt}{a(t)}. \quad (1.21)$$

Subtracting the first of these from the second, and expanding the result to first order in the small quantities δt , δT δL leads to the following relation

$$\delta L - \delta \ell = \int_{-\delta t}^{T-\delta T} \frac{dt}{a(t)} - \int_0^T \frac{dt}{a(t)} \approx \frac{\delta t}{a_0} - \frac{\delta T}{a(T)}, \quad (1.22)$$

where $a_0 = a(0)$. Dividing by δT then gives

$$\frac{\delta L}{\delta T} - \frac{\delta \ell}{\delta t} \left(\frac{\delta t}{\delta T} \right) = \frac{1}{a_0} \left(\frac{\delta t}{\delta T} \right) - \frac{1}{a(T)}. \quad (1.23)$$

This may now be solved for $\delta t/\delta T$ as a function of a_0 , $a(T)$ and the emitter and observer's peculiar velocities, $V_{\text{pec}} = a(T)[\delta L/\delta T]$ and $v_{\text{pec}} = a_0[\delta \ell/\delta t]$ to give

$$\frac{\delta t}{\delta T} = \frac{a_0}{a(T)} \left(\frac{1 + V_{\text{pec}}}{1 + v_{\text{pec}}} \right). \quad (1.24)$$

The redshift, z , of the light is defined in terms of its wavelength at emission, λ_{em} , and at observation, λ_{obs} , by $z = (\lambda_{\text{obs}} - \lambda_{\text{em}})/\lambda_{\text{em}}$ and so

$$\begin{aligned} 1 + z &= \frac{\lambda_{\text{obs}}}{\lambda_{\text{em}}} = \frac{\delta \tau_{\text{obs}}}{\delta \tau_{\text{em}}} = \frac{\delta t}{\delta T} \left[\frac{1 - v_{\text{pec}}^2}{1 - V_{\text{pec}}^2} \right]^{1/2} \\ &= \frac{a_0}{a(T)} \left(\frac{1 + V_{\text{pec}}}{1 + v_{\text{pec}}} \right) \left[\frac{1 - v_{\text{pec}}^2}{1 - V_{\text{pec}}^2} \right]^{1/2}. \end{aligned} \quad (1.25)$$

This last expression uses eq. (1.17) to relate the proper time of the observer, $\delta \tau_{\text{obs}}$, and of the emitter, $\delta \tau_{\text{em}}$, to the corresponding coordinate time differences, δt and δT .

Eq. (1.25) is the main result. For negligible peculiar motions it reduces to a simple expression for the redshift due to the Hubble flow

$$1 + z = \frac{a_0}{a(T)}, \quad (1.26)$$

which is a *redshift* – *i.e.* $z > 0$ – if the universe expands – *i.e.* $a_0 > a(T)$.

This expression gives a good method for measuring the universe's scale factor, $a(t)$, since it shows that it is simply related to the redshift of the light received from distant galaxies.

For non-relativistic peculiar velocities this generalizes to the approximate formula

$$1 + z \approx \frac{a_0}{a(T)} \left[1 + (V_{\text{pec}} - v_{\text{pec}}) \right]. \quad (1.27)$$

Notice that (as expected) relative peculiar motion also generates a redshift – $z > 0$ – if $V_{\text{pec}} > v_{\text{pec}}$ – that is, if the emitting galaxy is receding from the observing one.

In principle, the dependence of z on peculiar velocity complicates the inference of the universal scale factor from measurements of redshift, since in principle it requires knowledge of the peculiar velocity of the distant emitting galaxy. In practice, however, this complication is only important for relatively nearby galaxies, for which the redshift due to the peculiar velocities are not dominated by that due to the universal expansion.

1.3 Distance vs Redshift

In FRW cosmology the expansion of the universe is characterized by the time dependence of the scale factor, $a(t)$, which we shall see is in most circumstances a monotonic function of t . In principle, predictions for $a(t)$ can be tested by measuring the proper distances, $D(L, -T)$, to distant celestial objects and comparing this with the look-back time, T , to these objects. Measurements of $D(L, -T)$ vs T allow the inference of $a(t)$ because of the connection between L and T — *i.e.* the relation $L(T)$ given implicitly by eq. (1.21) — which expresses the fact that all of our observations about the distant universe lie along our past light cone, because they rely on our detecting photons which have come to us from the far reaches of space.

In practice, however, it is much easier to directly measure a than it is to measure T because of the direct relationship between a and redshift. So inferences about the geometry of spacetime instead are founded on measuring the dependence of distance on redshift, z , for distant objects, rather than on look-back time, T . z and T carry the same information provided $a(t)$ is a monotonic function of time, and so it is more convenient to use z itself as an operational measure of the universe’s age and size.

The remainder of this section derives expressions for the dependence of various measures of distance on redshift, given a universal expansion history, $a(t)$.

1.3.1 Proper Distance

Consider, then, a galaxy which at event $(-T, L, \theta_0, \phi_0)$ sends light to us which we receive at the origin at $t = 0$. Writing $a_0 = a(0)$, the present-day proper distance to this galaxy is given by

$$D(T) = D(L(T), -T) = a_0 L = \int_0^T \left(\frac{a_0}{a(t)} \right) dt. \quad (1.28)$$

This may be changed into an expression in terms of redshift by changing integration variable from t to z using the relations

$$1 + z = \frac{a_0}{a(t)} \quad \text{and so} \quad dz = - \left(\frac{a_0 \dot{a}}{a^2} \right) dt = -(1 + z) H dt, \quad (1.29)$$

where as before $H = \dot{a}/a$. This leads to the desired result

$$D(z) = \int_0^z \frac{dz'}{H(z')}. \quad (1.30)$$

Unfortunately, proper distance is also not particularly convenient since it is not easily obtained from observations. There are two other notions of distance which are more practical, whose dependence on z is now derived.

1.3.2 Luminosity Distance

One way of inferring how far away a distant object becomes possible if the object's intrinsic rate of energy release per unit time — *i.e.* luminosity, \mathcal{L} — is known. If \mathcal{L} is known then it may be compared with the observed energy flux, f , which is received at Earth from the object, with the distance to the object obtained by assuming that the flux is related to \mathcal{L} only by the geometrical solid angle which the Earth subtends at the source. For instance in Euclidean space the flux received by a source of luminosity \mathcal{L} situated a distance D away is given by

$$f = \frac{\mathcal{L}}{4\pi D^2}, \quad (1.31)$$

provided the source sends its energy equally in all directions and that there is no absorption or scattering of the light while it is *en route* from the source. The *luminosity distance*, D_L , to the object may then be defined in terms of \mathcal{L} and f by $D_L = [\mathcal{L}/(4\pi f)]^{1/2}$. This is the distance measure which is used, for example, in recent measurements of the universal expansion using distant Type I supernovae.

Suppose, then, that the source emits a packet of light having energy, δE_{em} , in a time, δt_{em} , and so has luminosity $\mathcal{L} = \delta E_{\text{em}}/\delta t_{\text{em}}$. In an FRW universe the relation between \mathcal{L} and the flux, f , we observe depends differently on distance, in the following ways.

- Because the wavelength of the light is stretched by the universal expansion, and the energy of a light wave is inversely proportional to its wavelength ($E = h\nu = hc/\lambda$) this packet of energy arrives to us having a red-shifted energy $\delta E_{\text{obs}} = \delta E_{\text{em}}/(1 + z)$.

- Because of the expansion of space the wavelength of the light stretches as space expands while it is en route. As a result the spatial extent of the packet also stretches by a factor $1 + z$ during its passage between the source and us. This means that on its arrival the time taken for the packet to deliver its energy is $\delta t_{\text{obs}} = \delta t_{\text{em}}(1 + z)$.
- The total energy from the source is sent in all directions, and so (using the FRW metric) it is spread over a sphere having surface area $\mathcal{A} = 4\pi r^2(L)a^2$ at a proper distance $D = La$ from the source, where $r(L)$ is given by eq. (1.2).

The flux observed at Earth is therefore given by

$$\begin{aligned}
 f &= \frac{1}{4\pi r^2(L) a_0^2} \left(\frac{\delta E_{\text{obs}}}{\delta t_{\text{obs}}} \right) \\
 &= \frac{1}{4\pi r^2(L) a_0^2} \left(\frac{\delta E_{\text{em}}/(1+z)}{\delta t_{\text{em}}(1+z)} \right) \\
 &= \left(\frac{\mathcal{L}}{4\pi r^2(L) a_0^2} \right) \frac{1}{(1+z)^2},
 \end{aligned} \tag{1.32}$$

and so the luminosity distance becomes

$$D_L(z) \equiv \left[\frac{\mathcal{L}}{4\pi f} \right]^{1/2} = a_0 r(L(z)) (1+z). \tag{1.33}$$

Notice that the present-day proper distance to the same galaxy would be $D = L a_0$. Since in the special case of a spatially-flat universe, $\kappa = 0$, we have $r(\ell) = \ell$, in this case D_L is related to this proper distance by

$$D_L(z) = D(z) (1+z) \quad (\text{if } \kappa = 0). \tag{1.34}$$

1.3.3 Angular-Diameter Distance

A second measure of distance becomes possible if an object of known proper length is observed at a distance, since the angle which the object subtends as seen from Earth is geometrically related to its distance from us. In Euclidean geometry an object of length ds placed a distance $D \gg ds$ from us subtends an angle

$$d\theta = \frac{ds}{D} \text{ (radians)}, \tag{1.35}$$

which motivates defining the angular-diameter distance by $D_A = ds/d\theta$ in terms of the (assumed) known length ds and measured angle $d\theta$. This notion of distance comes up

in a later section, where it arises in the study of the temperature fluctuations of the cosmic microwave background radiation.

The connection between ds and $d\theta$ differs in the FRW geometry in the following ways.

- At any given time, within an FRW geometry the proper length of an object which subtends an angle $d\theta$ when placed a proper distance $D = a\ell$ away is given by $ds = ar(\ell) d\theta$, with $r(\ell)$ given by eq. (1.2).
- When an object is observed from a great distance it is the proper distance at the time its light was emitted which appears in the previous argument. Due to the overall expansion of space this corresponds to a proper distance at present which is a factor $a_0/a(-T) = 1 + z$ larger.

With these two effects in mind, the angle subtended by an object having proper length ds when observed from a present-day proper distance $D = a_0L$ away is given by

$$d\theta = \frac{ds}{a(-T)r(L)} = \frac{ds}{a_0 r(L)/(1+z)}, \quad (1.36)$$

and so the angular-diameter distance of such an object is

$$D_A(z) \equiv \frac{ds}{d\theta} = \frac{a_0 r(L(z))}{1+z} = \frac{D_L(z)}{(1+z)^2}, \quad (1.37)$$

where the last equality uses eq. (1.33).

Notice that in the special case of a spatially-flat universe ($\kappa = 0$), we have $r(\ell) = \ell$ and so the angular-diameter distance to an object situated a proper distance $D = a_0L$ away is

$$D_A(z) = \frac{D(z)}{1+z} \quad (\text{if } \kappa = 0). \quad (1.38)$$

This is equivalent to the object's proper distance as measured at the time of the light's emission rather than its present proper distance.

1.4 Examples

For later purposes it is useful to evaluate the above distance-redshift expressions for various choices for the time-dependence of the universal expansion, $a(t)$. For simplicity (and because this appears to be a good description of the present-day universe) in the case of D_L and D_A we provide formulae for the special case $\kappa = 0$.

1.4.1 The Recent Universe

A great many cosmological observations are restricted to the comparatively nearby universe, for which the observed red-shifts are small. For such small red-shifts it is useful to evaluate the distance-redshift expressions by expanding about the present epoch, for which $z = 0$. Consider, therefore, a scale factor of the form

$$a(t) = a_0 + \dot{a}_0 (t - t_0) + \frac{1}{2} \ddot{a}_0 (t - t_0)^2 + \dots, \quad (1.39)$$

where $t = t_0$ denotes the present epoch. In what follows it is convenient to measure the time difference in units of H_0^{-1} , where $H_0 = \dot{a}_0/a_0$ by defining $\zeta = -H_0 (t - t_0)$, in which case the above expansion is expected to furnish a good approximation for $|\zeta| \lesssim 1$. (Notice that as defined $\zeta \geq 0$ when applied to $a(t)$ in the past universe, for which $t \leq t_0$.)

In terms of this expansion the redshift of light becomes

$$1 + z = \frac{a_0}{a(t)} = 1 + \zeta + \left(1 + \frac{q_0}{2}\right) \zeta^2 + \dots, \quad (1.40)$$

where $q_0 \equiv -a_0 \ddot{a}_0 / \dot{a}_0^2 = -\ddot{a} / (a_0 H_0^2)$, with the sign chosen so that $q_0 > 0$ for a decelerating universe (for which $\ddot{a}_0 < 0$).

The distance-redshift relations are governed by $H(z)$, which is given by

$$\begin{aligned} H &= H_0 \left[1 + \left(\frac{\ddot{a}_0}{\dot{a}_0} - \frac{\dot{a}_0}{a_0} \right) (t - t_0) + \dots \right] \\ &= H_0 \left[1 + (1 + q_0) z + \dots \right]. \end{aligned} \quad (1.41)$$

Using this in eq. (1.30) leads to the following expression for $D(z)$ near $z = 0$

$$D(z) = H_0^{-1} \left[z - \frac{1}{2} (1 + q_0) z^2 + \dots \right], \quad (1.42)$$

which for $\kappa = 0$ also imply the following small- z expansions for the luminosity and angular-diameter distances

$$\begin{aligned} D_L(z) &= H_0^{-1} \left[z + \frac{1}{2} (1 - q_0) z^2 + \dots \right] \\ D_A(z) &= H_0^{-1} \left[z - \frac{1}{2} (3 + q_0) z^2 + \dots \right]. \end{aligned} \quad (1.43)$$

Clearly a precise determination of distance vs redshift for objects having moderate redshifts permits the extraction of both the present-day Hubble constant (H_0) and the deceleration parameter (q_0).

1.4.2 Power-Law Expansion

Another situation of considerable practical interest is the case where the expansion varies as a power of t , as in

$$1 + z = \frac{a_0}{a(t)} = \left(\frac{t_0}{t}\right)^\alpha, \quad (1.44)$$

for some choices for the parameters a_0 , t_0 and α . In later sections we shall find this law is produced (if $\kappa = 0$) with $\alpha = 1/2$ for a universe full of radiation, and with $\alpha = 2/3$ for a universe consisting dominantly of non-relativistic matter (like atoms or stars). For such a universe the Hubble and deceleration parameters become

$$H(t) = \frac{\dot{a}}{a} = \frac{\alpha}{t} = H_0 \left(\frac{t_0}{t}\right) = H_0 (1+z)^{1/\alpha} \quad \text{and} \quad q(t) = -\frac{a\ddot{a}}{\dot{a}^2} = \frac{1-\alpha}{\alpha}. \quad (1.45)$$

Notice that this kind of power law implies that a vanishes for $t = 0$ provided only that $\alpha > 0$ (and so in particular does so for the cases $\alpha = 1/2$ and $2/3$ mentioned above). This is the Big Bang which underlies much of modern cosmology. In terms of $q = q_0$ and the present value of the Hubble parameter, H_0 , this occurs a time

$$t_0 = \alpha H_0^{-1} = \frac{H_0^{-1}}{q_0 + 1} \quad (1.46)$$

in the past.

Using the above expressions for q and $H(z)$ in eq. (1.30) gives the following expression for the proper distance

$$D(z) = H_0^{-1} \int_0^z \frac{dz'}{(1+z')^{1/\alpha}} = \frac{H_0^{-1}}{q} \left[1 - \frac{1}{(1+z)^q} \right], \quad (1.47)$$

which with $D_L(z) = D(z)(1+z)$ and $D_A(z) = D(z)/(1+z)$ give the luminosity and angular-diameter distances when $\kappa = 0$.

Radiation-Dominated Universe (if $\kappa = 0$): As mentioned above, the special case where the universe is dominated by radiation with $\kappa = 0$ turns out to correspond to a power-law expansion with $\alpha = 1/2$, and so we have $H(z) = H_0(1+z)^2$ and $q(z) = q_0 = 1$. This leads to the following proper distance

$$D(z) = H_0^{-1} \left(\frac{z}{1+z} \right) = \begin{cases} H_0^{-1} [z - z^2 + \dots] & \text{if } z \ll 1 \\ H_0^{-1} [1 - \frac{1}{z} + \dots] & \text{if } z \gg 1 \end{cases}. \quad (1.48)$$

Since $\kappa = 0$ the luminosity and angular-diameter distances become

$$D_L(z) = H_0^{-1} z, \quad D_A(z) = H_0^{-1} \left[\frac{z}{(1+z)^2} \right] = \begin{cases} H_0^{-1} [z - 2z^2 + \dots] & \text{if } z \ll 1 \\ \frac{H_0^{-1}}{z} [1 - \frac{2}{z} + \dots] & \text{if } z \gg 1 \end{cases}. \quad (1.49)$$

Matter-Dominated Universe (if $\kappa = 0$): The special case where $\kappa = 0$ and the universe is dominated by non-relativistic matter corresponds to power-law expansion with $\alpha = 2/3$, and so $H(z) = H_0(1+z)^{3/2}$ and $q(z) = q_0 = 1/2$. This leads to the proper distance

$$D(z) = 2H_0^{-1} \left[\frac{(1+z)^{1/2} - 1}{(1+z)^{1/2}} \right] = \begin{cases} H_0^{-1} [z - \frac{3}{4}z^2 + \dots] & \text{if } z \ll 1 \\ 2H_0^{-1} \left[1 - (\frac{1}{z})^{1/2} + \dots \right] & \text{if } z \gg 1 \end{cases}. \quad (1.50)$$

Because $\kappa = 0$ the luminosity and angular-diameter distances are

$$D_L(z) = 2H_0^{-1} [(1+z) - \sqrt{1+z}] = \begin{cases} 2H_0^{-1} [z + \frac{1}{4}z^2 + \dots] & \text{if } z \ll 1 \\ 2H_0^{-1} z \left[1 - (\frac{1}{z})^{1/2} + \dots \right] & \text{if } z \gg 1 \end{cases}, \quad (1.51)$$

and

$$D_A(z) = 2H_0^{-1} \left[\frac{(1+z)^{1/2} - 1}{(1+z)^{3/2}} \right] = \begin{cases} H_0^{-1} [z - \frac{7}{4}z^2 + \dots] & \text{if } z \ll 1 \\ \frac{2H_0^{-1}}{z} \left[1 - (\frac{1}{z})^{1/2} + \dots \right] & \text{if } z \gg 1 \end{cases}. \quad (1.52)$$

Notice that for both matter and radiation domination the present-day proper distance approaches a limiting value of order H_0^{-1} when $z \rightarrow \infty$. This implies that we do not learn about arbitrarily large distances when we look into the past at objects having larger and larger redshifts. A related observation is the fact that the angular-diameter distance is not a monotonic function of z , since it grows like z for small z but vanishes asymptotically for large z , proportional to $1/z$. Since (when $\kappa = 0$) angular-diameter distance is the proper distance to the source measured at the time the light is emitted rather than observed, this vanishing of D_A for large z shows that our observations are limited to a vanishingly small local region in the very distant past. This limitation to our view is called our local *particle horizon*. It arises because for these geometries the universe becomes vanishingly small at a finite time in our past and the universal expansion can be fast enough to permit objects to be sufficiently distant that light cannot reach us from them given the limited age of the universe.

1.4.3 Exponential Expansion

The next special case of interest corresponds to exponential expansion

$$1 + z = \frac{a_0}{a(t)} = \exp[-H_0(t - t_0)], \quad (1.53)$$

which may be regarded as the limiting case of a power law for which $\alpha \rightarrow \infty$. We shall find this kind of expansion can be produced when the universal energy density is dominated by the energy of the vacuum.

In this case the Hubble and deceleration parameters are time-independent, with

$$H(t) = \frac{\dot{a}}{a} = H_0 \quad \text{and} \quad q(t) = q_0 = -1, \quad (1.54)$$

and the redshift-dependence of the proper distance is $D(z) = H_0^{-1} \int_0^z dz' = H_0^{-1}z$. The luminosity and angular-diameter distances (when $\kappa = 0$) then become.

$$D_L(z) = H_0^{-1}z(1 + z) \quad \text{and} \quad D_A(z) = H_0^{-1} \left(\frac{z}{1 + z} \right). \quad (1.55)$$

Unlike the case of matter- and radiation-dominated considered earlier, in this case the present-day proper distance grows without bound but the proper-distance at emission approaches a fixed limit, $D_A \rightarrow H_0^{-1}$, as $z \rightarrow \infty$. This distance represents an apparent *horizon* beyond which we are unable to penetrate with observations, and differs from the particle horizon considered above because it is not tied to there only having been a finite proper time since the universe had zero size. For the exponentially-expanding universe only a finite proper distance in the past is accessible to us even though t can run back to $-\infty$. The existence of this horizon can be traced to the enormous speed of the exponential expansion, with which light waves travelling at finite speed cannot keep up.

2 Dynamics of an Expanding Universe

The previous section described the kinematics of how various distance-redshift relationships depend the universal expansion history, $a(t)$. The present section instead addresses the question of how this expansion history depends on the energy content of the matter which lives inside the universe. This connection has its roots in the Einstein field equations which relate the curvature of spacetime to its energy-momentum content — *i.e.* “matter tells space how to curve”.

2.1 Relating Cosmic Expansion to Matter Content

According to Einstein's field equations the curvature of spacetime — and so also the expansion history of the universe — are governed by the total distribution of stress-energy. This motivates a brief pause to consider the kinds of stress-energy which are possible in a homogeneous and isotropic universe.

2.1.1 Homogeneous and Isotropic Stress-Energy

The conditions of homogeneity and isotropy strongly restrict the distribution of matter and energy within the universe, in the same way that they restrict the metric to take the Friedmann-Robertson-Walker form, given by eq. (1.1). For the stress-energy tensor, $T_{\mu\nu}$, the analogous conditions have the following form.

- Isotropy permits the energy density, $\rho = T^{tt}$, to be an arbitrary function of time, t , and radial position, ℓ , but homogeneity forbids any dependence on the position ℓ . The most general energy density can therefore only be time dependent: $T^{tt} = \rho(t)$.
- Isotropy permits a net energy flux, $s^i = T^{ti}$ with $i = 1, 2, 3$, so long as it points purely in the radial direction.¹ In FRW coordinates this implies $T^{t\theta} = T^{t\phi} = 0$ while $T^{t\ell}$ can be a nonzero function of t and ℓ . Homogeneity, however, requires $T^{t\ell} = 0$ because having a nonzero energy flux would necessarily allow one to distinguish between the directions from which and to which the energy is flowing. The same conclusions equally apply to the momentum density: $\pi^i = T^{it} = 0$.
- Isotropy permits the 3-dimensional stress tensor, T^{ij} , to be nonzero provided it is built from the metric tensor itself, or from the radial direction vector, x^i . That is, isotropy allows $T^{ij} = p g^{ij} + q x^i x^j$, where p and q can be functions of both t and ℓ . However homogeneity precludes p from depending on ℓ , and does not permit a nonzero q at all, since the radial vector picks out a preferred place as its origin. It follows that the stress tensor must have the diagonal form $T_i^j = g_{ik} T^{kj} = p(t) \delta_i^j$.

We are led to the conclusion that homogeneity and isotropy only permit a stress-energy of the form

$$T^{tt} = \rho(t), \quad T^{ti} = T^{it} = 0 \quad \text{and} \quad T^{ij} = p(t) g^{ij}, \quad (2.1)$$

¹This can be removed by changing the radial coordinate, but we do not do so in order not to lose the simple connection between proper distance and coordinate distance, $D = a(t)\Delta\ell$.

which is characterized by two functions of time: $\rho(t)$ and $p(t)$. As is clear from the definition of $T^{\mu\nu}$, ρ represents the (average) energy density as seen by co-moving observers who are situated at fixed values of (ℓ, θ, ϕ) . The interpretation of T^{ij} as a momentum flux together with stress-energy conservation implies that the net rate of change in momentum of a volume V — *i.e.* the net force acting on V — is given by the flux of momentum current through the boundary, ∂V :

$$F^i \equiv \frac{dP^i}{dt} = \int_V \frac{\partial \pi^i}{\partial t} d^3V = - \int_{\partial V} T^{ij} n_j d^2S = - \int_{\partial V} p n^i d^2S, \quad (2.2)$$

which shows that p represents the total (average) *pressure* of the matter whose stress energy is under consideration.

Our goal now is to see how Einstein's equations relate these quantities to $a(t)$.

2.1.2 Einstein's Equations

Specializing the Einstein field equations, eq. (A.5), to homogeneous and isotropic geometries leads to two independent differential equations which relate $a(t)$ to $\rho(t)$ and $p(t)$. These may be chosen to be the *Friedmann equation*,

$$\left(\frac{\dot{a}}{a}\right)^2 + \frac{\kappa}{(R_0 a)^2} = \frac{8\pi G}{3} \rho, \quad (2.3)$$

as well as the equation describing the *Conservation of Stress-Energy*

$$\dot{\rho} + 3 \left(\frac{\dot{a}}{a}\right) (\rho + p) = 0. \quad (2.4)$$

The interpretation of this last equation as energy conservation is more easily seen if it is rewritten as

$$\frac{d}{dt} (\rho a^3) + p \frac{d}{dt} (a^3) = 0, \quad (2.5)$$

which relates the rate of change of the total energy, ρa^3 , to the work done by the pressure as the universe expands. For matter in thermal equilibrium, a comparison of this last equation with the 1st Law of Thermodynamics shows that the expansion of the universe is adiabatic, inasmuch as the total entropy of the matter in the universe does not change in a homogeneous and isotropic expansion.

2.1.3 Cosmic Acceleration and Matter

These two expressions allow the derivation of an equation which governs the acceleration, \ddot{a} , of the universe's expansion. Differentiating eq. (2.3) and using eq. (2.4) to

eliminate $\dot{\rho}$ gives the following result (if $\dot{a} \neq 0$)

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}(\rho + 3p). \quad (2.6)$$

Notice that this last equation implies that $\ddot{a} < 0$ for most forms of matter, since for these ρ and p are typically positive. This corresponds physically to the statement that gravity is always attractive, and so the mutual attraction of the galaxies in the universe always acts to slow down the universal expansion. As we shall see there can be exceptions to this general rule, for which $\rho + 3p < 0$, and so whose presence could cause the universal expansion to accelerate rather than decelerate.

Another application of eq. (2.6) is to use it to see what may be learned about the present-day values of ρ and p from measurements of the present-day expansion rate, H_0 , and deceleration parameter, q_0 . To this end notice that the Friedmann equation evaluated at the present epoch implies

$$H_0^2 + \frac{\kappa}{(a_0 R_0)^2} = \frac{8\pi G}{3} \rho_0 \quad \text{or} \quad 1 + \frac{\kappa}{(a_0 H_0 R_0)^2} = \frac{\rho_0}{\rho_c} \equiv \Omega_0, \quad (2.7)$$

where the *critical density* is defined by $\rho_c \equiv 3H_0^2/(8\pi G)$ and the last equality defines Ω_0 to be the energy density in units of this critical density, $\Omega_0 = \rho_0/\rho_c$. Given the current measurement $H_0 = 70 \pm 10$ km/sec/Mpc, the critical density's numerical value becomes $\rho_c = 5200 \pm 1000$ MeV m⁻³ = $(9 \pm 2) \times 10^{-30}$ g cm⁻³.

ρ_c is defined in the way it is because if $\rho_0 = \rho_c$ then $\kappa = 0$. Similarly if $\kappa = +1$ then we must have $\rho_0 > \rho_c$ and if $\kappa = -1$ then $\rho_0 < \rho_c$. Evaluating the acceleration equation, eq. (2.6), at the present epoch similarly gives

$$q_0 = -\frac{\ddot{a}_0}{a_0 H_0^2} = \frac{4\pi G}{3H_0^2}(\rho_0 + 3p_0) = \frac{\rho_0 + 3p_0}{2\rho_c} = \frac{\Omega_0}{2}(1 + 3w_0), \quad (2.8)$$

where we define $w_0 = p_0/\rho_0$. Clearly a measurement of H_0 and q_0 allows the inference of both ρ_0 and p_0 , and knowledge of ρ_0 also allows the determination of κ , since $\kappa = +1$ if and only if $\Omega_0 > 1$ and $q_0 > \frac{1}{2}$ while $\kappa = -1$ requires both $\Omega_0 < 1$ and $q_0 < \frac{1}{2}$.

2.2 Equations of State

Mathematically speaking, finding the evolution of the universe as a function of time requires the integration of eqs. (2.3) and (2.4), but in themselves these two equations are inadequate to determine the evolution of the three unknown functions, $a(t)$, $\rho(t)$ and $p(t)$. Another condition is required in order to make the problem well-posed.

The missing condition is furnished by the equation of state for the matter in question, which for the present purposes may be regarded as being an expression for the pressure as a function of energy density, $p = p(\rho)$. As we shall see this expression is typically characteristic of the microscopic constituents of the matter whose stress energy is of interest. Such an equation of state naturally arises for matter which is in local thermodynamic equilibrium, since this often allows both p and ρ to be expressed in terms of a single quantity like the local temperature, T . But it may also arise for matter which was only in equilibrium in the past, even if it is no longer in equilibrium at present.

Most of the equations of state of interest in cosmology have the general form

$$p = w \rho, \tag{2.9}$$

where w is a t -independent constant. Given an equation of state of this form it is possible to integrate eqs. (2.3) and (2.4) to determine how a , ρ and p vary with time, as we now see.

The first step is to determine how p and ρ depend on a , since this is dictated by energy conservation. Using eq. (2.9) to eliminate p allows eq. (2.4) to be written

$$\frac{\dot{\rho}}{\rho} + 3(1+w) \frac{\dot{a}}{a} = 0, \tag{2.10}$$

which may be integrated to obtain

$$\rho = \rho_0 \left(\frac{a_0}{a} \right)^\sigma \quad \text{with} \quad \sigma = 3(1+w). \tag{2.11}$$

The pressure satisfies an identical dependence on a by virtue of the equation of state: $p = w\rho$.

If eq. (2.11) is now used to eliminate ρ from eq. (2.3), the following differential equation for $a(t)$ is obtained

$$\dot{a}^2 = \frac{8\pi G \rho_0 a_0^2}{3} \left(\frac{a_0}{a} \right)^{\sigma-2} - \frac{\kappa}{R_0^2}, \tag{2.12}$$

In the special case that $\kappa = 0$ this equation is easily integrated to give

$$a(t) = a_0 \left(\frac{t}{t_0} \right)^\alpha \quad \text{with} \quad \alpha = \frac{2}{\sigma} = \frac{2}{3(1+w)}. \tag{2.13}$$

We now apply the above expressions to a few examples of the equations of state which are known to be relevant to cosmology.

2.2.1 Empty Space

The simplest cosmology possible is obtained in the absence of matter, in which case $\rho = p = 0$. In this case we have $\dot{a}^2 = -\kappa$, from which we see that $\kappa \neq +1$. Two distinct solutions are possible, depending on whether $\kappa = 0$ or $\kappa = -1$.

If $\kappa = 0$ we have $\dot{a} = 0$ and so we may choose $a = 1$ for all t . In this case the FRW metric simply reduces to the flat metric of Minkowski space, written in polar coordinates.

If $\kappa = -1$ then we have $\dot{a} = \pm 1$ and so $a = \pm(t - t_0) + a_0$. This negatively-curved geometry is known as the *Milne Universe*, but so far as we know it does not play any role in Big Bang cosmology.

2.2.2 Radiation

A gas of relativistic particles, like photons or neutrinos (or other particles for sufficiently high temperatures), when in thermal equilibrium has an energy density and pressure given by

$$\rho = a_B T^4 \quad \text{and} \quad p = \frac{1}{3} a_B T^4, \quad (2.14)$$

where $a_B = \pi^2/15 = 0.6580$ is the Stefan-Boltzmann constant (in units where $k_B = c = \hbar = 1$) and T is the temperature. These two expressions ensure that ρ and p satisfy the relation

$$p = \frac{1}{3} \rho \quad \text{and so} \quad w = \frac{1}{3}. \quad (2.15)$$

Since $w = 1/3$ we see that $\sigma = 3(1 + w) = 4$ and so $\rho \propto a^{-4}$. This has a simple physical interpretation for a gas of noninteracting photons, since for these the total number of photons is fixed (and so $n_\gamma \propto a^{-3}$), but each photon energy also redshifts like $1/a$ as the universe expands, leading to $\rho_\gamma \propto a^{-4}$.

Since $\sigma = 4$ we have $\alpha = 2/\sigma = 1/2$, and so if $\kappa = 0$ then $a(t) \propto t^{1/2}$. Explicit expressions are given in previous sections for the proper, luminosity and angular-diameter distance as functions of redshift for this type of expansion.

2.2.3 Non-relativistic Matter

An ideal gas of non-relativistic particles in thermal equilibrium has a pressure and energy density given by²

$$p = nT \quad \text{and} \quad \rho = nm + \frac{nT}{\gamma - 1}, \quad (2.16)$$

²Units are used for which Boltzmann's constant is unity: $k_B = 1$.

where n is the number of particles per unit volume, m is the particle's rest mass and $\gamma = c_p/c_v$ is its ratio of specific heats, with $\gamma = 5/3$ for a gas of monatomic atoms. For non-relativistic particles the total number of particles is usually also conserved, which implies that

$$\frac{d}{dt} [n a^3] = 0. \quad (2.17)$$

Since $m \gg T$ (or else the atoms would be relativistic) the equation of state for this gas may be taken to be

$$p/\rho \approx 0 \quad \text{and so} \quad w \approx 0. \quad (2.18)$$

If $w = 0$ then energy conservation implies $\sigma = 3(1 + w) = 3$ and so ρa^3 is a constant. This is appropriate for non-relativistic matter for which the energy density is dominated by the particle rest-masses, $\rho \approx n m$, because in this case energy conservation is equivalent to conservation of particle number, which we've seen is equivalent to $n \propto a^{-3}$ (since this leaves the total number of particles, $N \sim n a^3$, fixed).

Given that $\sigma = 3$ we have $\alpha = 2/\sigma = 2/3$ and so if $\kappa = 0$ then the universal scale factor expands like $a \propto t^{2/3}$. Explicit expressions for the proper, luminosity and angular-diameter distances for this type of expansion are all given in earlier sections.

Solutions for General κ :

When $\sigma = 3$ it is also possible to solve eq. (2.3) analytically even when $\kappa \neq 0$. We pause here to display these solutions in some detail because most of the history of the universe from $z \sim 10^4$ down to $z \sim 1$ appears to have been governed by a universe whose energy density was dominated by non-relativistic matter.

As was described in earlier sections, we may expect the solutions for general κ to be described by two integration constants, which we may take to be Ω_0 and H_0 , or equivalently to be $q_0 = \Omega_0/2$ and H_0 . The value of κ is related to these parameters because $\Omega_0 = 2q_0 = 1$ if and only if $\kappa = 0$, and $\kappa = +1$ if $\Omega_0 > 1$ and $\kappa = -1$ if $\Omega_0 < 1$.

For $\kappa = +1$ (and so $\rho_0 > \rho_c$) the solution for $a(t)$ is most compactly given in parametric form, as the formula for a cycloid:

$$\begin{aligned} \frac{a(\zeta)}{a_0} &= \frac{q_0}{2q_0 - 1} (1 - \cos \zeta) = \frac{1}{2} \left(\frac{\Omega_0}{\Omega_0 - 1} \right) (1 - \cos \zeta) \\ H_0 t(\zeta) &= \frac{q_0}{(2q_0 - 1)^{3/2}} (\zeta - \sin \zeta) = \frac{1}{2} \left(\frac{\Omega_0}{(\Omega_0 - 1)^{3/2}} \right) (\zeta - \sin \zeta). \end{aligned} \quad (2.19)$$

Here the initial conditions which parameterize this solution are given in terms of the physically measurable parameters, $q_0 = \Omega_0/2$ and H_0 .

As ζ increases from 0 to 2π , t increases monotonically from an initial value of 0 to $t_{\text{end}} = \pi\Omega_0 H_0^{-1}/(\Omega_0 - 1)^{3/2}$, but a/a_0 rises from 0 at $t = 0$ to a maximum value, $\Omega_0/(\Omega_0 - 1)$ when $t = t_{\text{max}} = t_{\text{end}}/2$. After this point a/a_0 decreases monotonically until it again vanishes at $t = t_{\text{end}}$. This describes a universe which begins in a Big Bang at $t = 0$, stops expanding at $t = t_{\text{max}}$ and then finally recollapses and ends in a Big Crunch at $t = t_{\text{end}}$.

For $\kappa = -1$ (and so $\Omega_0 < 1$ and $q_0 < \frac{1}{2}$) the solution for $a(t)$ is given by a very similar expression

$$\begin{aligned} \frac{a(\zeta)}{a_0} &= \frac{q_0}{1 - 2q_0} (\cosh \zeta - 1) = \frac{1}{2} \left(\frac{\Omega_0}{1 - \Omega_0} \right) (\cosh \zeta - 1) \\ H_0 t(\zeta) &= \frac{q_0}{(1 - 2q_0)^{3/2}} (\sinh \zeta - \zeta) = \frac{1}{2} \left(\frac{\Omega_0}{(1 - \Omega_0)^{3/2}} \right) (\sinh \zeta - \zeta). \end{aligned} \quad (2.20)$$

This time both t and a increase monotonically with ζ , whose range runs from 0 to infinity. In this case the universe begins in a Big Bang at $t = 0$ and then continues expanding (and cooling) forever, leading to a Big Chill in the remote future.

2.2.4 The Vacuum

If the vacuum is Lorentz invariant, as the success of special relativity seems to indicate, then its stress energy must satisfy $T_{\mu\nu} = \rho g_{\mu\nu}$. This implies the vacuum pressure must satisfy the only possible Lorentz-invariant equation of state:

$$p = -\rho \quad \text{and so} \quad w = -1. \quad (2.21)$$

Clearly either p or ρ must be negative with this equation of state, and unlike for other equations of state there is no reason of principle for choosing either sign for ρ *a priori*.

Because $w = -1$ when the vacuum energy is dominant, we see that $\sigma = 3(1+w) = 0$ and so energy conservation implies that ρ is a constant, independent of a or t . This kind of constant energy density is often called, for historical reasons, the *cosmological constant*.

In this situation $\alpha = 2/\sigma \rightarrow \infty$, which shows that the power-law solutions, $a \propto t^\alpha$, are not appropriate. Returning directly to the Friedmann equation, eq. (2.3), shows that if $\kappa = 0$ then $\dot{a} \propto \pm a$ and so the solutions are given by exponentials: $a \propto \exp[\pm H_0(t - t_0)]$. Explicit expressions for the proper, luminosity and angular-diameter distances as functions of z are given for this expansion in earlier sections.

Notice also that in this case $\rho + 3p = -2\rho$, which is negative if ρ is positive. As such this furnishes an explicit example of an equation of state for which the universal acceleration, $\ddot{a}/a = -\frac{4}{3}\pi G(\rho + 3p) = +\frac{8}{3}\pi G\rho$, can be positive if $\rho > 0$.

2.3 Multi-Component Fluids

In general the universe contains more than one kind of matter, with some relativistic particles (like photons) mixed with non-relativistic particles (like atoms) plus possibly other more exotic forms, each of which satisfies its own equation of state and interacts fairly weakly with the others. This section summarizes what is known about the universe's contents now, and what may be said about the expansion of the universe in the presence of a mixture of matter of this sort.

2.3.1 The Present-Day Energy Content

Indeed, at present there is evidence that the universe contains at least 4 independent types of matter, whose present-day abundances are now summarized. This section summarizes what is known about the abundance of various types of matter in our present best understanding of the universe.

Radiation

The universe is awash with radiation, with the following components.

Cosmic Photons: The sky is full of photons which are distributed in a thermal distribution whose temperature is $T_\gamma = 2.725$ K, called the Cosmic Microwave Background (CMB). These photons were first directly detected using a microwave horn on the Earth's surface, and their thermal properties have subsequently been precisely measured using balloon- and satellite-borne instruments.

The number density of these CMB photons is determined by its temperature, and turns out to be

$$n_{\gamma 0} = 4.11 \times 10^8 \text{ m}^{-3}, \quad (2.22)$$

which is very high, much higher than the number density of ordinary atoms. The energy density carried by these photons is also determined by their temperature, and turns out to be

$$\rho_{\gamma 0} = 0.261 \text{ MeV m}^{-3} \quad \text{or} \quad \Omega_{\gamma 0} = 5.0 \times 10^{-5}, \quad (2.23)$$

where the critical density, $\rho_c = 5200 \pm 1000 \text{ MeV}^{-3}$ is used.

Starlight: The CMB photons turn out to be somewhat more abundant and carry more energy in them than is the integrated number of photons emitted by stars since stars first formed, and represent the dominant contribution of photons to the universal energy density. For instance, a very rough estimate of the density in starlight is obtained by multiplying the present-day luminosity density of galaxies, $nL = 2 \times 10^8 L_\odot \text{ Mpc}^{-3}$ by

the approximate age of the universe, $H_0^{-1} = 14$ Gy, which gives $\rho_\star = 7 \times 10^{-3}$ MeV m⁻³, or $\Omega_\star = 1 \times 10^{-6}$.

Relict Neutrinos: It is believed on theoretical grounds (more about these grounds in subsequent sections) that there is also an almost equally large population of cosmic relict neutrinos filling the universe, although these neutrinos have never been detected. They are expected to be relativistic and to be thermally distributed, as are the photons. The neutrinos are expected to have a slightly lower temperature, $T_{\nu 0} = 1.9$ K, and are fermions and so have a slightly different energy-density/temperature relation than do neutrinos.

Their contribution to the present-day cosmological energy budget is therefore not negligible, and is predicted to be

$$\rho_{\nu 0} = 0.18 \text{ MeV m}^{-3} \quad \text{or} \quad \Omega_{\nu 0} = 3.4 \times 10^{-5}, \quad (2.24)$$

leading to a total radiation density, $\Omega_{R0} = \Omega_{\gamma 0} + \Omega_{\nu 0}$, of size

$$\rho_{R0} = 0.44 \text{ MeV m}^{-3} \quad \text{or} \quad \Omega_{r0} = 8.4 \times 10^{-5}. \quad (2.25)$$

Baryons

The main constituents of the matter we see around us on Earth are atoms, which are themselves made up of protons, neutrons and electrons, and these are predominantly non-relativistic at the present epoch. Furthermore the abundance of electrons is very likely to precisely equal that of protons, since these carry opposite electrical charge, and a precise equality of abundance is required to ensure that the universe carries no net charge.

The mass of the proton and neutron is 940 MeV, which is about 1840 times more massive than the electron, and so the energy density in ordinary non-relativistic particles is likely to be well approximated by the total energy in protons and neutrons. This is also called the total energy in baryons, since protons and neutrons carry an approximately conserved charge called baryon number. For reasons which will become clear in later sections, it is possible to determine the total number of baryons in the universe (regardless of whether or not they are presently visible) from the success of the predictions of the abundances of light elements due to primordial nucleosynthesis during the very early universe. This leads to the following determination of the total energy density in baryons (*i.e.* ordinary protons, neutrons and electrons)

$$\rho_{B0} = 210 \text{ MeV m}^{-3} \quad \text{or} \quad \Omega_{B0} = 0.04. \quad (2.26)$$

For purposes of comparison, the amount of *luminous* matter is considerably smaller than this. Using the previously-quoted luminosity density for galaxies, $nL = 2 \times 10^8 L_\odot \text{ Mpc}^{-3}$, together with a mass-to-luminosity ratio of $M/L = 4M_\odot/L_\odot$, gives an energy density in luminous baryons which is roughly 10% of the total amount in baryons

$$\rho_{L0} = 20 \text{ MeV m}^{-3} \quad \text{or} \quad \Omega_{L0} = 0.004. \quad (2.27)$$

It should be emphasized that although there is more energy in baryons than in CMB photons, the *number density* of baryons is much smaller. That is

$$n_{B0} = \frac{210 \text{ MeV m}^{-3}}{940 \text{ MeV}} = 0.22 \text{ m}^{-3} = 5 \times 10^{-10} n_{\gamma 0}. \quad (2.28)$$

Dark Matter

There several lines of evidence that point to the existence of another form of non-relativistic matter besides baryons, which carry much more energy than do the baryons. The evidence for this so-called *Dark Matter* comes from several independent measures of the total amount of gravitating mass in galaxies and in clusters of galaxies. The rotation rates of galaxies indicate that there is considerably more gravitating mass present than would be inferred by counting the luminous matter which can be seen. A similar result holds for the total mass in galaxy clusters, as estimated from the motions of their constituent galaxies, from the temperature of their hot inter-galactic gas and from the amounts of gravitational lensing which they produce. Furthermore this matter should be non-relativistic since it takes part in the gravitational collapse which gives rise to galaxies and their clusters. (Why this suggests a non-relativistic equation of state is explained in a later section.)

All of these estimates appear to be consistent with one another, and indicate a non-relativistic matter density which is of order

$$\rho_{DM0} = 1350 \text{ MeV m}^{-3} \quad \text{or} \quad \Omega_{DM0} = 0.26. \quad (2.29)$$

Provided this has the same equation of state, $p \approx 0$, as have the baryons, this leads to a total energy density in non-relativistic matter, $\Omega_{M0} = \Omega_{B0} + \Omega_{DM0}$, which is of order

$$\rho_{M0} = 1600 \text{ MeV m}^{-3} \quad \text{or} \quad \Omega_{m0} = 0.30. \quad (2.30)$$

Dark Energy

Finally, there are two lines of evidence which point to a second form of unknown matter in the universe. One line is based on the recent observations that the universal

expansion is accelerating, and so requires the universe must now be dominated by a form of matter for which $\rho + 3p < 0$. The second line of argument is based on the evidence in favor of the universe being spatially flat, $\kappa = 0$ and so $\Omega_0 = 1$, coming from measurements of the angular fluctuations in the temperature of the CMB photon distributions. These two lines of evidence are consistent with one another (within sizeable errors) and point to a *Dark Energy* density which is of order

$$\rho_{DE0} = 3600 \text{ MeV m}^{-3} \quad \text{or} \quad \Omega_{DE0} = 0.70. \quad (2.31)$$

The equation of state for the Dark Energy is not known, apart from the remark that the observations indicate both that at present $\rho_{DE0} \sim 0.7 \rho_c > 0$ and $w \lesssim -0.7$. If w is constant, it is likely on theoretical grounds that $w = -1$ and the Dark Energy is simply the Lorentz-invariant vacuum energy density. Although it is not yet known whether the vacuum need be Lorentz invariant to the precision required to draw cosmological conclusions of sufficient accuracy, in what follows it will be assumed that the Dark Energy equation of state is $w = -1$.

2.3.2 Earlier Epochs

Given the present-day cosmic ingredients of the previous section, this section uses the equations of state for each type of ingredient to extrapolate the relative abundances into the past in order to estimate what can be said about the cosmic environment during earlier epochs. The main assumption for this extrapolation is that the various components of the cosmic fluid are weakly coupled to one another, and so cannot transfer energy directly to one another.

Under these circumstances the equation of energy conservation, eq. (2.4), applies separately to each component of the fluid. The relative energy densities then change as these components respond differently to the expansion of the universe, as follows.

- **Radiation:** For photons, starlight and relict neutrinos of sufficiently small mass we have $w = 1/3$ and so $\rho(a)/\rho_0 = (a_0/a)^4$;
- **Non-relativistic Matter:** For both ordinary matter (baryons and electrons) and for the Dark Matter we have $w = 0$ and so $\rho(a)/\rho_0 = (a_0/a)^3$;
- **Vacuum Energy:** Assuming the Dark Energy has the equation of state $w = -1$ we have $\rho(a) = \rho_0$ for all a .

This implies the total energy density and pressure have the form

$$\begin{aligned}\rho(a) &= \rho_{DE0} + \rho_{M0} \left(\frac{a_0}{a}\right)^3 + \rho_{R0} \left(\frac{a_0}{a}\right)^4 \\ p(a) &= -\rho_{DE0} + \frac{1}{3} \rho_{R0} \left(\frac{a_0}{a}\right)^4.\end{aligned}\tag{2.32}$$

As the universe is run backwards to smaller sizes it is clear that these results imply that the Dark Energy becomes less and less important, while relativistic matter becomes more and more important. Although the Dark Energy now dominates, non-relativistic matter is the next most abundant contribution, and when extrapolated backwards would have satisfied $\rho_M(a) > \rho_{DE}(a)$ relatively recently, at a redshift

$$1 + z = \frac{a_0}{a} > \left(\frac{\Omega_{DE0}}{\Omega_{M0}}\right)^{1/3} = \left(\frac{0.7}{0.3}\right)^{1/3} = 1.3.\tag{2.33}$$

The energy density in baryons alone becomes larger than the Dark Energy density at a slightly earlier epoch

$$1 + z > \left(\frac{\Omega_{DE0}}{\Omega_{B0}}\right)^{1/3} = \left(\frac{0.7}{0.04}\right)^{1/3} = 2.6.\tag{2.34}$$

For times earlier than this the dominant component of the energy density is due to non-relativistic matter, and this remains true back until the epoch when the energy density in radiation became comparable with that in non-relativistic matter. Since $\rho_R \propto a^{-4}$ and $\rho_M \propto a^{-3}$ radiation-matter equality occurs when

$$1 + z > \frac{\Omega_{M0}}{\Omega_{R0}} = \frac{0.3}{8.4 \times 10^{-5}} = 3600.\tag{2.35}$$

This crossover would have occurred much later in the absence of Dark Matter, since the radiation energy density equals the energy density in baryons when

$$1 + z > \frac{\Omega_{B0}}{\Omega_{R0}} = \frac{0.04}{8.4 \times 10^{-5}} = 480.\tag{2.36}$$

Knowing how ρ depends on a immediately gives, with the Friedmann equation, H as a function of a , and so also an explicit form for the proper, luminosity and angular-diameter distances. For example, eq. (2.32) implies

$$H(a) = H_0 \left[\Omega_{DE0} + \Omega_{\kappa 0} \left(\frac{a_0}{a}\right)^2 + \Omega_{M0} \left(\frac{a_0}{a}\right)^3 + \Omega_{R0} \left(\frac{a_0}{a}\right)^4 \right]^{1/2},\tag{2.37}$$

where we define

$$\Omega_{\kappa 0} \equiv -\frac{\kappa}{(H_0 R_0 a_0)^2}.\tag{2.38}$$

Using $1 + z = a_0/a$ to eliminate a in favour of z then allows the present-day proper distance in such a universe to be written

$$D(z) = H_0^{-1} \int_0^z dz' \left[\Omega_{DE0} + \Omega_{\kappa 0}(1 + z')^2 + \Omega_{M0}(1 + z')^3 + \Omega_{R0}(1 + z')^4 \right]^{-1/2}, \quad (2.39)$$

with D_L and D_A being related to this by powers of $(1 + z)$ if $\kappa = 0$. It is clear from this expression how measurements of $D_L(z)$ or $D_A(z)$ for a range of z 's can allow an inference of the relative present-day density abundances, Ω_{i0} , for $i = DE, M, R$ and κ .

Given the dependence, eq. (2.37) of H on a , it is possible to integrate to obtain the t -dependence of a . Although in general this dependence must be obtained numerically, many of its features may be understood on simple analytic grounds based on the recognition that for most epochs there is only a single component of the cosmic fluid which is dominating the total energy density. We expect, then, that for redshifts larger than several thousand $a(t)$ should be well approximated by the expansion in a universe which is filled purely by radiation. Once a/a_0 rises to above $1/3600$ there should be a brief transition to the time dependence which describes the universal expansion in a universe dominated by non-relativistic matter. This should apply right up to the very recent past, when a/a_0 is around 0.8, after which there is a transition to vacuum-energy domination, during which the universal expansion accelerates to become exponential with t . In all likelihood we are at present still living in the transition period from matter to vacuum-energy domination.

Although the detailed relationship of a on t in principle depends on the value taken by κ , in practice the contribution of κ is only important in the very recent past. This is because the best present-day information indicates that $\Omega_0 = \Omega_{DE0} + \Omega_{m0} + \Omega_{r0} = 1$, which is consistent with $\kappa = 0$. But even if $\kappa \neq 0$, since the curvature term in eq. (2.3) varies like a^{-2} , it falls more slowly than does either the contribution of matter ($\rho_m \propto a^{-3}$) or radiation ($\rho_r \propto a^{-4}$). So given that the curvature term is at best only comparable to the other energy densities at present, it becomes more and more negligible the further one looks into the universe's past.

As a result it is a very good approximation to use $\kappa = 0$ in the expression for $a(t)$ during the matter-dominated and the earlier radiation-dominated epoch, in which case it has the very simple form $a(t) = a_0(t/t_0)^\alpha$, with $\alpha = \frac{1}{2}$ during radiation domination and $\alpha = \frac{2}{3}$ during matter domination. It may not be valid to neglect κ for the more recent periods of matter domination, and so in this case the more detailed expressions given in the previous section should instead be used. For the present-day epoch it is best to include both $\kappa \neq 0$ and $\rho_{DE} \neq 0$, although the best evidence remains consistent (within largish errors) with $\kappa = 0$.

3 Thermal Evolution of the Universe

The equations of state for radiation and non-relativistic matter used in the previous discussion are based on those which arise for radiation and atoms which are in thermal equilibrium, and for the case of CMB photons the photons can be seen explicitly to have a thermal distribution. This all points to matter being hot and dense at some point in the universe's past. As we shall see there is also other evidence that the matter in the universe was once as hot as 10^{10} K or more, at which time nuclei were once synthesized from a hot soup of protons, neutrons and electrons.

3.1 Big Bang Cosmology

The Big Bang theory of cosmology starts with the idea that the universe was once small and hot enough that it contained just a soup of elementary particles, in order to see if this leads to a later universe that we recognize in cosmological observations. This picture turns out to describe well many of the features we see around us, which are otherwise harder to understand. This section starts the discussion of the Big Bang theory by exploring the properties of a thermal bath of particles in an expanding universe, in order to understand the conditions under which equilibrium might be expected to hold, and to see what happens as such a bath cools as the universe expands.

3.1.1 The Known Particle Content

The starting point of any such description is a summary of the various types of elementary particles which are known, and their properties. These are well-known from decades of experimental and theoretical study over more than 40 years.

As mentioned earlier, the highest temperature there is direct observational evidence the universe has attained in the past is $T \sim 10^{10}$ K, which corresponds to thermal energies of order 1 MeV. The elementary particles which might be expected to be found within a soup having this temperature are the following.

- **Photons (γ):** are bosons and have no electric charge or mass, and can be singly emitted and absorbed by any electrically-charged particles.
- **Electrons and Positrons (e^\pm):** are fermions and have charge $\pm e$, where e denotes the proton charge,³ and their masses are the same size as one another, and equal numerically to $m_e = 0.511$ MeV. Because the positron, e^+ , is the

³Superscripts ' \pm ' or context should keep the use of e both as the symbol of the electron and to denote its charge.

antiparticle for the electron, e^- , (and vice versa), these particles can completely annihilate into photons through the reaction

$$e^+ + e^- \leftrightarrow 2\gamma. \quad (3.1)$$

- **Protons (p):** are fermions which have charge $+e$ and a mass $m_p = 938$ MeV. Unlike all of the other particles described here, the proton and neutron can take part in the *strong interactions*, which are what hold nuclei together. For example, this permits reactions like

$$p + n \leftrightarrow D + \gamma, \quad (3.2)$$

in which a proton and neutron combine to produce a *deuterium* nucleus, which is heavy Hydrogen and is a bound state of one proton and one neutron. The photon which appears in this expression simply carries off any excess energy which is released by the reaction.

- **Neutrons (n):** are fermions having no electric charge and a mass $m_n = 940$ MeV. Like protons, neutrons participate in the strong interactions. Isolated neutrons are unstable, and left to themselves decay through the *weak interactions* into a proton, an electron and an electron-antineutrino (see below).

$$n \rightarrow p + e^- + \bar{\nu}_e. \quad (3.3)$$

- **Neutrinos and Anti-neutrinos ($\nu_e, \bar{\nu}_e, \nu_\mu, \bar{\nu}_\mu, \nu_\tau, \bar{\nu}_\tau$):** are fermions which are electrically neutral, and have been found to have nonzero masses whose precise values are not known, but which are known to be smaller than 1 eV.
- **Gravitons (G):** are bosons which are not electrically charged and are massless. Gravitons are the quanta which mediate the gravitational force in the same way that photons do the same for the electromagnetic force. Gravitons only interact with other particles with gravitational strength, which is very weak compared to the strength of the other interactions. As a result they will turn out never to be in thermal equilibrium for any of the temperatures to which we have observational access in cosmology.

The next sections first ask how the temperature of a bath of particles would evolve on thermodynamic grounds as the universe expands, and then ask the same question in more detail by investigating the statistical mechanics of a thermal gas within an expanding universe.

3.2 Temperature Evolution - Thermodynamics

We have found (for several choices for the equation of state) how the energy density in different forms of matter varies with a as the universe expands, and we have seen how to find from this how a varies with time, t . The present section is devoted to learning how thermodynamics relates the temperature to a (and so also t), in order to quantify the rate with which a hot bath cools due to the universal expansion.

3.2.1 Relativistic Particles

The energy density and pressure appropriate to a gas of relativistic particles (like photons) when in thermal equilibrium at temperature T_R are given by

$$\rho_R = a_B T_R^4 \quad \text{and} \quad p_R = \frac{1}{3} a_B T_R^4, \quad (3.4)$$

where a_B is $g/2$ times the Stefan-Boltzmann constant, where g counts the number of internal (spin) states of the particles of interest (and so $g = 2$ for a gas of photons).

The evolution of T_R as the universe expands is simply determined by these expressions together with energy conservation, which for relativistic particles we have seen implies ρa^4 does not change as a increases. It is clear that these imply aT is constant, and so

$$T_R = T_{R0} \left(\frac{a_0}{a} \right) = T_{R0}(1+z). \quad (3.5)$$

Notice that this assumes only that $\rho_R \propto T_R^4 \propto a^{-4}$, and so (unlike the expression for a vs t) it does *not* assume that the total energy density is radiation-dominated. One way to see why this is so is to recognize that eq. (3.5) is equivalent to the statement that the expansion is *adiabatic*, since the entropy per unit volume of a relativistic gas is $s_R \propto T_R^3$, and so the total entropy in this gas is

$$S_R \propto s_R a^3 \propto (T_R a)^3 = \text{constant}. \quad (3.6)$$

3.2.2 Nonrelativistic Particles

An ideal gas of non-relativistic particles in thermal equilibrium has a pressure and energy density given by

$$p_M = n T_M \quad \text{and} \quad \rho_M = n m + \frac{n T_M}{\gamma - 1}, \quad (3.7)$$

where the first of these is the usual Ideal Gas Law, with n being the number of particles per unit volume.⁴ m is the particle's rest mass and $\gamma = c_p/c_v$ is its ratio of specific

⁴Units are used for which Boltzmann's constant is unity: $k_B = 1$.

heats, with $\gamma = 5/3$ for a gas of monatomic atoms. In the above expressions n is not regarded as being a function of temperature, with the total number of particles being conserved, which implies that

$$\frac{d}{dt} [n a^3] = 0. \quad (3.8)$$

Inserting the expressions into the equation of energy conservation, eq. (2.4), written in the form $d(\rho a^3)/dt + 3p a^2 da/dt = 0$, leads to the expression

$$\frac{\dot{T}_M}{T_M} + 3(\gamma - 1) \frac{\dot{a}}{a} = 0, \quad (3.9)$$

and so

$$T_M = T_{M0} \left(\frac{a_0}{a} \right)^{3(\gamma-1)} = T_{M0} (1+z)^{3(\gamma-1)}. \quad (3.10)$$

For example, for a monatomic gas with $\gamma = 5/3$ this implies $T_M \propto (1+z)^2$, or $T_M a^2 = \text{constant}$, as would be expected for an adiabatic expansion given that the entropy density for such a fluid varies with T_M like $s_M \propto (mT_M)^{3/2}$.

Once again, these expressions assume only that the non-relativistic matter does not exchange energy with any other components of the universe, and not that the matter dominates the fluid energy.

3.2.3 Multi-Component Fluids

We have seen that different components of the cosmic fluid cool with differing rates as the universe expands, as might be expected given that the different components have different equations of state. This happens due to the assumption that there is negligible energy exchange between these different components since this also implies that they cannot be in thermal equilibrium with one another. This leaves their respective temperatures free to evolve independently of one another.

But what happens when several components of the fluid *are* in thermal equilibrium with one another? After all, this situation actually happens in the early universe, with non-relativistic protons and neutrons (or nuclei) in equilibrium with relativistic photons, electrons and neutrinos. To see how this works, we now repeat the previous arguments for a fluid which consists of both relativistic and non-relativistic components, coexisting in mutual thermal equilibrium at a common temperature, T . In this case the energy density and pressure are given by $\rho = \rho_M + \rho_R$ and $p = p_M + p_R$, or

$$p = nT + \frac{1}{3} a_B T^4 \quad \text{and} \quad \rho = nm + \frac{nT}{\gamma - 1} + a_B T^4. \quad (3.11)$$

Inserting this into the energy conservation equation, as above, leads to the result

$$\frac{\dot{T}}{T} + \left[\frac{1 + \sigma}{\sigma + \frac{1}{3}(\gamma - 1)^{-1}} \right] \frac{\dot{a}}{a} = 0, \quad (3.12)$$

where

$$\sigma \equiv \frac{4a_B T^3}{3n} = 74.0 \left[\frac{(T/\text{deg})^3}{n/\text{cm}^{-3}} \right], \quad (3.13)$$

is the relativistic entropy per non-relativistic gas particle. For example, if the relativistic gas consists of photons, then the number of photons per unit volume is $n_\gamma = [30 \zeta(3)/\pi^4] a_B T^3 = 3.7 a_B T^3$, and so $\sigma = 0.37(n_\gamma/n)$.

Eq. (3.12) shows how T varies with a , and reduces to the pure radiation result, $T a = \text{constant}$, when $\sigma \gg 1$ and to the non-relativistic matter result, $T a^{3(\gamma-1)} = \text{constant}$, when $\sigma \ll 1$. In general, however, this equation has more complicated solutions because σ need not be a constant. Given that particle conservation implies $n \propto a^{-3}$, we see that the time-dependence of σ is given by $\sigma \propto (T a)^3$.

We are led to the following limiting behaviour. If, initially, $\sigma = \sigma_0 \gg 1$ then at early times $T \propto a^{-1}$ and so σ remains approximately constant (and large). For such a gas the common temperature of the relativistic and non-relativistic fluids continues to fall like $T \propto a^{-1}$. In this case the high-entropy relativistic fluid controls the temperature evolution and drags the non-relativistic temperature along with it. Interestingly, it can do so even if $\rho_M \approx n m$ is larger than $\rho_R = a_B T^4$, as can easily happen when $m \gg T$. In practice this happens until the two fluid components fall out of equilibrium with one another, after which their two temperatures continue to evolve separately according to the expressions given previously.

On the other hand if $\sigma = \sigma_0 \ll 1$ initially, then $T \propto a^{-3(\gamma-1)}$ and so $\sigma \propto a^{3(4-3\gamma)}$. This falls as a increases provided $\gamma > 4/3$, and grows otherwise. For instance, the particularly interesting case $\gamma = 5/3$ implies $T \propto a^{-2}$ and so $\sigma \propto a^{-3}$. We see that if $\gamma > 4/3$, then an initially small σ gets even smaller still as the universe expands, implying the temperature of both radiation and matter continues to fall like $T \propto a^{-3(\gamma-1)}$. If, however, $1 < \gamma < 4/3$, an initially small σ can grow even as the temperature falls, until the fluid eventually crosses over into the relativistic regime for which $T \propto a^{-1}$ and σ stops evolving.

3.3 Temperature Evolution - Statistical Mechanics

We now turn to a more microscopic picture of the universe's temperature evolution. The goal in forming this microscopic picture is to allow a more detailed understanding

about when thermal equilibrium should apply, and under what circumstances more complicated out-of-equilibrium physics might be important. Before exploring these issues we first review the thermal distributions of weakly-interacting particles, starting from first principles in order to ascertain how things might differ between relativistic applications and the non-relativistic ones usually encountered when studying statistical mechanics.

3.3.1 Equilibrium Distributions

The statistical description of a weakly-interacting gas of particles is determined by the distribution function, \mathcal{N} , which gives the average number of particles per unit spatial volume, d^3x , and per unit volume of momentum space, d^3p . This distribution is known very robustly when the particles are in thermal equilibrium, since it is then determined by detailed balance and so is largely independent of the microscopic details of the particles and the interactions through which equilibrium is maintained.

For example, if the equilibrating interactions have the form

$$A + B \leftrightarrow C + D, \quad (3.14)$$

where A , B , C and D represent distinct particle types, then the differential reaction rates per unit volume for running the reaction forward and backwards have the form

$$\begin{aligned} d\mathcal{R}(A + B \rightarrow C + D) &= \mathcal{N}_A \mathcal{N}_B (1 \pm \mathcal{N}_C) (1 \pm \mathcal{N}_D) d\sigma_{A+B \rightarrow C+D} v_{\text{rel}} \\ d\mathcal{R}(C + D \rightarrow A + B) &= \mathcal{N}_C \mathcal{N}_D (1 \pm \mathcal{N}_A) (1 \pm \mathcal{N}_B) d\sigma_{C+D \rightarrow A+B} v_{\text{rel}}, \end{aligned} \quad (3.15)$$

where v_{rel} denotes the relative velocity of the two incident particles. Here $d\sigma_{i \rightarrow f}$ denotes the differential cross section for the reaction $i \rightarrow f$, which is expressible in terms of an underlying scattering amplitude, $\langle f | \mathcal{S} | i \rangle$, by

$$d\sigma_{i \rightarrow f} \propto |\langle f | \mathcal{S} | i \rangle|^2 \prod_{k \in f} d^3 \mathbf{p}_k. \quad (3.16)$$

The two factors of \mathcal{N}_k appearing in eqs. (3.15) for the initial-state particles show how the reaction rate depends on the number of reactants which are available. The \mathcal{N} -dependence of the final-state factors, $1 \pm \mathcal{N}$, instead arises due to particle statistics, with the ‘+’ sign applying to bosons and describing stimulated emission and the ‘−’ sign applying to fermions and describing Pauli blocking. For dilute systems $\mathcal{N}_k \ll 1$ and these final-state factors reduce to unity.

The principle of detailed balance states that in equilibrium the particle distributions, \mathcal{N}_k , adjust themselves to ensure that the rate for the reaction $A + B \rightarrow C + D$

precisely equals the rate for its inverse $C + D \rightarrow A + B$, for all choices of momenta for the initial and final particle types. Given that the general principles — *i.e.* the unitarity of the scattering matrix, \mathcal{S} — imply that $d\sigma_{A+B \rightarrow C+D} = d\sigma_{C+D \rightarrow A+B}$, it follows that equality of these rates requires the equilibrium distributions must satisfy

$$\mathcal{N}_A \mathcal{N}_B (1 \pm \mathcal{N}_C)(1 \pm \mathcal{N}_D) = \mathcal{N}_C \mathcal{N}_D (1 \pm \mathcal{N}_A)(1 \pm \mathcal{N}_B), \quad (3.17)$$

or, equivalently

$$\prod_{k \in i} \left(\frac{1 \pm \mathcal{N}_k}{\mathcal{N}_k} \right) = \prod_{l \in f} \left(\frac{1 \pm \mathcal{N}_l}{\mathcal{N}_l} \right), \quad (3.18)$$

where the product over k is over all particle types in the initial state and the product over l is over all particle types in the final state.

The implication of condition (3.18) is most easily understood by taking its logarithm, in which case it states that the sum, $S = \sum_k \ln[(1 \pm \mathcal{N}_k)/\mathcal{N}_k]$, is *conserved* inasmuch as it takes the same value in equilibrium when summed over the initial and final particles in any microscopic collision. This can be possible only if S is a linear combination of the particle quantum numbers which are conserved during these collisions, such as energy E , momentum, \mathbf{p} , and any other conserved charges, Q^a (like electric charge, baryon number, and so on). That is, equilibrium requires

$$\ln \left(\frac{1 \pm \mathcal{N}_k}{\mathcal{N}_k} \right) = \beta_\mu P_k^\mu + \sum_a \xi_a q_k^a, \quad (3.19)$$

where $P_k^\mu = (\epsilon_k, \mathbf{p}_k)$ denotes the particle energy-momentum 4-vector and q_k^a denotes the value of the conserved charge Q^a for particle type ' k '. The different kinds of equilibrium which are possible are parameterized by the coefficients β_μ and ξ_a , and so these are the only intrinsic quantities on which any macroscopic physics (like thermodynamics) can depend in equilibrium.

Specializing to a fluid whose center of mass is not moving allows the choice $\beta_i = 0$, and solving for \mathcal{N}_k , gives the equilibrium distribution in its usual form⁵

$$\mathcal{N}_k(\mathbf{p}) = \frac{1}{e^{(\epsilon_k - \mu_k)/T} \mp 1}, \quad (3.20)$$

where now the $+$ sign applies for fermions and the $-$ sign for bosons. Here we write $\beta = \beta_t = 1/T$ and $\mu_k = \sum_a \mu_a q_k^a$, where $\xi_a = \mu_a/T$. An examination of the average change of energy with entropy and particle number exposes T as the thermodynamic

⁵Recall we use units for which $k_B = c = 1$.

temperature and μ_a as the chemical potential for the quantity, Q^a . For non-relativistic applications it is often true that the only conserved quantity of interest is simply particle number for each type of particle, $Q_a = N_a$, in which case there is a separate chemical potential for each particle type and $q_k^a = \delta_k^a$. As is described in more detail below, this choice is not possible for relativistic systems.

Since the Q^a are unchanged by microscopic interactions, the net value for Q^a taken for the whole bath is preserved by the scattering processes responsible for equilibrium. In this case the total charge density in the bath,

$$n^a = \sum_k q_k^a \int \frac{d^3p}{(2\pi)^3} \mathcal{N}_k(\mathbf{p}), \quad (3.21)$$

is constant and independent of T , and so may be specified as an externally chosen constraint on the bath. Here the sum is over all particle species in the bath. In equilibrium the chemical potentials simply adjusts itself as a function of temperature, $\mu^a = \mu^a(T)$, in order to ensure that n^a does not change.

A particle distribution is said to be *statistically degenerate* if $\epsilon_k - \mu_k \lesssim T$ because in this case the exponential does not dominate in the denominator of eq. (3.20), and so the difference between Bose and Fermi statistics becomes important. In the case of bosons the distribution function can then become singular if there are momenta for which $\epsilon_k = \mu_k$, since for these momenta the denominator passes through zero. This reflects the physical process of Bose-Einstein condensation. For fermions degeneracy instead reflects the formation of a *Fermi sea*, wherein particles fill the lowest energies available consistent with the Pauli exclusion principle.

For most cosmological applications the temperatures and chemical potential are such that the particles of interest are not degenerate. In this case the final-state densities drop out of the reaction rates, \mathcal{R} , and the equilibrium distributions reduce to those of Maxwell-Boltzmann statistics: $\mathcal{N}_k \approx e^{-(\epsilon_k - \mu_k)/T}$.

3.3.2 Statistical Mechanics in Special Relativity

There are two ways in which statistical mechanics with special relativity differs from statistical mechanics in a non-relativistic setting. The simplest difference arises through the dependence on momentum in eq. (3.20), which arises only through the particle energy, ϵ . Special relativity dictates this must be given by eq. (??):

$$\epsilon_k(p) = \sqrt{p^2 + m_k^2}, \quad (3.22)$$

where $p = |\mathbf{p}|$. Here m is the particle rest mass, defined by its energy at zero momentum, $\epsilon(0) = m$. A dependence of the energy on the other quantum numbers the particle carries can enter through the dependence of m on these other variables. Since statistical degeneracy requires $\epsilon_k(p) - \mu_k \lesssim T$, it can only happen when $\mu_k \gtrsim \epsilon_k(p) - T \geq m_k - T$, where m_k denotes the rest mass of particle type ‘ k ’. In particular, degeneracy is impossible for any value of momentum if it happens that $\delta\mu_k \equiv \mu_k - m_k \ll -T$, as is usually the case in the cosmological applications which follow.

The second main difference of relativistic statistical mechanics is the presence of antiparticles. It is a basic fact that special relativity and quantum mechanics together imply that every particle has an antiparticle whose statistics and mass are precisely the same as those of the original particle, but whose additive charges (like electric charge, or baryon number) are precisely opposite. If a particle carries no additive charges it can be its own antiparticle, such as is the case for the photon for example. It is also a fact that all interactions in a relativistic theory necessarily change the number of particles (given sufficient available energy), including in particular the possibility of having particles annihilate with their antiparticles into other degrees of freedom (or to be created by the inverse process).

The presence of antiparticles has several important implications. Any thermal bath whose temperature, T , is large compared with a particle rest mass, m , necessarily contains large numbers of particles and antiparticles and the existence of annihilation reactions precludes these from having separate chemical potentials since the very reactions which keep them in equilibrium can also change their number. In particular, if there are equal numbers of particles and antiparticles, then should the average thermal energy, T , fall below m the inability to pair-produce particles (due to there being insufficient energy) allows the annihilation reactions to predominate. These annihilations force the abundance of particles and antiparticles to become Boltzmann-suppressed by factors of order $\exp(-m/T)$, so long as the underlying reactions remain in equilibrium. In a cooling universe it can happen that the universe cools quickly enough that the last few particles cannot find antiparticles with which to annihilate, leaving a few relicts which (if they are stable) can survive into the present day. As we shall see, this may provide an explanation of the origins of the dark matter.

Since annihilation tends to remove particles from a thermal bath if $T > m$, something else must be going on to explain the presence of non-relativistic matter like baryons, which persist in the present-day cosmic soup without an appreciable abundance of antiparticles. Their presence indicates they carry a conserved quantum number (like baryon number) which does not change during particle scattering, and so which

can ensure an excess of particles over antiparticles if the total charge is nonzero in the cosmic thermal bath. If this is so then it corresponds to there being a nonzero chemical potential associated with this charge.

Since particles and antiparticles must carry opposite charges, $\bar{q}_k^a = -q_k^a$, the chemical potential enters oppositely into their distribution functions,

$$\mathcal{N}_k = \frac{1}{e^{(\epsilon_k - \mu_k)/T} \pm 1} \quad \text{and} \quad \bar{\mathcal{N}} = \frac{1}{e^{(\epsilon_k + \mu_k)/T} \pm 1} \quad (3.23)$$

where we use $\bar{\mu}_k \equiv \sum_k \mu_a \bar{q}_k^a = -\sum_a \mu_a q_k^a = -\mu_k$. It is the presence of such a chemical potential which can ensure a net excess of particles over antiparticles, and so thereby ensure the survival of a net number of particles at low temperatures, $T \ll m$, where annihilation can efficiently remove the antiparticles.

The average number of particles per unit volume and the average energy per unit volume of a gas of such particles is given by marginalizing over their unmeasured momenta, to give

$$\begin{aligned} n_k(T) &= \int \frac{d^3p}{(2\pi)^3} \frac{1}{e^{(\epsilon_k - \mu_k)/T} \pm 1} = T^3 \mathcal{C}_0^\pm \left(\frac{m_k}{T}, \frac{\mu_k}{T} \right) \\ \rho_k(T) &= \int \frac{d^3p}{(2\pi)^3} \frac{\epsilon}{e^{(\epsilon_k - \mu_k)/T} \pm 1} = T^4 \mathcal{C}_1^\pm \left(\frac{m_k}{T}, \frac{\mu_k}{T} \right), \end{aligned} \quad (3.24)$$

where the functions $\mathcal{C}_k^\pm(y, z)$ are given as functions of $y = m_k/T$ and $z = \mu_k/T$ by the integrals

$$\mathcal{C}_k^\pm(y, z) = \frac{1}{2\pi^2} \int_0^\infty dx \frac{x^2 (x^2 + y^2)^{k/2}}{e^{\sqrt{x^2 + y^2 - z}} \pm 1}. \quad (3.25)$$

Relativistic Particles with $\mu = 0$:

For the present purposes a relativistic particle is defined to be one for which $T \gg m$, and so for which the typical thermal energy, $E_{\text{th}} \sim T$, is much greater than the rest energy, m . In this case $y \approx 0$ and so $C_k^\pm(y) \approx c_k^\pm \equiv C_k^\pm(0)$. For such a gas eqs. (3.24) become

$$\begin{aligned} n(T) &= \int \frac{d^3p}{(2\pi)^3} \frac{1}{e^{p/T} \pm 1} = c_0^\pm T^3 \\ \rho(T) &= \int \frac{d^3p}{(2\pi)^3} \frac{p}{e^{p/T} \pm 1} = c_1^\pm T^4, \end{aligned} \quad (3.26)$$

with the constants c_k^\pm given by the integrals

$$c_k^\pm(y) = \frac{1}{2\pi^2} \int_0^\infty dx \frac{x^{k+2}}{e^x \pm 1}. \quad (3.27)$$

In particular $c_1^+ = \frac{7}{8} c_1^- = 7\pi^2/240$ and $c_0^+ = \frac{3}{4} c_0^- = (3/4\pi^2)\zeta(3)$, where $\zeta(z)$ is Riemann's zeta function. Numerically, since $\zeta(3) = 1.202056903\dots$ we have $c_0^+ = 0.0913$, $c_0^- = 0.1218$ while $c_1^+ = 0.2879$ and $c_1^- = 0.3290$. For N_b bosonic and N_f fermionic degrees of freedom satisfying $m, \mu \ll T$ we therefore have

$$\begin{aligned} n(T) &= \left[N_b + \frac{3}{4} N_f \right] (0.1218 T^3) \\ \rho(T) &= \left[N_b + \frac{7}{8} N_f \right] (0.3290 T^4). \end{aligned} \quad (3.28)$$

For example, for photons $N_b = 2$ (2 spin states) and $N_f = 0$ and so $n_\gamma = [2\zeta(3)/\pi^2] T^3 = 0.2436 T^3$ and $\rho_\gamma = [\pi^2/15] T^4 = 0.6580 T^4$.

These expressions show that so long as $\mu = 0$ both n and ρ are precisely the same for both particles and antiparticles, since these share exactly the same mass. For relativistic particles the density of both falls with decreasing T like powers of T , with $n \propto T^3$ and $\rho \propto T^4$, precisely as was used in previous sections for the equations of state for such particles.

Non-relativistic Particles with $\mu = 0$:

A non-relativistic particle similarly satisfies $T \ll m$, and so $y \gg 1$. Using $\epsilon \approx m + p^2/(2m)$ in this case gives

$$C_k^\pm(y) \approx \frac{e^{-y}}{2\pi^2} \int_0^\infty dx x^2 [y + x^2/2y]^k e^{-x^2/2y} = \left(\frac{y}{2\pi}\right)^{3/2} e^{-y} \times \begin{cases} 1 & \text{if } k = 0; \\ y + \frac{3}{2} & \text{if } k = 1 \end{cases}. \quad (3.29)$$

Using this (and $\mu = 0$) in eq. (3.24) then implies

$$n = \left(\frac{mT}{2\pi}\right)^{3/2} e^{-m/T} \quad \text{and} \quad \rho = n \left[m + \frac{3T}{2} \right], \quad (3.30)$$

which would vanish exponentially quickly at low temperatures if thermal equilibrium were to continue to be maintained. Physically, this exponential decline arises because when $\mu = 0$ there is no conservation law which prevents particles and their antiparticles from mutually annihilating once the temperature falls through $T \sim m$.

Notice also that these expressions do not agree with those used earlier for a non-relativistic gas, due to the exponential T -dependence of the particle density.

Relativistic Particles with $\mu \neq 0$:

Notice that the density of particles does not equal that of antiparticles if $\mu_a \neq 0$ for a conserved charge Q^a carried by both. In this case the net charge density becomes,

$$\begin{aligned} n^a &= \sum_k q_k^a T^3 \left[\mathcal{C}_0^\pm \left(\frac{m_k}{T}, \frac{\mu_k}{T} \right) - \mathcal{C}_0^\pm \left(\frac{m_k}{T}, -\frac{\mu_k}{T} \right) \right] \\ &\approx \sum_k 2q_k^a \mu_k T^2 \left[\frac{\partial \mathcal{C}_0^\pm}{\partial z}(y, z) \right]_{y=m_k/T, z=0} \left[1 + \mathcal{O} \left(\frac{\mu}{T} \right) \right]. \end{aligned} \quad (3.31)$$

where the approximate equality assumes $\mu_k \ll T$. In the relativistic limit $m_k \ll T$ the integral may be performed explicitly to give $(\partial \mathcal{C}_0^+ / \partial z)_{y=z=0} = \frac{1}{2} (\partial \mathcal{C}_0^- / \partial z)_{y=z=0} = 1/12$. In the presence of a known excess, n^a , for a conserved charge Q^a this equation is to be solved for the equilibrium choice, $\mu_a(T)$, which is required to produce the given n^a .

Non-relativistic Particles with $\mu \neq 0$:

In a cooling universe even a small excess of particles over antiparticles can survive to dominate at temperatures $T \ll m$, after annihilation takes place. The chemical potential required to maintain a residual particle abundance, n , obtained at low temperatures in this way is $\mu = m + \delta\mu$, with $\delta\mu \ll m$ required to satisfy

$$n_k = \frac{1}{2\pi^2} (m_k T)^{3/2} \int_0^\infty dx \frac{x^2}{e^{-\delta z_k + x^2/2} \pm 1}, \quad (3.32)$$

where $\delta z_k = \delta\mu_k/T$. For non-degenerate particles, for which $\epsilon_k - \mu_k \ll T$ (the case of practical interest for cosmology) this implies $\delta\mu_k$ is determined by

$$n_k = \frac{1}{2\pi^2} (m_k T)^{3/2} e^{\delta z_k} \int_0^\infty dx x^2 e^{-x^2/2} = \left(\frac{m_k T}{2\pi} \right)^{3/2} e^{\delta\mu_k/T}. \quad (3.33)$$

The equilibrium energy density tied up in these particles similarly becomes

$$\rho_k = n_k \left[m_k + \frac{3T}{2} \right], \quad (3.34)$$

which agrees with the results for a monatomic gas ($\gamma = 5/3$) used in earlier sections.

3.3.3 Statistical Mechanics in an Expanding Universe

How does this change in an expanding universe? The dependence on $a(t)$ enters through the generalization of the expression for particle energy as a function of momentum, $\epsilon(p)$, whose form is easily seen by repeating for an expanding universe the arguments which led to eq. (??). This is most easily done when this equation is written in the form $g_{\mu\nu} p^\mu p^\nu = -m^2$, where $p^\mu = m dx^\mu/d\tau$ defines the energy-momentum 4-vector.

In terms of the peculiar velocity, $v_{\text{pec}}^i = a(t) dx^i/dt$ we then have $p^i = m\gamma dx^i/dt = m\gamma v_{\text{pec}}^i/a(t)$, where $\gamma = dt/d\tau = [1 - \mathbf{v}_{\text{pec}}^2]^{-1/2}$, and so eq. (3.22) becomes

$$\epsilon^2 = (p^t)^2 = a^2(t) \mathbf{p}^2 + m^2. \quad (3.35)$$

Physically, the expansion of a particle's wavelength with the expansion of the universe implies its momentum varies with scale factor as $p^i \propto a^{-1}$. The dispersion relation of eq. (3.35) then follows because the energy and rest-mass do not change as the momentum scales in this way.

In the case of a relativistic particle eq. (3.35) becomes $\epsilon \approx a(t) p$, and the particle distribution function is well approximated by $\mathcal{N}(p) \approx [\exp[a(t) p/T - \mu/T] \pm 1]^{-1}$. This result is exactly what would be obtained by replacing the temperature, T , and chemical potential, μ , within the distribution function, eq. (3.20), with quantities which fall with the expansion of the universe according to $T_{\text{eff}}(t) = T[a_0/a(t)]$ and $\mu_{\text{eff}}(t) = \mu[a_0/a(t)]$.⁶ We see that the thermal distribution function for relativistic particles retains its detailed shape as the universe expands, provided that the temperature of the bath is taken to fall inversely with a . For instance, repeating the arguments for the average number of particles and the average energy per unit volume of such a gas of particles shows that eqs. (3.24) now become

$$n(T) = \mathcal{C}_0^\pm \left(0, \frac{\mu_{\text{eff}}}{T_{\text{eff}}}\right) T_{\text{eff}}^3 \quad \text{and} \quad \rho(T) = \mathcal{C}_1^\pm \left(0, \frac{\mu_{\text{eff}}}{T_{\text{eff}}}\right) T_{\text{eff}}^4, \quad (3.36)$$

with $T_{\text{eff}} \equiv T[a_0/a(t)]$.

Recall that this cooling law for T_{eff} vs a is precisely what was obtained in previous sections from the requirements of energy conservation for a relativistic gas. This shows that once a gas of relativistic particles acquires a thermal distribution, the expansion of the universe preserves this distribution, even if the gas itself were to be no longer in equilibrium. This will remain true until T_{eff} falls to the point that it becomes comparable with the particle mass, m .

3.4 Equilibrium and Decoupling

As we shall now see, a proper understanding of the universe's thermal history depends on an important way on precisely when thermal equilibrium is lost.

⁶Notice that the ratio $\mu_{\text{eff}}/T_{\text{eff}} = \mu/T$ is independent of t , and so the condition which this ratio must satisfy to ensure a nonzero number of particles, N , is not changed as the universe expands.

Thermal equilibrium requires there to be sufficient scattering amongst the particles involved to ensure that the total energy is equally distributed amongst the various degrees of freedom which share a common energy. The detailed conditions for equilibrium may be computed within a non-equilibrium formulation which follows the changes in the average abundance and energies of particles as a result of their mutual scattering. For example, if the number density of particles of type ‘ i ’ is denoted n_i , then the rate of change of the total number of such particles is given by

$$\frac{d}{dt} [n_i a^3] = \sum_j [\Gamma(j \rightarrow i) - \Gamma(i \rightarrow j)], \quad (3.37)$$

where $\Gamma(i \rightarrow j)$ represents the average rate (per particle) of scattering from particle type ‘ j ’ to type ‘ i ’, and the sum represent both a sum over particle types and over momenta. The two terms on the right-hand-side describe the contribution of reactions which produce type-‘ i ’ particles and those which remove them. What makes this equation complicated to solve in general is that these reaction rates themselves are functionals of the densities, n_k , of particle types involved.

Dividing this equation by a^3 gives the following Boltzmann-like equation which governs the evolution of \mathcal{N}_i :

$$\dot{n}_i + 3 \frac{\dot{a}}{a} n_i = \sum_j [\lambda(j \rightarrow i) - \lambda(i \rightarrow j)], \quad (3.38)$$

where $\lambda(i \rightarrow j) \equiv \Gamma(i \rightarrow j)/a^3$ are the reaction rates per unit spatial volume. The distribution in thermal equilibrium may be derived from the condition of detailed balance, which states that the distribution functions, $\mathcal{N}_i(\mathbf{p})$ must be chosen to ensure that the contributions on the right-hand side cancel for arbitrary momenta and particle types.

In the absence of the universal expansion this would be the end of the story, since eq. (3.38) then guarantees that $\sum_j [\Gamma(j \rightarrow i) - \Gamma(i \rightarrow j)] = 0$ also ensures self-consistently that these equilibrium solutions do not change in time. Things are different in an expanding universe, since in this case the vanishing of the right-hand-side of eq. (3.38) instead implies that n_i simply falls with the universal expansion according to $n_i \propto a^{-3}$. This expresses that the total number of such particles is conserved (on average) by microscopic scattering processes. This means that there are now potentially two contradictory ways to compute the time-evolution of a thermal distribution: (1) use eq. (3.38) with vanishing right-hand-side; or (2) compute n_i from $\mathcal{N}_i(\mathbf{p})$ as a function of T and μ , and use the time-dependence of T and μ using energy-conservation arguments along the lines of those used in previous sections.

For relativistic particles these two ways agree with one another because on one hand $n_i \propto a^{-3}$ and on the other hand $n_i \propto T^3$. These give consistent answers since $T \propto a^{-1}$ if the entropy is dominated by relativistic particles.

The two answers are different, however, for non-relativistic particles or when $\mu \gtrsim T$, since in this case using the a -dependence of T in the equilibrium result $n_i(T)$ does *not* give $n_i \propto a^{-3}$. In this case one of the two calculation methods must fail, and the one which does is the one which assumes \mathcal{N}_i takes the thermal-equilibrium form. This shows that the abundances for this type of relativistic particle typically fall out of equilibrium once the universal expansion cools the gas past the *freeze-out* temperature, T_f . Below T_f its abundance simply scales with the universal expansion as $n \propto a^{-3}$.

For example, if $\mu = 0$ an initially thermal equilibrium distribution fails once temperatures fall to $T_f \sim m$, after which the particle abundance becomes

$$n = n_f \left(\frac{a_f}{a} \right)^3 \quad \text{with} \quad n_f = n_{\text{th}}(T_f) = \left(\frac{mT_f}{2\pi} \right)^{3/2} e^{-m/T_f}, \quad (3.39)$$

where $a_f = a(T_f)$ and $n_{\text{th}}(T_f)$ denotes the appropriate thermal distribution at temperature T_f . Clearly this abundance is extremely sensitive to the precise value of T_f/m , which is determined by the more precise equilibrium conditions given below.

Because the relict abundance described by eq. (3.39) varies like $n \propto a^{-3}$ below the freeze-out temperature, it scales with a in precisely the same way as does a thermal gas of relativistic particles (like photons), for which $n_{\text{rel}} \propto T^3 \propto a^{-3}$. For this reason it is convenient to express these abundances relative to the abundance of photons, $\eta = n/n_\gamma$, since this is time-independent and the present-day relic photon abundance has been measured from the temperature of the cosmic microwave background, $T_\gamma = 2.7 \text{ K} \sim 10^{-4} \text{ eV}$.

3.4.1 Scattering Rate vs Expansion Rate

Relativistic particles can also fall out of equilibrium, and the above discussion leads to a very useful approximate criterion for understanding when a loss of equilibrium occurs as the universe cools. The criterion is built on the observation that it is the universal expansion which makes it impossible to satisfy eq. (3.38) using an equilibrium distribution. This shows that equilibrium will start to fail once the rates which contribute to the right-hand-side of this equation become comparable to the expansion rate, $H = \dot{a}/a$.

To see what this entails suppose that the interaction which is responsible for maintaining thermal equilibrium within the cosmic soup is a two-body reaction $A + B \rightarrow$

$C + D$, with a cross section σ . Given such a cross section, the rate per initial particle A , $\Gamma_A(A + B \rightarrow C + D)$, for 2-body particle scattering is obtained by averaging the cross section times the initial particle flux,

$$\Gamma_A(A + B \rightarrow C + D) \sim \langle n_B \sigma v_{\text{rel}} \rangle, \quad (3.40)$$

where n_B is the density of initial B particles involved, v_{rel} is the relative speed of the initial particles and $\langle \dots \rangle$ denotes the average over the ensemble of particles which are present. (A similar expression holds for the rate per particle B , Γ_B .) Equilibrium demands these rates must be much larger than H because if $\Gamma < H$ then the reactions occur too rarely to enforce the equipartition of energy on which thermal equilibrium relies. If $\Gamma > H$ then scattering can be efficient enough to maintain equilibrium.

For example, for a thermal gas of relativistic particles we would have $n \sim T^3$, $v_{\text{rel}} \sim 1$ and $E \sim T$ and so

$$\Gamma \sim \sigma(T) T^3. \quad (3.41)$$

As the universe expands and T falls this interaction rate typically gets smaller and smaller. It does so because the density of scattering particles falls comparatively quickly with T , making it harder and harder for particles to find one another to scatter. This either reinforces or overwhelms the dependence of the interaction cross section with energy, which shall see can either fall or rise as E decreases.

The temperature below which equilibrium fails can be obtained given the dependence of the total energy, ρ , on T because the Friedmann equation — $H^2 = 8\pi G\rho$ — turns this into an expression for $H(T) = [8\pi G \rho(T)]^{1/2}$. As we have seen, this rate also falls as T falls and the question is whether Γ falls faster than does H . Combining these results for their T dependence implies that the equilibrium condition, $\Gamma(T) > H(T)$, becomes

$$\sigma(T) T^3 > \frac{\rho^{1/2}(T)}{M_p}, \quad (3.42)$$

where $8\pi G = 1/M_p^2$ defines the *Planck mass*, $M_p \sim 10^{18}$ GeV. Depending on how strongly σ and ρ vary with T this implies that thermal equilibrium fails whenever T is above or below a critical temperature, T_{eq} , with T_{eq} obtained by replacing the inequality in eq. (3.42) with equality.

A more explicit determination of the temperatures for which equilibrium occurs can be made once the energy-dependence of the interaction cross sections are known. For the known interactions these cross sections typically vary as a power of the particle energy, at least in the energy ranges ($E \lesssim 100$ MeV) of interest. The resulting power

law has the form

$$\sigma(E) = \frac{\lambda^2}{M^2} \left(\frac{E}{M} \right)^s, \quad (3.43)$$

where M is some characteristic energy scale, λ is a dimensionless coupling factor and s is a characteristic power. We return to the values taken by the parameters M , λ and s for various interactions in the next section. Using this general power-law expression for the energy-dependence of the cross sections for a thermal gas of relativistic particles leads to the following reaction rate

$$\Gamma \sim \frac{\lambda^2 T^3}{M^2} \left(\frac{T}{M} \right)^s. \quad (3.44)$$

On the other hand, we have seen that the universe is radiation dominated, and so $\rho \propto T^4$, for $z \gtrsim 3600$. As we shall see, because relict photons are observed having temperatures of order 3K, redshifts this large turn out to correspond to temperatures $T \gtrsim 10^4$ K, or energies $T \gtrsim 1$ eV. The Friedmann equation in this case states

$$H_{\text{rad}} = (8\pi G\rho)^{1/2} \sim \frac{T^2}{M_p}, \quad (3.45)$$

for these temperatures, where $M_p = (8\pi G)^{-1/2} \sim 10^{18}$ GeV is the Planck mass introduced earlier.

Requiring $\Gamma > H_{\text{rad}}$ therefore leads to the following condition on T :

$$\left(\frac{T}{M} \right)^{s+1} > \frac{M}{\lambda^2 M_p}, \quad (3.46)$$

which shows that if $s > -1$ equilibrium fails for temperatures $T < T_{\text{eq}}$, where

$$T_{\text{eq}} \sim M \left(\frac{M}{\lambda^2 M_p} \right)^{1/(s+1)}. \quad (3.47)$$

On the other hand, if $s < -1$ then equilibrium fails for $T > T_{\text{eq}}$. Self-consistency of this calculation requires T_{eq} to be larger than the minimum temperature, $T \sim 1$ eV, above which radiation dominates (as was assumed when taking $H \propto T^2$).

3.4.2 Energy Dependence of Interactions

We now pause to list the values for the parameters M , λ and s which are relevant for the scattering of relativistic particles through the 4 known interactions. There are four known interactions in Nature, and for the energy range of interest these obey the form of eq. (3.43). For the scattering of relativistic particles the following properties hold (up to order-unity factors):

- **Electromagnetic Interactions:** $\lambda^2 \sim \alpha^2$ for photon scattering from charged particles, where $\alpha \sim 0.01$ denotes the fine-structure constant. Also $s = -2$ for relativistic processes, ensuring M drops out. For non-relativistic particles of mass m , we instead have $s = 0$ and $M = m$.⁷ This leads to the following energy dependence for the scattering cross section (neglecting logarithmic dependence on E)

$$\sigma_{\text{em}}(E) \sim \frac{\alpha^2}{E^2} \quad (\text{relativistic}) \quad \sigma_{\text{em}}(E) \sim \frac{\alpha^2}{m^2 v_{\text{rel}}} \quad (\text{non-relativistic}); \quad (3.48)$$

- **Low-Energy Strong Interactions ($E < \Lambda_s \sim 100$ MeV):** In the energy range of interest only protons and neutrons take part in the strong interactions, and both of these are non-relativistic. For energies and temperatures smaller than $\Lambda_s \sim 100$ MeV we have $\lambda^2 \sim 1$ and $s = 0$ while $M \sim \Lambda_s$, leading to;

$$\sigma_{\text{strong}}(E)v_{\text{rel}} \sim \frac{1}{\Lambda_s^2}; \quad (3.49)$$

- **Low-Energy Weak Interactions ($E < M_w \sim 100$ GeV):** $\lambda^2 \sim \alpha_w^2$ where $\alpha_w \sim 0.1$. For relativistic scattering it happens that $s = +2$, with $M = M_w$. For scattering involving non-relativistic particles of mass m we instead have $s = +1$ and $M^3 = M_w^4/m$, and so

$$\begin{aligned} \sigma_{\text{wk}}(E) &\sim \frac{\alpha_w^2 E^2}{M_w^4} \sim G_F^2 E^2 \quad (\text{relativistic}) \\ \sigma_{\text{wk}}(E) &\sim \frac{\alpha_w^2 m E}{M_w^4 v_{\text{rel}}} \sim \frac{G_F^2 m E}{v_{\text{rel}}} \quad (\text{non-relativistic}), \end{aligned} \quad (3.50)$$

where the *Fermi constant* is given by $G_F \sim \alpha_w/M_w^2 \sim 10^{-5}$ GeV⁻²;

- **Gravitational Interactions:** $\lambda^2 \sim 1$ and $s = +2$ with $M \sim M_p \sim 10^{18}$ GeV. The gravitational cross section may also be written

$$\sigma_{\text{grav}}(E) \sim \frac{E^2}{M_p^4} \sim (8\pi G E)^2, \quad (3.51)$$

where $8\pi G = 1/M_p^2$ relates M_p to Newton's constant, G . (Notice the very large value for M_p is a reflection of the very weak nature of the gravitational force relative to the other forces.)

⁷Since it is the interaction rate, Γ , rather than the cross section, σ , which is finite in the limit $v_{\text{rel}} \rightarrow 0$, it is the product σv_{rel} to which the dimensional estimates apply.

3.4.3 Some Decoupling Examples

It is instructive to use the above considerations to determine more explicitly the temperatures for which various kinds of particles fall out of equilibrium.

Gravitons

The only interaction experienced by gravitons is the (extremely weak) gravitational interaction. For gravity we have seen $\lambda \sim 1$, $M \sim M_p$ and $s = 2$ which leads to equilibrium for

$$T > T_{\text{eq}} \sim M_p \sim 10^{18} \text{ GeV}. \quad (3.52)$$

This shows that gravitational interactions amongst relativistic particles are never in equilibrium for the entire range of temperatures, $1 \text{ eV} < T \lesssim 100 \text{ MeV}$, for which the universe is radiation-dominated temperatures and which are of interest for the observational tests of Big Bang cosmology discussed later. Since gravitons experience only this interaction, they never need be in equilibrium with any other particles in this temperature range.

Neutrinos

Neutrinos take part in both gravitational and weak interactions, of which it is the weak interactions which are by far the strongest. Since neutrinos are relativistic particles for the weak interactions we have $\lambda \sim \alpha_w \sim 0.1$, $M \sim M_w \sim 100 \text{ GeV}$ and $s = +2$, leading to equilibrium when

$$T > T_{\text{eq}} \sim M_w \left(\frac{M_w}{\alpha_w^2 M_p} \right)^{1/3} \sim 2 \text{ MeV}. \quad (3.53)$$

This shows that for relativistic particles the weak interactions fall out of equilibrium for temperatures of a few MeV. Since these are the strongest interaction which neutrinos experience, they may be expected to thermally decouple from all other particles at this temperature.

Electromagnetic Interactions

In this case for relativistic particles we take $\lambda \sim \alpha \sim 0.01$ and $s = -2$ and so $\Gamma \sim \alpha^2 T$. This shows that electromagnetic interactions are in equilibrium for

$$T < T_{\text{eq}}^{\text{rel}} = \alpha^2 M_p \sim 10^{14} \text{ GeV}, \quad (3.54)$$

which holds throughout the entire radiation-dominated temperature range of interest. This shows that electrons and photons may be expected to remain in equilibrium with one another, at least up to the point where the electrons become non-relativistic.

For non-relativistic charged particles with mass $m \lesssim T$ the appropriate electromagnetic cross section is $\sigma v_{\text{rel}} \sim \alpha^2/m^2$, and so $\Gamma \sim \alpha^2 n/m^2 \sim \eta \alpha^2 T^3/m^2$, where $\eta = n/n_\gamma \sim n/T^3$. Using $\eta \sim 10^{-10}$ for both nucleons and electrons shows that such particles can be kept in equilibrium by electromagnetic processes for

$$T > T_{\text{eq}}^{\text{nonrel}} = \frac{m^2}{\eta_B \alpha^2 M_p} \sim 10^{-4} \text{ GeV} \left(\frac{m}{\text{GeV}} \right)^2 \sim \begin{cases} 0.1 \text{ MeV} & (\text{for protons}) \\ 0.03 \text{ eV} & (\text{for electrons}) \end{cases}. \quad (3.55)$$

Notice that equilibrium breaks down first between photons and protons, and only later does equilibrium fail between photons and electrons. As we shall see, for electrons other physics can intervene to decouple electromagnetic interactions before temperatures as low as $T \sim 0.03 \text{ eV} \sim 300 \text{ K}$ are reached, due to congregation of charged particles into electrically-neutral bound states (atoms) beforehand.

Strong Interactions

Only protons and neutrons interact strongly for the present purposes, and these interactions provide neutrons with a way of maintaining equilibrium with protons, electrons and photons, which are electromagnetically coupled. Since protons and neutrons are non-relativistic for the temperatures of interest, their interaction rate is $\Gamma \sim n \sigma v_{\text{rel}} \sim n/\Lambda_s^2$, where $\Lambda_s \sim 100 \text{ MeV}$. Relating n to the photon density, $n = \eta_B n_\gamma \sim \eta_B T^3$, the interaction rate becomes $\Gamma \sim \eta_B T^3/\Lambda_s^2$, and so equilibrium requires

$$T > T_{\text{eq}} \sim \frac{\Lambda_s^2}{\eta_B M_p} \sim 0.1 \text{ eV}. \quad (3.56)$$

Nucleons

Nucleons (*i.e.* protons and neutrons) are sufficiently massive that they are always non-relativistic for the temperatures of interest: $T < 100 \text{ MeV}$. The fact that the universe is nonetheless observed to have lots of nucleons in it (the baryons of the previous section) but none of their antiparticles shows that they have a net chemical potential, $\mu_B = m_B + \delta\mu_B$, where the excess, $\delta\mu_B \ll m_B$, ensures that the present-day nucleon abundance satisfies $n_{B0} \sim 5 \times 10^{-10} n_{\gamma 0} \sim 10^{-10} T_{\gamma 0}^3$. Here $T_{\gamma 0}$ is the present temperature of the observed cosmic microwave background photons: 2.7 K .

An estimate of how large $\delta\mu$ must be to do so may be found from eq. (3.33) for $n(T, \delta\mu)$, which gives

$$\frac{\delta\mu}{T} \approx \ln \left(\frac{n}{(2\pi m T)^{3/2}} \right) = \ln \left[\frac{n}{T^3} \left(\frac{T}{2\pi m} \right)^{3/2} \right]. \quad (3.57)$$

Since at present there is roughly 1 baryon per 10^{10} CMB photons, we have $\eta_{B0} \sim n_{B0}/T_{\gamma 0}^3 \sim 10^{-10}$. Using also $m_B \sim 1$ GeV and $T = T_{\gamma 0} = 2.7\text{ K} \sim 10^{-4}$ eV, we have $2\pi m_B/T_{\gamma 0} \sim 10^{14}$ leading to $\delta\mu_{B0}/T_{\gamma 0} \sim \ln(10^{-31}) \sim -71$, and so $|\delta\mu_{B0}/T_{\gamma 0}| \gg 1$. By contrast, when $T \sim 100$ MeV it is still true that $n_B/T^3 \sim 10^{-10}$, but $2\pi m_B/T \sim 100$ and so $\delta\mu_B/T \sim \ln(10^{-13}) \sim -30$.

So the density of baryons froze out at temperatures $T \sim 1$ GeV when baryons and anti-baryons annihilated, much higher than the energies of present interest. The relic density which survived is now preserved from further change by conservation of baryon number. However since neutrons and protons both carry baryon number $B = +1$, baryon number conservation does not determine the *relative* number of protons and neutrons in the cosmic soup, and the present-day relative abundance is also largely determined by the physics of freezing out,⁸ as is now described.

We have seen that the proton and neutron fluids share a common temperature because they are maintained in thermal equilibrium for temperatures greater than 0.1 MeV due to their strong interactions amongst themselves and because of their electromagnetic scattering from photons and electrons. The two fluids can also exchange particles with each other, due to the weak-interaction reaction

$$p + e^- \leftrightarrow n + \nu, \quad (3.58)$$

which we've also seen remains in equilibrium down to about 1 MeV. So long as this interaction remains in equilibrium the neutron and proton abundances are given by eq. (3.33),

$$n_i = \left(\frac{m_i T}{2\pi} \right)^{3/2} e^{\delta\mu_i/T}, \quad (3.59)$$

where $\delta\mu_p = \mu_B - m_p$ and $\delta\mu_n = \mu_B - m_n$, and protons and neutrons share the same baryon-number chemical potential, $\mu_n = \mu_p = \mu_B$, because they share the same baryon number: $B = +1$.

The relative abundance therefore is

$$\frac{n_n}{n_p} = \left(\frac{m_n}{m_p} \right)^{3/2} e^{-\Delta m/T}, \quad (3.60)$$

where $\Delta m = m_n - m_p \sim 1.3$ MeV. Since $\Delta m/m_n \sim 1 \times 10^{-3}$ we may take $m_n/m_p \sim 1$ and so $n_n \approx n_p \approx \frac{1}{2} n_B$ so long as $T \gg \Delta m$. For temperatures $T \lesssim \Delta m$, the neutron abundance starts to become exponentially suppressed as the reaction $p + e^- \rightarrow n + \nu$

⁸As we shall see, neutron decay also plays a role in the relic neutron abundance.

becomes rarer than the inverse reaction due to the heat bath thermal energy being insufficient to produce the energy needed to make neutrons from protons.

This suppression continues until the neutron/proton ratio freezes out, when reaction (3.58) falls out of equilibrium. For relativistic electrons we have $n_e \sim T^3$ and so the reaction rate per proton for the process $p + e^- \rightarrow n + \nu$ is $\Gamma_p \sim G_F^2 T^5$, which we've seen implies that equilibrium occurs (within a radiation-dominated universe) when $T \gtrsim 1$ MeV. Coincidentally this is numerically close to Δm . A more careful treatment shows that the freeze-out temperature is $T_f \sim 0.8$ MeV = 9×10^9 K, after which the ratio $n_n/n_p = 0.2$ is approximately constant, because both n_p and n_n scale with the universal expansion proportional to a^{-3} .

Strictly speaking, this ratio is only approximately constant because free neutrons are unstable, and decay into protons through the weak decay⁹

$$n \rightarrow p + e^- + \bar{\nu}. \quad (3.61)$$

Taking $m_n/m_p \approx 1$, this ensures that $n_n/n_B = \frac{1}{2} e^{-t/\tau_n}$ and $n_p/n_B = 1 - \frac{1}{2} e^{-t/\tau_n}$, where $\tau_n = 890$ sec denotes the neutron mean lifetime. This has a non-negligible effect because τ_n is comparable to the age of the universe when temperatures are of order $T \sim 1$ MeV. To see this, recall that in a radiation-dominated universe we have $H(t) = 1/(2t)$ and $H(T) = 8\pi G\rho/3 \sim 27T^4/M_p^2$. Combining these implies $t(T) = \frac{1}{2} H^{-1}(T) \sim 0.1 M_p/T^2$ and so — given $1 \text{ GeV}^{-1} = 10^{-24} \text{ sec}$ — when $T \sim 1$ MeV we have $t \sim 10^{23} \text{ GeV}^{-1} \sim 0.1$ sec. By contrast, by the time T falls to 0.1 MeV (or 0.01 MeV) the universe is $t \sim 10$ sec old (or 1000 sec old), during which time about 1% (or 77%) of the available neutrons decay.

Combining the neutron decay rate with the freeze-out result gives the modified expression

$$\frac{n_n}{n_p} \approx e^{-\Delta m/T_f} \left(\frac{e^{-t/\tau_n}}{2 - e^{-t/\tau_n}} \right). \quad (3.62)$$

This result is the starting point for the discussion of nucleosynthesis in a later section.

Electrons

For temperatures $T > m_e = 0.5$ MeV electrons are relativistic and are kept in equilibrium (see above) through their electromagnetic interactions with photons. In particular, reactions of the form $e^\pm + \gamma \leftrightarrow e^\pm + \gamma$ and $e^+ + e^- \leftrightarrow \gamma + \gamma$ ensure an abundance of

⁹Neutrons in nuclei are usually stable because the Coulomb interactions with other nuclear protons make it energetically too expensive for the decay to take place.

both electrons and positrons remain in equilibrium with each other and with photons in the thermal bath.

On the other hand, the overall electrical neutrality of the universe requires that $n_e - \bar{n}_e = n_p$, where n_e (\bar{n}_e) represents the electron (positron) particle density. So the presence of a chemical potential, μ_B , for baryon number (and the residual baryon density this implies) implies the necessity of there also being a nonzero chemical potential, μ_Q , for electric charge. Because there are two chemical potentials, the equilibrium distribution function for particle type ‘ i ’ becomes

$$\mathcal{N}_i = \frac{1}{e^{(\epsilon - \mu_i)/T} \pm 1}, \quad (3.63)$$

where $\mu_i = \mu_B b_i + \mu_Q q_i$, with b_i and q_i respectively being this particle’s eigenvalue for baryon number, B , and electric charge, Q . That is, for protons, neutrons, electrons and positrons we have

$$\mu_p = \mu_B + \mu_Q, \quad \mu_n = \mu_B, \quad \mu_e = -\mu_Q \quad \text{and} \quad \bar{\mu}_e = \mu_Q. \quad (3.64)$$

The conditions $n_B \neq 0$ and $n_e = n_p$ determine the two chemical potentials μ_B and μ_Q , and for $T \lesssim m_e \sim 0.5$ MeV both electrons and protons are non-relativistic and so $\mu_B \approx m_B$ and $\mu_Q \approx m_e \ll m_B$.

Given these choices it is possible to compute the residual abundance of positrons which survive once their abundance freezes out as T falls below the electron mass. Using eq. (3.33) we have the equilibrium relative abundance

$$\frac{\bar{n}_e}{n_e} = e^{(\mu_e - \bar{\mu}_e)/T} = e^{-2\mu_Q/T} \approx e^{-2m_e/T}, \quad (3.65)$$

which expresses the expected Boltzmann suppression for producing positrons through the creation of e^+e^- pairs. As a result the number density of positrons falls sharply once $T < m_e$, because below this temperature e^+e^- annihilation into photons cannot be compensated by the reverse reaction because on average the photons at these temperatures have too little energy.

Because the equilibrium positron density drops so dramatically with T , the equilibrating reaction rates inevitably become too small to keep \bar{n}_e in equilibrium. Since the reaction rate per positron is $\Gamma \sim n_e \sigma v_{\text{rel}} \sim \eta_B \alpha^2 T^3 / m_e^2$, the condition $\Gamma \gtrsim H$ implies $T \gtrsim T_f$ with $T_f / m_e \sim m_e / (\eta_B \alpha^2 M_p) \sim 5 \times 10^{-8}$, at which point $\bar{n}_e / n_e(T_f)$ is vanishingly small.

4 Cosmic Relics

What is spectacular about the study of cosmology now is the ability to test cosmological ideas with observations. Since these tests largely rely on detecting particles which persist to the present day as residual relics of the early universe, this section provides a brief history of the early universe with a focus on describing the various types of relics which arise.

4.1 A Thermal History of the Universe

In order to provide a context for the discussion of cosmic relics, it is worth first briefly stating a brief chronology of the main events which play a role in producing these relics. Our starting point is the epoch when the universe has a temperature of about 10 MeV, at which point it consists of a hot soup of non-relativistic protons and neutrons, in equilibrium with a population of relativistic electrons, positrons, photons and three species of neutrino.

We have seen that at these temperatures we have approximately equal numbers of protons and neutrons. Since all of the other particles satisfy $m \ll T$ at these temperatures, equipartition ensures that there are roughly equal numbers of electrons, positrons, photons and each species of neutrino. Furthermore, the relativistic particles are considerably more numerous, with $\eta_B = n_B/n_\gamma = (n_n + n_p)/n_\gamma \sim 10^{-10}$, although there is a slight excess of electrons over positrons so that $n_e - \bar{n}_e = n_p$ in order to ensure the electrical neutrality of the cosmic environment. This enormous excess of relativistic particles over non-relativistic ones ensures that the entropy of the equilibrium bath which they all share is dominated by the relativistic particles, and so the temperature of the bath falls like $T \propto a^{-1}$, as discussed above. The excess of relativistic matter over non-relativistic matter also ensures that the energy density is radiation-dominated, and so $\rho_{\text{tot}} \propto T^4 \propto a^{-4}$.

We now list a number of landmarks in the thermal history of the universe, which make an important impact on the relics we see today which are left over from this earlier and hotter time.

1. **Neutrino Freeze-out:** Once the temperatures fall below a few MeV, the weak interactions are not sufficiently strong to keep the 3 neutrino species in thermal equilibrium. After this point these neutrinos continue to run around the universe without scattering, and are still present during the present epoch as a *Cosmic Neutrino Background*. Since the neutrinos are relativistic, however, their

number density remains in its equilibrium form with the temperature simply redshifting, $T_\nu \propto a^{-1}$, as the universe expands. Since this is precisely the same time-dependence as for the thermal bath containing the rest of the particles, T_ν continues to track the temperature of the thermal bath as the universe expands. Although these neutrinos are in principle all around us, they have so far escaped detection due to their extremely small interaction cross sections.

2. **Electron-Positron Annihilation:** Once the temperature falls below twice the electron mass, $2m_e = 1.2$ MeV, the abundance of electrons and positrons begins to decline relative to photons due to the reaction $e^+e^- \rightarrow \gamma\gamma$ beginning to predominate over the inverse process of pair creation. As discussed above, this ends with the removal of essentially all of the positrons, leaving the same number of residual electrons as there are protons. This has an important consequence for the later universe, because this process of annihilation dumps a considerable amount of energy which reheats the equilibrium bath of photons, neutrons and charged particles relative to the neutrino temperature, which continues to redshift.
3. **Formation of Nuclei:** The thermal evolution at temperatures lower than 1 MeV is richer than would be believed from previous sections due to the possibility which arises of forming bound states. In particular, nuclear interactions can bind a neutron and proton into deuterium, with a binding energy of 2.22 MeV, and so once temperatures reach this energy range nuclei begin to form and so change the chemical composition of the cosmic fluid. The residual abundance of these nuclei predicted by this process agrees well with the observed primordial abundances, which provides strong evidence for the validity of the Big Bang picture of cosmology, and gives important information about the total abundance of baryons, η_B .
4. **Formation of Atoms:** Electromagnetic interactions furnish another important set of bound states which complicate the picture of the universe at lower temperatures. In particular, electrons can bind with nuclei to form neutral atoms once the temperature falls below the relevant binding energies, $T \sim 1$ eV. At this point the equilibrium conditions for charged particles and photons changes dramatically, since at this point the cosmic fluid becomes electrically neutral and so largely transparent to photons. The cosmic microwave background consists of those photons which last scattered from matter at this point, and have survived unscathed to be observed during the present epoch. The observation of these

photons gives a direct measure of the temperature of the heat bath from which the photons eventually decoupled.

The implications of these landmarks are now fleshed out in somewhat more detail.

4.2 Relict Neutrinos

Once neutrinos decouple they remain their temperature, T_ν , is free to evolve separately from the temperature, T , of the equilibrium bath. Although both temperatures continue to evolve together, with $T = T_\nu \propto a^{-1}$, the dumping of energy into the thermal bath by electron-positron annihilation has the effect of raising T while not changing T_ν . The amount of this reheating can be computed by keeping track of the entropy during the annihilation process, since this is conserved.

The key idea is that the entropy density of a thermal bath consisting of relativistic particles is $s \propto g_* T^3$ where T is the temperature and¹⁰ $g_* = N_b + \frac{7}{8} N_f$ counts the total number of relativistic bosonic (N_b) and fermionic (N_f) degrees of freedom which are in equilibrium. For instance $g_* = 2$ for a bath consisting only of photons because each photon has $N_b = 2$ corresponding to its two separate spin states. Since spin- $\frac{1}{2}$ particles also have two spin states each fermion also contributes $2(\frac{7}{8}) = \frac{7}{4}$ to the total value of g_* . If the energy release during electron-positron annihilation is adiabatic, then we know that the temperature increase which it causes may be found by equating the entropy density before and after the annihilation is complete

$$1 = \frac{s_{\text{before}}}{s_{\text{after}}} = \frac{g_{*\text{before}}}{g_{*\text{after}}} \left(\frac{T_{\text{before}}}{T_{\text{after}}} \right)^3. \quad (4.1)$$

The ratio $T_{\text{after}}/T_{\text{before}}$ may be read off from this expression if the ratio $g_{*\text{before}}/g_{*\text{after}}$ is known.

Before the electrons and positrons annihilate the total number of relativistic particles which are in equilibrium is $g_{*\text{before}} = 2 + (2 + 2) \frac{7}{8} = \frac{11}{2}$, corresponding to the contributions of photons, electrons and positrons. (The non-relativistic protons and neutrons do not contribute to this estimate because we have seen that their entropy is much smaller than that for the relativistic particles, because $\eta_B \sim 10^{-10} \ll 1$.) After e^+e^- annihilation we instead have $g_{*\text{after}} = 2$ consisting of photons only. (Again the small residual number of electrons contributes a negligible entropy in comparison.) We see from this that $g_{*\text{before}}/g_{*\text{after}} = (11/2)/2 = 11/4$ and so $T_{\text{before}}/T_{\text{after}} = (4/11)^{1/3}$.

¹⁰The factor of 7/8 comes from the difference between integrating over Bosonic and Fermionic distribution functions - *c.f.* the result $c_1^+ = \frac{7}{8} c_1^-$, derived in an earlier section.

The neutrino temperature is not similarly raised by this process because the neutrinos have dropped out of equilibrium by the time electrons and positrons annihilate, and are simply red-shifting along as the universe expands. So $T_\nu = T_{\text{before}}$ immediately after electron-positron annihilation is complete. Combining this with the above reheating calculation shows that electron-positron annihilation changes the neutrino-photon temperature ratio to

$$\frac{T_\nu}{T_\gamma} = \left(\frac{4}{11}\right)^{1/3} \approx 0.71. \quad (4.2)$$

After this time this temperature ratio remains unchanged, since both temperatures continue to redshift proportional to a^{-1} .

Given that the cosmic microwave background photons are now observed to have a temperature of $T_{\gamma 0} = 2.7$ K, it follows that the cosmic neutrino background should have a temperature $T_{\nu 0} = 1.9$ K, as was used in earlier sections. Although these neutrinos are in principle all around us, they have so far escaped detection due to their extremely small interaction cross sections.

These arguments assume the neutrinos remain relativistic right down to the present epoch, when $T_\nu \sim 10^{-4}$ eV. In fact, this is unlikely to be the case for at least some of the neutrinos since the recent detection of neutrino oscillations implies they cannot all be massless. Unfortunately, oscillations only measure neutrino mass *differences* rather than absolute masses, and the present evidence is that one pair of neutrinos has a squared-mass difference of $\Delta m_{\text{atm}}^2 = 3 \times 10^{-3} \text{ eV}^2$, and another pair has a difference $\Delta m_{\text{solar}}^2 = 5 \times 10^{-5} \text{ eV}^2$. Even assuming the lightest neutrino is massless, this implies the heaviest neutrino cannot be lighter than $(\Delta m_{\text{atm}}^2)^{1/2} = 0.05$ eV. At the very least this neutrino should have fairly recently become non-relativistic and so no longer have the standard relativistic equilibrium distribution. This is motivating the study as to whether such a massive relict neutrino can have measurably different cosmological effects than do massless relicts.

4.3 Nucleosynthesis

The formation of nuclei at temperatures near $T = 1$ MeV defines the epoch of *Big Bang Nucleosynthesis* (BBN), and represents the first epoch for which direct evidence exists that the universe was once very small and very hot.

Because there are relatively few protons and neutrons in the cosmic soup ($\eta_B \sim 10^{-10}$) their collisions are relatively rare, making the formation of nuclei relatively inefficient. Since two-body collisions are more probable than three- (or higher-) body ones in a dilute fluid, it is two-body reactions which dominate the formation of nuclei

from the hot proton-neutron gas. The two body reactions which can form nuclei directly from protons and neutrons are

$$p + p \rightarrow D + e^+ + \nu, \quad n + n \rightarrow D + e^- + \bar{\nu} \quad p + n \rightarrow D + \gamma, \quad (4.3)$$

where $D = {}^2H$ denotes the deuterium nucleus, which is a bound state of a proton and a neutron whose binding energy is 2.2 MeV. Since the first two of these reactions require the conversion of a proton into a neutron (or vice versa) they require the weak interactions in addition to the strong interactions and so proceed with cross sections which are much smaller than the $p - n$ collision process.

Once deuterium forms more two-body reactions become possible, such as

$$D + p \leftrightarrow {}^3He + \gamma \quad \text{and} \quad D + n \leftrightarrow {}^3H + \gamma, \quad (4.4)$$

and once sufficient deuterium accumulates even more possibilities arise, including

$$D + D \leftrightarrow {}^4He + \gamma \quad D + D \leftrightarrow {}^3H + p \quad \text{and} \quad D + D \leftrightarrow {}^3He + n. \quad (4.5)$$

These reaction products do not accumulate because they can also react with particles in the bath to produce 4He through the strong-interaction reactions

$${}^3H + p \leftrightarrow {}^4He + \gamma \quad {}^3He + n \leftrightarrow {}^4He + \gamma \quad {}^3H + D \leftrightarrow {}^4He + n \quad {}^3He + D \leftrightarrow {}^4He + p. \quad (4.6)$$

Helium-4 proves to be something of a bottleneck, however, for two reasons. First of all 4He is a particularly strongly-bound nucleus, with a total binding energy of 28.3 MeV (or 7.1 MeV per nucleon). Second, there are no stable nuclei involving 5 nucleons (both 5He and 5Li are unstable). Further progress up to heavier nuclei therefore requires collisions with the relatively rare D , 3H and 3He nuclei, such as through the reactions

$${}^4He + D \leftrightarrow {}^6Li + \gamma \quad {}^4He + {}^3H \leftrightarrow {}^7Li + \gamma \quad \text{or} \quad {}^4He + {}^3He \leftrightarrow {}^7Be + \gamma. \quad (4.7)$$

Beyond this point yet another bottleneck arises due to the absence of stable nuclei containing 8 nucleons, precluding the productions of still heavier nuclei. These heavier elements must wait to get formed within stars during later epochs of the universe.

Consequently once deuterium starts to form, essentially all of the neutrons which are available get eventually cooked into 4He . The fractional abundance by mass (compared to the total number of nucleons), Y_p , of Helium-4 which is produced is therefore simply related to the neutron-to-proton ratio at the time of Deuterium formation.

Keeping in mind that each ${}^4\text{He}$ nucleus consists of 2 neutrons and 2 protons it follows that N neutrons must combine with N protons to form $N_{\text{He4}} = \frac{1}{2} N$ Helium-4 nuclei, each of which is 4 times heavier than a proton or neutron. We therefore have

$$Y_p \equiv \frac{\rho_{\text{He4}}}{\rho_B} \approx \frac{4 n_{\text{He4}}}{n_p + n_n} = \frac{2 n_n}{n_p + n_n} = \frac{2f}{1+f}, \quad (4.8)$$

where $f \equiv n_n/n_p$ at the time of deuterium nucleosynthesis.

As we have seen, the neutron/proton abundance freezes out of equilibrium to the value $f \approx 1/5$ when $T = T_f \approx 0.8$ MeV. If this ratio were frozen in time then we would have $Y_p = 1/3$, however we shall see that neutron decays further deplete the neutrons, and so lower f , before deuterium synthesis occurs.

Determining f requires the neutron/proton ratio at the precise instant when deuterium nucleosynthesis occurs, and it is impossible to be overly precise about this time because the nuclear reactions play themselves out over an interval of time. On the other hand, since the reactions which cook Helium-4 occur relatively quickly once deuterium starts to form, a reasonable estimate can approximate all deuterium formation as taking place at a particular time, which may be taken to be the time when the deuterium and neutron abundances are equal: $n_D = n_n$. An estimate of when this occurs may be found by computing these particle densities using equilibrium abundances. During equilibrium, the abundance of deuterium may be computed using the non-relativistic density distribution

$$n_D = g_D \left(\frac{m_D T}{2\pi} \right)^{3/2} \exp[\delta\mu_D/T], \quad (4.9)$$

where $g_D = 3$ counts the 3 spin states of the deuteron. Here $\delta\mu_D = \mu_D - m_D$, and the deuterium chemical potential is given in the usual way in terms of its baryon number ($B = 2$) and electric charge ($Q = 1$): $\mu_D = 2\mu_B + \mu_Q$. Taking the ratio of this result to the corresponding result for protons and neutrons implies

$$\begin{aligned} \frac{n_D}{n_p n_n} &= \frac{g_D}{g_p g_n} \left(\frac{2\pi m_D}{m_p m_n T} \right)^{3/2} \exp[(\delta\mu_D - \delta\mu_p - \delta\mu_n)/T] \\ &= \frac{g_D}{g_p g_n} \left(\frac{2\pi m_D}{m_p m_n T} \right)^{3/2} \exp[B_D/T] \end{aligned} \quad (4.10)$$

$$= 6 \left(\frac{\pi}{m_B T} \right)^{3/2} \exp[B_D/T] \quad (4.11)$$

where $g_p = g_n = 2$ and the second equality uses $\delta\mu_p = \mu_p - m_p$ and $\delta\mu_n = \mu_n - m_n$ where the nucleon baryonic and electric charges dictate that $\mu_p = \mu_B + \mu_Q$ and $\mu_n = \mu_B$.

Finally, the last equality uses $m_D \approx 2m_n \approx 2m_p \approx 2m_B$ in the pre-factor which multiplies the exponential. B_D denotes the deuteron binding energy, $B_D \equiv m_p + m_n - m_D = 2.22$ MeV.

Using $n_p \approx (1 - f)n_B = (1 - f)\eta_B n_\gamma \approx 0.8 \eta_B (0.243 T^3)$, we find

$$\frac{n_D}{n_n} \approx 6.5 \eta_B \left(\frac{T}{m_B} \right)^{3/2} \exp[B_D/T], \quad (4.12)$$

which shows how the deuterium abundance rises exponentially relative to that of neutrons as the universe cools below the deuterium binding energy. This abundance change happens for the usual reasons: below these temperatures the heat bath does not provide sufficient energy on average to dissociate deuterium into its constituents during collisions. The nucleosynthesis temperature is obtained from eq. (4.12) by setting $n_D/n_n = 1$. This leads to $T_{\text{nuc}} = 0.066$ MeV $= 7.6 \times 10^8$ K, which is quite low compared with the binding energy, $B_D = 2.22$ MeV, because of the very small number of baryons available per photon, $\eta_B = 5 \times 10^{-10}$.

Because deuterium is formed at such a low temperature, there is a comparatively long time available during which the free neutrons can decay. In a radiation-dominated universe a temperature of 0.066 MeV is reached when $t \approx 200$ sec, at which point $\exp[-t/\tau_n] \approx \exp[-200/890] \approx 0.8$ of the original neutrons have not decayed. This lowers the neutron/proton ratio from $f_0 = 1/5 = 0.2$ to $f_{\text{nuc}} = 0.8/5.2 = 0.15$, leading in turn to a Helium mass fraction of $Y_p = 0.27$. More careful calculations which keep track of deviations from equilibrium and which follow the whole reaction chain forming heavier nuclei lead instead to the result $Y_p = 0.24$.

Notice that these numerical results are fairly sensitive to the total nucleon abundance, $\eta_B = n_B/n_\gamma$, since this controls the likelihood of two nucleons finding one another to collide and react. The sensitivity to η_B is even stronger in the abundances of the trace nuclei, like D , 3H , 3He and 7Li because the production of these nuclei involves the collision of secondary reaction products. A remarkable feature of the detailed calculations is that the same value of η_B is required to obtain agreement with *all* of the observed primordial nuclear abundances — *i.e.* for the abundances of all of the nuclei discussed above — demonstrating the consistency of the overall picture. This is one of the ways that the baryon/photon ratio, η_B , is determined observationally.

4.4 The Cosmic Microwave Background

Photons are abundant in the early universe, and we saw in earlier sections that they would be quite efficiently kept in thermal equilibrium with a bath of electrons right

down to temperatures $T \sim 0.03 \text{ eV} \sim 300 \text{ K}$, and so to redshifts $z \sim 100$. But it turns out that this picture breaks down because the electron bath itself does not survive down to such low temperatures since electrons instead first combine with protons and other nuclei to form electrically-neutral atoms. Since photons scatter much less efficiently from neutral atoms than from charged free electrons this changes the precise epoch when photons decouple from equilibrium.

4.4.1 Recombination

The formation of atoms occurs once the temperature falls below the atom's electronic binding energy, since then collisions are typically not energetic enough to dissociate atoms once they form. Neglecting for simplicity the Helium-4 content of the universe allows this binding process to be understood purely in terms of the formation of neutral Hydrogen from protons and electrons, through the reaction



The equilibrium abundance of H which is produced in this way may be understood along the same lines as was done in the previous section for the abundance of deuterium nuclei, starting from the equilibrium result

$$n_H = g_H \left(\frac{m_H T}{2\pi} \right)^{3/2} \exp[\delta\mu_H]. \quad (4.14)$$

Here $g_H = 4$ counts the Hydrogen spin states, and $\delta\mu_H \equiv \mu_H - m_H$ where the baryon and electric charges of the hydrogen atom ($B = +1$ and $Q = 0$) imply its chemical potential is $\mu_H = \mu_B$. The relative abundance of H to free protons and electrons then is given by the *Saha* equation

$$\begin{aligned} \frac{n_H}{n_p n_e} &= \frac{g_H}{g_p g_e} \left(\frac{2\pi m_H}{m_p m_e T} \right)^{3/2} \exp[B_H/T] \\ &\approx \left(\frac{2\pi}{m_e T} \right)^{3/2} \exp[B_H/T], \end{aligned} \quad (4.15)$$

where $B_H = \delta\mu_H - \delta\mu_p - \delta\mu_e = m_H - m_p - m_e = 13.6 \text{ eV}$ is the binding energy for neutral Hydrogen, and the pre-exponential factor is simplified using $g_H = g_p g_e$ and $m_H \approx m_p$.

Multiplying eq. (4.15) through by n_e and using

$$n_e = n_p \approx X n_B = X \eta_B (0.243 T^3), \quad (4.16)$$

where the ionization fraction, X , is defined by $X = n_p/(n_p + n_H)$, leads to the following expression relating X and T

$$\frac{1 - X}{X^2} = 3.84 \eta_B \left(\frac{T}{m_e} \right)^{3/2} \exp[B_H/T] \equiv S(\eta_B, T). \quad (4.17)$$

Equivalently, solving for X gives $X = [-1 + (1 + 4S)^{1/2}]/(2S)$. These equations allow a more precise determination of the temperature where neutral Hydrogen forms. Defining the *recombination* temperature, T_{rec} , by the condition $X_{\text{rec}} = \frac{1}{2}$ leads to $T_{\text{rec}} = 0.323 \text{ eV} \sim 3740 \text{ K}$, which corresponds to a redshift $z_{\text{rec}} = 1370$. The exponential temperature dependence makes the development of ionization fairly rapid: X falls from 0.9 to 0.1 between redshifts $z = 1475$ and $z = 1255$.

4.4.2 Photon Decoupling

Given the rapid loss of ionization, we may now recompute the temperature below which photons drop out of equilibrium. This occurs once the rate (per photon) for photon-electron scattering, $e^- \gamma \leftrightarrow e^- \gamma$,

$$\Gamma_\gamma \sim n_e \sigma v_{\text{rel}} \sim X \eta_B T^3 \left(\frac{\alpha^2}{m_e^2} \right) = \alpha^2 X \eta_B (1 + z)^3 \frac{T_{\gamma 0}^3}{m_e^2}, \quad (4.18)$$

drops below the Hubble rate H . Here $T_{\gamma 0} = 2.7 \text{ K} = 3 \times 10^{-4} \text{ eV}$ denotes the present-day CMB photon temperature. Since z is smaller than 3600 it is the matter-dominated form for H which is appropriate, $(H/H_0)^2 = \Omega_{m0}(1+z)^3$, and so the condition $\Gamma_\gamma = H$ leads to the condition

$$1 + z_{\text{dec}} = \frac{43}{X(z_{\text{dec}})^{2/3}}. \quad (4.19)$$

Using $H_0 = 70 \text{ km/sec/Mpc}$ and the expression for $X(T)$ (and so also $X(z)$) given above, this leads to $z_{\text{dec}} = 1130$. A more accurate treatment gives a somewhat smaller value, $z_{\text{dec}} = 1100$, because the abundance of ionized protons persists longer than would be indicated by the Saha equation, eq. (4.15), because the reaction $p + e^- \leftrightarrow H + \gamma$ also begins to drop out of equilibrium as decoupling takes place.

4.4.3 Last Scattering

Since relic photons have been detected, there is one further important transition epoch which must be understood. This is the epoch of last scattering, during which the relic photons scattered for the last time from the ambient cosmic matter. *A priori* this need not be precisely the same time as the time for decoupling (which is when electron-photon scattering fell out of equilibrium) because it can happen that photons continue

to scatter, but do so too infrequently to equilibrate the photon temperature with that of the electrons. That this can be possible is evident from everyday experience, for instance when light passes through glass or water (and is reflected or refracted and so undergoes scattering) but in so doing does not equilibrate with the temperature of the glass or water.

This section therefore studies the emission and absorption of light by a gas of atoms, without assuming the atoms and light are in equilibrium. The purpose is to identify precisely what the epoch of last scattering is, and it is shown to occur very close to the decoupling time. Besides having applications to the universe immediately after photon decoupling, these kinds of interactions are also of interest because they can also occur in the relatively recent universe, such as for redshifts of order 1-20, because of the re-ionization of interstellar matter due to stars and other energetic processes.

The treatment here follows that of ref. [6]. Our interest is in the number of photons, \mathcal{N}_γ , per unit angular frequency interval, $d\omega$, per unit volume which have the present-day angular frequency ω_0 . The change with time in this density arising as the photons pass through a medium in an expanding universe is given by

$$\frac{d}{dt} \left[\mathcal{N}_\gamma(\omega_0, t) a^3(t) \right] = R_{\text{sp}} + R_{\text{st}} - R_{\text{ab}}, \quad (4.20)$$

where the three terms on the right-hand side respectively describe the rates for *spontaneous emission*, *stimulated emission* and *absorption*. The absorption rate may be expressed as

$$R_{\text{ab}} = \Lambda[\omega(t), t] \mathcal{N}_\gamma(\omega_0, t), \quad (4.21)$$

where $\omega(t) = \omega_0 [a_0/a(t)]$ and $\Lambda(\omega, t)$ denotes the absorption rate of photons of angular frequency ω at time t , per photon per unit proper volume. The stimulated-emission rate is similarly given by

$$R_{\text{st}} = \Omega[\omega(t), t] \mathcal{N}_\gamma(\omega_0, t), \quad (4.22)$$

where $\Omega(\omega, t)$ is the emission rate of photons having angular frequency ω at time t , per photon per unit proper volume. Finally, the rate for spontaneous emission is given by

$$R_{\text{sp}} = \Gamma[\omega(t), t] a^3(t) \left[\frac{a_0}{a(t)} \right], \quad (4.23)$$

where $\Gamma(\omega, t) = (\omega/\pi)^2 \Omega(\omega, t)$ is the emission rate of photons having angular frequency ω at time t per unit proper volume *per unit angular frequency interval* (not per photon). The last factor of eq. (4.23) expresses the cosmological redshift from the observed frequency interval, $d\omega_0$, to the one relevant at time t : $d\omega = d\omega_0 [a_0/a(t)]$.

The solution to eq. (4.20) is

$$\mathcal{N}_\gamma(\omega_0, t) a^3(t) = e^{-\tau(t, t_1)} \mathcal{N}_\gamma(\omega_0, t_1) a^3(t') + \left(\frac{\omega_0}{\pi}\right)^2 a_0^3 \int_{t_1}^t dt' e^{-\tau(t, t')} \Omega[\omega(t'), t'], \quad (4.24)$$

where

$$\tau(t, t') \equiv \int_{t'}^t dt'' \left\{ \Lambda[\omega(t''), t''] - \Omega[\omega(t''), t''] \right\}, \quad (4.25)$$

is called the medium's *optical depth* and t_1 is an arbitrary time. Physically, the first term in eq. (4.24) describes the contributions to $\mathcal{N}_\gamma(\omega_0, t)$ from photons which were already present before the time t_1 and the second term describes effects occurring at times later than t_1 . The transfer function, $P(t, t') = e^{-\tau(t, t')}$, describes the effects of absorption and stimulated emission by the medium when passing from t' to t .

For applications to CMB photons we may take $t_1 \approx 0$ — *i.e.* well before recombination so that the universe is opaque — and so $P(t, t_1) = 0$, allowing the neglect of the first term in eq. (4.24). We may also use that the cosmic fluid through which the photons move is in equilibrium even if the photons are not, in which case detailed balance requires

$$\Omega(\omega, t) = \Lambda(\omega, t) e^{-\omega/T_m(t)}, \quad (4.26)$$

where $T_m(t)$ is the medium's temperature dependence (which varies with t as the universe expands). With these assumptions we have

$$\mathcal{N}_\gamma(\omega_0, t) a^3(t) = \left(\frac{\omega_0}{\pi}\right)^2 a_0^3 \int_0^t dt' e^{-\tau(t, t')} e^{-\omega(t')/T_m(t')} \Lambda[\omega(t'), t'], \quad (4.27)$$

where

$$\tau(t, t') \equiv \int_{t'}^t dt'' \left\{ 1 - e^{-\omega(t'')/T_m(t'')} \right\} \Lambda[\omega(t''), t'']. \quad (4.28)$$

The physical implications of these expressions are most easily seen if eq. (4.27) is re-written in the form

$$\mathcal{N}_\gamma(\omega_0, t) a^3(t) = \left(\frac{\omega_0}{\pi}\right)^2 a_0^3 \int_0^t dt' \frac{1}{e^{\omega(t')/T_m(t')} - 1} \left[\frac{d}{dt'} e^{-\tau(t, t')} \right]. \quad (4.29)$$

This is a expression because the emission and absorption rates for photons drop sharply as the protons and electrons combine into neutral atoms, in a way which parallels the similar drop in the reactions which keep the photons in thermal equilibrium. As a result the function $e^{-\tau(t, t')}$ is effectively a step function which vanishes for times earlier

than the recombination time, t_{dec} , and is close to unity thereafter. When this is so its derivative is a delta-function and so eq. (4.29) further simplifies to

$$\mathcal{N}_\gamma(\omega_0, t) a^3(t) = \left(\frac{\omega_0}{\pi}\right)^2 a_0^3 \left[\frac{1}{e^{\omega(t_{\text{dec}})/T_m(t_{\text{dec}})} - 1} \right]. \quad (4.30)$$

Under these circumstances the last-scattering and decoupling times are effectively equal, $t_{\text{ls}} \approx t_{\text{dec}}$, at redshift $z_{\text{ls}} \approx z_{\text{dec}} = 1100$. (More detailed calculations reveal that $e^{-\tau}$ falls to zero over a small range of redshifts, whose influence can leave a small imprint on the detailed temperature distribution which is observed at late times.) The photons thenceforth retain a thermal distribution (provided $e^{-\tau}$ remains small) governed by the *matter* temperature, $T_m(t_{\text{rec}})$. The late-time photon distribution thereby provides a snapshot of the matter temperature at the epoch of last scattering.

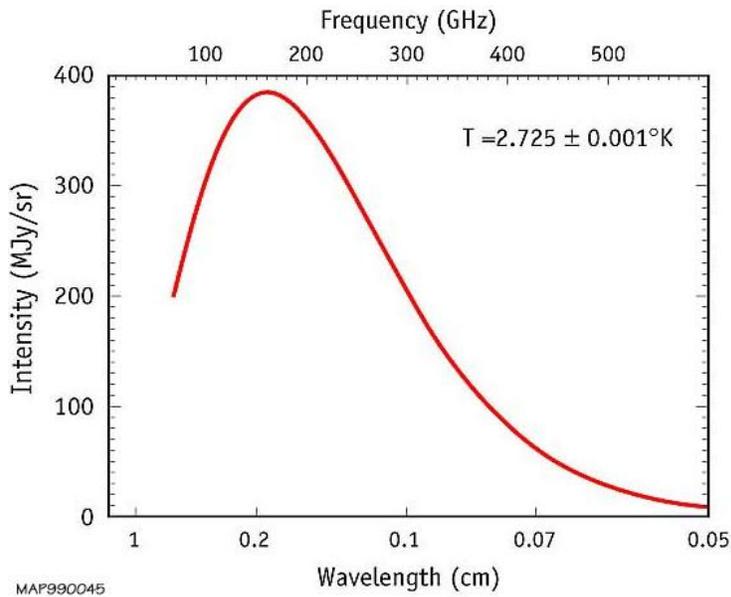


Figure 1. The FIRAS measurement of the thermal distribution of the CMB photons. The experimental points lie on the theoretical curve, with errors which are smaller than the width of the curve.

These photons have been observed, and their distribution has a beautiful thermal form as a function of the present-day photon angular frequency, ω_0 , as shown in Fig. 1. The temperature of this distribution has been measured as a function of direction in the sky, $T_\gamma(\theta, \phi)$, and it is the angular average of this measured temperature,

$$T_{\gamma 0} = \langle T_\gamma \rangle = \frac{1}{4\pi} \int T_\gamma(\theta, \phi) \sin \theta \, d\theta \, d\phi = 2.2725 \text{ K}, \quad (4.31)$$

which we use above as the present temperature of the relic photons.

There is also considerable information in the direction-dependence of this temperature, including a measurement of the Earth's overall motion relative to the average rest-frame of these photons due to the Doppler effect. The speed of this motion is of order $v_{\oplus} \sim 300$ km/sec, or $v_{\oplus}/c \sim 10^{-3}$, and so causes a part-per-mille deviation of a few mK in the effective photon temperature which is seen in opposite hemispheres of the sky. Even more interesting are the one-part-in-100,000 (*i.e.* ~ 10 μ K) angular-dependent deviations in the temperature about the mean, which survive once the effects of the Earth's motion are removed. These very small primordial fluctuations carry considerable information about the very early universe, and are the topic of a later section.

Should $e^{-\tau}$ increase at later times, such as when the later universe becomes re-ionized by star formation, eq. (4.29) shows how this increase modifies the observed photon distribution. Evidence for such a deviation has been detected through measurements by the WMAP collaboration of the correlation of the polarization of CMB photons with their temperature, taken as a function of direction in the sky [2].

4.5 WIMP Dark Matter

Observations support about 25% of the universe's present-day energy content consisting of non-relativistic matter, but the agreement between measured primordial abundances of light nuclei and the predictions of Big Bang Nucleosynthesis indicate that at most about 4% of the total energy content can consist of known forms of matter (nucleons, electrons, photons and neutrinos). What is the rest of this non-relativistic Dark Matter made of?

Although a definitive answer to this is not yet known, one class of explanations posits that it consists of the relic abundance of a hitherto-unknown species of particle, χ , whose mass and interactions resembles the other exotic particles we already know, such as the W or Z bosons of the weak interactions, or the t quark. This kind of explanation is well-motivated from our understanding of very-short-distance physics, because known flaws in our best theories tell us that a collection of new particles is very likely to be discovered having masses and interactions which are similar to those of the heaviest particles currently known: $m \sim M_w \sim 100$ GeV, and $\sigma(E = m) \sim \alpha_w^2/m^2$. Indeed, many of the expected particles can be electrically neutral, as they must be if they are to be 'dark', in the sense that their cosmological presence can only be detected by gravity.

Most heavy particles are unstable and would rapidly decay into lighter, known particles even if they were to be produced at some time in the earlier universe (or

at present). However, should any such a particle *not* decay, such as would happen if they were the lightest particle carrying a conserved quantum number, then their relic abundance would very naturally provide an explanation for the Dark Matter. Any such a particle is known as a WIMP, which is an acronym for a *Weakly Interacting Massive Particle*.

This section computes the relic dependence which would be expected for such a particle, to show why the masses and interactions which are expected from particle physics also naturally provide the desired relic abundance. To do so we adopt two conservative assumptions about the abundance of these particles in the early universe. First, we assume that these new particles are initially in thermal equilibrium with all of the other particles in the universe for temperatures much larger than the new-particle mass, $T \gg m$. Second, we assume equal numbers of these new particles and their antiparticles, so that even if their stability is explained by their carrying a conserved quantum number, there is no net excess of this quantum number in the universe at large. This amounts to assuming that their chemical potential vanishes, $\mu = 0$.

Under these assumptions their relic abundance is a standard ‘freeze-out’ calculation. That is, we know that equilibrium ties the abundance of these particles to that of all the known particles until the temperature falls below $T \sim 2m$. Once this temperature is reached χ particles and their antiparticles start to annihilate, since on average the reverse creation reactions cannot proceed due to there being insufficient available energy. Once the abundance is sufficiently low the χ particles drop out of equilibrium, and after this point their number density and energy density simply scales with the universal expansion like a^{-3} . Since the number density of photons also scales like $n_\gamma \sim a^{-3}$, it is convenient to compute the ratio $\eta_\chi \equiv n_\chi/n_\gamma$, since this is independent of time and so takes the same value now as it does at freeze-out: $n_{\chi 0} = \eta_\chi n_{\gamma 0}$. Furthermore, using eq. (3.30) to compute the freeze-out density of χ particles implies

$$\eta_\chi = \frac{n_\chi(T_f)}{n_\gamma(T_f)} = \frac{n_\chi(T_f)}{0.2 T_f^3} = 5 \left(\frac{m}{2\pi T_f} \right)^{3/2} e^{-m/T_f}, \quad (4.32)$$

so all is known given the ratio T_f/m .

The freeze-out temperature is estimated in terms of the cross section, σv_{rel} , of the equilibrating interactions evaluated for $E \sim m$, by requiring the reaction rate per χ particle satisfy $\Gamma_\chi(T_f) \sim H(T_f) \sim T_f^2/M_p$, where $\Gamma_\chi(T_f) \sim n(T_f)\sigma v_{\text{rel}}$. Using eq. (3.30) for $n(T_f)$ then implies T_f is found by solving

$$\left(\frac{m T_f}{2\pi} \right)^{3/2} \sigma v_{\text{rel}} e^{-m/T_f} \sim \frac{T_f^2}{M_p}, \quad (4.33)$$

or

$$\left(\frac{m}{2\pi T_f}\right)^{1/2} e^{-m/T_f} \sim \frac{2\pi}{\sigma v_{\text{rel}} M_p m}. \quad (4.34)$$

Since the exponential varies much faster than the pre-exponential factor, a first approximation to T_f/m is

$$\frac{m}{T_f} \sim \ln\left(\frac{\sigma v_{\text{rel}} M_p m}{2\pi}\right) \quad (4.35)$$

and so

$$\eta_\chi = \frac{5}{\sigma v_{\text{rel}} M_p m} \ln\left(\frac{\sigma v_{\text{rel}} M_p m}{2\pi}\right). \quad (4.36)$$

Using now the WIMP values, $m \sim 100$ GeV and $\sigma v_{\text{rel}} \sim 0.1 \alpha_w^2/m^2 \sim 10^{-7}$ GeV $^{-2}$, gives $\eta_\chi \sim 5 \times 10^{-13} \ln(2 \times 10^{12}) \sim 10^{-11} \sim 0.1 \eta_B$. This predicts a present-day relic density of χ particles of order $n_{\chi 0} \sim (\eta_\chi/\eta_B) n_{B0}$ and so the energy density in χ particles is at present

$$\Omega_{\chi 0} \sim \left(\frac{\eta_\chi}{\eta_B}\right) \left(\frac{m}{m_B}\right) \Omega_{B0}, \quad (4.37)$$

or $\Omega_{\chi 0} \sim 10 \Omega_{B0} \sim 0.4$, compared to the observed Dark Matter abundance $\Omega_{m0} = 0.26$.

This abundance is clearly close enough for government work, with the difference between 0.4 and 0.26 easily being fixed by making minor adjustments to the mass or cross section. This observation that stable particles with weak-interaction masses and cross sections naturally have a relic abundance comparable to the observed Dark Matter, is one of the reasons that makes the WIMP explanation of Dark Matter so attractive.

4.6 Baryogenesis

Another relic abundance which cries out for explanation is the baryon abundance, $\eta_B = n_B/n_\gamma = 5 \times 10^{-10}$. In the Big Bang this arises purely as an initial condition because baryon number is conserved and so any initial excess of baryons over anti-baryons survives through the ages to become the net baryon abundance at late times.

The small size of this excess is particularly striking once one entertains, as in the previous section, temperatures which are larger than the nucleon mass, $T \gg m_B \sim 1$ GeV. For instance, for the temperatures $T \sim 100$ GeV of interest for WIMP dark matter, equilibrium ensures that the total number density of baryons, n , and anti-baryons, \bar{n} , are as abundant as all other relativistic particles, and so $n \sim \bar{n} \sim n_\gamma \sim 0.243 T^3$. At these temperatures $\eta_B = 10^{-10}$ implies $(n - \bar{n})/n \sim 10^{-10}$, and so for every 10,000,000,000 baryons there are 9,999,999,999 anti-baryons. It is only when T falls to

the nucleon mass that the anti-baryons each find a baryon with which to annihilate, leaving the one lucky left-over nucleon to survive into the later universe.

Why should the universe start off with such an unlikely imbalance between baryons and anti-baryons? Nobody knows for sure, but suspicions lurk that the net baryon abundance may have a physical explanation. This is because for baryon number, unlike for electric charge, there is no fundamental reason why conservation should be exact. (For these purposes the important difference between baryon number and electric charge is that electric charge is the source of a long-range force, while baryon number seems not to be.) Indeed, numerous sensible proposals exist for physics on very short distances which predict that baryon number is only approximately conserved, but is violated by some hitherto undiscovered interactions which happen to act only extremely weakly during the present epoch. If this were the way that nature works, then it could become possible to generate a net excess of baryons over anti-baryons starting from an initial universe for which no such excess existed.

It turns out that there are three properties, first articulated by Zel'dovich, which the laws of physics must possess if they are to hope to explain the present-day baryon abundance in this way. The following three properties are necessary (and not sufficient) prerequisites for generating a nonzero baryon excess from an initially baryon/anti-baryon symmetric universe.

1. **Baryon Number Violation:** It is necessary that baryon number not be conserved if it is to be nonzero now but is to have vanished in the remote past;
2. **CP Violation:** In order to generate more baryons than anti-baryons it is necessary that there be a baryon-number-changing process for which the rate, $\Gamma(A \rightarrow B)$, for producing baryons from an initially baryon-neutral state A . But to get a net excess also requires that this rate differ from the reverse rate, $\Gamma(B \rightarrow A)$, which converts the baryon number back to the baryon-neutral state again. This can only be possible if the underlying interactions are not themselves invariant under time-reversal, T , which takes $t \rightarrow -t$. Since in local relativistic theories it is always a symmetry to simultaneously reverse the direction of time, and reflect all of the directions in space (*i.e.* *parity*, P) and interchange all particles with anti-particles (or charge conjugation, C), the condition of T invariance is usually stated as a condition that the interactions be invariant under CP , which is a combined action of both C and P .
3. **Loss of Equilibrium:** As was seen earlier, a fundamental feature of the equi-

librium distributions is that they preserve detailed balance. In equilibrium the distribution functions adjust themselves to ensure that the rate for any reaction is equal to the rate for the same reaction when it is run backwards. This is incompatible with the requirement that a reaction, $A \rightarrow B$, produce more baryon number than is destroyed by the reverse reaction, $B \rightarrow A$. So a net baryon number can only be generated if equilibrium does not hold for all particles.

Models may be built which satisfy these properties, and some may be contrived to generate a net baryon asymmetry which survives to the late universe. Such models necessarily involve physics beyond the particles and interactions for which we have direct experimental evidence. They must do so because the known interactions preserve baryon number¹¹ and although they violate CP they do so in a way which is much too small to produce a sufficiently large baryon asymmetry.

5 An early accelerated epoch

This section now switches from a general description of the Λ CDM model to a discussion about the peculiar initial conditions on which its success seems to rely. This is followed by a summary of the elements of some simple single-field inflationary models, and why their proposal is motivated as explanations of the initial conditions for the later universe.

5.1 Peculiar initial conditions

The Λ CDM model describes well what we see around us, provided that the universe is started off with a very specific set of initial conditions. There are several properties of these initial conditions that seem peculiar, as is now summarized.

Flatness problem

The first problem concerns the observed spatial flatness of the present-day universe. As described earlier, observations of the CMB indicate that the quantity κ/a^2 of the Friedmann equation, eq. (2.3), is at present consistent with zero. What is odd about this condition is that this curvature term tends to grow in relative importance as the universe expands, so finding it to be small now means that it must have been *extremely* small in the remote past.

¹¹Strictly speaking baryon number is not exactly conserved in the presently-successful Standard Model, but its violation is far from large enough to produce any appreciable baryon asymmetry in cosmology.

More quantitatively, it is useful to divide the Friedmann equation by $H^2(t)$ to give

$$1 + \frac{\kappa}{(aH)^2} = \frac{8\pi G\rho}{3H^2} =: \Omega(a), \quad (5.1)$$

where (as before) the final equality defines $\Omega(a)$. The problem arises because the product aH decreases with time during both matter and radiation domination. For instance, observations indicate that at present $\Omega = \Omega_0$ is unity to within about 10%, and since during the matter-dominated era the product $(aH)^2 \propto a^{-1}$ it follows that at the epoch $z_{\text{eq}} \simeq 3600$ of radiation-matter equality we must have had

$$\Omega(z_{\text{eq}}) - 1 = (\Omega_0 - 1) \left(\frac{a}{a_0} \right) = \frac{\Omega_0 - 1}{1 + z_{\text{eq}}} \simeq \frac{0.1}{3600} \simeq 2.8 \times 10^{-5}. \quad (5.2)$$

So $\Omega - 1$ had to be smaller than a few tens of a millionth at the time of radiation-matter equality in order to be of order 10% now.

And it only gets worse the further back one goes, provided the extrapolation back occurs within a radiation- or matter-dominated era (as seems to be true at least as far back as the epoch of nucleosynthesis). Since during radiation-domination we have $(aH)^2 \propto a^{-2}$ and the redshift of nucleosynthesis is $z_{\text{BBN}} \sim 10^{10}$ it follows that at this epoch one must require

$$\Omega(z_{\text{BBN}}) - 1 = \left[\Omega(z_{\text{eq}}) - 1 \right] \left(\frac{1 + z_{\text{eq}}}{1 + z_{\text{BBN}}} \right)^2 = \frac{0.1}{3600} \left(\frac{3600}{10^{10}} \right)^2 \approx 3.6 \times 10^{-18}, \quad (5.3)$$

requiring Ω to be unity with an accuracy of roughly a part in 10^{18} . The discomfort of having the success of a theory hinge so sensitively on the precise value of an initial condition in this way is known as the Big Bang's *Flatness Problem*.

Horizon problem

Perhaps a more serious question asks why the initial universe can be so very homogeneous. In particular, the temperature fluctuations of the CMB only arise at the level of 1 part in 10^5 , and the question is how this temperature can be so incredibly uniform across the sky.

Why is this regarded as a problem? It is not uncommon for materials on earth to have a uniform temperature, and this is usually understood as a consequence of thermal equilibrium. An initially inhomogeneous temperature distribution equilibrates by having heat flow between the hot and cold areas, until everything eventually shares a common temperature.

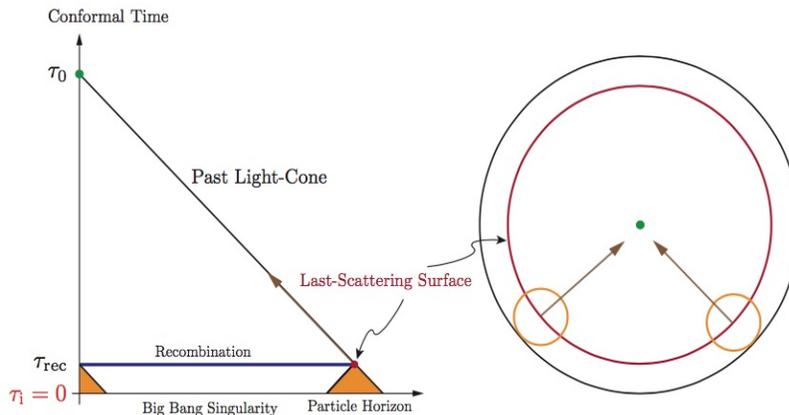


Figure 2. A conformal diagram illustrating how there is inadequate time in a radiation-dominated universe for there to be a causal explanation for the correlation of temperature at different points of the sky in the CMB. (Figure taken from [?], with permission of the author.)

The same argument is harder to make in cosmology because in the Hot Big Bang model the universe generically expands so quickly that there has not been enough time for light to travel across the entire sky to bring everyone the news as to what the common temperature is supposed to be. This is easiest to see using conformal coordinates, as in (??), since in these coordinates it is simple to identify which regions can be connected by light signals. In particular, radially directed light rays travel along lines $d\ell = \pm d\tau$, which can be drawn as straight lines of slope ± 1 in the $\tau - \ell$ plane, as in Figure 2. The problem is that $a(\tau)$ reaches zero in a finite conformal time (which we can conventionally choose to happen at $\tau = 0$), since $a(\tau) \propto \tau$ during radiation domination and $a(\tau) \propto \tau^2$ during matter domination. Redshift $z_{\text{rec}} \simeq 1100$ (the epoch of recombination, at which the CMB photons last sampled the temperature of the Hydrogen gas with which they interact) is simply too early for different directions in the sky to have been causally connected in the entire history of the universe up to that point.

To pin this down quantitatively, let us assume that the universe is radiation-dominated for all points earlier than the epoch of radiation-matter equality, t_{eq} , so the complete evolution of $a(t)$ until recombination is

$$a(t) \simeq \begin{cases} a_{\text{eq}}(t/t_{\text{eq}})^{1/2} & \text{for } 0 < t < t_{\text{eq}} \\ a_{\text{eq}}(t/t_{\text{eq}})^{2/3} & \text{for } t_{\text{eq}} < t < t_{\text{rec}} . \end{cases} \quad (5.4)$$

(The real evolution does not have a discontinuous derivative at $t = t_{\text{eq}}$, but this inaccuracy is not important for the argument that follows.) The maximum proper distance, measured at the time of recombination, that a light signal could have travelled by the time of recombination, t_{rec} , then is

$$\begin{aligned} D_{\text{rec}} &= a_{\text{rec}} \left[\int_0^{t_{\text{eq}}} \frac{d\hat{t}}{a(\hat{t})} + \int_{t_{\text{eq}}}^{t_{\text{rec}}} \frac{d\hat{t}}{a(\hat{t})} \right] = \frac{a_{\text{rec}} t_{\text{eq}}}{a_{\text{eq}}} \left[3 \left(\frac{t_{\text{rec}}}{t_{\text{eq}}} \right)^{1/3} - 1 \right] \\ &= \frac{2}{H_{\text{eq}}^+} \left(\frac{a_{\text{rec}}}{a_{\text{eq}}} \right)^{3/2} \left[1 - \frac{1}{3} \left(\frac{a_{\text{eq}}}{a_{\text{rec}}} \right)^{1/2} \right] \simeq \frac{1.6}{H_{\text{rec}}}, \end{aligned} \quad (5.5)$$

where $H_{\text{eq}}^+ = 2/(3t_{\text{eq}})$ denotes the limit of the Hubble scale as $t \rightarrow t_{\text{eq}}$ on the matter-dominated side. The approximate equality in this expression uses $H \propto a^{-3/2}$ during matter domination as well as using the redshifts $z_{\text{rec}} \simeq 1100$ and $z_{\text{eq}} \simeq 3600$ (as would be true in the Λ CDM model) to obtain $a_{\text{eq}}/a_{\text{rec}} \simeq 1100/3600 \simeq 0.31$.

To evaluate this numerically we use the present-day value for the Hubble constant, $H_0 \simeq 70$ km/sec/Mpc — or (keeping in mind our units for which $c = 1$), $H_0^{-1} \simeq 13$ Gyr $\simeq 4$ Gpc. This then gives $H_{\text{rec}}^{-1} \simeq H_0^{-1} (a_{\text{rec}}/a_0)^{3/2} \simeq 3 \times 10^{-5} H_0^{-1} \simeq 0.1$ Mpc, if we use $a_0/a_{\text{rec}} = 1 + z_{\text{rec}} \simeq 1100$, and so $D_{\text{rec}} \simeq 0.2$ Mpc.

Now CMB photons arriving to us from the surface of last scattering left this surface at a distance from us that is now of order

$$R_0 = a_0 \int_{t_{\text{rec}}}^{t_0} \frac{d\hat{t}}{a(\hat{t})} = 3t_0 - 3t_0^{2/3} t_{\text{rec}}^{1/3} = \frac{2}{H_0} \left[1 - \left(\frac{a_{\text{rec}}}{a_0} \right)^{1/2} \right], \quad (5.6)$$

again using $a \propto t^{2/3}$ and $H \propto a^{-3/2}$, and so $R_0 \simeq 2/H_0 \simeq 8$ Gpc. So the angle subtended by D_{rec} placed at this distance away (in a spatially-flat geometry) is really $\theta \simeq D_{\text{rec}}/R_{\text{rec}}$ where $R_{\text{rec}} = (a_{\text{rec}}/a_0)R_0 \simeq 7$ Mpc is its distance *at the time of last scattering*, leading to¹² $\theta \simeq 0.2/7 \simeq 1^\circ$. Any two directions in the sky separated by more than this angle (about twice the angular size of the Moon, seen from Earth) are so far apart that light had not yet had time to reach one from the other since the universe's beginning.

How can all the directions we see have known they were all to equilibrate to the same temperature? It is very much as if we were to find a very uniform temperature distribution, *immediately* after the explosion of a very powerful bomb.

Defect problem

¹²This estimate is related to the quantity θ_{MC} in the table of Fig. ??.

Historically, a third problem — called the ‘Defect’ (or ‘Monopole’) Problem is also used to motivate changing the extrapolation of radiation domination into the remote past. A defect problem arises if the physics of the much higher energy scales relevant to the extrapolation involves the production of topological defects, like domain walls, cosmic strings or magnetic monopoles. Such defects are often found in Grand Unified theories; models proposed to unify the strong and electroweak interactions as energies of order 10^{15} GeV.

These kinds of topological defects can be fatal to the success of late-time cosmology, depending on how many of them survive down to the present epoch. For instance if the defects are monopoles, then they typically are extremely massive and so behave like non-relativistic matter. This can cause problems if they are too abundant because they can preclude the existence of a radiation dominated epoch, because their energy density falls more slowly than does radiation as the universe expands.

Defects are typically produced with an abundance of one per Hubble volume, $n_d(a_f) \sim H_f^3$, where $H_f = H(a_f)$ is the Hubble scale at their epoch of formation, at which time $a = a_f$. Once produced, their number is conserved, so their density at later times falls like $n_d(a) = H_f^3(a_f/a)^3$. Consequently, at present the number surviving within a Hubble volume is $n_d(a_0)H_0^{-3} = (H_f a_f/H_0 a_0)^3$.

Because the product aH is a falling function of time, the present-day abundance of defects can easily be so numerous that they come to dominate the universe well before the nucleosynthesis epoch.¹³ This could cause the universe to expand (and so cool) too quickly as nuclei were forming, and so give the wrong abundances of light nuclei. Even if not sufficiently abundant during nucleosynthesis, the energy density in relict defects can be inconsistent with measures of the current energy density.

This is clearly more of a hypothetical problem than are the other two, since whether there is a problem depends on whether the particular theory for the high-energy physics of the very early universe produces these types of defects or not. It can be fairly pressing in Grand Unified models since in these models the production of magnetic monopoles can be fairly generic.

5.2 Acceleration to the rescue

The key observation when trying to understand the above initial conditions is that they only seem unreasonable because they are based on extrapolating into the past assuming the universe to be radiation (or matter) dominated (as would naturally be true if the

¹³Whether they do also depends on their dimension, with magnetic monopoles tending to be more dangerous in this regard than are cosmic strings, say.

Λ CDM model were the whole story). This section argues that these initial conditions can seem more reasonable if a different type of extrapolation is used; in particular if there were an earlier epoch during which the universal expansion were to accelerate: $\ddot{a} > 0$ [? ?].

Why should acceleration help? The key point is that the above initial conditions are a problem because the product aH is a falling function as a increases, for both matter and radiation domination. For instance, for the flatness problem the evolution of the curvature term in the Friedmann equation is $\Omega_\kappa \propto (aH)^{-2}$ and this grows as a grows only because aH decreases with a . But if $\ddot{a} > 0$ then $\dot{a} = aH$ *increases* as a increases, and this can help alleviate the problems. For example, finding Ω_κ to be very small in the recent past would be less disturbing if the more-distant past contained a sufficiently long epoch during which aH grew.

How long is long enough? To pin this down suppose there were an earlier epoch during which the universe were to expand in the same way as during Dark Energy domination, $a(t) \propto e^{Ht}$, for constant H . Then $aH = a_0 H e^{Ht}$ grows exponentially with time and so even if Ht were of order 100 or less it would be possible to explain why Ω_κ could be as small as 10^{-18} or smaller.

Having aH grow also allows a resolution to the horizon problem. One way to see this is to notice that $a(t) \propto e^{Ht}$ implies $\tau = -H^{-1}e^{-Ht}$ plus a constant (with the sign a consequence of having τ increase as t does), and so

$$a(\tau) = -\frac{1}{H(\tau - \tau_0)}, \quad (5.7)$$

with $0 < a < \infty$ corresponding to the range $-\infty < \tau < \tau_0$. Exponentially accelerated expansion allows τ to be extrapolated to arbitrarily negative values, and so allows sufficient time for the two causally disconnected regions of the conformal diagram of Figure 2 to have at one point been in causal contact.

Another way to visualize this is to plot physical distance $\lambda(t) \propto a(t)$ and the Hubble radius, H^{-1} , against t , as in Figure 3. Focus first on the right-hand side of this figure, which compares these quantities during radiation or matter domination. During these epochs the Hubble length evolves as $H^{-1} \propto t$ while the scale factor satisfies $a(t) \propto t^p$ with $0 < p < 1$. Consequently H^{-1} grows more quickly with t than do physical length scales $\lambda(t)$. During radiation or matter domination systems of any given size eventually get caught by the growth of H^{-1} and so ‘come inside the Hubble scale’ as the universe expands. Systems involving larger $\lambda(t)$ do so later than those with smaller λ . The largest sizes currently visible have only recently crossed inside of the Hubble length,

Evolution of Scales

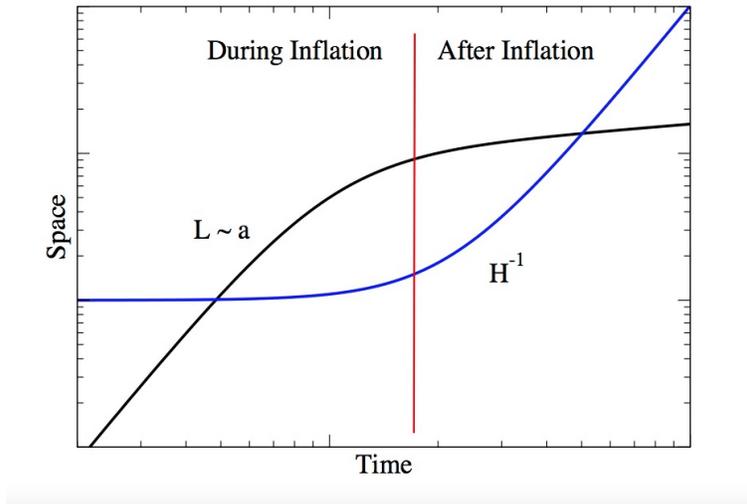


Figure 3. A sketch of the relative growth of physical scales, $L(t)$, (in black) and the Hubble length, H^{-1} , (in blue) during and after inflation. Horizon exit happens during inflation where the blue and black curves first cross, and this is eventually followed by horizon re-entry where the curves cross again during the later Hot Big Bang era.

having spent their entire earlier history larger than H^{-1} (assuming always a radiation- or matter-dominated universe).

Having $\lambda > H^{-1}$ matters because physical quantities tend to freeze when their corresponding length scales satisfy $\lambda(t) > H^{-1}$. This then precludes physical processes from acting over these scales to explain things like the uniform temperature of the CMB. The freezing of super-Hubble scales can be seen, for example, in the evolution of a massless scalar field in an expanding universe, since the field equation $\square\phi = 0$ becomes in FRW coordinates

$$\ddot{\phi}_k + 3H\dot{\phi}_k + \left(\frac{k}{a}\right)^2 \phi_k = 0, \quad (5.8)$$

where we Fourier expand the field $\phi(x) = \int d^3k \phi_k \exp[i\mathbf{k} \cdot \mathbf{x}]$ using co-moving coordinates, \mathbf{x} . For modes satisfying $2\pi/\lambda = p = k/a \ll H$ the field equation implies $\dot{\phi}_k \propto a^{-3}$ and so $\phi_k = C_0 + C_1 \int dt/a^3$ is the sum of a constant plus a decaying mode.

Things are very different during exponential expansion, however, as is shown on the left-hand side of Figure 3. In this regime $\lambda(t) \propto a(t) \propto e^{Ht}$ grows exponentially

with t while H^{-1} remains constant. This means that modes that are initially smaller than the Hubble length get stretched to become larger than the Hubble length, with the transition for a specific mode of length $\lambda(t)$ occurring at the epoch of ‘Hubble exit’, $t = t_{\text{he}}$, defined by $2\pi/\lambda(t_{\text{he}}) = p_{\text{he}} = k/a(t_{\text{he}}) = H$. In this language it is because the criterion for Hubble exit and entry is $k = aH$ that the growth or shrinkage of aH is relevant to the horizon problem.

How much expansion is required to solve the horizon problem? Choosing a mode ϕ_k that is only now crossing the Hubble scale tells us that $k = a_0 H_0$. This same mode would have crossed the horizon during an exponentially expanding epoch when $k = a_{\text{he}} H_I$, where H_I is the constant Hubble scale during exponential expansion. So clearly $a_0 H_0 = a_{\text{he}} H_I$ where t_{he} is the time of Hubble exit for this particular mode. To determine how much exponential expansion is required solve the following equation for $N_e := \ln(a_{\text{end}}/a_{\text{he}})$, where a_{end} is the scale factor at the end of the exponentially expanding epoch:

$$1 = \frac{a_{\text{he}} H_I}{a_0 H_0} = \left(\frac{a_{\text{he}} H_I}{a_{\text{end}} H_I} \right) \left(\frac{a_{\text{end}} H_I}{a_{\text{eq}} H_{\text{eq}}} \right) \left(\frac{a_{\text{eq}} H_{\text{eq}}}{a_0 H_0} \right) = e^{-N_e} \left(\frac{a_{\text{eq}}}{a_{\text{end}}} \right) \left(\frac{a_0}{a_{\text{eq}}} \right)^{1/2}. \quad (5.9)$$

This assumes (for the purposes of argument) that the universe is radiation dominated right from t_{end} until radiation-matter equality, and uses $aH \propto a^{-1}$ during radiation domination and $aH \propto a^{-1/2}$ during matter domination. $N_e = H_I(t_{\text{end}} - t_{\text{he}})$ is called the number of e -foldings of exponential expansion and is proportional to how long exponential expansion lasts

Using, as above, $(a_{\text{eq}} H_{\text{eq}})/(a_0 H_0) = (a_0/a_{\text{eq}})^{1/2} \simeq 60$, and $(a_{\text{eq}} H_{\text{eq}})/(a_{\text{end}} H_{\text{end}}) = a_{\text{end}}/a_{\text{eq}} = T_{\text{eq}}/T_M$ with $T_{\text{eq}} \sim 3$ eV, and assuming the energy density of the exponentially expanding phase is transferred perfectly efficiently to produce a photon temperature T_M then leads to the estimate

$$N_e \sim \ln [(3 \times 10^{23}) \times 60] + \ln \left(\frac{T_M}{10^{15} \text{ GeV}} \right) \approx 58 + \ln \left(\frac{T_M}{10^{15} \text{ GeV}} \right). \quad (5.10)$$

Roughly 60 e -foldings of exponential expansion can provide a framework for explaining how causal physics might provide the observed correlations that are observed in the CMB over the largest scales, even if the energy densities involved are as high as 10^{15} GeV. We shall see below that life is even better than this, because in addition to providing a *framework* in which a causal understanding of correlations could be solved, inflation itself can provide the *mechanism* for explaining these correlations (given an inflationary scale of the right size).

5.3 Inflation or a bounce?

An early epoch of near-exponential accelerated expansion has come to be known as an ‘inflationary’ early universe. Acceleration within this framework speeds up an initially expanding universe to a higher expansion rate. However, an attentive reader may notice that although acceleration is key to helping with Λ CDM’s initial condition issues, there is no *a priori* reason why the acceleration must occur in an initially expanding universe, as opposed (say) to one that is initially contracting. Models in which one tries to solve the problems of Λ CDM by having an initially contracting universe accelerate to become an expanding one are called ‘bouncing’ cosmologies.

Since it is really the acceleration that is important, bouncing models should in principle be on a similar footing to inflationary ones. In what follows only inflationary models are considered, for the following reasons:

Validity of the semiclassical methods

Predictions in essentially all cosmological models are extracted using semiclassical methods: one typically writes down the action for some system and then explores its consequences by solving its classical equations of motion. So a key question for all such models is the identification of the small parameter (or parameters) that suppresses quantum effects and so controls the underlying semiclassical approximation. In the absence of such a control parameter classical predictions need not capture what the system really does. Such a breakdown of the semiclassical approximation really means that the ‘theory error’ in the model’s predictions could be arbitrarily large, making comparisons to observations essentially meaningless.

A reason sometimes given for not pinning down the size of quantum corrections when doing cosmology is that gravity plays a central role, and we do not yet know the ultimate theory of quantum gravity. Implicit in this argument is the belief that the size of quantum corrections is incalculable without such an ultimate theory, perhaps because of the well-known divergences in quantum predictions due to the non-renormalizability of General Relativity [?]. But experience with non-renormalizable interactions elsewhere in physics tells us that quantum predictions can sometimes be made, provided one recognizes they involve an implicit low-energy/long-distance expansion relative to the underlying physical scale set by the dimensionful non-renormalizable couplings. Because of this the semiclassical expansion parameter in such theories is usually the ratio between this underlying short-distance scale and the distances of interest in cosmology (which, happily enough, aims at understanding the largest distances on offer). Effective

field theories provide the general tools for quantifying these low-energy expansions, and this is why EFT methods are so important for any cosmological studies.

As is argued in more detail in §??, the semiclassical expansion in cosmology is controlled by small quantities like $(\lambda M_p)^{-2}$ where λ is the smallest length scale associated with the geometry of interest. In practice it is often $\lambda \sim H^{-1}$ that provides the relevant scale in cosmology, particularly when all geometrical dimensions are similar in size. So a rule of thumb generically asks the ratio H^2/M_p^2 to be chosen to be small:

$$\frac{H^2}{M_p^2} \propto \frac{\rho}{M_p^4} \ll 1, \quad (5.11)$$

as a necessary condition¹⁴ for quantum cosmological effects to be suppressed.

For inflationary models H is usually at its largest during the inflationary epoch, with geometrical length scales only increasing thereafter, putting one deeper and deeper into the semiclassical domain. It is a big plus for these models that they can account for observations while wholly remaining within the regime set by (5.11), and this is one of the main reasons why they receive so much attention.

For bouncing cosmologies the situation can be more complicated. The smallest geometrical scale λ usually occurs during the epoch near the bounce, even though H^{-1} itself usually tends to infinity there. In models where λ becomes comparable to M_p^{-1} (or whatever other scale – such as the string length scale, $M_s^{-1} \gg M_p^{-1}$ – that governs short-distance gravity), quantum effects during the bounce need not be negligible and the burden on proponents is to justify why semiclassical predictions actually capture what happens during the bounce.

Difficulty of achieving a semiclassically large bounce

Another issue arises even if the scale λ during a bounce does remain much larger than the more microscopic scales of gravity. In this regime the bounce can be understood purely within the low-energy effective theory describing the cosmology, for which General Relativity should be the leading approximation. But (when $\kappa = 0$) the Friedmann equation for FRW geometries in General Relativity states that $H^2 = \rho/3M_p^2$, and so ρ must pass through zero at the instant where the contracting geometry transitions to

¹⁴The semiclassical criterion can be stronger than this, though this can often only be quantified within the context of a specific proposal for what quantum gravity is at the shortest scales. For instance, if it is string theory that takes over at the shortest scales then treatment of cosmology using a field theory – rather than fully within string theory – requires (5.11) be replaced by the stronger condition $H^2/M_s^2 \ll 1$, where $M_s \ll M_p$ is the string scale, set for example by the masses of the lightest string excited states.

expansion (since $H = \dot{a}/a$ vanishes at this point). Furthermore, using (2.3) and (?), it must also be true that

$$\dot{H} = \frac{\ddot{a}}{a} - H^2 = -\frac{1}{2M_p^2}(\rho + p) > 0, \quad (5.12)$$

at this point in order for H to change sign there, which means the dominant contributions to the cosmic fluid must satisfy $\rho + p < 0$ during the bounce.¹⁵

Although there are no definitive no-go theorems, it has proven remarkably difficult to find a convincing physical system that both satisfies the condition $\rho + p < 0$ and does not also have other pathologies, such as uncontrolled runaway instabilities. For instance within the class of multiple scalar field models for which the lagrangian density is $\mathcal{L} = \sqrt{-g} \left[\frac{1}{2} G_{ij}(\phi) \partial_\mu \phi^i \partial^\mu \phi^j + V(\phi) \right]$ we have $\rho + p = G_{ij}(\phi) \dot{\phi}^i \dot{\phi}^j$ and so $\rho + p < 0$ requires the matrix of functions $G_{ij}(\phi)$ to have a negative eigenvalue. But if this is true then there is always a combination of fields for which the kinetic energy is negative (what is called a ‘ghost’), and so is unstable towards the development of arbitrarily rapid motion. Such a negative eigenvalue also implies the gradient energy $\frac{1}{2} G_{ij} \nabla \phi^i \cdot \nabla \phi^j$ is also unbounded from below, indicating instability towards the development of arbitrarily short-wavelength spatial variations.

Phenomenological issues

In addition to the above conceptual issues involving the control of predictions, there are also potential phenomenological issues that bouncing cosmologies must face. Whereas expanding geometries tend to damp out spatially varying fluctuations – such as when gradient energies involve factors like $(k/a)^2$ that tend to zero as $a(t)$ grows – the opposite typically occurs during a contracting epoch for which $a(t)$ shrinks. This implies that inhomogeneities tend to grow during the pre-bounce contraction — even when the gradient energies are bounded from below — and so a mechanism must be provided for why we emerge into the homogeneous and isotropic later universe seen around us in observational cosmology.

It is of course important that bouncing cosmologies be investigated, not least in order to see most fully what might be required to understand the flatness and horizon problems. Furthermore it is essential to know whether there are alternative observational implications to those of inflation that might be used to marshal evidence about what actually occurred in the very early universe. But within the present state of the

¹⁵This is usually phrased as a violation of the ‘null-energy’ condition, which states that $T_{\mu\nu} n^\mu n^\nu \geq 0$ for all null vectors n^μ .

art inflationary models have one crucial advantage over bouncing cosmologies: they provide concrete semiclassical control over the key epoch of acceleration on which the success of the model ultimately relies. Because of this inflationary models are likely to remain the main paradigm for studying pre- Λ CDM extrapolations, at least until bouncing cosmologies are developed to allow similar control over how primordial conditions get propagated to the later universe through the bounce.

5.4 Simple inflationary models

So far so good, but what kind of physics can provide both an early period of accelerated expansion and a mechanism for ending this expansion to allow for the later emergence of the successful Hot Big Bang cosmology?

Obtaining the benefits of an accelerated expansion requires two things: (i) some sort of physics that hangs the universe up for a relatively long period with an accelerating equation of state, $p < -\frac{1}{3}\rho < 0$; and (ii) some mechanism for ending this epoch to allow the later appearance of the radiation-dominated epoch within which the usual Big Bang cosmology starts. Although a number of models exist that can do this, none yet seems completely compelling. This section describes some of the very simplest such models.

The central requirement is to have some field temporarily dominate the universe with potential energy, and for the vast majority of models this new physics comes from the dynamics of a scalar field, $\varphi(x)$, called the ‘inflaton’. This field can be thought of as an order parameter characterizing the dynamics of the vacuum at the very high energies likely to be relevant to inflationary cosmology. Although the field φ can in principle depend on both position and time, once inflation gets going it rapidly smooths out spatial variations, suggesting the study of homogeneous configurations: $\varphi = \varphi(t)$.

5.4.1 Higgs field as inflaton

No way is known to obtain a viable inflationary model simply using the known particles and interactions, but a minimal model [?] does use the usual scalar Higgs field already present in the Standard Model as the inflaton, provided it is assumed to have a nonminimal coupling to gravity of the form $\delta\mathcal{L} = -\xi\sqrt{-g}(\mathcal{H}^\dagger\mathcal{H})R$, where \mathcal{H} is the usual Higgs doublet and R is the Ricci scalar. Here ξ is a new dimensionless coupling, whose value turns out must be of order 10^4 in order to provide a good description of cosmological observations. Inflation in this case turns out to occur when the Higgs field takes trans-Planckian values, $\mathcal{H}^\dagger\mathcal{H} > M_p^2$, assuming V remains proportional to $(\mathcal{H}^\dagger\mathcal{H})^2$ at such large values.

As argued in [? ?], although the large values required for both ξ and $\mathcal{H}^\dagger\mathcal{H}$ needn't invalidate the validity of the EFT description, they do push the envelope for the boundaries of its domain of validity. In particular, semiclassical expansion during inflation turns out to require the neglect of powers of $\sqrt{\xi} H/M_p$, which during inflation is to be evaluated with $H \sim M_p/\xi$. This means both that the semiclassical expansion is in powers of $1/\sqrt{\xi}$, and that some sort of new physics (or ‘UV completion’) must intervene at scales $M_p/\sqrt{\xi} \sim \sqrt{\xi} H$, not very far above inflationary energies. Furthermore, it must do so in a way that also explains why the lagrangian should have the very particular large-field form that is required for inflation. In particular, V must be precisely proportional to the square, f^2 , of the coefficient of the nonminimal Ricci coupling, $f(\mathcal{H}^\dagger\mathcal{H})R$, at trans-Planckian field values, since this is ultimately what ensures the potential is flat when expressed in terms of canonically normalized variables in this regime. There are no known proposals for UV completions that satisfy all of the requirements, although conformal or scale invariance seems likely to be relevant [?].

This example raises a more general point that is worth noting in passing: having trans-Planckian fields during inflation need not in itself threaten the existence of a controlled low-energy EFT description. The reason for this — as is elaborated in more detail in §?? below — is that the EFT formulation is ultimately a low-energy expansion and so large fields are only dangerous if they also imply large energy densities. Using an EFT to describe trans-Planckian field evolution need not be a problem so long as the evolution satisfies $H \ll M$ at the field values of interest, where $M \lesssim M_p$ is the scale of the physics integrated out to obtain the EFT in question. The condition $H \ll M$ becomes $V \ll M_p^4$ if it happens that $M \sim M_p$. (In any explicit example the precise conditions for validity of EFT methods are obtained using power-counting arguments along the lines of those given in §?? below.)

5.4.2 New field as inflaton

The simplest models instead propose a single new relativistic scalar field, φ , and design its dynamics through choices made for its potential energy, $V(\varphi)$. Taking

$$\mathcal{L} = \sqrt{-g} \left[\frac{1}{2} \partial_\mu \varphi \partial^\mu \varphi + V(\varphi) \right], \quad (5.13)$$

the inflaton field equation becomes $\square\varphi = V'(\varphi)$, which for homogeneous configurations $\varphi(t)$ reduces in an FRW geometry to

$$\ddot{\varphi} + 3H\dot{\varphi} + V' = 0, \quad (5.14)$$

where $V' = dV/d\varphi$.

The Einstein field equations are as before, but with new φ -dependent contributions to the energy density and pressure: $\rho = \rho_{\text{rad}} + \rho_{\text{m}} + \rho_{\varphi}$ and $p = \frac{1}{3}\rho_{\text{rad}} + p_{\varphi}$, where

$$\rho_{\varphi} = \frac{1}{2}\dot{\varphi}^2 + V(\varphi) \quad \text{and} \quad p_{\varphi} = \frac{1}{2}\dot{\varphi}^2 - V(\varphi). \quad (5.15)$$

The Dark Energy of the present-day epoch is imagined to arise by choosing V so that its minimum satisfies $\rho_{DE} = V(\varphi_{\text{min}})$. Inflation is imagined to occur when φ evolves slowly through a region where $V(\varphi) \gg V(\varphi_{\text{min}})$ is very large, and ends once φ rolls down towards its minimum.

With these choices energy conservation for the φ field — $\dot{\rho}_{\varphi} + 3(\dot{a}/a)(\rho_{\varphi} + p_{\varphi}) = 0$ follows from the field equation, eq. (5.14). Some couplings must also exist between the φ field and ordinary Standard Model particles in order to provide a channel to transfer energy from the inflaton to ordinary particles, and so reheat the universe as required for the later Hot Big Bang cosmology. But φ is not imagined to be in thermal equilibrium with itself or with the other kinds of matter during inflation or at very late times, and this can be self-consistent if the coupling to other matter is sufficiently weak and if the φ particles are too heavy to be present once the cosmic fluid cools to the MeV energies and below (for which we have direct observations).

5.4.3 Slow-Roll Inflation

To achieve an epoch of near-exponential expansion, we seek a solution to the above classical field equations for $\varphi(t)$ in which the Hubble parameter, H , is approximately constant. This is ensured if the total energy density is dominated by ρ_{φ} , with ρ_{φ} also approximately constant. As we have seen, energy conservation implies the pressure must then satisfy $p_{\varphi} \approx -\rho_{\varphi}$. Inspection of eqs. (5.15) shows that both of these conditions are satisfied if the φ kinetic energy is negligible compared with its potential energy:

$$\frac{1}{2}\dot{\varphi}^2 \ll V(\varphi), \quad (5.16)$$

since then $p_{\varphi} \simeq -V(\varphi) \simeq -\rho_{\varphi}$. So long as $V(\varphi)$ is also much larger than any other energy densities, it would dominate the Friedmann equation and $H^2 \simeq V/(3M_p^2)$ would then be approximately constant.

What properties must $V(\varphi)$ satisfy in order to allow (5.16) to hold for a sufficiently long time? This requires a long period of time where φ moves slowly enough to allow *both* the neglect of $\frac{1}{2}\dot{\varphi}^2$ relative to $V(\varphi)$ in the Friedmann equation, (2.3), *and* the neglect of $\ddot{\varphi}$ in the scalar field equation, (5.14).

The second of these conditions allows eq. (5.14) to be written in the approximate *slow-roll* form,

$$\dot{\varphi} \approx - \left(\frac{V'}{3H} \right). \quad (5.17)$$

Using this in (5.16) then shows V must satisfy $(V')^2/(9H^2V) \ll 1$, leading to the condition that slow-roll inflation requires φ must lie in a region for which

$$\epsilon := \frac{1}{2} \left(\frac{M_p V'}{V} \right)^2 \ll 1. \quad (5.18)$$

Physically, this condition requires H to be approximately constant over any given Hubble time, inasmuch as $3M_p^2 H^2 \simeq V$ implies $6M_p^2 H \dot{H} \simeq V' \dot{\varphi} \simeq -(V')^2/3H$ and so

$$-\frac{\dot{H}}{H^2} \simeq \frac{(V')^2}{18H^4 M_p^2} \simeq \frac{M_p^2 (V')^2}{2V^2} = \epsilon \ll 1. \quad (5.19)$$

Self-consistency also demands that if eq. (5.17) is differentiated to compute $\ddot{\varphi}$ it should be much smaller than $3H\dot{\varphi}$. Performing this differentiation and demanding that $\ddot{\varphi}$ remain small (in absolute value) compared with $3H\dot{\varphi}$, then implies $|\eta| \ll 1$ where

$$\eta := \frac{M_p^2 V''}{V}, \quad (5.20)$$

defines the second slow-roll parameter. The slow-roll parameters ϵ and η are important [?] because (as shown below) the key predictions of single-field slow-roll inflation for density fluctuations can be expressed in terms of the three parameters ϵ , η and the value, H_I , of the Hubble parameter during inflation.

Given an explicit shape for $V(\varphi)$ one can directly predict the amount of inflation that occurs after the epoch of Hubble exit (where currently observable scales become larger than the Hubble length). This is done by relating the amount of expansion directly to the distance φ traverses in field space during this time. To this end, rewriting eq. (5.17) in terms of $\varphi' \equiv d\varphi/da$, leads to

$$\frac{d\varphi}{da} = \frac{\dot{\varphi}}{\dot{a}} = - \frac{V'}{3aH^2} = - \frac{M_p^2 V'}{aV}, \quad (5.21)$$

which when integrated between horizon exit, φ_{he} , and final value, φ_{end} , gives the amount of expansion during inflation as $a_{\text{end}}/a_{\text{he}} = e^{N_e}$, with

$$N_e = \int_{a_{\text{he}}}^{a_{\text{end}}} \frac{da}{a} = \int_{\varphi_{\text{end}}}^{\varphi_{\text{he}}} d\varphi \left(\frac{V}{M_p^2 V'} \right) = \frac{1}{M_p} \int_{\varphi_{\text{end}}}^{\varphi_{\text{he}}} \frac{d\varphi}{\sqrt{2\epsilon}}. \quad (5.22)$$

In these expressions φ_{end} can be defined by the point where the slow-roll parameters are no longer small, such as where $\epsilon \simeq \frac{1}{2}$. Then this last equation can be read as defining $\varphi_{\text{end}}(N_e)$, as a function of the desired number of e -foldings between the the epoch of horizon exit and the end of inflation, since this is this quantity constrained to be large by the horizon and flatness problems.

Notice also that if ϵ were approximately constant during inflation, then eq. (5.22) implies that $N_e \approx (\varphi_{\text{he}} - \varphi_{\text{end}})/(\sqrt{2\epsilon} M_p)$. In such a case φ must traverse a range of order $N_e M_p \sqrt{2\epsilon}$ between φ_{he} and φ_{end} . This is larger than order M_p provided only that $1 \gg \epsilon \gtrsim 1/N_e^2$, showing why Planckian fields are often of interest for inflation [?].

5.4.4 Some illustrative examples

It is worth working through what these formulae mean in a few concrete choices for the shape of the scalar potential.

Example I: Quadratic model

The simplest example of an inflating potential [? ?] chooses φ to be a free massive field, for which

$$V = \frac{1}{2} m^2 \varphi^2, \quad (5.23)$$

and so $V' = m^2 \varphi$ and $V'' = m^2$, leading to slow-roll parameters of the form

$$\epsilon = \frac{1}{2} \left(\frac{2M_p}{\varphi} \right)^2 \quad \text{and} \quad \eta = \frac{2M_p^2}{\varphi^2}, \quad (5.24)$$

and so $\epsilon = \eta$ in this particular case, and slow roll requires $\varphi \gg M_p$. The scale for inflation in this field range is $V = \frac{1}{2} m^2 \varphi^2$ and so $H_I^2 \simeq m^2 \varphi^2 / (6 M_p^2)$. We can ensure $H_I^2 / M_p^2 \ll 1$ even if $\varphi \gg M_p$ by choosing m/M_p sufficiently small. Observations will turn out to require $\epsilon \sim \eta \sim 0.01$ and so the regime of interest is $\varphi_{\text{he}} \sim 10M_p$, and so $H_I/M_p \ll 1$ requires $m/M_p \ll 0.1$.

In this large-field regime φ (and so also V and H) evolves only very slowly despite there being no nearby stationary point for V because Hubble friction slows φ 's slide down the potential. Since φ evolves towards smaller values, eventually slow roll ends once η and ϵ become $O(1)$. Choosing φ_{end} by the condition $\epsilon(\varphi_{\text{end}}) = \eta(\varphi_{\text{end}}) = \frac{1}{2}$ implies $\varphi_{\text{end}} = 2M_p$. The number of e -foldings between horizon exit and $\varphi_{\text{end}} = 2M_p$ is then given by eq. (5.22), which in this instance becomes

$$N_e = \int_{2M_p}^{\varphi_{\text{he}}} d\varphi \left(\frac{\varphi}{2M_p^2} \right) = \left(\frac{\varphi_{\text{he}}}{2M_p} \right)^2 - 1, \quad (5.25)$$

and so obtaining $N_e \sim 63$ e -foldings (say) requires choosing $\varphi_{\text{he}} \sim 16 M_p$. In particular $\epsilon_{\text{he}} := \epsilon(\varphi_{\text{he}})$ and $\eta_{\text{he}} := \eta(\varphi_{\text{he}})$ can be expressed directly in terms of N_e , leading to

$$\epsilon_{\text{he}} = \eta_{\text{he}} = \frac{1}{2(N_e + 1)}, \quad (5.26)$$

which are both of order 10^{-2} for $N_e \simeq 60$. As seen below, the prediction $\epsilon = \eta$ is beginning to be disfavoured by cosmological observations.

Example II: pseudo-Goldstone axion

The previous example shows how controlled inflation requires the inflaton mass to be small compared with the scales probed by φ . Small masses arise because the condition $|\eta| \ll 1$ implies the inflaton mass satisfies $m^2 \sim |V''| \sim |\eta V/M_p^2| \ll V/M_p^2 \simeq 3H^2$. Consequently m must be very small compared with H , which itself must be Planck suppressed compared with other scales (such as $v \sim V^{1/4}$) during inflation. From the point of view of particle physics such small masses pose a puzzle because it is fairly uncommon to find interacting systems with very light spinless particles in their low-energy spectrum.¹⁶

The main exceptions to this statement are Goldstone bosons for the spontaneous breaking of continuous global symmetries since these are guaranteed to be massless by Goldstone's theorem. This makes it natural to suppose the inflaton to be a pseudo-Goldstone boson (*i.e.* a would-be Goldstone boson for an approximate symmetry, much like the pions of chiral perturbation theory). In this case Goldstone's theorem ensures the scalar's mass (and other couplings in the scalar potential) must vanish in the limit the symmetry becomes exact, and this 'protects' it from receiving generic UV-scale contributions. For abelian broken symmetries this shows up in the low-energy EFT as an approximate shift symmetry under which the scalar transforms inhomogeneously: $\varphi \rightarrow \varphi + \text{constant}$.

If the approximate symmetry arises as a $U(1)$ phase rotation for some microscopic field, and if this symmetry is broken down to discrete rotations, $Z_N \subset U(1)$, then the inflaton potential is usually trigonometric [?]:

$$V = V_0 + \Lambda^4 \left[1 - \cos \left(\frac{\varphi}{f} \right) \right] = V_0 + 2\Lambda^4 \sin^2 \left(\frac{\varphi}{2f} \right), \quad (5.27)$$

for some scales V_0 , Λ and f . Here V_0 is chosen to agree with ρ_{DE} and because ρ_{DE} is so small the parameter V_0 is dropped in what follows. The parameter Λ represents

¹⁶From an EFT perspective having a light scalar requires the coefficients of low-dimension effective interactions like ϕ^2 to have unusually small coefficients like m^2 rather than being as large as the (much larger) UV scales M^2 .

the scale associated with the explicit breaking of the underlying $U(1)$ symmetry while f is related to the size of its spontaneous breaking. The statement that the action is approximately invariant under the symmetry is the statement that Λ is small compared with UV scales like f . Expanding about the minimum at $\varphi = 0$ reveals a mass of size $m = \Lambda^2/f \ll \Lambda \ll f$, showing the desired suppression of the scalar mass.

With this choice $V' = (\Lambda^4/f) \sin(\varphi/f)$ and $V'' = (\Lambda^4/f^2) \cos(\varphi/f)$, leading to slow-roll parameters of the form

$$\epsilon = \frac{M_p^2}{2f^2} \cot^2\left(\frac{\varphi}{2f}\right) \quad \text{and} \quad \eta = \frac{M_p^2}{2f^2} \left[\cot^2\left(\frac{\varphi}{2f}\right) - 1 \right], \quad (5.28)$$

and so $\eta = \epsilon - (M_p^2/2f^2)$. Notice that in the limit $M_p \lesssim \varphi \ll f$ these go over to the $m^2\varphi^2$ case examined above, with $m = \Lambda^2/f$.

Slow roll in this model typically requires $f \gg M_p$. This can be seen directly from (5.28) for generic $\varphi \simeq f$, but also follows when $\varphi \ll f$ because in this case the potential is close to quadratic and slow roll requires $M_p \ll \varphi \ll f$. The scale for inflation is $V \simeq \Lambda^4$ and so $H_I \sim \Lambda^2/M_p$. This ensures $H_I^2/M_p^2 \ll 1$ follows from the approximate-symmetry limit which requires $\Lambda \ll M_p$. The condition $\epsilon \sim 0.01$ is arranged by choosing $f \sim 10M_p$, but once this is done the prediction $\epsilon \simeq \eta$ is in tension with recent observations.

The number of e -foldings between horizon exit and φ_{end} is again given by eq. (5.22), so

$$N_e = \frac{2f}{M_p^2} \int_{\varphi_{\text{end}}}^{\varphi_{\text{he}}} d\varphi \tan\left(\frac{\varphi}{2f}\right) = \left(\frac{2f}{M_p}\right)^2 \ln \left| \frac{\sin(\varphi_{\text{he}}/2f)}{\sin(\varphi_{\text{end}}/2f)} \right|, \quad (5.29)$$

which is only logarithmically sensitive to φ_{he} , but which can easily be large due to the condition $f \gg M_p$.

While models such as this do arise generically from UV completions like string theory [?], axions in string theory typically arise with $f \ll M_p$ [?], making the condition $f \gg M_p$ tricky to arrange [?].

Example III: pseudo-Goldstone dilaton

Another case where the inflaton mass is protected by an approximate shift symmetry arises when it is a pseudo-Goldstone boson for a scaling symmetry of the underlying UV theory. Such ‘accidental’ scale symmetries turn out to be fairly common in explicit examples of UV completions because scale invariances are automatic consequences of higher-dimensional supergravities [?]. Because it is a scaling symmetry the same arguments leading to trigonometric potentials for the compact $U(1)$ rotations instead in this case generically lead to exponential potentials [?].

In this case the form expected for the scalar potential during the inflationary regime would be

$$V = V_0 - V_1 e^{-\varphi/f} + \dots, \quad (5.30)$$

for some scales V_0 , V_1 and f . Our interest is in the regime $\varphi \gg f$ and in this regime V_0 dominates, and so is chosen as needed for inflationary cosmology, with $H_I^2 \simeq V_0/(3M_p^2)$. Control over the semiclassical limit requires $V_0 \ll M_p^4$.

With this choice the relevant potential derivatives are $V' \simeq (V_1/f) e^{-\varphi/f}$ and $V'' \simeq -(V_1/f^2) e^{-\varphi/f}$ leading to slow-roll parameters of the form

$$\epsilon \simeq \frac{1}{2} \left(\frac{M_p V_1}{f V_0} \right)^2 e^{-2\varphi/f} \quad \text{and} \quad \eta \simeq - \left(\frac{M_p^2 V_1}{f^2 V_0} \right) e^{-\varphi/f}, \quad (5.31)$$

and so

$$\epsilon = \frac{1}{2} \left(\frac{f}{M_p} \right)^2 \eta^2. \quad (5.32)$$

The number of e -foldings between horizon exit and φ_{end} is again given by eq. (5.22), so

$$N_e = \left(\frac{f V_0}{M_p^2 V_1} \right) \int_{\varphi_{\text{end}}}^{\varphi_{\text{he}}} d\varphi e^{\varphi/f} = \left(\frac{f^2 V_0}{M_p^2 V_1} \right) [e^{\varphi_{\text{he}}/f} - e^{\varphi_{\text{end}}/f}], \quad (5.33)$$

which can easily be large so long as $\varphi_{\text{he}} \gg f$ and φ_{end}/f is order unity.

Notice that ϵ and η are generically small whenever $\varphi \gg f$, even if $V_1 \sim V_0$, so there is no need to require f be larger than M_p to ensure a slow roll. Typical examples of underlying UV theories (see below) give $f \sim M_p$, in which case $\epsilon \simeq \eta^2$. It turns out that this prediction provides better agreement with experiment than $\epsilon \simeq \eta$ does, and (as seen below) the generic expectation that $\epsilon \sim \eta^2$ has potentially interesting observational consequences for measurements of primordial gravitational waves because it relates the as-yet-unmeasured tensor-to-scalar ratio, $r \lesssim 0.07$, to the observed spectral tilt, $n_s \simeq 0.96$, giving the prediction $r_{\text{th}} \simeq (n_s - 1)^2 \simeq 0.002$.

Interestingly, many successful inflationary models can be recast into this exponential form, usually with specific values predicted for f . The earliest instance using an exponential potential [?] came from a supergravity example with $f = \sqrt{\frac{1}{6}} M_p$, with a nonlinearly realized $SU(1, 1)$ symmetry. Such symmetries are now known to arise fairly commonly when dimensionally reducing higher-dimensional supersymmetric models [?]. This early supergravity example foreshadows the results from a class of explicit higher-dimensional UV completions within string theory [?], which reduce to the above with $f = \sqrt{3} M_p$, while the first extra-dimensional examples of this type [?] gave $f = \sqrt{2} M_p$.

In fact, the Higgs-inflation model described earlier can also be recast to look like a scalar field with an exponential potential of the form considered here, once it is written with canonically normalized fields. The prediction in this case is $f = \sqrt{\frac{3}{2}} M_p$. The same is true for another popular model that obtains inflation using curvature-squared interactions [?], for which again $f = \sqrt{\frac{3}{2}} M_p$. Although both of these models are hard to obtain in a controlled way from UV completions directly, their formulation in terms of exponential potentials may provide a way to do so through the back door.

5.5 Flies in the ointment

Although not really the main line of development of these lecture notes, it would be wrong to leave the impression that inflationary theories must be the last word in early universe cosmology. Indeed they have problems that motivate some to seek out better alternatives. Here are a few of the main complaints.

Initial-condition problems

A major motivation for inflation comes from trying to understand the peculiar initial conditions required for the success of the late-time cosmology of the Λ CDM model. But this cannot be regarded as being a success if inflation itself also requires contrived initial conditions. In particular there are concerns that inflation might not start unless the universe is initially prepared in a sufficiently homogeneous configuration over several Hubble scales.

Although the fragility of the required initial conditions is in dispute it is true that there are not many explicit calculations done with more generic initial conditions. There are calculations involving more generic random potentials that do indicate that inflation can be a rare occurrence but it is still being explored how much these calculations depend on the assumptions being made.

Fine-tuning problems

Slow-roll inflation requires relatively shallow potentials, and these are relatively difficult to obtain within the low-energy limit of explicit UV completions. Not all models are equally bad in this regard, with those based on pseudo-Goldstone bosons being able to arrange shallow potentials in more controlled ways.

But even with these models inflationary predictions are notoriously sensitive to small effects. Because the inflationary effect being sought is gravitational it is Planck suppressed and so can be threatened even by other Planck-suppressed effective interactions that in any other circumstances would have been regarded as negligible.

The multiverse and the landscape

Once it gets going inflation can be hard to stop. And even if it ends in some parts of the universe, if it survives in others these inflating regions expand so quickly that they can come to dominate the volume of the universe. These kinds of effects are made even worse if even physical constants are really controlled by the expectation values of fields that can be different in different parts of the universe, as seems to be the case in theories like string theory.

For such theories it becomes hard to see how to make definite predictions in a traditional way. A great variety of universes might arise within any given framework, and how does one falsify such a theory if the options available become too numerous? One way out is to use anthropic reasoning, but it is not yet clear what the proper rules should be for doing so.

One point that is worth making is that problems with the multiverse (if problems they prove to be) are actually common to pretty much any cosmological framework in theories that admit a complicated landscape of solutions. That is because even if you think early-universe cosmology is described by something besides inflation (such as a bouncing cosmology) all the above predictivity problems in any case arise if inflation nevertheless should unintentionally get going in any remote corner of the landscape.

My two cents

My own opinion is to accept that inflationary models are a work in progress, leaving many things to be desired. But even so they seem at this point to be in better shape than all of their alternatives, mostly because of the control they allow over all of the approximations being made during their use. This situation could change as alternatives get better explored (as they should certainly be), but the shortage of convincing alternatives shows that inflation already sets a fairly high bar for other theories to pass.

Although it may be premature to speculate about issues of the multiverse that are hard to compare with observations, inflationary models do seem to give a clean answer to the more limited practical question of what kind of extrapolation could be useful into our relatively immediate pre-Big-Bang past. Their predictions seem to be under good theoretical control and to agree well with the properties of the primordial fluctuations that have so far been revealed.

6 Density Perturbations

Previous sections show that the universe was very homogeneous at the time of photon last scattering, since the temperature fluctuations observed in the distribution of CMB photons have an amplitude $\delta T/T \sim 10^{-5}$. On the other hand the universe around us is full of stars and galaxies and so is far from homogeneous. How did the one arise from the other?

The basic mechanism for this is based on gravitational instability: the gravitational force towards an initially over-dense region acts to attract even more material towards this region, thereby making it even more dense. This process can feed back on itself until an initially small density perturbation becomes dramatically amplified, such as into a star. This section describes the physics of this instability, in the very early universe when the density contrasts are small enough to be analyzed perturbatively in the fluctuation amplitude. The discussion follows that of ref.

6.1 Nonrelativistic Density Perturbations

We start with the discussion of gravitational instability in the non-relativistic gravitating limit, both for simplicity and since this limit provides a good description of the behaviour of density fluctuations in a matter-dominated universe (which is the one relevant for almost all of cosmology after radiation-matter decoupling occurs at $z_{\text{dec}} = 1100$).

The following equations of motion describe the dynamics of a simple non-relativistic fluid with energy density, ρ , pressure, p , entropy density, s , and local fluid velocity \mathbf{v} . The equations express local conservation laws, and are

$$\begin{aligned}
 \frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) &= 0 && \text{(energy conservation)} \\
 \rho \left[\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla) \mathbf{v} \right] + \nabla p + \rho \nabla \phi &= 0 && \text{(momentum conservation)} \\
 \frac{\partial s}{\partial t} + \nabla \cdot (s \mathbf{v}) &= 0 && \text{(entropy conservation)} \\
 \nabla^2 \phi - 4\pi G \rho &= 0 && \text{(universal gravitation)},
 \end{aligned} \tag{6.1}$$

as well as the equation of state, $p = p(\rho, s)$. Here ϕ denotes the local gravitational potential.

6.1.1 Perturbations About a Static Background

A simple solution to the above equations corresponds to a homogeneous and static fluid, with constant values $\mathbf{v} = 0$, $\rho = \rho_0$, $p = p_0$ and $s = s_0$. We also choose ϕ_0 to be con-

stant, although this is inconsistent with Poisson's equation (which expresses the Law of Universal Gravitation above) given the above choices for the other background quantities, ρ_0 , p_0 , *etc.*. Choosing ϕ_0 to be constant corresponds to studying the dynamics of fluctuations in a small part of a larger gravitating system, for which the background gravitational potential is dominated by external sources and is approximately constant over the region of interest.

We are interested in how small perturbations about this solution evolve, $\rho = \rho_0 + \delta\rho$, $p = p_0 + \delta p$, $s = s_0 + \delta s$, $\phi = \phi_0 + \delta\phi$ and \mathbf{v} . To linear order in the perturbations the equations of motion are

$$\begin{aligned} \frac{\partial \delta\rho}{\partial t} + \rho_0 \nabla \cdot \mathbf{v} &= 0 \\ \rho_0 \frac{\partial \mathbf{v}}{\partial t} + \nabla \delta p + \rho_0 \nabla \delta\phi &= 0 \\ \frac{\partial \delta s}{\partial t} &= 0 \\ \nabla^2 \delta\phi - 4\pi G \delta\rho &= 0. \end{aligned} \tag{6.2}$$

Perturbing the equation of state allows the elimination of δp from these equations, $\delta p = c_s^2 \delta\rho + \xi \delta s$, where

$$c_s^2 = \left(\frac{\partial p}{\partial \rho} \right)_{s|_0} \quad \text{and} \quad \xi = \left(\frac{\partial p}{\partial s} \right)_{\rho|_0}, \tag{6.3}$$

and so $\nabla^2 \delta p = c_s^2 \nabla^2 \delta\rho + \xi \nabla^2 \delta s$. Notice that for radiation we have $c_{sr}^2 = \frac{1}{3}$ and for non-relativistic matter $c_{sm}^2 = O(T/m) \ll 1$, where m is the particle rest mass. (For instance, for air (N_2) at room temperature we have $m \sim 14$ GeV and $T \sim 300$ K ~ 0.03 eV, and so $c_{sm}^2 \sim 3 \times 10^{-12}$ which gives the fairly accurate estimate $c_{sm} \sim 10^{-6} c \sim 300$ m/s.)

The velocity perturbation, \mathbf{v} , can be eliminated by subtracting the divergence of the momentum-conservation equation from the time derivative of the energy-conservation equation, leading to perturbation equations which involve only the two independent fluid perturbations, $\delta\rho$ and δs :

$$\frac{\partial^2 \delta\rho}{\partial t^2} - c_s^2 \nabla^2 \delta\rho - 4\pi G \rho_0 \delta\rho = \xi \nabla^2 \delta s \quad \text{and} \quad \frac{\partial \delta s}{\partial t} = 0. \tag{6.4}$$

Fourier transforming, $\delta\rho(\mathbf{r}, t) = \delta\rho_k(t) e^{i\mathbf{k}\cdot\mathbf{r}}$, gives the master result which governs perturbations in the fluid

$$\frac{d^2 \delta\rho_k}{dt^2} + c_s^2 k^2 \delta\rho_k - 4\pi G \rho_0 \delta\rho_k = -\xi k^2 \delta s_k, \tag{6.5}$$

and $d\delta s_k/dt = 0$, where $k^2 = \mathbf{k} \cdot \mathbf{k}$.

These equations show that entropy (heat) perturbations do not propagate in time, since $d\delta s_k/dt = 0$, and they act as sources for density perturbations. They also show that the dynamics of density perturbations depends keenly on the wavelength of the perturbation. Defining the Jean's wave-number, k_J , and Jean's length, $\ell_J = 2\pi/k_J$, by

$$k_J^2 = \frac{4\pi G \rho_0}{c_s^2}, \quad (6.6)$$

we see that short-wavelength perturbations, $k \gg k_J$, satisfy

$$\frac{d^2 \delta \rho_k}{dt^2} + c_s^2 k^2 \delta \rho_k \approx -\xi k^2 \delta s_k, \quad (6.7)$$

which when $\delta s_k = 0$ has oscillatory solutions, $\delta \rho_k \propto \exp(\pm i\omega_k t)$, with $\omega_k = c_s k$. These solutions describe ordinary sound waves, for which the fluid pressure provides the 'restoring force' for the oscillation.

Alternatively, long-wavelength modes, $k \ll k_J$, satisfy

$$\frac{d^2 \delta \rho_k}{dt^2} - 4\pi G \rho_0 \delta \rho_k = -\xi k^2 \delta s_k, \quad (6.8)$$

which for $\delta s_k = 0$ has solutions $\delta \rho_k \propto \exp(\pm \lambda_k t)$, with $\lambda_k = (4\pi G \rho_0)^{1/2}$ (for all k). The solution with the positive sign in the exponential grows without limit as t increases, indicating the *Jeans instability* towards the gravitational amplification of an initially-small density contrast.

Physically, this instability arises because gravitational collapse occurs on a time scale — the *free-fall time*, $t_{\text{fall}} \sim (G\rho_0)^{-1/2}$ — which is set purely by the density of the collapsing object. A fluid tries to resist this collapse by adjusting its pressure accordingly, but for a region of size ℓ this can only be done on a time scale of order $t_{\text{pr}} \sim \ell/c_s$, where c_s is the speed of sound. It follows that for sufficiently large objects gravitational collapse can occur faster than the fluid pressure can adjust to resist it, *i.e.* instability occurs once $t_{\text{pr}} < t_{\text{fall}}$, or $\ell > c_s/(G\rho_0)^{1/2} \sim \ell_J$.

This is the instability which is at the root of the formation of the structure we see in the universe around us, and it only arises for sufficiently long-wavelength perturbations. Because ℓ_J depends on c_s the minimum length scale of an unstable perturbation depends on the equation of state of the matter which is being perturbed. In particular, $k_J \sim H/c_s$ and so $k_J \sim H$ for relativistic systems (for which $c_s \sim 1$) and $k_J \gg H$ for non-relativistic matter (for which $c_s \sim T/m \ll 1$).

6.1.2 Perturbations About an Expanding Background

For cosmological applications it is instructive to repeat the previous exercise, but this time expanding about a homogeneously and radially expanding background fluid configuration. For these purposes consider instead a fluid background for which $\mathbf{v}_0 = H(t) \mathbf{r}$, where $H(t)$ is assumed a given function of t . In this case $\nabla \cdot \mathbf{v}_0 = 3H(t)$. This flow is motivated by the observation that it corresponds to the proper velocity if particles within the fluid were moving apart from one another according to the law $\mathbf{x}(t) = a(t) \mathbf{y}$, with \mathbf{y} being a time-independent co-moving coordinate. In this case $\mathbf{v}_0 \equiv d\mathbf{x}/dt = \dot{a} \mathbf{y} = H(t) \mathbf{x}(t)$, where $H = \dot{a}/a$. In this sense $H(t)$ describes the non-relativistic analog of the Hubble parameter for the background fluid's expansion.

Background Quantities

We now ask what the rest of the background quantities, $\rho_0(t)$, $p_0(t)$ and $\phi_0(t)$ must satisfy in order to be consistent with this flow. The equation of energy conservation implies ρ_0 must satisfy

$$0 = \dot{\rho}_0 + \nabla \cdot (\rho_0 \mathbf{v}_0) = \dot{\rho}_0 + 3H \rho_0, \quad (6.9)$$

and so, given $H = \dot{a}/a$, it follows that $\rho_0 \propto a^{-3}$. That is, the non-relativistic expanding fluid necessarily requires the background density to fall with expansion as would the density in a matter-dominated universe.

Using this density in the law for universal gravitation requires the gravitational potential, ϕ_0 , take the form

$$\phi_0 = \frac{2\pi G \rho_0}{3} \mathbf{r}^2, \quad (6.10)$$

and so $\nabla \phi_0 = \frac{4}{3} \pi G \rho_0 \mathbf{r}$. This describes the radially-directed gravitational potential which acts to decelerate the overall universal expansion.

Given this gravitational force, the momentum conservation equation, using $\dot{\mathbf{v}}_0 + (\mathbf{v}_0 \cdot \nabla) \mathbf{v}_0 = [H + \dot{H}/H] \mathbf{v}_0$ and $\mathbf{v}_0 = H \mathbf{r}$, becomes

$$\left[\dot{H} + H^2 + \frac{4\pi G \rho_0}{3} \right] \mathbf{r} = 0. \quad (6.11)$$

This is equivalent to the Friedmann equation, as is now shown. Notice that if we take $a \propto t^\alpha$ then $H = \alpha/t$ and $\dot{H} = -\alpha/t^2 = -H^2/\alpha$. This, together with $\rho_0 \propto a^{-3} \propto t^{-3\alpha}$, is consistent with eq. (6.11) only if $\alpha = 2/3$, as expected for a matter-dominated universe. Furthermore, with this choice for α we also have $\dot{H} + H^2 = -\frac{1}{2} H^2$, and so eq. (6.11) is equivalent to

$$H^2 = \frac{8\pi G}{3} \rho_0, \quad (6.12)$$

which is the Friedmann equation, as claimed.

When studying perturbations we solve the entropy equation, $\dot{s}_0 = 0$, by taking $s_0 = 0$.

Perturbations during matter domination

To study perturbations about this background take $\mathbf{v} = \mathbf{v}_0 + \delta\mathbf{v}$, $\rho = \rho_0 + \delta\rho$, $p = p_0 + \delta p$, $s = \delta s$ and $\phi = \phi_0 + \delta\phi$, and expand as before the equations of motion to first order in the perturbations. Defining $D_t = \partial/\partial t + \mathbf{v}_0 \cdot \nabla$, the linearized equations in this case become

$$\begin{aligned} D_t \delta\rho + 3H \delta\rho + \rho_0 \nabla \cdot \delta\mathbf{v} &= 0 \\ \rho_0 (D_t \delta\mathbf{v} + H \delta\mathbf{v}) + \nabla \delta p + \rho_0 \nabla \delta\phi &= 0 \\ D_t \delta s &= 0 \\ \nabla^2 \delta\phi - 4\pi G \delta\rho &= 0. \end{aligned} \tag{6.13}$$

To obtain this form for the momentum conservation equation requires using the equations of motion for the background quantities.

Performing the same manipulations as for the static case allows these equations to be recast in terms of the two basic fluid perturbations, $\delta\rho$ and δs . The equations become $D_t \delta s = 0$ and

$$D_t^2 \left(\frac{\delta\rho}{\rho_0} \right) + 2H D_t \left(\frac{\delta\rho}{\rho_0} \right) - c_s^2 \nabla^2 \left(\frac{\delta\rho}{\rho_0} \right) - 4\pi G \rho_0 \left(\frac{\delta\rho}{\rho_0} \right) = \frac{\xi}{\rho_0} \delta s. \tag{6.14}$$

In order to analyze the solutions to this equation, it is convenient to change variables to a co-moving coordinate, \mathbf{y} , defined by $\mathbf{r} = a(t) \mathbf{y}$. In this case, for any function $f = f(\mathbf{r}, t)$ we have $(\partial f/\partial t)_{\mathbf{y}} = (\partial f/\partial t)_{\mathbf{r}} + H \mathbf{r} \cdot \nabla f = D_t f$, and $\nabla f = (1/a) \nabla_{\mathbf{y}} f$. Fourier transforming the perturbations in co-moving coordinates, $\delta\rho/\rho_0 = \delta_k(t) \exp[i\mathbf{k} \cdot \mathbf{y}]$, leads to the following master equation governing density perturbations

$$\ddot{\delta}_k + 2H \dot{\delta}_k + \left(\frac{c_s^2 k^2}{a^2} - 4\pi G \rho_0 \right) \delta_k = \left(\frac{\xi}{\rho_0} \right) \delta s, \tag{6.15}$$

where the over-dot denotes d/dt .

These equations have solutions whose character depends on the relative size of k/a and the Jeans wave-number,

$$k_J^2(t) = \frac{4\pi G \rho_0(t)}{c_s^2(t)} = \frac{3H^2(t)}{2c_s^2(t)}, \tag{6.16}$$

with instability occurring once $k/a \ll k_J$. Notice that so long as $c_s \sim O(1)$ the Jeans length is comparable in size to the Hubble length, $\ell_J \sim H^{-1}$. For adiabatic fluctuations ($\delta s_k = 0$) the above equation implies that the short-wavelength fluctuations ($k/a \gg k_J$) undergo damped oscillations of the form

$$\delta_k(t) \propto a^{-1/2} \exp \left[\pm i k c_s \int^t \frac{dt'}{a(t')} \right]. \quad (6.17)$$

The new feature here relative to the non-expanding case is the damping of the oscillations due to the universal expansion. This kind of damping is sometimes known as *Hubble friction*.

Long-wavelength adiabatic oscillations ($k/a \ll k_J$) again exhibit an instability, but in this case the overall expansion dilutes the instability into a power law in t (compared to the exponential encountered earlier for perturbations of a static background). This dilution occurs because the overall expansion reduces the density, and this effect fights the density increase due to gravitational collapse. The approximate solutions in this case are

$$\delta_k(t) \propto t^{2/3} \propto a(t) \quad \text{and} \quad \delta_k(t) \propto t^{-1} \propto a^{-3/2}(t), \quad (6.18)$$

with the $\delta_k(t) \sim t^{2/3}$ solution describing the instability to gravitational collapse.

Because both the red-shifted wave-number, k/a , and the Jeans wave-number, k_J , depend on time, the overall expansion of the background can convert modes from stable to unstable (or vice versa). Whether this conversion is towards stability or instability depends on the time dependence of ak_J , which is governed by the time-dependence of the combination aH/c_s . If $a \propto t^\alpha$ then $aH \propto t^{\alpha-1} \propto a^{1-1/\alpha}$, and so aH increases with t if $\alpha > 1$ and decreases with t if $\alpha < 1$. Since $\alpha = 2/3$ for the radiation-dominated universe of interest here, it follows that $aH \propto t^{-1/3} \propto a^{-1/2}$, and so *decreases* with t . Provided that c_s does not change much, this ensures that in the absence of other influences modes having fixed k pass from being unstable to stable as a increases due to the overall expansion.

Perturbations during radiation and vacuum domination

A completely relativistic treatment of density perturbations requires following fluctuations in the matter stress energy as well as in the metric itself (since these are related by Einstein's equations relating geometry and stress-energy). The details of such calculations go beyond the scope of these notes, although some of the main features are described below. But the above considerations suffice to address a result that is an important part of the structure-formation story: the stalling of perturbation growth for nonrelativistic matter during radiation- or vacuum-dominated epochs.

To contrast how fluctuations grow during radiation and matter domination it is instructive to examine the transition from radiation to matter domination. To this end we again track the growth of density fluctuations for non-relativistic matter, $\delta\rho_{m0}/\rho_{m0}$, and do so using the same Fourier-transformed equation as before,

$$\ddot{\delta}_{\mathbf{k}} + 2H \dot{\delta}_{\mathbf{k}} + \left(\frac{c_s^2 \mathbf{k}^2}{a^2} - 4\pi G \rho_{m0} \right) \delta_{\mathbf{k}} = 0, \quad (6.19)$$

but with $H^2 = 8\pi G \rho_0/3$ where $\rho_0 = \rho_{m0} + \rho_{r0}$ includes both radiation and matter. In particular, during the transition between radiation and matter domination the Hubble scale satisfies

$$H^2(a) = \frac{8\pi G \rho_0}{3} = \frac{H_{\text{eq}}^2}{2} \left[\left(\frac{a_{\text{eq}}}{a} \right)^3 + \left(\frac{a_{\text{eq}}}{a} \right)^4 \right], \quad (6.20)$$

where radiation-matter equality occurs when $a = a_{\text{eq}}$, at which point $H(a = a_{\text{eq}}) = H_{\text{eq}}$.

As described above, any departure from the choice $a(t) \propto t^{2/3}$ — such as occurs when radiation is non-negligible in $\rho(a)$ — means that the background momentum-conservation equation, eq. (6.11), is no longer satisfied. Instead the expression for H comes from solving the fully relativistic radiation-dominated Friedmann equation, eq. (6.20). But this does not mean that the nonrelativistic treatment of the fluctuations must fail, since the important kinematics and gravitational interactions amongst these perturbations remain the Newtonian ones. To first approximation the leading effect of the radiation domination for these fluctuations is simply to change the expansion rate, as parameterized by $H(a)$ in (6.20).

For all modes for which the pressure term, $c_s^2 \mathbf{k}^2/a^2$, is negligible, (6.19) implies $\delta(x)$ satisfies

$$2x(1+x)\delta'' + (3x+2)\delta' - 3\delta = 0, \quad (6.21)$$

where the rescaled scale factor, $x = a/a_{\text{eq}}$, is used as a proxy for time and primes denote differentiation with respect to x . As is easily checked, this is solved by $\delta^{(1)} \propto (x + \frac{2}{3})$, and so the growing mode during matter domination (*i.e.* $x \gg 1$) does not also grow during radiation domination ($x \ll 1$). Furthermore, the solution linearly independent to this one can be found using the Frobenius method, and this behaves for $x \ll 1$ (*i.e.* deep in the radiation-dominated regime) as $\delta^{(2)} \propto \delta^{(1)} \ln x + (\text{analytic})$, where ‘analytic’ denotes a simple power series proportional to $1 + c_1 x + \dots$. These solutions show how density perturbations for non-relativistic matter grow at most logarithmically during the radiation-dominated epoch.

A similar analysis covers the case where Dark Energy (modelled as a cosmological constant) dominates in an $\Omega = 1$ universe. In this case $4\pi G \rho_{m0} \sim \Omega_m H^2 \ll H^2$ and so

the instability term becomes negligible relative to the first two terms of (6.19). This leads to

$$\ddot{\delta} + 2H \dot{\delta} \simeq 0, \quad (6.22)$$

which has as solution $\dot{\delta} \propto a^{-2}$. Integrating again gives a frozen mode, $\delta \propto a^0$, and a damped mode that falls as $\delta \propto a^{-2}$ when H is constant (as it is when Dark Energy dominates and $a \propto e^{Ht}$). This shows that non-relativistic density perturbations also stop growing once matter domination ends.

We are now in a position to summarize how inhomogeneities grow in the late universe, assuming the presence of an initial spectrum of very small primordial density fluctuations. The key observation is that several conditions all have to hold in order for there to be appreciable growth of density inhomogeneities. These conditions are:

1. Fluctuations of any type do not grow for super-Hubble modes, for which $k/a \ll H$, regardless of what type of matter dominates the background evolution.
2. Fluctuations in nonrelativistic matter can be unstable, growing as $\delta_k \propto a(t)$, but only in a matter-dominated universe and for those modes in the momentum window $H \ll (k/a) \ll H/c_s$.
3. No fluctuations in relativistic matter ever grow appreciably, either inside or outside the Hubble scale. (Although this is not shown explicitly above for relativistic matter, and requires the fully relativistic treatment, the instability window $H \ll k/a \ll H/c_s$ for nonrelativistic fluctuations is seen to close as they become relativistic — *i.e.* as $c_s \rightarrow 1$.)

Before pursuing the implications of these conditions for instability, a pause is in order to describe what properties of fluctuations are actually measured.

6.1.3 Multi-Component Fluids

We have seen that there are at least three types of matter present whose fluctuations we might explore during most of the matter-dominated universe. These are the dominant non-relativistic Dark Matter, the relativistic radiation and (at least for redshifts $3600 \gtrsim z \gtrsim 1100$) electrically-charged non-relativistic matter, like electrons and nuclei. It is necessary to examine the dynamics of multi-component fluids like this in order to apply the previous considerations to questions of structure formation in the later universe.

In this case we take the density, pressure and entropy to be the sum of a contribution from each fluid component,

$$\rho = \sum_i \rho_i, \quad s = \sum_i s_i \quad \text{and} \quad p = \sum_i p_i(\rho_i, s_i), \quad (6.23)$$

where the index ‘ i ’ runs over the values d, r, \dots , representing Dark Matter, radiation and any other fluid components, and so $\delta p = \sum_i [c_{si}^2 \delta \rho_i + \xi_i \delta s_i]$. In what follows we assume that each fluid component is sufficiently decoupled from the others so that there is negligible energy and entropy transfer between them, so that the equations of energy and entropy conservation apply to each component separately. We also take the only momentum interchange between the fluids to be due to their gravitational fields, with the background gravitational field assumed to be dominated by the non-relativistic Dark Matter contribution. Notice that this assumption of negligible energy exchange necessitates treating as one fluid any group of relativistic and non-relativistic particles which are in thermal equilibrium.

Repeating the previous analysis leads to the following coupled system of equations governing the fluctuations in each component,

$$\begin{aligned} D_t \delta \rho_i + 3H \delta \rho_i + \rho_{0i} \nabla \cdot \delta \mathbf{v}_i &= 0 \\ \rho_{0i} (D_t \delta \mathbf{v}_i + H \delta \mathbf{v}_i) + \nabla \delta p_i + \rho_{0i} \nabla \delta \phi &= 0 \\ D_t \delta s_i &= 0 \\ \nabla^2 \delta \phi - 4\pi G \sum_i \delta \rho_i &= 0. \end{aligned} \quad (6.24)$$

Defining the Fourier transform (in co-moving coordinates) of the relative density fluctuation of fluid component ‘ i ’ by $\delta \rho_i / \rho_{0i} = \delta_{ki}(t) \exp[i\mathbf{k} \cdot \mathbf{y}]$, leads to the following coupled set of equations

$$\ddot{\delta}_{ki} + 2H \dot{\delta}_{ki} + \left(\frac{c_{si}^2 k^2}{a^2} \right) \delta_{ki} - 4\pi G \sum_j \rho_{0j} \delta_{kj} = \left(\frac{\xi_i}{\rho_{0i}} \right) \delta s_i. \quad (6.25)$$

Specialize now to two fluids, consisting of the dominant Dark Matter ($i = d$) plus a second component ($i = r$). In this case we may use the following approximations, $\rho_0 \approx \rho_{0d}$, $\rho_{0d} \delta_{kd} \gg \rho_{0r} \delta_{kr}$ and $c_{sd}^2 \approx 0$, to get

$$\begin{aligned} \ddot{\delta}_{kd} + 2H \dot{\delta}_{kd} - 4\pi G \rho_0 \delta_{kd} &= \left(\frac{\xi_d}{\rho_0} \right) \delta s_d \\ \ddot{\delta}_{kr} + 2H \dot{\delta}_{kr} + \left(\frac{c_{sr}^2 k^2}{a^2} \right) \delta_{kr} &= 4\pi G \rho_0 \delta_{kd} + \left(\frac{\xi_r}{\rho_{0r}} \right) \delta s_r. \end{aligned} \quad (6.26)$$

The first of these shows that perturbations in the Dark Matter evolve just as if the other components did not exist. Assuming no entropy fluctuation, $\delta s_{kd} = 0$, implies the density modes grow like $\delta_{kd} \propto a(t) \propto t^{2/3}$ for all $k \ll k_J$. Furthermore, since $k_J \sim H/c_{sd}$ with $c_{sd}^2 \ll 1$ we see that very many Dark Matter modes are unstable.

On the other hand, the second fluid component satisfies an oscillatory equation, but with oscillations which are driven by the growing dark matter density perturbations. What this implies for this second fluid depends on what this fluid is. For instance if the second fluid consists of the equilibrated fluid of coupled photons, electrons and nuclei which survives for $z \gtrsim 1100$ — then $c_{sr}^2 = \frac{1}{3}$ and so the fluid response to the growing Dark Matter perturbations depends on the mode's wave-number k . For $k/a \ll k_J \sim H$, δ_{kr} the $c_{sr}^2 k^2/a^2$ term may be neglected in comparison with the Dark Matter driving term, leading to the solution (if $\delta s_{kr} = 0$) $\delta_{kr}(t) = \delta_{kd}(t)$. This shows that the Dark Matter can drag the super-Hubble radiation modes along with it. By contrast, if $k/a \gg k_J \sim H$ then the Dark Matter driving term may be neglected, leading to the damped oscillatory solutions of eq. (6.17).

On the other hand, if the second fluid is the purely non-relativistic gas of neutral atoms expected for redshifts $z < 1100$, then $c_{sr}^2 \ll 1$ and for a much wider range of modes the appropriate solution is $\delta_{kr}(t) = \delta_{kd}(t)$. All of these modes are dragged along by the Dark Matter instability.

6.1.4 The Power Spectrum

The presence of unstable density fluctuations implies the universe does not remain precisely homogeneous and isotropic once matter domination begins, and so the view seen by observers like us depends on their locations in the universe relative to the fluctuations. For this reason, when comparing with observations it is less useful to try to track the detailed form of a specific fluctuation and instead better to characterize fluctuations by their statistical properties, since these can be more directly applied to observers without knowing their specific place in the universe. In particular we imagine there being an ensemble of density fluctuations, whose phases we assume to be uncorrelated and whose amplitudes are taken to be random variables.

On the observation side statistical inferences can be made about the probability distribution governing the distribution of fluctuation amplitudes by measuring statistical properties of the matter distribution observed around us. For instance, a useful statistic measures the mass-mass auto-correlation function

$$\xi(\mathbf{r} - \mathbf{r}') \equiv \frac{\langle \delta\rho(\mathbf{r}) \delta\rho(\mathbf{r}') \rangle}{\langle \rho \rangle^2}, \quad (6.27)$$

which might be measured by performing surveys of the positions of large samples of galaxies.¹⁷ When using (6.27) with observations the average $\langle \dots \rangle$ is interpreted as integration of one of the positions (say, \mathbf{r}') over all directions in the sky.¹⁸

When making predictions $\langle \dots \rangle$ instead is regarded as an average over whatever ensemble is thought to govern the statistics of the fluctuations δ_k . Fourier transforming $\delta\rho(\mathbf{r})/\langle\rho\rangle = \int d^3k \delta_k \exp[i\mathbf{k} \cdot \mathbf{r}]$ in comoving coordinates, as before, allows $\xi(\mathbf{r})$ to be related to the following ensemble average over the Fourier mode amplitudes, δ_k .

$$\xi(r) = \int \frac{d^3k}{(2\pi)^3} \langle |\delta_k|^2 \rangle \exp[i\mathbf{k} \cdot \mathbf{r}] = \frac{1}{2\pi^2} \int_0^\infty \frac{dk}{k} k^3 P_\rho(k) \left(\frac{\sin kr}{kr} \right), \quad (6.28)$$

which defines the density *power spectrum*: $P_\rho(k) := \langle |\delta_k|^2 \rangle$.

For homogeneous and isotropic backgrounds $P_\rho(k)$ depends only on the magnitude $k = |\mathbf{k}|$ and not on direction, and this is used above to perform the angular integrations. The average in these expressions is over the ensemble, and it is this average which collapses the right-hand side down to a single Fourier integral. The last equality motivates the definition

$$\Delta_\rho^2(k) := \frac{k^3}{2\pi^2} P_\rho(k). \quad (6.29)$$

A variety of observations over the years give the form of $P_\rho(k)$ as inferred from the distribution of structure around us, with results summarized in Figure ???. As this figure indicates, inferences about the shape of $P_\rho(k)$ for small k come from measurements of the temperature fluctuations in the CMB; those at intermediate k come from galaxy distributions as obtained through galaxy surveys and those at the largest k come from measurements of the how quasar light is absorbed by intervening Hydrogen gas clouds, the so-called Lyman- α ‘forest’. The reasons why different kinds of measurements control different ranges of k are illustrated in Figure 5, which shows how the distance accessible to observations is correlated with how far back one looks into the universe: measurements of distant objects in the remote past (*e.g.* the CMB) determine the shape of $P_\rho(k)$ for small k while measurements of more nearby objects in the more recent past (*e.g.* the Lyman- α forest) constrain $P_\rho(k)$ for larger k .

¹⁷A practical complication arises because although galaxies are relatively easy to count, most of the mass density is actually Dark Matter. Consequently assumptions are required to relate these to one another; the usual choice being that the galaxy and mass density functions are related to one another through a phenomenologically defined ‘bias’ factor.

¹⁸The density correlation function can also be measured using the temperature fluctuations of the CMB, because these fluctuations can be interpreted as redshifts acquired by CMB photons as they climb out of the gravitational potential wells formed by density fluctuations in nonrelativistic matter.

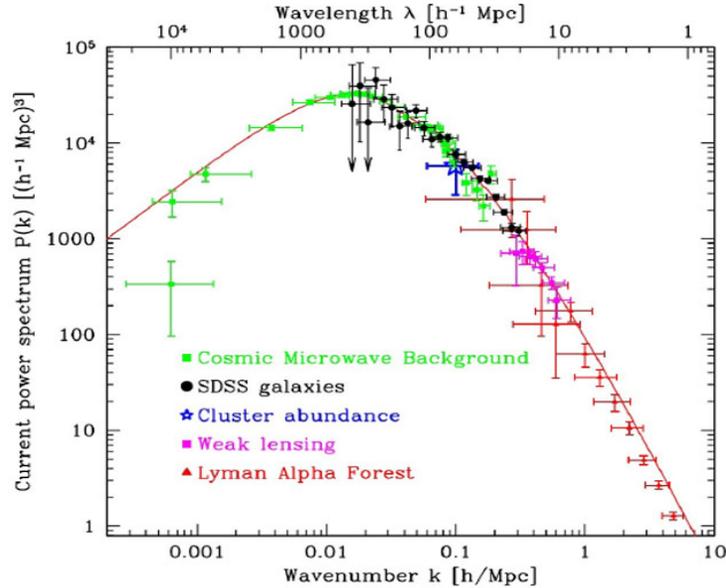


Figure 4. The power spectrum as obtained from measurements of the CMB spectrum, together with the SDSS Galaxy Survey, observations of abundances of galaxy clusters and Lyman- α measurements (taken from [?]).

The observations summarized in Figure 4 are well approximated by the phenomenological formula,

$$P(k) = \frac{A k^{n_s}}{(1 + \alpha k + \beta k^2)^2}, \quad (6.30)$$

where

$$\alpha = 16 \left(\frac{0.5}{\Omega h^2} \right) \text{ Mpc} \quad \text{and} \quad \beta = 19 \left(\frac{0.5}{\Omega h^2} \right)^2 \text{ Mpc}^2 \quad \text{and} \quad n_s = 0.97. \quad (6.31)$$

Here $h = H_0/(100 \text{ km/sec/Mpc}) \approx 0.7$, and $\Omega \approx 1$ denotes the present value of ρ/ρ_c . Given that $n_s \approx 1$ the observations suggest the power spectrum is close to linear, $P(k) \propto k$ for $k \ll k_*$ $\sim 0.07 \text{ Mpc}^{-1}$, and $P(k) \propto k^{-3}$ for $k \gg k_*$. The value k_* here is simply defined to be the place where $P_\rho(k)$ turns over and makes the transition from $P_\rho \propto k$ to $P_\rho \propto k^{-3}$.

As described below, there are good reasons to believe that the shape of $P_\rho(k)$ for $k \ll k_*$ represents the pattern of primordial fluctuations inherited from the very early universe, while the shape for $k > k_*$ reflects how fluctuations evolve in the later universe. Consequently observations are consistent with primordial fluctuations being close to¹⁹

¹⁹Close to but not equal to. Fits to Λ CDM cosmology establish n_s is significantly different from 1.

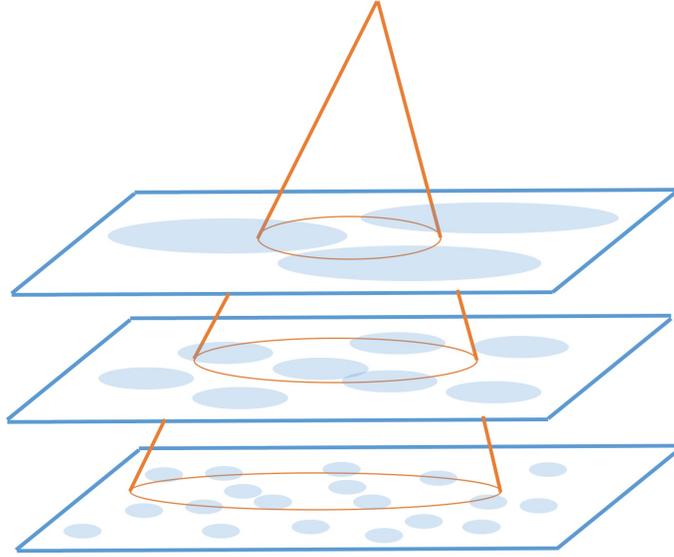


Figure 5. A sketch of several spatial slices intersecting the past light cone of an astronomer on Earth. The orange ovals indicate how the light cone has larger intersections with the spatial slices the further back one looks. The pale blue ovals indicate regions the size of the Hubble distance on each spatial slice. Correlations outside of these ovals (such as the uniformity of the CMB temperature) represent a puzzle for Λ CDM cosmology. The figure shows how later times (higher slices) have larger Hubble distances, as well as how observations only sample the largest distance scales on the most remote spatial slices. This illustrates why CMB measurements tend to constrain the power spectrum for small k while observations of more nearby objects (like galaxy distributions or the distribution of foreground Lyman- α Hydrogen gas clouds) constrain larger k .

a *Zel'dovich* spectrum, $P_\rho(k) = Ak$, corresponding to $n_s = 1$. As is seen below, the parameter n_s is predicted to be close to, but not equal to, unity by inflationary models.

For later purposes it proves more convenient to work with the power spectrum for the Newtonian gravitational potential, $\delta\phi$, that is related to $\delta\rho$ by Poisson's equation — the last of eqs. (??) — and so $\delta\phi_k \propto \delta_k/k^2$. Because of this relation their power spectra are related by $P_\phi(k) = P_\rho(k)/k^4$ as well as

$$\Delta_\phi^2(k) := \frac{k^3}{2\pi^2} P_\phi(k) = \frac{P_\rho(k)}{2\pi^2 k} \propto \begin{cases} k^{n_s-1} & \text{if } k \ll k_\star \\ k^{n_s-5} & \text{if } k \gg k_\star \end{cases}. \quad (6.32)$$

This last expression also clarifies why the choice $n_s = 1$ is called scale invariant. When

$n_s = 1$ the primordial ($k \ll k_*$) spectrum for $\Delta_\phi^2(k)$ becomes k -independent, as would be expected for a scale-invariant process.

6.1.5 Late-time structure growth

Before trying to explain the properties of the primordial part of the power spectrum — $\Delta_\phi^2(k) \propto Ak^{n_s-1}$ — a further digression is in order to explain the explanation for why the measured distribution has the peculiar hump-shaped form, bending at $k \simeq k_*$. This shape arises due to the processing of density fluctuations by their evolution in the subsequent universe, as is now described.

The key observations go back to the three criteria, given at the end of §??, for when fluctuating modes can grow. These state that the fluctuations that are most important are those involving nonrelativistic matter, although these remain frozen unless the universe is matter dominated and the mode number lies within the interval $H \ll k/a \ll H/c_s$. These conditions for growth superimpose a k -dependence on $P_\rho(k)$, for the following reasons.

The important wave-number k_* corresponds to the wave-number, k_{eq} , for which modes satisfy $k/a \sim H$ at the epoch of radiation-matter equality (which occurs at $z_{\text{eq}} = 3600$). Numerically, k_{eq} corresponds to a co-moving wave-number of order $k_{\text{eq}} \sim 0.07 \text{ Mpc}^{-1}$. What is important about this scale is that it divides modes (with $k > k_{\text{eq}}$) that re-enter the Hubble scale during radiation domination and those (with $k < k_{\text{eq}}$) that re-enter during matter domination.

Because they re-enter during matter domination, all Dark Matter fluctuation modes with $k < k_{\text{eq}}$ are free to begin growing immediately on re-entry and have done so ever since, at least until they either become nonlinear — when $\delta_k \sim \mathcal{O}(1)$ — or the universe reaches the very recent advent of Dark Energy domination. So the present-day power spectrum for these modes reflects the primordial one which was frozen into these modes long ago when they left the Hubble scale in the pre- Λ CDM era. It is these modes that reveal the primordial distribution

$$P(k) \propto k^{n_s} \quad (\text{for } k \ll k_{\text{eq}}). \quad (6.33)$$

By contrast, those modes with $k \gg k_{\text{eq}}$ re-enter the Hubble scale during the radiation-dominated epoch that precedes matter-radiation equality. The amplitude of these modes therefore remain frozen at their values at the time of re-entry, because they are unable to grow while the universe is radiation dominated. Consequently they remain stunted in amplitude relative to their longer-wavelength counterparts while

waiting for the universe to become matter-dominated, leading to a suppression of $P_\rho(k)$ for $k \gg k_{\text{eq}}$.

The *relative* stunting of large- k modes compared to small- k modes can be computed from the information that the unstable modes grow with amplitude $\delta_k(a) \propto a$ during matter-domination. For $k < k_{\text{eq}}$ this growth applies as soon as they cross the Hubble scale, while for $k > k_{\text{eq}}$ the modes cannot grow in this way until the transition from radiation to matter domination. As a result the relative size of two modes, one with $k_0 \ll k_{\text{eq}}$ and one with $k \gg k_{\text{eq}}$, is

$$\frac{\delta_k(a)}{\delta_{k_0}(a)} \propto \frac{\delta_k(a_k)(a/a_{\text{eq}})}{\delta_{k_0}(a_{k_0})(a/a_{k_0})} \propto \frac{\delta_k(a_k)(a/a_k)}{\delta_{k_0}(a_{k_0})(a/a_{k_0})} \left(\frac{k_{\text{eq}}}{k}\right)^2, \quad (6.34)$$

where a_k denotes the scale factor at the (k -dependent) epoch of re-entry, defined by $k = a_k H_k$. The first relation in (6.34) uses that modes in the numerator all start growing at the same time (radiation-matter equality), while those in the denominator grow for a k_0 -dependent amount a/a_{k_0} . The second relation then makes the k -dependence of the suppression a_k/a_{eq} in the numerator explicit, using the matter-domination evolution $aH \propto a^{-1/2}$ in the re-entry condition to conclude $k = a_k H_k \propto a_k^{-1/2}$ and so $a_k \propto k^{-2}$.

This leads to the expectation that the power spectrum has the form $P(k) = P_{\text{prim}}(k) \mathcal{T}(k)$, where $P_{\text{prim}}(k) = \langle |\delta_k(a)|^2 \rangle = \langle |\delta_k(a_k)|^2 \rangle (a/a_k)^2$ is the primordial power spectrum and $\mathcal{T}(k)$ is the transfer function that expresses the relative stunting of modes for $k \gg k_{\text{eq}}$. Keeping in mind that $P(k) \propto |\delta_k|^2$ the above discussion shows we expect $\mathcal{T}(k) \simeq 1$ for $k \ll k_{\text{eq}}$ and $\mathcal{T}(k) \simeq (k_{\text{eq}}/k)^4$ for $k \gg k_{\text{eq}}$. Given a primordial distribution $P_{\text{prim}}(k) \simeq A k^{n_s}$ this leads to

$$P_\rho(k) \propto \begin{cases} k^{n_s} & \text{if } k \ll k_\star \\ k^{n_s-4} & \text{if } k \gg k_\star \end{cases}, \quad (6.35)$$

much as is observed.

It is noteworthy that the success of the above argument contains more evidence for the existence of Dark Matter. For many modes $\delta_k \simeq \mathcal{O}(1)$ occurs before the present epoch, at which point nonlinear gravitational physics is expected to produce the large-scale structure actually seen in galaxy surveys. But the observed isotropy of the CMB implies the amplitude of $\delta_k(a_{\text{rec}})$ must have started off very small at the time photons last scattered from the Hydrogen gas at redshift $z_{\text{rec}} \sim 1100$. Given this small start; given the fact that modes cannot grow before matter-radiation equality; and given that instability growth is proportional to a , a minimum amount of time is required for fluctuations to become nonlinear early enough to account for the observed distribution

of nonlinear structure (like galaxies). Crucially, if Dark Matter did not exist then growth could not start until redshift $z_{\text{eq}}(\text{baryons only}) \simeq 480$ — *c.f.* (??) — which does not leave enough time. But the presence of Dark Matter moves back the epoch of radiation-matter equality to $z_{\text{eq}} \simeq 3600$ — *c.f.* (??) — giving sufficient time for nonlinear structure to form at the required scales.

The story of late-time fluctuations is even much richer than the above would lead one to believe, with detailed comparisons between observations and theory. A spectacular example of this is provided by the observation of ‘baryon acoustic oscillations’ (BAO), which are observed correlations between the distribution of galaxies and the distribution of CMB temperature fluctuations. The physical origin of these correlations lies in the coupled late-time evolution of perturbations in the Dark Matter and baryon-radiation fluid. Once fluctuations in the baryon-photon fluid begin to be free to oscillate the local dark Dark Matter evolution acts as a forcing term. This sends out a sound wave in the density of the baryon-photon fluid that initially propagates at a significant fraction of the speed of light, due to the dominance of the photon entropy in this fluid. But the speed of sound for the baryons drops like a rock once the baryons and photons decouple from one another at recombination, causing the sound wave to stall. The resulting correlation has been observed, and its properties again confirm the Λ CDM model with values for the model parameters consistent with other determinations.

6.1.6 Hot Dark Matter

The comparison between the power spectrum and observations rules out the possibility of the Dark Matter consisting of particles which were relativistic at the times when they decoupled (such as was the case for neutrinos for example). This is because the relativistic motion of such a particle would wipe out any density perturbations on scales shorter than the distance these particles can have travelled by the time they decouple — a process called *free streaming*. This corresponds to the proper distance $D_{\text{min}} \sim t_d \sim \frac{1}{2} H_d^{-1}$, where H_d is the Hubble parameter at the time this particle decouples. Equivalently, this corresponds to the co-moving length scale $\ell_{\text{min}} = D_{\text{min}}/a(t_d)$, leading to the prediction that $P(k) \approx 0$ for all $k \gtrsim 2\pi/\ell_{\text{min}}$. The absence of such a suppression of $P(k)$ in observations precludes there being more than about a 10% contribution of relativistic particles to the Dark Matter.

6.2 Primordial fluctuations from inflation

The previous discussion shows that fluctuations in the Λ CDM model also provide a successful description of structure in the universe, but only given the initial condition of a primordial spectrum of fluctuations having a specific power-law form: $P_\rho(k) \simeq A_s k^{n_s}$ (or $\Delta_\phi^2(k) \simeq A_s k^{n_s-1}$). It again falls to the earlier universe to explain why primordial fluctuations should have this specific form, and why it should be robust against the many poorly understood details governing the physics of this earlier epoch.

It is remarkable that there is evidence that an earlier period of inflationary expansion can also explain this initial distribution of fluctuations. This section provides a sketch of this evidence. Since the modes of interest start off during Λ CDM outside the Hubble length, $k \ll aH$, and are known to be small, their evolution can be tracked into earlier epochs using linear perturbation theory. Because the modes are super-Hubble in size the treatment must be relativistic, and so involves linearizing the coupled Einstein-matter field equations. The first part of this section sketches how this super-Hubble evolution works, and shows how to relate the primordial fluctuations that re-enter the Hubble scale to those that exit the Hubble scale during the inflationary epoch (see Figure 3).

At first sight this just pushes the problem back to an earlier time, requiring an explanation why a particular pattern of fluctuations should exist during inflation. Even worse, within the classical approximation there is good reason to believe there should be no fluctuations at all leaving at horizon exit during inflation. This is because the exponential growth of the scale factor, $a \propto e^{Ht}$, during inflation is absolutely ruthless in ironing out any spacetime wrinkles since momentum-dependent terms like $(k/a)^2$ in the field equations go to zero so quickly.

But the key words in the above are “within the classical approximation”. Quantum fluctuations are *not* ironed away during inflation, and persist at a level proportional to the Hubble scale. Because this Hubble scale is approximately constant the resulting fluctuations are largely scale-independent, providing a natural explanation for why primordial fluctuations seem to be close to the Zel’dovich spectrum. But H during inflation also cannot be exactly constant since inflation must end eventually. In the explicit models examined earlier the time-dependence of H arises at a level suppressed by the slow-roll parameters ϵ and η and so deviations from scale invariance should arise at the few percent level. Because of this we shall find below that the prediction for n_s in inflationary models is a bit smaller than unity, naturally agreeing with the observed value $n_s \simeq 0.97$.

6.2.1 Linear evolution of metric-inflaton fluctuations

The first task is to evolve fluctuations forward from the epoch of inflationary horizon exit until they re-enter during the later Hot Big Bang era. In particular our focus is on the perturbations of the metric, $\delta g_{\mu\nu}$, since these include perturbations of the Newtonian potential and so also the density fluctuations whose power spectrum is ultimately measured. The discussion here follows that of

The symmetry of the FRW background allows the fluctuations of the metric to be classified by their rotational properties, with fluctuations of different spin not mixing at linear order in the field equations. Fluctuations of the metric come in three such kinds: *scalar*, *vector* and *tensor*. Specializing to a spatially flat FRW background and transforming to conformal time, $\tau = \int dt/a$, the scalar perturbations may be written

$$\delta_S g_{\mu\nu} = a^2 \begin{pmatrix} 2\phi & \partial_j \mathcal{B} \\ \partial_i \mathcal{B} & 2\psi \delta_{ij} + \partial_i \partial_j \mathcal{E} \end{pmatrix}, \quad (6.36)$$

while the vector and tensor ones are

$$\delta_V g_{\mu\nu} = a^2 \begin{pmatrix} 0 & \mathcal{V}_j \\ \mathcal{V}_i & \partial_i \mathcal{W}_j + \partial_j \mathcal{W}_i \end{pmatrix} \quad \text{and} \quad \delta_T g_{\mu\nu} = a^2 \begin{pmatrix} 0 & 0 \\ 0 & h_{ij} \end{pmatrix}. \quad (6.37)$$

Here all vectors are divergence-free, as is the tensor (which is also traceless). To these are added the fluctuations in the inflaton field, $\delta\varphi$.

There is great freedom to modify these functions by performing infinitesimal coordinate transformations, so it is useful to define the following combinations that are invariant at linearized order:

$$\begin{aligned} \Phi &= \phi - \frac{1}{a} [a(\mathcal{B} - \mathcal{E}')]', & \Psi &= \psi + \frac{a'}{a} (\mathcal{B} - \mathcal{E}') \\ \delta\chi &= \delta\varphi - \varphi' (\mathcal{B} - \mathcal{E}'), & V_i &= \mathcal{V}_i - \mathcal{W}_i \quad \text{and} \quad h_{ij}, \end{aligned} \quad (6.38)$$

in terms of which all physical inferences can be drawn. Here primes denote differentiation with respect to conformal time, τ . Notice that Φ , Ψ and V_i reduce to ϕ , ψ and \mathcal{V}_i in the gauge choice where $\mathcal{B} = \mathcal{E} = \mathcal{W}_i = 0$, and so Φ is the relativistic generalization of the Newtonian potential.

These functions are evolved forward in time by linearizing the relevant field equations:

$$\square\varphi - V'(\varphi) = 0 \quad \text{and} \quad R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} = \frac{T_{\mu\nu}}{M_p^2}, \quad (6.39)$$

and provided we use the invariant stress-energy perturbations,

$$\begin{aligned}
\delta\mathcal{T}^0_0 &= \delta T^0_0 - [t^0_0]' (\mathcal{B} - \mathcal{E}'), \\
\delta\mathcal{T}^0_i &= \delta T^0_i - \left[t^0_0 - \frac{1}{3} t^k_k \right] \partial_i (\mathcal{B} - \mathcal{E}'), \\
\delta\mathcal{T}^i_j &= \delta T^i_j - [t^i_j]' (\mathcal{B} - \mathcal{E}'),
\end{aligned} \tag{6.40}$$

(where t^μ_ν denotes the background stress-energy), the results can be expressed purely in terms of the gauge-invariant quantities, eqs. (6.38).

The equations which result show that in the absence of vector stress-energy perturbations (*i.e.* if $\delta\mathcal{T}^0_i$ is a pure gradient - as would be the case for perturbed inflaton), then vector perturbations, V_i , are not sourced, and decay very rapidly in an expanding universe, allowing them to be henceforth ignored. Similarly, in the absence of off-diagonal stress-energy perturbations (*i.e.* if $\delta\mathcal{T}^i_j = \delta p \delta^i_j$) it is also generic that $\Psi = \Phi$.

Switching back to FRW time, the equations which govern the evolution of tensor modes then become (after Fourier transforming)

$$\ddot{h}_{ij} + 3H \dot{h}_{ij} + \frac{k^2}{a^2} h_{ij} = 0, \tag{6.41}$$

showing that these evolve independent of all other fluctuations. Such primordial tensor fluctuations can be observable if they survive into the later universe, since the differential stretching of spacetime that they predict can contribute observably to the polarization of CMB photons that pass through them. The search for evidence for this type of primordial tensor fluctuations is active and ongoing, and (as is shown below) is expected in inflation to be characterized by a near scale-invariant tensor power spectrum,

$$P_h(k) = A_T k^{n_T}, \tag{6.42}$$

with n_T close to zero.

The equations evolving the scalar fluctuations are more complicated and similarly reduce to

$$\begin{aligned}
\delta\ddot{\chi} + 3H\delta\dot{\chi} + \frac{k^2}{a^2}\delta\chi + V''(\varphi)\delta\chi - 4\dot{\varphi}\dot{\Phi} + 2V'(\varphi)\Phi &= 0 \\
\text{and} \quad \dot{\Phi} + H\Phi &= \frac{\dot{\varphi}}{2M_p^2} \delta\chi.
\end{aligned} \tag{6.43}$$

The homogeneous background fields themselves satisfy the equations

$$\ddot{\varphi} + 3H\dot{\varphi} + V'(\varphi) = 0 \quad \text{and} \quad 3M_p^2 H^2 = \frac{1}{2}\dot{\varphi}^2 + V(\varphi). \tag{6.44}$$

These expressions show that although Φ and $\delta\chi$ would decouple from one another if expanded about a static background (for which $\dot{\varphi} = V' = 0$), they do not when the background is time-dependent.

6.2.2 Slow-roll evolution of scalar perturbations

The character of the solutions of these equations depends strongly on the size of k/a relative to H , since this dictates the extent to which the frictional terms can compete with the spatial derivatives. As usual the two independent solutions for $\delta\chi$ that apply when $k/a \gg H$ describe damped oscillations

$$\delta\chi_k \propto \frac{1}{a\sqrt{k}} \exp \left[\pm ik \int^t \frac{dt'}{a(t')} \right]. \quad (6.45)$$

Our interest during inflation is in the limit $k/a \ll H$ in a slow-roll regime for which $\delta\ddot{\chi}$, $\ddot{\varphi}$ and $\dot{\Phi}$ can all be neglected. In this limit the scalar evolution equations simplify to

$$3H\delta\dot{\chi} + V''(\varphi)\delta\chi + 2V'(\varphi)\Phi \simeq 0 \quad \text{and} \quad 2M_p^2 H \Phi \simeq \dot{\varphi} \delta\chi, \quad (6.46)$$

and have approximate solutions (after Fourier transformation) of the form

$$\delta\chi_k \simeq C_k \frac{V'(\varphi)}{V(\varphi)} \quad \text{and} \quad \Phi_k \simeq -\frac{C_k}{2} \left(\frac{V'(\varphi)}{V(\varphi)} \right)^2. \quad (6.47)$$

where C_k is a (potentially k -dependent) constant of integration. Since the background fields satisfy $M_p V'/V = \sqrt{2\epsilon}$ these equations show how the amplitude of $\delta\chi_k$ and Φ_k during inflation track the evolution of the slow-roll parameter, ϵ , for super-Hubble modes, and therefore tend to grow in amplitude as inflation eventually draws to a close.

We have two remaining problems: (i) What is the origin of the initial fluctuations at horizon exit? (ii) How do we evolve fluctuations from the end of inflation through to the later epoch of horizon re-entry? The latter of these seems particularly vicious since it *a priori* might be expected to depend on the many details involved in getting the universe from its inflationary epoch to the later Hot Big Bang.

6.2.3 Post-Inflationary evolution

For the case of single-field inflation discussed here, post-inflationary evolution of the fluctuation Φ actually turns out to be quite simple. This is because it can be shown that when $k \ll aH$ the quantity

$$\zeta = \Phi + \frac{2}{3} \left(\frac{\Phi + \dot{\Phi}/H}{1+w} \right) = \frac{1}{3(1+w)} \left[(5+3w)\Phi + \frac{2\dot{\Phi}}{H} \right], \quad (6.48)$$

is *conserved*, inasmuch as $\dot{\zeta} \simeq 0$ for $k \rightarrow 0$.

This result follows schematically because the perturbed metric can be written as proportional to $e^{\zeta} g_{ij}$ and so spatially constant ζ is indistinguishable from the background scale factor, $a(t)$. Conservation has been proven under a wide variety of assumptions but the form used here assumes that the background cosmology satisfies an equation of state $p = w\rho$, but w is *not* assumed to be constant. The same result is known not to be true if there were more than a single scalar field evolving.

Conservation of ζ is a very powerful result because it can be used to evolve fluctuations using $\zeta(t_i) = \zeta(t_f)$, assuming only that they involve a single scalar field, and that the modes in question are well outside the horizon: $k/a \ll H$. Furthermore, although $\dot{\Phi}$ in general becomes nonzero at places where w varies strongly with time, this time dependence quickly damps due to Hubble friction for modes outside the Hubble scale.

We may therefore for most of the universe's history also neglect the dependence of ζ on $\dot{\Phi}$ provided we restrict t_i and t_f to epochs during which w is roughly constant. This allows the expression $\zeta(t_i) = \zeta(t_f)$ to be simplified to

$$\Phi_f = \frac{1 + w_f}{1 + w_i} \left(\frac{5 + 3w_i}{5 + 3w_f} \right) \Phi_i, \quad (6.49)$$

where $w_i = w(t_i)$ and $w_f = w(t_f)$, implying in particular $\Phi_f = \Phi_i$ whenever $w_i = w_f$. Similarly, the values of Φ deep within radiation and matter dominated phases are related by $\Phi_{\text{mat}} \simeq \frac{9}{10} \Phi_{\text{rad}}$.

To infer the value of Φ in the later Hot Big Bang era we choose t_i just after horizon exit (where a simple calculation shows $w_i \simeq -1 + \frac{2}{3} \epsilon_{\text{he}}$, with ϵ_{he} the slow-roll parameter at horizon exit). t_f is then chosen in the radiation dominated universe (where $w_f = \frac{1}{3}$), either just before horizon re-entry for the mode of interest, or just before the transition to matter domination, whichever comes first. Eqs. (6.47) and (6.49) then imply

$$\Phi_f \simeq \left(\frac{2\Phi}{3\epsilon} \right)_{\text{he}}. \quad (6.50)$$

It remains to grapple with what should be expected for the initial condition for Φ at horizon exit.

6.2.4 Quantum origin of fluctuations

The primordial fluctuation amplitude derived in this way depends on the integration constants C_k , which are themselves set by the initial conditions for the fluctuation at horizon exit, during inflation. But why should this amplitude be nonzero given

that all previous evolution is strongly damped, as in eq. (6.45)? The result remains nonzero (and largely independent of the details of earlier evolution) because quantum fluctuations in $\delta\chi$ continually replenish the perturbations long after any initial classical configurations have damped away.

The starting point for the calculation of the amplitude of scalar perturbations is the observation that the inflaton and metric fields whose dynamics we are following are quantum fields, not classical ones. For instance, for spatially-flat spacetimes the linearized inflaton field, $\delta\chi$, is described by the operator

$$\delta\chi(x) = \int \frac{d^3k}{(2\pi)^3} \left[\mathbf{c}_k u_k(t) e^{i\mathbf{k}\cdot\mathbf{r}/a} + \mathbf{c}_k^* u_k^*(t) e^{-i\mathbf{k}\cdot\mathbf{r}/a} \right], \quad (6.51)$$

where the expansion is in a basis of eigenmodes of the scalar field equation in the background metric, $u_k(t) e^{i\mathbf{k}\cdot\mathbf{x}}$, labelled by the co-moving momentum \mathbf{k} . For constant H the time-dependent mode functions are

$$u_k(t) \propto \frac{H}{k^{3/2}} \left(i + \frac{k}{aH} \right) \exp\left(\frac{ik}{aH} \right), \quad (6.52)$$

which reduces to the standard flat-space form, $u_k(t) \propto a^{-1} k^{-1/2} e^{-ik \int dt/a}$, when $k/a \gg H$. [This is perhaps easiest to see using conformal time, for which $\exp(ik/aH) = \exp(-ik\tau)$, or more directly by using $\exp(-ik \int dt/a) = \exp(ik/aH)$ when $a \propto e^{Ht}$.] The quantities \mathbf{c}_k and their adjoints \mathbf{c}_k^* are *annihilation* and *creation operators*, which define the adiabatic vacuum state, $|\Omega\rangle$, through the condition $\mathbf{c}_k|\Omega\rangle = 0$ (for all \mathbf{k}).

The $\delta\chi$ auto-correlation function in this vacuum, $\langle \delta\chi(x)\delta\chi(x') \rangle$, describes the quantum fluctuations of the field amplitude in the quantum ground state, and the key assumption is that the quantum statistics of the mode leaving the horizon during inflation agrees with the classical fluctuations of the field $\delta\chi$ after evolving outside of the Hubble scale. This assumes the quantum fluctuations to be decohered (for preliminary discussions see ref. [? ?]) into classical distribution for $\delta\chi$ sometime between horizon exit and horizon re-entry.

It turns out that during inflation interactions with the bath of short-wavelength, sub-Hubble modes is extremely efficient at decohering the quantum fluctuations of long-wavelength, super-Hubble modes. As is usual when a system is decohered through interactions with an environment, the resulting classical distribution is normally defined for the ‘pointer basis’, that diagonalizes the interactions with the environment. It turns out that the freezing of super-Hubble modes has the effect of making them very classical (WKB-like), and so ensure the fields canonical momenta become functions of the fields

themselves. This ensures that it is always the field basis that diagonalizes any local interactions, and so guarantees that quantum fluctuations become classical fluctuations for the fields (like $\delta\chi$) rather than (say) their canonical momenta.

The upshot is that after several e -foldings even very weak interactions (like gravitational strength ones) eventually convert quantum fluctuations into classical statistical fluctuations for the classical field, φ , about its spatial mean. For practical purposes, this means in the above calculations we can simply use the initial condition $|\delta\chi_k| \sim [\langle\delta\chi_k\delta\chi_{-k}\rangle]^{1/2} \propto |u_k(t)|$. For observational purposes what matters is that the classical variance of these statistical fluctuations is well-described by the corresponding quantum auto-correlations – a property that relies on the kinds of ‘squeezed’ quantum states that arise during inflation

Evaluating $\delta\chi_k \sim u_k$ at t_{he} (where $k = aH$) and equating the result to the fluctuation of eq. (6.47) allows the integration constant in this equation to be determined to be

$$C_k = u_k(t_{\text{he}}) \left(\frac{V}{V'} \right)_{\varphi_{\text{he}}}, \quad (6.53)$$

where both t_{he} and $\varphi_{\text{he}} = \varphi(t_{\text{he}})$ implicitly depend on k . Using this to compute Φ_k in eq. (6.47) then gives

$$\Phi_k(t) = -\frac{1}{2}u_k(t_{\text{he}}) \left(\frac{V}{V'} \right)_{\varphi_{\text{he}}} \left(\frac{V'}{V} \right)_{\varphi(t)}^2 = -\epsilon(t) \left(\frac{u_k}{\sqrt{2\epsilon} M_p} \right)_{t_{\text{he}}}. \quad (6.54)$$

In particular, evaluating at $t = t_{\text{he}}$ then gives

$$\Phi_k(t_{\text{he}}) = - \left(\frac{u_k}{M_p} \sqrt{\frac{\epsilon}{2}} \right)_{t_{\text{he}}}. \quad (6.55)$$

6.2.5 Predictions for the scalar power spectrum

We are now in a situation to pull everything together and compute in more detail the inflationary prediction for the properties of the primordial fluctuation spectrum. Using (6.55) in (6.50) gives

$$\Phi_k(t_f) \simeq \left(\frac{2\Phi}{3\epsilon} \right)_{\text{he}} = - \left(\frac{2u_k}{3\sqrt{2\epsilon} M_p} \right)_{t_{\text{he}}}. \quad (6.56)$$

Using this in the definition of the dimensionless power spectrum for Φ , $\Delta_{\Phi}^2 = k^3 P_{\Phi}/(2\pi^2)$, then leads to

$$\Delta_{\Phi}^2(k) = \frac{k^3 |\Phi_k(t_f)|^2}{2\pi^2} \propto \frac{k^3 |u_k(t_{\text{he}})|^2}{\pi^2 \epsilon(\varphi_{\text{he}}) M_p^2}. \quad (6.57)$$

Once the order-unity factors are included one finds

$$\Delta_{\Phi}^2(k) = \frac{k^3 P_{\Phi}(k)}{2\pi^2} = \left(\frac{H^2}{8\pi^2 M_p^2 \epsilon} \right)_{\text{he}} = \left(\frac{V}{24\pi^2 M_p^4 \epsilon} \right)_{\text{he}}, \quad (6.58)$$

It is the quantity V/ϵ evaluated at Hubble exit that controls the amplitude of density fluctuations, and so is to be compared with the observed power spectrum of scalar density fluctuations,

$$\Delta_{\Phi}^2(k) = \Delta_{\Phi}^2(\hat{k}) \left(\frac{k}{\hat{k}} \right)^{n_s}, \quad (6.59)$$

where $n_s = 0.968 \pm 0.006$ and

$$\Delta_{\Phi}^2(\hat{k}) = 2.28 \times 10^{-9}, \quad (6.60)$$

is the amplitude evaluated at the reference ‘pivot’ point $k = \hat{k} \sim 7.5 a_0 H_0$. In terms of V this implies

$$\left(\frac{V}{\epsilon} \right)_{\text{pivot}}^{1/4} = 6.6 \times 10^{16} \text{ GeV}, \quad (6.61)$$

for the epoch when the pivot scale underwent Hubble exit. The smaller ϵ becomes, the smaller the required potential energy during inflation. For $\epsilon \sim 0.01$ we have $V \sim 2 \times 10^{16}$ GeV. This is titillatingly close to the scale where the couplings of the three known interactions would unify in Grand Unified models, which may indicate a connection between the physics of Grand Unification and inflation.²⁰

Notice also that the size of $\Delta_{\Phi}^2(k)$ is set purely by H and ϵ at horizon exit, and these only depend weakly on k (through their weak dependence on time) during near-exponential inflation. This is what ensures the approximate scale-invariance of the primordial power spectrum which inflation predicts for the later universe. To pin down the value of n_s more precisely notice that the power-law form of (6.59) implies

$$n_s - 1 \equiv \left. \frac{d \ln \Delta_{\Phi}^2}{d \ln k} \right|_{\text{he}}. \quad (6.62)$$

To evaluate this during slow-roll inflation use the condition $k = aH$ (and the approximate constancy of H during inflation) to write $d \ln k = H dt$. Since the right-hand side of eq. (6.58) depends on k and t only through its dependence on φ , it is

²⁰Of course, V can be much smaller if ϵ is smaller as well, or if primordial fluctuations actually come from another source.

convenient to use the slow-roll equations, eq. (5.17) to further change variables from t to φ : $dt = d\varphi/\dot{\varphi} \simeq -(3H/V') d\varphi$, and so

$$\frac{d}{d \ln k} = -M_p^2 \left(\frac{V'}{V} \right) \frac{d}{d\varphi} = \sqrt{2\epsilon} M_p \frac{d}{d\varphi}. \quad (6.63)$$

Performing the φ derivative using (6.58) finally gives the following relation between n_s and the slow-roll parameters, ϵ and η

$$n_s - 1 = -6\epsilon + 2\eta, \quad (6.64)$$

where the right-hand side is evaluated at $\varphi = \varphi_{\text{he}}$. For single-field models the right-hand side is negative and typically of order 0.01, agreeing well with the measured value $n_s \simeq 0.97$.

6.2.6 Tensor fluctuations

A similar story goes through for the tensor fluctuations, though without the complications involving mixing between $\delta\chi$ and Φ . Tensor modes are also directly generated by quantum fluctuations, in this case where the vacuum is the quantum state of the graviton part of the Hilbert space. Although tensor fluctuations have not yet been observed, they are potentially observable through the polarization effects they produce as CMB photons propagate through them to us from the surface of last scattering.

Just like for scalar fluctuations, for each propagating mode the amplitude of fluctuations in the field h_{ij} is set by $H/(2\pi)$, but because there is no longer a requirement to mix with any other field (unlike Φ , which because it does not describe a propagating particle state has to mix with the fluctuating field $\delta\chi$), the power spectrum for tensor perturbations depends only on H^2 rather than on H^2/ϵ . Repeating the above arguments leads to the following dimensionless tensor power spectrum

$$\Delta_h^2(k) = \frac{8}{M_p^2} \left(\frac{H}{2\pi} \right)^2 = \frac{2V}{3\pi^2 M_p^4}. \quad (6.65)$$

This result is again understood to be evaluated at the epoch when observable modes leave the horizon during inflation, $\varphi = \varphi_{\text{he}}$.

Should both scalar and tensor modes be measured, a comparison of their amplitudes provides a direct measure of the slow-roll parameter ϵ . This is conventionally quantified in terms of a parameter r , defined as a ratio of the scalar and tensor power spectra

$$r := \frac{\Delta_h^2}{\Delta_\Phi^2} = 16\epsilon. \quad (6.66)$$

The absence of evidence for these perturbations to date places an upper limit: $r \lesssim 0.07$ and so $\epsilon \lesssim 0.004$. Because ϵ appears to be so small, the measured value for n_s used with (6.64) permits an inference of how large η can be. Fitting to a global data set gives a less stringent bound on r leading to

$$\epsilon < 0.012 \text{ (95\% CL)}; \quad \text{and} \quad \eta = -0.0080^{+0.0080}_{-0.0146} \text{ (68\% CL)}, \quad (6.67)$$

and it is this incipient evidence that $\epsilon \neq \eta$ that drives the tension with some of the model predictions described earlier. This information is given pictorially in Figure 6.

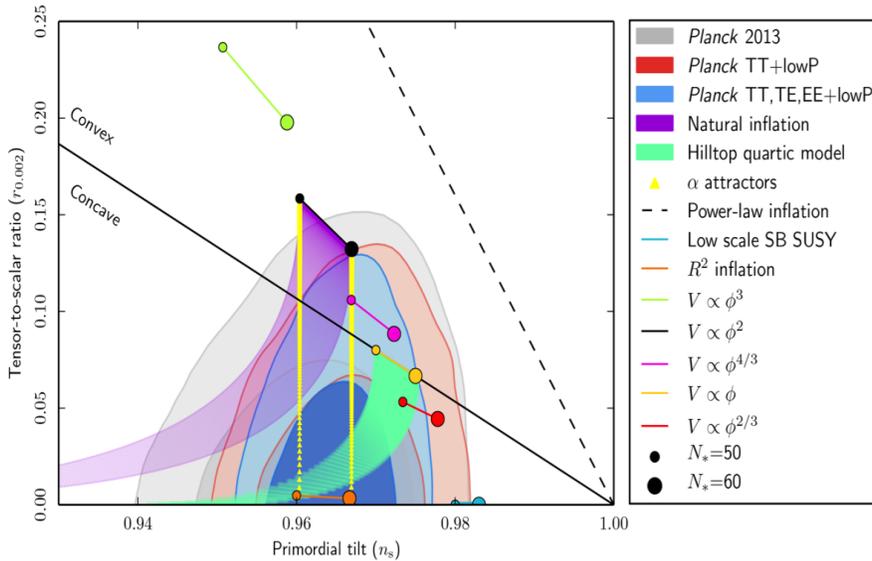


Figure 6. A comparison of a variety of inflationary models to a suite of cosmological measurements (taken from [?]), with the ellipses showing the observationally preferred values for the scalar-to-tensor ratio, r , and primordial ‘tilt’, n_s . Each model is portrayed as giving a range of values rather than a single point, with N_e chosen for a variety of assumptions about the nature of reheating (*c.f.* eq. (5.10)). The model labelled ‘Natural Inflation’ corresponds to the ‘pseudo-Goldstone axion’ model described in the text, while the ones called ‘ α -attractors’ represent the text’s ‘pseudo-Goldstone dilaton’ models.

The detection of tensor modes in principle also allows a measurement of the k dependence of their power spectrum. This is usually quantified in terms of a tensor spectral index, n_T , defined by eq. (6.42) and so

$$n_T \equiv \frac{d \ln \Delta_h^2}{d \ln k} = -2\epsilon = -\frac{r}{8}, \quad (6.68)$$

where the second-last equality evaluates the derivative within inflation as before by changing variables from k to φ .

Ultimately single-field models have three parameters: ϵ , η and the Hubble scale during inflation, H_I . But the scalar and tensor fluctuation spectra provide four observables: A_s , A_T , n_s and n_T . The ability to describe four observables using just three parameters implies a predicted relation amongst the observables: $n_T = -r/8$ (as seen from (6.68)). This is a robust prediction shared by all single-field slow-roll inflationary models.

A General Relativity

This appendix collects some useful results from General Relativity.

Within General Relativity the laws of nature are expressed in a coordinate-invariant way and so naturally lend themselves to being described in terms of tensors and the notions of differential geometry described in previous sections. In particular, Einstein postulated that all of the effects of gravity are completely encoded within the metric not of space, but of spacetime.

A.1 Metrics

The interpretation of such a metric, $g_{\mu\nu}$, is similar to the interpretation given above for the flat Minkowski metric in the spacetime formulation of special relativity. That is, the invariant proper distance spanned by an infinitesimal coordinate difference dx^μ is given by

$$ds^2 = g_{\mu\nu}(x) dx^\mu dx^\nu, \quad (\text{A.1})$$

where at every point the symmetric matrix $g_{\mu\nu}$ has one negative eigenvalue and three positive eigenvalues. This line element can be used just as was done in special relativity, to compute the proper distance along space-like curves (for which $g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu > 0$); the proper time, $d\tau^2 = -ds^2$, along time-like curves (for which $g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu < 0$); and to identify the trajectories of light-like curves (for which $g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu = 0$).

Since at any given point, x_0 , $g_{\mu\nu}(x_0)$ is just a symmetric matrix, it is always possible to change coordinates in such a way as to ensure that $g_{\mu\nu}(x_0) = \eta_{\mu\nu}$. This expresses the principle of equivalence, inasmuch as it ensures that there is always an observer (typically a freely-falling observer) for whom all of the non-gravitational laws of nature near a given point may be expressed just as if gravity did not exist (as in special relativity).

In general it is *not* possible to find a similar class of observers for whom $g_{\mu\nu} = \eta_{\mu\nu}$ simultaneously for all of the points throughout an entire region of spacetime, rather than just at a particular point. Such observers may only be found if the spacetime is flat, in the precise sense that its Riemann curvature tensor vanishes, $R^\mu{}_{\nu\lambda\rho} = 0$. This tensor is defined in terms of the metric $g_{\mu\nu}$ in precisely the same way as the Riemann tensor $R^i{}_{jkl}$ was related to the spatial metric g_{ij} in the previous sections. Since this means that it is curvature which keeps a family of observers from all experiencing only non-gravitational interactions, it is clearly curvature which is the physical expression of gravity in General Relativity.

In order to make this connection precise Einstein had to specify two new laws of nature, which have been phrased (by John Wheeler) as the statements that “Spacetime tells Matter how to move” and “Matter tells Spacetime how to curve”. The remainder of this section describes these connections more explicitly.

A.2 Particle Motion

Within General Relativity the postulate which tells how particle motions respond to the curvature of space-time states that particle trajectories follow the geodesics in the absence of all non-gravitational forces. This implies the equation of motion which defines the trajectory, $x^\mu(s)$, of a freely-falling particle is the geodesic equation

$$\frac{d^2x^\mu}{ds^2} + \Gamma^\mu{}_{\nu\lambda}[x(s)] \left(\frac{dx^\nu}{ds}\right) \left(\frac{dx^\lambda}{ds}\right) = 0, \quad (\text{A.2})$$

where as usual the Christoffel symbols, $\Gamma^\mu{}_{\nu\lambda}$, are defined by

$$\Gamma^\mu{}_{\nu\lambda} = \frac{1}{2} g^{\mu\sigma} (\partial_\nu g_{\sigma\lambda} + \partial_\lambda g_{\sigma\nu} - \partial_\sigma g_{\nu\lambda}) . \quad (\text{A.3})$$

Notice that a first integral of these equations may be obtained by taking the inner product of eq. (A.2) with the velocity vector, dx^μ/ds , and using eq. (A.3) to simplify the result:

$$\begin{aligned} 0 &= g_{\mu\nu} \left(\frac{dx^\mu}{ds}\right) \left[\frac{d^2x^\nu}{ds^2} + \Gamma^\nu{}_{\alpha\beta} \left(\frac{dx^\alpha}{ds}\right) \left(\frac{dx^\beta}{ds}\right) \right] \\ &= g_{\mu\nu} \left(\frac{dx^\mu}{ds}\right) \left(\frac{d^2x^\nu}{ds^2}\right) + \frac{1}{2} \partial_\alpha g_{\mu\beta} \left(\frac{dx^\mu}{ds}\right) \left(\frac{dx^\alpha}{ds}\right) \left(\frac{dx^\beta}{ds}\right) \\ &= \frac{1}{2} \frac{d}{ds} \left[g_{\mu\nu} \left(\frac{dx^\mu}{ds}\right) \left(\frac{dx^\nu}{ds}\right) \right] . \end{aligned} \quad (\text{A.4})$$

This shows that the quantity $g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu$ is a constant along a geodesic²¹ and so in particular its sign does not change. As a result it follows that if a particle initially starts out moving at the local speed of light, $g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu = 0$, then this is always true. Similarly, if a particle initially moves more slowly than light, $g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu < 0$, then this is always true. Furthermore, in this case it is also always possible to re-scale the parameter, $s \rightarrow \lambda\tau$, to ensure that $g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu = -1$, in which case τ has the physical interpretation of counting the proper time along the trajectory of the falling particle.

A.3 Einstein’s Field Equations

The second ingredient to General Relativity is the law which expresses how sources of mass and energy give rise to gravitational fields. A clue to how this is done comes from the observation that special relativity may be regarded as the motion of particles within the geometry of flat space, described by the Minkowski metric, eq. (??). In this way of thinking a gravitational field is what makes it impossible to find observers for whom physics everywhere is described by special relativity, and gravity must be represented by the curvature of spacetime. This leads to the expectation that the law of gravity must relate spacetime curvature to the local distribution of energy, as expressed by the stress-energy tensor, $T_{\mu\nu}$.

The Field Equations

We are now in a position to state Einstein’s law of gravity. Einstein proposed that the spacetime curvature tensor, $R^\mu{}_{\nu\lambda\rho}$, is related to the local distribution, $T_{\mu\nu}$, of stress-energy by the following field equations:

$$R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} = -8\pi G T_{\mu\nu}, \quad (\text{A.5})$$

where $R_{\mu\nu} = R^\lambda{}_{\mu\lambda\nu}$ is the spacetime’s Ricci tensor and G is Newton’s gravitational constant. This represents the relativistic generalization of the Newtonian field equation for the gravitational potential, ϕ :

$$\nabla^2\phi = 4\pi G \rho. \quad (\text{A.6})$$

References

- [1] D. N. Spergel *et al.* [WMAP Collaboration], “First Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Determination of Cosmological Parameters,” *Astrophys. J. Suppl.* **148** (2003) 175 [arXiv:astro-ph/0302209].

²¹Strictly speaking, this is only true for an affinely-parameterized geodesic.

- [2] A. Kogut *et al.*, “Wilkinson Microwave Anisotropy Probe (WMAP) First Year Observations: TE Polarization,” *Astrophys. J. Suppl.* **148** (2003) 161 [arXiv:astro-ph/0302213].
- [3] H. V. Peiris *et al.*, “First year Wilkinson Microwave Anisotropy Probe (WMAP) observations: Implications for inflation,” *Astrophys. J. Suppl.* **148** (2003) 213 [arXiv:astro-ph/0302225].
- [4] R. H. Brandenberger, “Lectures on the theory of cosmological perturbations,” *Lect. Notes Phys.* **646** (2004) 127 [arXiv:hep-th/0306071].
- [5] B. Ryden, *Introduction to Cosmology*, Pearson Education 2003.
- [6] S. Weinberg, *Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity*, Wiley 1972.
- [7] P.J.E. Peebles, *Principles of Physical Cosmology*, Princeton University Press (1993).
- [8] A. Linde, *Particle Physics and Inflationary Cosmology*, Harwood Academic Publishers (1990).
- [9] E. W. Kolb and M. S. Turner, *The Early Universe*, Addison-Wesley (1990).
- [10] A. R. Liddle and D. H. Lyth, *Cosmological Inflation and Large-Scale Structure*, Cambridge University Press (2000).
- [11] S. Weinberg, *Rev. Mod. Phys.* **61** (1989) 1.