MPhys Advanced Cosmology 2012–2013

John Peacock

Room C20, Royal Observatory; jap@roe.ac.uk

http://www.roe.ac.uk/japwww/teaching/cos5.html

Synopsis This course is intended to act as an extension of the current 4th-year course on Astrophysical Cosmology, which develops the basic tools for dealing with observations in an expanding universe, and gives an overview of some of the central topics in contemporary research. The aim here is to revisit this material at a level of detail more suitable as a foundation for understanding current research. Cosmology has a standard model for understanding the universe, in which the dominant theme is the energy density of the vacuum. This is observed to be non-zero today, and is hypothesised to have been much larger in the past, causing the phenomenon of 'inflation'. An inflationary phase can not only launch the expanding universe, but can also seed irregularities that subsequently grow under gravity to create galaxies, superclusters and anisotropies in the microwave background. The course will present the methods for analysing these phenomena, leading on to some of the frontier issues in cosmology, particularly the possible existence of extra dimensions and many universes. It is intended that the course should be self contained; previous attendance at courses on cosmology or general relativity will be useful, but not essential.

Recommended books (in reserve section of ROE library)

Peacock: Cosmological Physics (CUP) Gives an overview of cosmology at the level of this course, but contains much more than will be covered here. More recent developments to be covered in the lectures are not in the book.

Dodelson: Modern Cosmology (Wiley) Concentrating on the details of relativistic perturbation theory, with applications to the CMB. Higher level than this course, but contains many useful things.

Other good books for alternative perspectives and extra detail:

Mukhanov: Physical Foundations of Cosmology (CUP)
Peebles: Principles of Physical Cosmology (Princeton)
Weinberg: Gravitation & Cosmology (Wiley)

Syllabus

(1) **Review of Friedmann models** FRW spacetime; Dynamics; Observables; Horizons

(2) **The hot big bang** Thermal history; Freezeout; Relics; Recombination and last scattering

(3) **Inflation − I** Initial condition problems; Planck era; Physics beyond the SM; Scalar fields; Noether's theorem

(4) **Inflation − II** The zoo of inflation models; Equation of motion; Slow-roll; Ending inflation

(5) **Fluctuations from inflation** Gauge issues; Power spectra; Basics of fluctuation generation; Tilt; Tensor modes; Eternal inflation

(6) **Structure formation − I** Newtonian analysis neglecting pressure; Perturbation modes; Coupled perturbations; matter transfer functions

(7) **Structure formation − II** Nonlinear development: Spherical model; Lagrangian approach; N-body simulations; Dark-matter haloes & mass function; Gas cooling; Brief overview of galaxy formation

(8) **Gravitational lensing** Basics of light deflection; strong lensing and mass measurement; weak lensing and mapping dark matter

(9) **CMB anisotropies - I** Anisotropy mechanisms; Overview of Boltzmann approach; Power spectrum; Properties of the temperature field

(10) **CMB anisotropies - II** Geometrical degeneracies; Reionization; Polarization and tensor modes; The cosmological standard model

(11) **Frontiers** Measuring dark energy; Extra dimensions and modified gravity; anthropics and the multiverse

1 Review of Friedmann models

Topics to be covered:

- Cosmological spacetime and RW metric
- Expansion dynamics and Friedmann equation
- Calculating distances and times

1.1 Cosmological spacetime

One of the fundamentals of a cosmologist's toolkit is to be able to assign coordinates to events in the universe. We need a large-scale notion of space and time that allows us to relate observations we make here and now to physical conditions at some location that is distant in time and space. The starting point is the relativistic idea that spacetime must have a **metric**: the equivalence principle says that conditions around our distant object will be as in special relativity (if it is freely falling), so there will be the usual idea of the **interval** or **proper time** between events, which we want to rewrite in terms of our coordinates:

$$-ds^2 = c^2 d\tau^2 = c^2 dt'^2 - dx'^2 - dy'^2 - dz'^2 = g_{\mu\nu} dx^\mu dx^\nu . \tag{1}$$

Here, dashed coordinates are local to the object, undashed are the global coordinates we use. As usual, the Greek indices run from 0 to 3. Note the ambiguity in defining the sign of the squared interval. The matrix $g_{\mu\nu}$ is the **metric tensor**, which is found in principle by solving Einstein's gravitational field equations. A simpler alternative, which fortunately matches the observed universe pretty well, is to consider the most symmetric possibilities for the metric.

ISOTROPIC EXPANSION Again according to Einstein, any spacetime with non-zero matter content must have some spacetime curvature, i.e. the metric cannot have the special relativity form $\mathrm{diag}(+1, -1, -1, -1)$. This curvature is something intrinsic to the spacetime, and does not need to be associated with extra spatial dimensions; these are nevertheless a useful intuitive way of understanding curved spaces such as the 2D surface of a 3D sphere. To motivate what is to come, consider the higher-dimensional analogue of this surface: something that is almost a 4D (hyper)sphere in Euclidean 5D space:

$$x^2 + y^2 + z^2 + w^2 - v^2 = \mathcal{R}^2 \tag{2}$$

where the metric is

$$ds^2 = dx^2 + dy^2 + dz^2 + dw^2 - dv^2. \tag{3}$$

Effectively, we have made one coordinate imaginary because we know we want to end up with the 4D spacetime signature.

This maximally symmetric spacetime is known as **de Sitter space**. It looks like a static spacetime, but relativity can be deceptive, as the interpretation depends on the coordinates you choose. Suppose we re-express things using the analogues of polar coordinates:

$$\begin{aligned} v &= \mathcal{R}\sinh\alpha \\ w &= \mathcal{R}\cosh\alpha\cos\beta \\ z &= \mathcal{R}\cosh\alpha\sin\beta\cos\gamma \\ y &= \mathcal{R}\cosh\alpha\sin\beta\sin\gamma\cos\delta \\ x &= \mathcal{R}\cosh\alpha\sin\beta\sin\gamma\sin\delta. \end{aligned} \tag{4}$$

This has the advantage that it is an orthogonal coordinate system: a vector such as $\mathbf{e}_\alpha = \partial(x, y, z, w, v)/\partial\alpha$ is orthogonal to all the other $\mathbf{e}_i$ (most simply seen by considering $\mathbf{e}_\delta$ and imagining continuing the process to still more dimensions). The squared length of the vector is just the sum of $|\mathbf{e}_{\alpha_i}|^2\, d\alpha_i^2$, which makes the metric into

$$ds^2 = -\mathcal{R}^2 d\alpha^2 + \mathcal{R}^2\cosh^2\alpha\left(d\beta^2 + \sin^2(\beta)[d\gamma^2 + \sin^2\gamma d\delta^2]\right), \tag{5}$$

which by an obvious change of notation becomes

$$c^2 d\tau^2 = c^2 dt^2 - \mathcal{R}^2\cosh^2(ct/\mathcal{R})\left(dr^2 + \sin^2(r)[d\theta^2 + \sin^2\theta d\phi^2]\right). \tag{6}$$

Now we have a completely different interpretation of the metric:

$$(\text{interval})^2 = (\text{time interval})^2 - (\text{scale factor})^2(\text{comoving interval})^2. \tag{7}$$

There is a universal **cosmological time**, which is the ticking of clocks at constant **comoving radius** r and constant angle on the sky. The spatial part of the metric expands with time, according to a universal **scale factor** $R(t) = \mathcal{R}\cosh(ct/\mathcal{R})$, so that particles at constant r recede from the

origin, and must thus suffer a Doppler redshift. This of course presumes that constant r corresponds to the actual trajectory of a free particle, which we have not proved – although it is true.

Historically, de Sitter space was extremely important in cosmology, although it was not immediately clear that the model is non-static. It was eventually concluded (in 1923, by Weyl) that one would expect a redshift that increased linearly with distance in de Sitter's model, but this was interpreted as measuring the constant radius of curvature of spacetime, $\mathcal{R}$. By this time, Slipher had already established that most galaxies were redshifted. Hubble's 1929 'discovery' of the expanding universe was explicitly motivated by the possibility of finding the 'de Sitter effect' (although we now know that his sample was too shallow to be able to detect it reliably).

In short, it takes more than just the appearance of $R(t)$ in a metric to prove that something is expanding. That this is the correct way to think about things only becomes apparent when we take a local (and thus Newtonian, thanks to the equivalence principle) look at particle dynamics. Then it becomes clear that a static distribution of test particles is impossible in general, so that it makes more sense to use an expanding coordinate system defined by the locations of such a set of particles.

THE ROBERTSON-WALKER METRIC The de Sitter model is only one example of an isotropically expanding spacetime, and we need to make the idea general. What we are interested in is a situation where, locally, all position vectors at time t are just scaled versions of their values at a reference time t_0:

$$\mathbf{x}(t) = R(t)\mathbf{x}(t_0), \tag{8}$$

where $R(t)$ is the **scale factor**. Differentiating this with respect to t gives

$$\dot{\mathbf{x}}(t) = \dot{R}(t)\mathbf{x}(t_0) = [\dot{R}(t)/R(t)]\,\mathbf{x}(t), \tag{9}$$

or a velocity proportional to distance, independent of origin, with

$$H(t) = \dot{R}(t)/R(t). \tag{10}$$

The characteristic time of the expansion is called the **Hubble time**, and takes the value

$$t_{\rm H} \equiv H^{-1} = 9.78\,{\rm Gyr}\ \times\ (H/100\,{\rm km\,s^{-1}Mpc^{-1}})^{-1}. \tag{11}$$

As with de Sitter space, we assume a **cosmological time** t, which is the time measured by the clocks of these observers – *i.e.* t is the proper time measured by an observer at rest with respect to the local matter distribution. It makes sense that such a universal time exists if we accept that we are looking for models that are **homogeneous**, so that there are no preferred locations. This is obvious in de Sitter space: because it derives from a 4-sphere, all spacetime points are manifestly equivalent: the spacetime curvature and hence the matter density must be a constant. The next step is to to weaken this so that conditions can change with time, but are uniform at a given time. A cosmological time coordinate can then be defined and synchronized by setting clocks to a reference value at some standard density.

By analogy with the de Sitter result, we now guess that the spatial metric will factorize into the scale factor times a comoving part that includes curvature. This overall Robertson–Walker metric (**RW metric**), can be written as:

$$c^2 d\tau^2 = c^2 dt^2 - R^2(t) \left[dr^2 + S_k^2(r)\, d\psi^2 \right]. \tag{12}$$

The angle $d\psi$ separates two points on the sky, so that $d\psi^2 = d\theta^2 + \sin^2\theta\, d\phi^2$ in spherical polars. The function $S_k(r)$ allows for positive and negative curvature of the comoving part of the metric:

$$S_k(r) \equiv \begin{cases} \sin r & (k=+1) \\ \sinh r & (k=-1) \\ r & (k=0). \end{cases} \tag{13}$$

We only saw the $k=+1$ case of this in the de Sitter example, but mathematically we can then generate the $k=-1$ case by letting R and r both become imaginary.

The comoving radius r is dimensionless, and the scale factor R really is the spatial radius of curvature of the universe. Both are required in order to give a comoving distance dimensions of length – e.g. the combination $R_0 S_k(r)$. Nevertheless, it is often convenient to make the scale factor dimensionless, via

$$a(t) \equiv \frac{R(t)}{R_0}, \tag{14}$$

so that $a = 1$ at the present.

LIGHT PROPAGATION AND REDSHIFT Light follows trajectories with zero proper time (**null geodesics**). The radial equation of motion therefore integrates to

$$r = \int c\,dt/R(t). \tag{15}$$

The comoving distance is constant, whereas the domain of integration in time extends from $t_{\rm emit}$ to $t_{\rm obs}$; these are the times of emission and reception of a photon. Thus $dt_{\rm emit}/dt_{\rm obs} = R(t_{\rm emit})/R(t_{\rm obs})$, which means that events on distant galaxies time-dilate. This dilation also applies to frequency, so

$$\frac{\nu_{\rm emit}}{\nu_{\rm obs}} \equiv 1 + z = \frac{R(t_{\rm obs})}{R(t_{\rm emit})}. \tag{16}$$

In terms of the normalized scale factor $a(t)$ we have simply $a(t) = (1+z)^{-1}$. So just by observing shifts in spectral lines, we can learn how big the universe was at the time the light was emitted. This is the key to performing observational cosmology.

1.2 Cosmological dynamics

THE FRIEDMANN EQUATION The equation of motion for the scale factor resembles Newtonian conservation of energy for a particle at the edge of a uniform sphere of radius R:

$$\dot{R}^2 - \frac{8\pi G}{3}\rho R^2 = -kc^2. \tag{17}$$

This is almost obviously true, since the Newtonian result that the gravitational field inside a uniform shell is zero does still hold in general relativity, and is known as **Birkhoff's theorem**. For the present course, we will accept this quasi-Newtonian 'derivation', and merely attempt to justify the form of the rhs.

This energy-like equation can be turned into a force-like equation by differentiating with respect to time:

$$\ddot{R} = -4\pi G R(\rho + 3p/c^2)/3. \tag{18}$$

To deduce this, we need to know $\dot{\rho}$, which comes from conservation of energy:

$$d[\rho c^2 R^3] = -p d[R^3]. \tag{19}$$

The surprising factor here is the occurrence of the **active mass density** $\rho + 3p/c^2$. This is here because the weak-field form of Einstein's gravitational field equations is

$$\nabla^2 \Phi = 4\pi G(\rho + 3p/c^2). \tag{20}$$

The extra term from the pressure is important. As an example, consider a **radiation-dominated fluid** – *i.e.* one whose equation of state is the same as that of pure radiation: $p = u/3$, where u is the energy density. For such a fluid, $\rho + 3p/c^2 = 2\rho$, so its gravity is twice as strong as we might have expected.

But the greatest astonishment in the Friedmann equation is the term on the rhs. This is related to the curvature of spacetime, and $k = 0, \pm 1$ is the same integer that is found in the RW metric. This cannot be completely justified without the Field Equations, but the **flat** $k = 0$ case is readily understood. Write the energy-conservation equation with an arbitrary rhs, but divide through by R^2:

$$H^2 - \frac{8\pi G}{3}\rho = \frac{\text{const}}{R^2}. \tag{21}$$

Now imagine holding the observables H and ρ constant, but let $R \to \infty$; this has the effect of making the rhs of the Friedmann equation indistinguishable from zero. Looking at the metric with $k \neq 0$, $R \to \infty$ with Rr fixed implies $r \to 0$, so the difference between $S_k(r)$ and r becomes negligible and we have in effect the $k = 0$ case.

There is thus a **critical density** that will yield a flat universe,

$$\rho_c = \frac{3H^2}{8\pi G}. \tag{22}$$

It is common to define a dimensionless **density parameter** as the ratio of density to critical density:

$$\Omega \equiv \frac{\rho}{\rho_c} = \frac{8\pi G\rho}{3H^2}. \tag{23}$$

The current value of such parameters should be distinguished by a zero subscript. In these terms, the Friedmann equation gives the present value of the scale factor:

$$R_0 = \frac{c}{H_0}\,[k/(\Omega_0 - 1)]^{1/2}, \tag{24}$$

which diverges as the universe approaches the flat state with $\Omega = 1$. In practice, Ω_0 is such a common symbol in cosmological formulae, that it is normal to omit the zero subscript. We can also define a dimensionless (current) Hubble parameter as

$$h \equiv \frac{H_0}{100\,\mathrm{km\,s^{-1}Mpc^{-1}}}, \tag{25}$$

in terms of which the current density of the universe is

$$\begin{aligned} \rho_0 &= 1.878 \times 10^{-26}\,\Omega h^2\ \mathrm{kg\,m^{-3}} \\ &= 2.775 \times 10^{11}\,\Omega h^2\ M_\odot\,\mathrm{Mpc^{-3}}. \end{aligned} \tag{26}$$

MODELS WITH GENERAL EQUATIONS OF STATE To solve the Friedmann equation, we need to specify the matter content of the universe, and there are two obvious candidates: pressureless nonrelativistic matter, and radiation-dominated matter. These have densities that scale respectively as a^{-3} and a^{-4}. The first two relations just say that the number density of particles is diluted by the expansion, with photons also having their energy reduced by the redshift. We can be more general, and wonder if the universe might contain another form of matter that we have not yet considered. How this varies with redshift depends on its equation of state. If we define the parameter

$$w \equiv p/\rho c^2, \tag{27}$$

then conservation of energy says

$$d(\rho c^2 V) = -p\,dV \Rightarrow d(\rho c^2 V) = -w\rho c^2\,dV \Rightarrow d\ln\rho/d\ln a = -3(w+1), \tag{28}$$

so

$$\rho \propto a^{-3(w+1)} \tag{29}$$

if w is constant. Pressureless nonrelativistic matter has $w = 0$ and radiation has $w = 1/3$.

But this may not be an exhaustive list, and the universe could contain substances with less familiar equations of state. Inventing new forms of matter may seem like a silly game to play, but cosmology can be the only way to learn if something unexpected exists. As we will see in more detail later, modern data force us to accept a contribution that is approximately independent of time with $w \simeq -1$: a **vacuum energy** that is simply an invariant property of empty space. A general name for this contribution is **dark energy**, reflecting our ignorance of its nature (although the name is not very good, since it is too similar to dark matter: 'dark tension' would better reflect its unusual equation of state with negative pressure).

In terms of observables, this means that the density is written as

$$\frac{8\pi G\rho}{3} = H_0^2(\Omega_v a^{-3(w+1)} + \Omega_m a^{-3} + \Omega_r a^{-4}) \tag{30}$$

(using the normalized scale factor $a = R/R_0$). We will generally set $w = -1$ without comment, except where we want to focus explicitly on this parameter. This expression allows us to write the Friedmann equation in a manner useful for practical solution. Start with the Friedmann equation in the form $H^2 = 8\pi G\rho/3 - kc^2/R^2$. Inserting the expression for $\rho(a)$ gives

$$H^2(a) = H_0^2 \left[\Omega_v + \Omega_m a^{-3} + \Omega_r a^{-4} - (\Omega - 1)a^{-2}\right]. \tag{31}$$

This equation is in a form that can be integrated immediately to get $t(a)$. This is not possible analytically in all cases, nor can we always invert to get $a(t)$, but there are some useful special cases worth knowing. Mostly these refer to the **flat universe** with total $\Omega = 1$. Curvature can always be neglected at sufficiently early times, as can vacuum density (except that the theory of inflation postulates that the vacuum density was very much higher in the very distant past). The solutions look simplest if we appreciate that normalization to the current era is arbitrary, so we can choose $a = 1$ to be at a convenient point where the densities of two main components cross over. Also, the Hubble parameter at that point (H_*) sets a characteristic time, from which we can make a dimensionless version $\tau \equiv tH_*$.

MATTER AND RADIATION Using dashes to denote $d/d(t/\tau)$, we have $a'^2 = (a^{-2} + a^{-1})/2$, which is simply integrated to yield

$$\tau = \frac{2\sqrt{2}}{3}\left(2 + (a-2)\sqrt{1+a}\right). \tag{32}$$

This can be inverted to yield $a(\tau)$, but the full expression is too ugly to be much use. It will suffice to note the limits:

$$\begin{aligned} \tau \ll 1: &\quad a = (\sqrt{2}\tau)^{1/2}. \\ \tau \gg 1: &\quad a = (3\tau/2\sqrt{2})^{2/3}, \end{aligned} \tag{33}$$

so the universe expands as $t^{1/2}$ in the radiation era, which becomes $t^{2/3}$ once matter dominates. Both these powers are shallower than t, reflecting the decelerating nature of the expansion.

RADIATION AND VACUUM Now we have $a'^2 = (a^{-2} + a^2)/2$, which is easily solved in the form $(a^2)'/\sqrt{2} = \sqrt{1 + (a^2)^2}$, and simply inverted:

$$a = \Big(\sinh(\sqrt{2}\tau)\Big)^{1/2}. \tag{34}$$

Here, we move from $a \propto t^{1/2}$ at early times to an exponential behaviour characteristic of vacuum-dominated **de Sitter space**. This would be an appropriate model for the onset of a phase of inflation following a big-bang singularity.

MATTER AND VACUUM Here, $a'^2 = (a^{-1} + a^2)/2$, which can be tackled via the substitution $y = a^{3/2}$, to yield

$$a = \Big(\sinh(3\tau/2\sqrt{2})\Big)^{2/3}. \tag{35}$$

This transition from the flat matter-dominated $a \propto t^{2/3}$ to de Sitter space seems to be the one that describes our actual universe (apart from the radiation era at $z \gtrsim 10^4$).

CURVED MODELS We will not be very strongly concerned with highly curved models in this course, but it is worth knowing some basic facts, as shown in figure 1 (neglecting radiation). On a plot of the $\Omega_m - \Omega_v$ plane, the diagonal line $\Omega_m + \Omega_v = 1$ always separates open and closed models. If $\Omega_v < 0$, recollapse always occurs – whereas a positive vacuum density does not always guarantee expansion to infinity, especially when the matter density is high. For closed models with sufficiently high vacuum density, there was no big bang in the past, and the universe must have emerged from a 'bounce' at some finite minimum radius. All these statements can be deduced quite simply from the Friedmann equation.

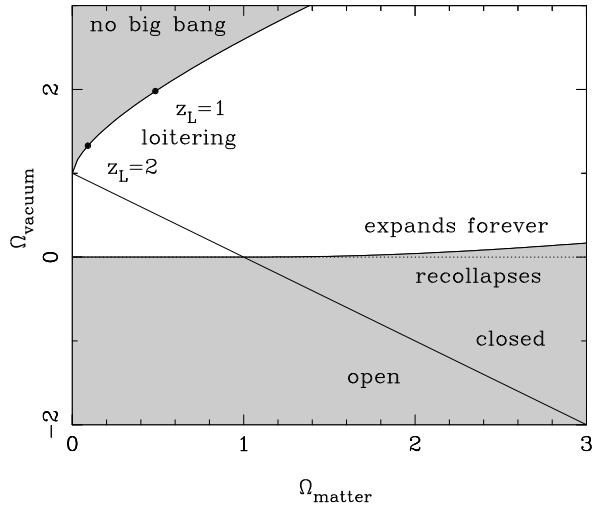

Figure 1. This plot shows the different possibilities for the cosmological expansion as a function of matter density and vacuum energy. Models with total $\Omega > 1$ are always spatially closed (open for $\Omega < 1$), although closed models can still expand to infinity if $\Omega_v \neq 0$. If the cosmological constant is negative, recollapse always occurs; recollapse is also possible with a positive Ω_v if $\Omega_m \gg \Omega_v$. If $\Omega_v > 1$ and Ω_m is small, there is the possibility of a 'loitering' solution with some maximum redshift and infinite age (top left); for even larger values of vacuum energy, there is no big bang singularity.

1.3 Observational cosmology

AGE OF THE UNIVERSE Since $1 + z = R_0/R(z)$, we have

$$\frac{dz}{dt} = -\frac{R_0}{R^2}\frac{dR}{dt} = -(1+z)H(z), \tag{36}$$

so $t(z) = \int_z^\infty H(z)^{-1}\, dz/(1+z)$, where

$$H^2(a) = H_0^2\left[\Omega_v + \Omega_m a^{-3} + \Omega_r a^{-4} - (\Omega - 1)a^{-2}\right]. \tag{37}$$

This can't be done analytically in general, but the following simple approximate formula is accurate to a few % for cases of practical interest:

$$H(z)t(z) \simeq \frac{2}{3}\,(0.7\Omega_m(z) - 0.3\Omega_v(z) + 0.3)^{-0.3}. \tag{38}$$

At $10 < z < 1000$, where matter dominates, this is

$$t \simeq (2/3)H^{-1} \simeq (2/3)H_0^{-1}\Omega_m^{-1/2}(1+z)^{-3/2}. \tag{39}$$

For a flat universe, the current age is $H_0 t_0 \simeq (2/3)\Omega_m^{-0.3}$. For many years, estimates of this product were around unity, which is hard to understand without vacuum energy, unless the density is very low ($H_0 t_0$ is exactly 1 in the limit of an empty universe). This was one of the first astronomical motivations for a vacuum-dominated universe.

DISTANCE-REDSHIFT RELATION The equation of motion for a photon is $R\,dr = c\,dt$, so $R_0 dr/dz = (1+z)c\,dt/dz$, or

$$R_0 r = \int \frac{c}{H(z)}\,dz. \tag{40}$$

Remember that non-flat models need the combination $R_0 S_k(r)$, so one has to divide the above integral by $R_0 = (c/H_0)|\Omega - 1|^{-1/2}$, apply the S_k function, and then multiply by R_0 again. Once more, this process is not analytic in general.

PARTICLE HORIZON If the integral for comoving radius is taken from $z = 0$ to ∞, we get the full distance a particle can have travelled since the big bang – the **horizon distance**. For flat matter-dominated models,

$$R_0 r_{\rm H} \simeq \frac{2c}{H_0}\Omega_m^{-0.4}. \tag{41}$$

At high redshift, where H increases, this tends to zero. The onset of radiation domination does not change this: even though the presently visible universe was once very small, it expanded so quickly

that causal contact was not easy. The observed large-scale near-homogeneity is therefore something of a puzzle.

ANGULAR DIAMETERS Recall the RW metric:

$$c^2 d\tau^2 = c^2 dt^2 - R^2(t) \left[dr^2 + S_k^2(r)\, d\psi^2 \right]. \tag{42}$$

The spatial parts of the metric give the *proper* transverse size of an object seen by us as its comoving size $d\psi\, S_k(r)$ times the scale factor at the time of emission:

$$d\ell_\perp = d\psi\ R(z) S_k(r) = d\psi\ R_0 S_k(r)/(1+z). \tag{43}$$

If we know r, we can therefore convert the angle subtended by an object into its physical extent perpendicular to the line of sight.

LUMINOSITY AND FLUX DENSITY Imagine a source at the centre of a sphere, on which we sit. The photons from the source pass though a proper surface area $4\pi[R_0 S_k(r)]^2$. But redshift still affects the flux density in four further ways: (1) photon energies are redshifted, reducing the flux density by a factor $1+z$; (2) photon arrival rates are time dilated, reducing the flux density by a further factor $1+z$; (3) opposing this, the bandwidth $d\nu$ is reduced by a factor $1+z$, which increases the energy flux per unit bandwidth by one power of $1+z$; (4) finally, the observed photons at frequency ν_0 were emitted at frequency $[1+z] \times \nu_0$. Overall, the flux density is the luminosity at frequency $[1+z]\nu_0$, divided by the total area, divided by $1+z$:

$$S_\nu(\nu_0) = \frac{L_\nu([1+z]\nu_0)}{4\pi R_0^2 S_k^2(r)(1+z)} = \frac{L_\nu(\nu_0)}{4\pi R_0^2 S_k^2(r)(1+z)^{1+\alpha}}, \tag{44}$$

where the second expression assumes a power-law spectrum $L \propto \nu^{-\alpha}$.

SURFACE BRIGHTNESS The flux density is the product of the **specific intensity** I_ν and the solid angle subtended by the source: $S_\nu = I_\nu\ d\Omega$. Combining the angular size and flux-density relations gives a relation that is independent of cosmology:

$$I_\nu(\nu_0) = \frac{B_\nu([1+z]\nu_0)}{(1+z)^3}, \tag{45}$$

where B_ν is **surface brightness** (luminosity emitted into unit solid angle per unit area of source). This $(1+z)^3$ dimming makes it hard to detect extended objects at very high redshift. The factor becomes $(1+z)^4$ if we integrate over frequency to get a bolometric quantity.

EFFECTIVE DISTANCES The angle and flux relations can be made to look Euclidean:

$$\begin{aligned} \textbf{angular}-\textbf{diameter distance}: &\quad D_{\rm A} = (1+z)^{-1} R_0 S_k(r) \\ \textbf{luminosity distance}: &\quad D_{\rm L} = (1+z)\, R_0 S_k(r). \end{aligned} \tag{46}$$

Some example distance-redshift relations are shown in figure 2. Notice how a high matter density tends to make high-redshift objects brighter: stronger deceleration means they are closer for a given redshift.

2 The hot big bang

Topics to be covered:

- Thermal history
- Freezeout & relics
- Recombination and last scattering

2.1 Thermal history

Although the timescale for expansion of the early universe is very short, the density is also very high, so it is normally sensible to assume that conditions are close to thermal equilibrium. Also the fluids of interest are simple enough that we can treat them as perfect gases. The thermodynamics of such a gas is derived staring with a box of volume $V = L^3$, and expanding the fields inside into periodic waves with **harmonic boundary conditions**. The density of states in k space is

$$dN = g\,\frac{V}{(2\pi)^3}\,d^3k \tag{47}$$

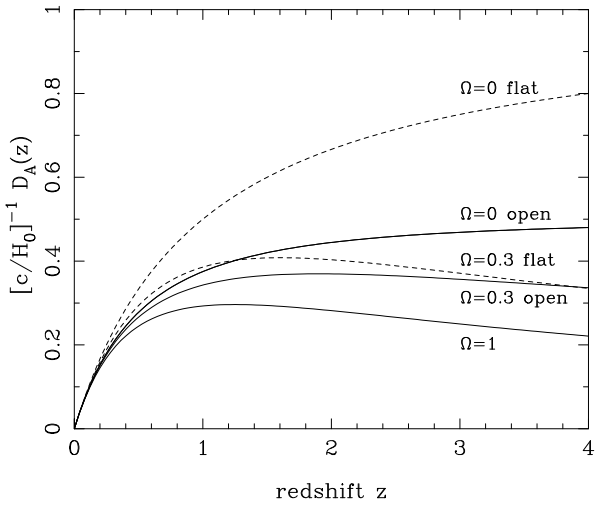

Figure 2. A plot of dimensionless angular-diameter distance versus redshift for various cosmologies. Solid lines show models with zero vacuum energy; dashed lines show flat models with $\Omega_m + \Omega_v = 1$. In both cases, results for $\Omega_m = 1, 0.3, 0$ are shown; higher density results in lower distance at high z, due to gravitational focusing of light rays.

(where g is a degeneracy factor for spin *etc.*). The equilibrium **occupation number** for a quantum state of energy ϵ is given generally by

$$\langle f \rangle = \left[e^{(\epsilon - \mu)/kT} \pm 1 \right]^{-1} \tag{48}$$

(+ for fermions, − for bosons). Now, for a thermal radiation background, the **chemical potential**, μ is always zero. The reason for this is quite simple: μ appears in the first law of thermodynamics as the change in energy associated with a change in particle number, $dE = TdS - PdV + \mu dN$. So,

as N adjusts to its equilibrium value, we expect that the system will be stationary with respect to small changes in N. The thermal equilibrium **background number density** of particles is

$$n = \frac{1}{V}\int f\, dN = g\,\frac{1}{(2\pi\hbar)^3}\int_0^\infty \frac{4\pi\, p^2 dp}{e^{\epsilon(p)/kT}\pm 1}, \tag{49}$$

where we have changed to momentum space; $\epsilon = \sqrt{m^2c^4 + p^2c^2}$ and g is the degeneracy factor. There are two interesting limits of this expression.

(1) Ultrarelativistic limit. For $kT \gg mc^2$ the particles behave as if they were massless, and we get

$$n = \left(\frac{kT}{c}\right)^3 \frac{4\pi g}{(2\pi\hbar)^3}\int_0^\infty \frac{y^2 dy}{e^y \pm 1}. \tag{50}$$

(2) Non-relativistic limit. Here we can neglect the ± 1 in the occupation number, in which case the number is suppressed by a dominant $\exp(-mc^2/kT)$ factor. This shows us that the background 'switches on' at about $kT \sim mc^2$; at this energy, known as a **threshold**, photons and other species in equilibrium will have sufficient energy to create particle-antiparticle pairs.

The above thermodynamics also gives the energy density of the background, since it is only necessary to multiply the integrand by a factor $\epsilon(p)$ for the energy in each mode:

$$u = \rho\, c^2 = g\,\frac{1}{(2\pi\hbar)^3}\int_0^\infty \frac{4\pi\, p^2\, dp}{e^{\epsilon(p)/kT}\pm 1}\,\epsilon(p). \tag{51}$$

In the ultrarelativistic limit, $\epsilon(p) = pc$, this becomes

$$u = \frac{\pi^2}{30(\hbar c)^3}\, g\,(kT)^4 \quad \text{(bosons)}. \tag{52}$$

The thermodynamic properties of Fermions can be obtained from those of Bosonic black-body radiation by the following trick: $1/(e^x+1) = 1/(e^x-1) - 2/(e^{2x}-1)$. Thus, a gas of fermions looks like a mixture of bosons at two different temperatures. Knowing that boson number density and energy density scale as $n \propto T^3$ and $u \propto T^4$, we find $n_{\rm F} = (3/4)\, n_{\rm B}$; $u_{\rm F} = (7/8)\, u_{\rm B}$.

It will also be useful to know the **entropy of the background**. This is not too hard to work out, because energy and entropy are extensive quantities for a thermal background. Thus, writing the first law for $\mu = 0$ and using $\partial S/\partial V = S/V$ *etc.* for extensive quantities,

$$dE = TdS - PdV \quad \Rightarrow \quad \left(\frac{E}{V}dV + \frac{\partial E}{\partial T}dT\right) = \left(T\frac{S}{V}dV + T\frac{\partial S}{\partial T}dT\right) - PdV. \tag{53}$$

Equating the dV and dT parts gives the familiar $\partial E/\partial T = T\,\partial S/\partial T$ and

$$S = \frac{E + PV}{T} \tag{54}$$

These results take an interesting and simple form in the ultrarelativistic limit. The energy density, u, obeys the usual black-body scaling $u \propto T^4$. In the ultrarelativistic limit, we also know that the pressure is $P = u/3$, so that the entropy density is

$$s = (4/3)u/T = \frac{2\pi^2 k}{45(\hbar c)^3}\, g\,(kT)^3 \quad \text{(bosons)}, \tag{55}$$

and 7/8 of this for fermions. Now, we saw earlier that the number density of an ultrarelativistic background also scales as T^3 – therefore we have the simple result that entropy just counts the number of particles. This justifies a common piece of terminology, in which the ratio of the number density of photons in the universe to the number density of **baryons** (protons plus neutrons) is called the **entropy per baryon**.

DEGREES OF FREEDOM Overall, the equilibrium relativistic density is

$$\rho c^2 = \frac{\pi^2}{30(\hbar c)^3}\, g_{\rm eff}\,(kT)^4; \quad g_{\rm eff} \equiv \sum_{\rm bosons} g_i + \frac{7}{8}\sum_{\rm fermions} g_j, \tag{56}$$

expressing the fermion contribution as an effective number of bosons. A similar relation holds for entropy density: $s = [2\pi^2 k/45(\hbar c)^3]\, h_{\rm eff}\,(kT)^3$. In equilibrium, $h_{\rm eff} = g_{\rm eff}$, but this ceases to be true at late times, when the neutrinos and photons have different temperatures. The $g_{\rm eff}$ functions are plotted against photon temperature in figure 3. They start at a number determined by the total number of distinct elementary particles that exist (of order 100, according to the standard model of particle physics), and fall as the temperature drops and more species of particles become nonrelativistic.

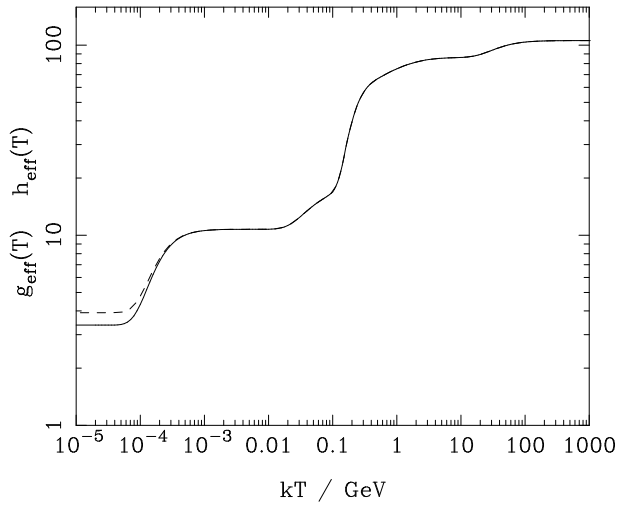

Figure 3. The number of relativistic degrees of freedom as a function of photon temperature. $g_{\rm eff}$ measures the energy density; $h_{\rm eff}$ the entropy (dashed line). The two depart significantly at low temperatures, when the neutrinos are cooler than the photons. For a universe consisting only of photons, we would expect $g = 2$. The main features visible are (1) The electroweak phase transition at 100 Gev; (2) The QCD phase transition at 200 MeV; (3) the $e^{\pm}$ annihilation at 0.3 MeV.

TIME AND TEMPERATURE This temperature-dependent equilibrium density sets the timescale for expansion in the early universe. Using the relation between time and density for a flat radiation–dominated universe, $t = (32\pi G\rho/3)^{-1/2}$, we can deduce the time–temperature relation:

$$t/\text{seconds} = g_{\rm eff}^{-1/2} \left(T/10^{10.257}\,\text{K}\right)^{-2}. \tag{57}$$

This is independent of the present-day temperature of the photon background, which manifests itself as the **cosmic microwave background** (CMB),

$$T = 2.725 \pm 0.002\,\text{K}. \tag{58}$$

This temperature was of course higher in the past, owing to the adiabatic expansion of the universe. Frequently, we will assume

$$T(z) = 2.725(1+z), \tag{59}$$

which is justified informally by arguing that photon energies scale as $E \propto 1/a$ and saying that the typical energy in black-body radiation is $\sim kT$. Being more careful, we should conserve entropy, so that $s \propto a^{-3}$. Since $s \propto T^3$ while $h_{\rm eff}$ is constant, this requires $T \propto 1/a$. But clearly this does *not* apply near a threshold. At these points, $h_{\rm eff}$ changes rapidly and the universe will expand at nearly constant temperature for a period.

The energy density in photons is supplemented by that of the neutrino background. Because they have a lower temperature, as shown below, they contribute an energy density 0.68 times that from the photons (if the neutrinos are massless and therefore relativistic). If there are no other contributions to the energy density from relativistic particles, then the total effective radiation density is $\Omega_r h^2 \simeq 4.2 \times 10^{-5}$ and the redshift of **matter–radiation equality** is

$$1 + z_{\rm eq} = 24\,074\,\Omega h^2\,(T/2.725\,{\rm K})^{-4}. \tag{60}$$

The time of this change in the global equation of state is one of the key epochs in determining the appearance of the present-day universe.

The following table shows some of the key events in the history of the universe. Note that, for very high temperatures, energy units for kT are often quoted instead of T. The conversion is $kT = 1$ eV for $T = 10^{4.06}$ K. Some of the numbers are rounded, rather than exact; also, some of them depend a little on Ω and H_0. Where necessary, a flat model with $\Omega = 0.3$ and $h = 0.7$ has been assumed.

Event	T	kT	$g_{\rm eff}$	redshift	time
Now	2.73 K	0.0002 eV	3.3	0	13 Gyr
Distant galaxy	16 K	0.001 eV	3.3	5	1 Gyr
Recombination	3000 K	0.3 eV	3.3	1100	$10^{5.6}$ years
Radiation domination	9500 K	0.8 eV	3.3	3500	$10^{4.7}$ years
Electron pair threshold	$10^{9.7}$ K	0.5 MeV	11	$10^{9.5}$	3 s
Nucleosynthesis	10^{10} K	1 MeV	11	10^{10}	1 s
Nucleon pair threshold	10^{13} K	1 GeV	70	10^{13}	$10^{-6.6}$ s
Electroweak unification	$10^{15.5}$ K	250 GeV	100	10^{15}	10^{-12} s
Grand unification	10^{28} K	10^{15} GeV	100(?)	10^{28}	10^{-36} s
Quantum gravity	10^{32} K	10^{19} GeV	100(?)	10^{32}	10^{-43} s

2.2 Freezeout and relics

So far, we have assumed that thermal equilibrium will be followed in the early universe, but this is far from obvious. Equilibrium is produced by reactions that involve individual particles, *e.g.* $e^+e^- \leftrightarrow 2\gamma$ converts between electron-positron pairs and photons. When the temperature is low, typical photon energies are too low for this reaction to proceed from right to left, so there is nothing to balance annihilations.

Nevertheless, the annihilations only proceed at a finite rate: each member of the pair has to find a partner to interact with. We can express this by writing a simple differential equation for the electron density, called the **Boltzmann equation**:

$$\dot{n} + 3Hn = -\langle \sigma v \rangle n^2 + S, \tag{61}$$

where σ is the reaction cross-section, v is the particle velocity, and S is a source term that represents thermal particle production. The $3Hn$ term just represents dilution by the expansion of the universe. Leaving aside the source term for the moment, we see that the change in n involves two timescales:

$$\begin{aligned} \textbf{expansion timescale} &= H(z)^{-1} \\ \textbf{interaction timescale} &= (\langle \sigma v \rangle n)^{-1} \end{aligned} \tag{62}$$

Both these times increase as the universe expands, but the interaction time usually changes fastest. The situation therefore changes from one of thermal equilibrium at early times to a state of **freezeout** or **decoupling** at late times. Once the interaction timescale becomes much longer than the age of the universe, the particle has effectively ceased to interact. It thus preserves a 'snapshot' of the properties of the universe at the time the particle was last in thermal equilibrium. This phenomenon of freezeout is essential to the understanding of the present-day nature of the universe. It allows for a whole set of **relics** to exist from different stages of the hot big bang.

To complete the Boltzmann equation, we need the source term S. This term can be fixed by a thermodynamic equilibrium argument: for a non-expanding universe, n will be constant at the equilibrium value for that temperature, n_T, showing that

$$S = \langle \sigma v \rangle n_T^2. \tag{63}$$

If we define comoving number densities $N \equiv a^3 n$ (effectively the ratio of n to the relativistic density for that temperature, $n_{\rm rel}$), the rate equation can be rewritten in the simple form

$$\frac{d\ln N}{d\ln a} = -\frac{\Gamma}{H}\left[1 - \left(\frac{N_{_T}}{N}\right)^2\right], \tag{64}$$

where $\Gamma = n\langle\sigma v\rangle$ is the interaction rate experienced by the particles.

Unfortunately, this equation must be solved numerically. The main features are easy enough to see, however. Suppose first that the universe is sustaining a population in approximate thermal equilibrium, $N \simeq N_{_T}$. If the population under study is relativistic, $N_{_T}$ does not change with time, because $n_{_T} \propto T^3$ and $T \propto a^{-1}$. This means that it is possible to keep $N = N_{_T}$ exactly, whatever Γ/H. It would however be grossly incorrect to conclude from this that the population stays in thermal equilibrium: if $\Gamma/H \ll 1$, a typical particle suffers no interactions even while the universe doubles in size, halving the temperature. A good example is the microwave background, whose photons last interacted with matter at $z \simeq 1100$.

Now consider the opposite case, where the thermal solution would be nonrelativistic, with $N_{_T} \propto T^{-3/2}\exp(-mc^2/kT)$. If the background stays at the equilibrium value, the lhs of the rate equation will therefore be negative and $\gg 1$ in magnitude. This is consistent if $\Gamma/H \gg 1$, because then the $(N_{_T}/N)^2$ term on the rhs can still be close to unity. However, if $\Gamma/H \ll 1$, there must be a deviation from equilibrium. When $N_{_T}$ changes sufficiently fast with a, the actual abundance cannot keep up, so that the $(N_{_T}/N)^2$ term on the rhs becomes negligible and $d\ln N/d\ln a \simeq -\Gamma/H$, which is $\ll 1$. There is therefore a critical time at which the reaction rate drops low enough that particles are simply conserved as the universe expands – the population has **frozen out**. This provides a more detailed justification for the intuitive rule-of-thumb used above to define decoupling,

$$N(a \to \infty) = N_{_T}(\Gamma/H = 1). \tag{65}$$

Exact numerical solutions of the rate equation almost always turn out very close to this simple rule, as shown in figure 4.

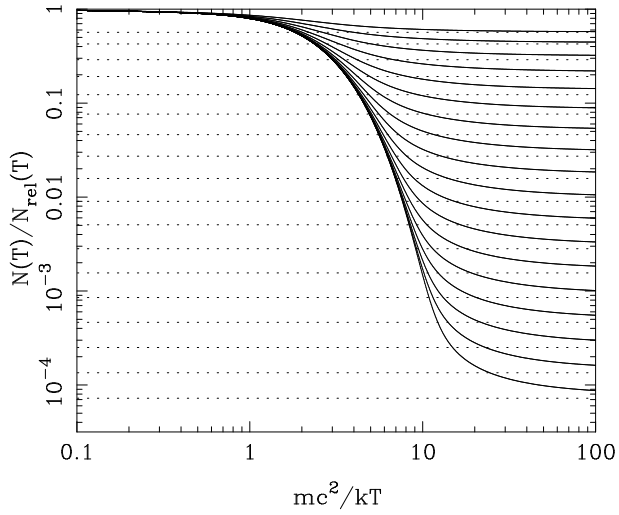

Figure 4. Solution of the Boltzmann equation for freezeout of a single massive fermion. We set $\Gamma/H = \epsilon(kT/mc^2)N/N_{\rm rel}$, as appropriate for a radiation-dominated universe in which $\langle\sigma v\rangle$ is assumed to be independent of temperature. The solid lines show the case $\epsilon = 1$ and increasing by powers of 2. A high value of ϵ leads to freezeout at increasingly low abundances. The dashed lines show the abundance predicted by the simple recipe of the thermal density for which $\Gamma/H = 1$.

THE RELIC DENSITY The above freezeout criterion can be used to deduce a simple and very important expression for the present-day density of a non-relativistic relic:

$$\Omega_{\rm relic}h^2 \simeq 0.03\,(\sigma/{\rm pb})^{-1}, \tag{66}$$

where the 'picobarn' is $1\,{\rm pb} = 10^{-40}$ m^2. Thus only a small range of annihilation cross-sections will be of observational interest. The steps needed to get this formula are as follows. (1) From $\Gamma/H = 1$, the number density of relics at freezeout is $n_f = H_f/\langle\sigma v\rangle$; (2) $H = (8\pi G\rho/3)^{1/2}$, where $\rho c^2 = (\pi^2/30\hbar^3c^3)g_{\rm eff}(kT)^4$; (3) $\Omega_{\rm relic} = 8\pi Gmn_0/3H_0^2$. The only missing ingredient here is how

to relate the present number density n_0 to the density n_f at temperature T_f. Since the relics are conserved, the number density must have fallen by the same factor as the entropy density:

$$n_f/n_0 = (h^f_{\rm eff}T_f^3)/(h^0_{\rm eff}T_0^3). \tag{67}$$

Today, $h^0_{\rm eff} = 43/11$, and $h^f_{\rm eff} = g_{\rm eff}$ at high redshift. This allows us to deduce the relic density, given the mass, cross-section and temperature of freezeout:

$$\Omega_{\rm relic}h^2 \simeq \frac{10^{-33.0}\ {\rm m}^2}{\langle\sigma v\rangle}\ \left(\frac{mc^2}{kT_f}\right)\ g_{\rm eff}^{-1/2}. \tag{68}$$

We see from figure 4 that $mc^2/kT_f \sim 10$ with only a logarithmic dependence on reaction rate, which roughly cancels the last factor on the rhs. Finally, since particles are nearly relativistic at freezeout, we set $\langle\sigma v\rangle = \sigma c$ to get our final estimate of the typical cross-section for an interesting relic abundance. The eventual conclusion makes sense: the higher the cross-section, the longer the particle can stay in equilibrium, and the more effective annihilations can be in suppressing the number density. Note that, in detail, we need to worry about whether the particle is a **Majorana particle** (i.e. its own antiparticle) or a **Dirac particle** where particles and antiparticles are distinct.

NEUTRINO DECOUPLING The best case for application of this freezeout apparatus is to relic neutrinos. At the later stages of the big bang, energies are such that only light particles survive in equilibrium: photons (γ), neutrinos (ν) and e^+e^- pairs. As the temperature falls below $T_e = 10^{9.7}$ K), the pairs will annihilate. Electrons can interact via either the electromagnetic or the weak interaction, so in principle the annihilations might yield pairs of photons or neutrinos. However, in practice the weak reactions freeze out earlier, at $T \simeq 10^{10}$ K.

The effect of the electron-positron annihilation is therefore to enhance the numbers of photons relative to neutrinos. Strictly, what is conserved in this process is the *entropy*. The entropy of an $e^\pm + \gamma$ gas is easily found by remembering that it is proportional to the number density, and that all three particle species have $g = 2$ (polarization or spin). The total is then

$$s(\gamma + e^+ + e^-) = \frac{11}{4}s(\gamma). \tag{69}$$

Equating this to photon entropy at a new temperature gives the factor by which the photon temperature is enhanced with respect to that of the neutrinos. Thus we infer the existence of a neutrino background with a temperature

$$T_\nu = \left(\frac{4}{11}\right)^{1/3} T_\gamma = 1.945\,\mathrm{K}, \tag{70}$$

for $T_\gamma = 2.725$ K. These relativistic relic neutrinos contribute an energy density that is a factor $(7/8) \times (4/11)^{4/3}$ times that of the photons. For three neutrino species, this enhances the energy density in relativistic particles by a factor 1.68 (there are three different kinds of neutrinos, just as there are three **leptons**: the μ and τ particles are heavy analogues of the electron).

MASSIVE NEUTRINOS Theoretical progress in understanding the origin of masses in particle physics means that there is no reason for the neutrino to be completely devoid of mass. Also, there is now clear experimental evidence that neutrinos have a small non-zero mass. The consequences of this for cosmology could be quite profound, as relic neutrinos are expected to be very abundant. The above section showed that $n(\nu + \overline{\nu}) = (3/4)n(\gamma;\, T = 1.945\,\mathrm{K})$. That yields a total of 113 relic neutrinos in every cm^3 for each species. Suppose these neutrinos were ultrarelativistic at decoupling: as the universe expands to $kT < m_\nu c^2$, the total number of neutrinos is preserved, so the present-day mass density in neutrinos is just the zero-mass number density times m_ν, and the consequence for the cosmological density in light neutrinos is easily worked out to be

$$\Omega_\nu h^2 = \frac{\sum m_i}{94.1\,\mathrm{eV}}. \tag{71}$$

The more complicated case of neutrinos that decouple when they are already nonrelativistic is studied below.

The current direct laboratory limits to the neutrino masses are

$$\nu_e \lesssim 2.2\,\mathrm{eV} \quad \nu_\mu \lesssim 0.17\,\mathrm{MeV} \quad \nu_\tau \lesssim 15\,\mathrm{MeV}. \tag{72}$$

Based on this, even the electron neutrino could be of great cosmological significance. But in practice, we will see later that studies of cosmological large-scale structure limit the sum of the masses to a maximum of about 0.5 eV. This is becoming interesting, since it is known that neutrino masses must be non-zero. In brief, this comes from studies of **neutrino mixing**, in which each neutrino type

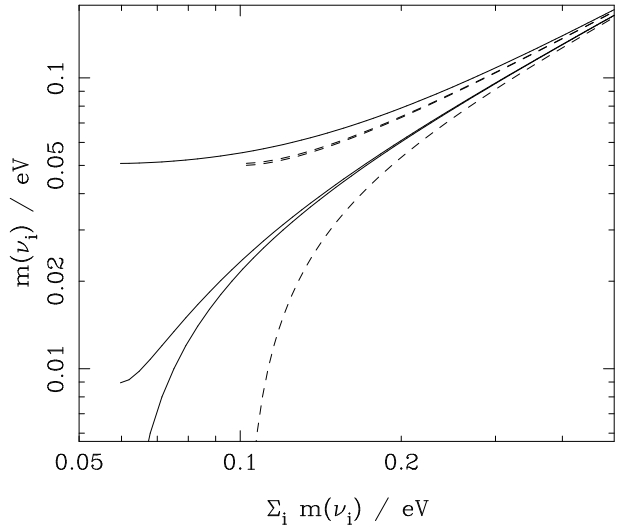

Figure 5. The masses of the individual neutrino mass eigenstates, plotted against the total neutrino mass for a normal hierarchy (solid lines) and an inverted hierarchy (dashed lines). Current cosmological data set an upper limit on the total mass of light neutrinos of around 0.5 eV.

is a mixture of energy eigenstates. The energy differences can be measured, which yields a measure of the difference in the square of the masses (consider the relativistic relation $E^2 = m^2 + p^2$, and expand to get $E \simeq p + m^2/2p$; neutrinos have well-defined momentum, but imprecise energy owing to mixing). These mixings are known from wonderfully precise experiments detecting neutrinos generated in the sun and the Earth's atmosphere:

$$\begin{aligned} \Delta(m_{21})^2 &= 8.0 \times 10^{-5}\ \mathrm{eV}^2 \\ \Delta(m_{32})^2 &= 2.5 \times 10^{-3}\ \mathrm{eV}^2, \end{aligned} \tag{73}$$

where m_1, m_2 and m_3 are the three mass eigenstates. This information does not give the absolute mass scale, nor does it tell us whether there is a **normal hierarchy** with $m_3 \gg m_2 \gg m_1$, or an **inverted hierarchy** in which states 1 & 2 are a close doublet lying well above state 3. Cosmology can settle both these issues by measuring the total density in neutrinos. The absolute minimum

situation is a normal hierarchy with m_1 negligibly small, in which case the mass is dominated by m_3, which is around 0.05 eV. The cosmological limits are within a power of 10 of this interesting point.

RELIC PARTICLES AS DARK MATTER Many other particles exist in the early universe, so there are a number of possible relics in addition to the massive neutrino. A common collective term for these particles is **WIMP** – standing for weakly interacting massive particle. There are really three generic types to consider, as follows.

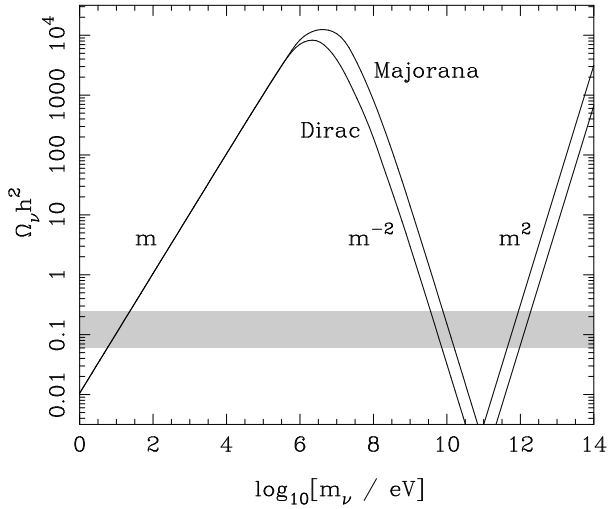

Figure 6. The contribution to the density parameter produced by relic neutrinos (or neutrino-like particles) as a function of their rest mass. The shaded band shows a factor of 2 either side of the observed CDM density. At low masses, the neutrinos are highly relativistic when they decouple: their abundance takes the zero-mass value, and the density is just proportional to the mass. Above about 1 MeV, the neutrinos are non-relativistic at decoupling, and their relic density is reduced by annihilation. Above the mass of the Z boson, the cross-section falls, so that annihilation is less effective and the relic density rises again.

(1) **Hot Dark Matter (HDM)** These are particles that decouple when relativistic, and which have a number density roughly equal to that of photons; eV-mass neutrinos are the archetype. The relic density scales linearly with the particle mass.

(2) **Warm Dark Matter (WDM)** If the particle decouples sufficiently early, the relative abundance of photons can then be boosted by annihilations other than just $e^{\pm}$. In modern particle physics theories, there are of order 100 distinct particle species, so the critical particle mass to make $\Omega = 1$ can be boosted to around $1-10$ keV.

(3) **Cold Dark Matter (CDM)** If the relic particles decouple while they are nonrelativistic, the number density can be exponentially suppressed. If the interactions are like those of neutrinos, then the freezeout temperature is about 1 MeV, and the relic mass density then falls with increasing mass (see figure 6). For weak interactions, cross-sections scale as (energy)^2, so that the relic density falls as $1/m^2$. Interesting masses then lie in the $\simeq 10$ GeV range, this cannot correspond to the known neutrinos, since such particles would have been seen in accelerators. But beyond about 90 GeV (the mass of the Z boson), the strength of the weak interaction is reduced, with cross-section going as (energy)$^{-2}$. The relic density now rises as m^2, so that the observed dark matter density is attained at $m \simeq 1$ TeV. Plausible candidates of this sort are found among so-called **supersymmetric** theories, which predict many new weakly-interacting particles. The favoured particle for a CDM relic is called the **neutralino**.

Since these particles exist to explain galaxy rotation curves, they must be passing through us right now. There is therefore a huge effort in the direct laboratory detection of dark matter, mainly via cryogenic detectors that look for the recoil of a single nucleon when hit by a DM particle (in deep mines, to shield from cosmic rays). Well-constructed experiments with low backgrounds are starting to set interesting limits, as shown in figure 7. There is no unique target to aim for, since even the simplest examples of supersymmetric models contain a variety of free parameters. These allow models that are optimistically close to current limits, but also some that will be hard to verify. The public-domain package **DarkSUSY** is available at `www.physto.se/~edsjo/darksusy` to make these detailed abundance calculations.

This subject saw a lot of publicity at the end of 2009, when the **CDMS** experiment announced events that were consistent with relic WIMPs (see `http://arxiv.org/abs/0912.3592`). In brief, cryogenic Ge and Si detectors are examined for evidence of nuclear recoil, which manifests itself in two distinct ways: heat (phonons) and ionization (electrons). The double signature allows rejection of many non-WIMP background events, although high-energy neutrons from cosmic ray events or radioactivity are a fundamental limit. CDMS estimate that these processes should cause on average 0.8 WIMP-like events during their 2 years of data; 2 events were actually seen. This is thus not so far inconsistent with background, but it is equally possible that there is a signal at a level of up

to about 5 times the background. If they run for more years, or increase the detector size, to the point of expecting around 10 background events, these possibilities will be distinguishable; we will will then have either a detection, or will be able to reduce the current upper limits.

What is particularly exciting is that the properties of these relic particles can also be observed via new examples manufactured in particle accelerators. The most wonderful outcome would be for the same particle to be found in these two different ways. The chances of success in this enterprise are hard to estimate, and some models exist in which detection would be impossible for many decades. But it would be a tremendous scientific achievement if dark matter particles were to be detected in this way, and a good part of the plausible parameter space will be covered over the next decade.

BARYOGENESIS It should be emphasised that these freezeout calculations predict equal numbers of particles and antiparticles. This makes a critical contrast with the case of normal or **baryonic** material. The number density of baryons is low (roughly 10^{-9} that of the CMB photons), so one's first thought might be that baryons are another frozen-out relic. But as far as is known, there is a negligible cosmic density of antibaryons; even if antimatter existed, freezeout applied to protons-antiproton pairs predicts a density far below what is observed. The inevitable conclusion is that the universe began with a very slight asymmetry between matter and antimatter: at high temperatures there were $1 + O(10^{-9})$ protons for every antiproton. If baryon number is conserved, this imbalance cannot be altered once it is set in the initial conditions; but what generates it? This is clearly one of the big challenges in cosmology, but our ideas are less well formed here than in many other areas.

2.3 Recombination

Moving closer to the present, and passing through matter-radiation equality at $z \sim 10^4$, the next critical epoch in the evolution of the universe is reached when the temperature drops to the point ($T \sim 1000$ K) where it is thermodynamically favourable for the ionized plasma to form neutral atoms. This process is known as **recombination**: a complete misnomer, as the plasma has always been completely ionized up to this time.

THE RATE EQUATION A natural first thought is that the ionization of the plasma may be treated by a thermal-equilibrium approach, but such an approach is almost always invalid. This is not because electromagnetic interactions are too slow to maintain equilibrium: rather, thay are too fast. Consider a single recombination; if this were to occur directly to the ground state, a photon with

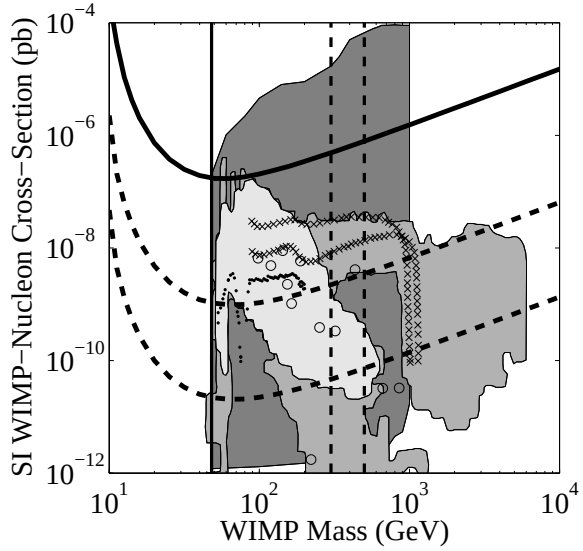

Figure 7. A plot of the dark-matter experimentalists' space: cross-section for scattering off nucleons (in wonderfully baroque units: the 'picobarn' is $10^{-40}\,\mathrm{m}^2$) against WIMP mass. The shaded areas and points indicate various supersymmetric models that match particle-physics constraints and have the correct relic density. The upper curve indicates current direct (non)detection limits, and dashed curves are where we might be in about a decade. Vertical lines are current collider limits, and predictions for the LHC and a future linear collider.

$\hbar\omega > \chi$ would be produced. Such photons are almost immediately destroyed by ionizing another neutral atom. Similarly, reaching the ground state requires the production of photons at least as energetic as the $2P \rightarrow 1S$ spacing (Lyman α, with $\lambda = 1216$Å), and these also are re-absorbed very efficiently. This is a common phenomenon in astrophysics: the Lyman α photons undergo **resonant scattering** and are very hard to get rid of (unlike a finite HII region, where the Lyα photons can escape).

There is a way out, however, using **two-photon emission**. The $2S \rightarrow 1S$ transition is strictly forbidden at first order and one can only conserve energy and angular momentum in the

transition by emitting a *pair* of photons. Because of this slow bottleneck, the ionization at low redshift is well above the equilibrium level.

A highly stripped-down analysis of events simplifies the hydrogen atom to just two levels ($1S$ and $2S$). Any chain of recombinations that reaches the ground state can be ignored through the above argument: these reactions produce photons that are immediately re-absorbed elsewhere, so they have no effect on the ionization balance. The main chance of reaching the ground state comes through the recombinations that reach the $2S$ state, since some fraction of the atoms that reach that state will suffer two-photon decay before being re-excited. The rate equation for the fractional ionization is thus

$$\frac{d(nx)}{dt} = -R\,(nx)^2\,\frac{\Lambda_{2\gamma}}{\Lambda_{2\gamma}+\Lambda_{\rm U}(T)}, \tag{74}$$

where n is the number density of protons, x is the fractional ionization, R is the recombination coefficient ($R \simeq 3{\times}10^{-17}T^{-1/2}\,{\rm m^3 s^{-1}}$), $\Lambda_{2\gamma}$ is the two-photon decay rate, and $\Lambda_{\rm U}(T)$ is the stimulated transition rate upwards from the $2S$ state. This equation just says that recombinations are a two-body process, which create excited states that cascade down to the $2S$ level, from whence a competition between the upward and downward transition rates determines the fraction that make the downward transition.

An important point about the rate equation is that it is only necessary to solve it once, and the results can then be scaled immediately to some other cosmological model. Consider the rhs: both R and $\Lambda_{\rm U}(T)$ are functions of temperature, and thus of redshift only, so that any parameter dependence is carried just by n^2, which scales $\propto (\Omega_b h^2)^2$, where Ω_b is the baryonic density parameter. Similarly, the lhs depends on $\Omega_b h^2$ through n; the other parameter dependence comes if we convert time derivatives to derivatives with respect to redshift:

$$\frac{dt}{dz} \simeq -3.09\times10^{17}(\Omega_m h^2)^{-1/2}\,z^{-5/2}\ {\rm s}, \tag{75}$$

for a matter-dominated model at large redshift. Putting these together, the fractional ionization must scale as

$$x(z) \propto \frac{(\Omega_m h^2)^{1/2}}{\Omega_b h^2}. \tag{76}$$

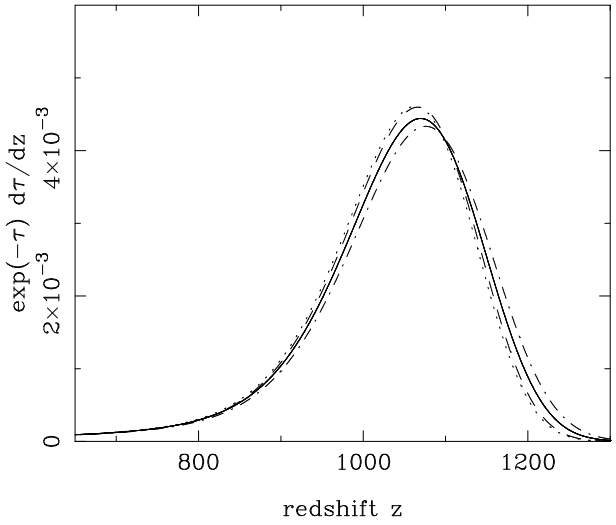

Figure 8. The 'visibility function' governing where photons in the CMB undergo their final scattering. This is very nearly independent of cosmological parameters, as illustrated by the effect of a 50% increase in Ω_b (dotted line), Ω_m (dot-dashed line) and h (dashed line), relative to the standard model (solid line).

LAST SCATTERING Recombination is important observationally because it marks the first time that photons can travel freely. When the ionization is high, Thomson scattering causes them to proceed in a random walk, so the early universe is opaque. The interesting thing from our point of view is to work out the maximum redshift from which we can receive a photon without it suffering scattering. To do this, we work out the optical depth to Thomson scattering,

$$\tau = \int n_e^{\rm tot} x \sigma_{\rm T} d\ell_{\rm proper}; \qquad d\ell_{\rm proper} = R(z)\, dr = R_0\, dr/(1+z). \tag{77}$$

For a fully ionized plasma with 25% He by mass, the total electron number density is

$$n_e^{\rm tot}(z) = 9.83\, \Omega_b h^2\, (1+z)^3\ {\rm m}^{-3}. \tag{78}$$

Also, $d\ell_{\rm proper} = c\,dt$, which brings in a factor of $(\Omega_m h^2)^{-1/2}$. These two density terms automatically cancel the principal dependence of $x(z)$, so we predict that the optical depth should be very largely a function of redshift only. For standard parameters, a good approximation around $\tau = 1$ is

$$\tau(z) \simeq \left(\frac{1+z}{1080}\right)^{13} \tag{79}$$

(*cf.* Jones & Wyse 1985; A&A 149, 144).

This approximation is not perfect, however, and very accurate work needs detailed numerical solution of the evolution equations, including the omitted processes. See Seager, Sasselov & Scott (2000; ApJS, 128, 407). Because τ changes rapidly with redshift, the **visibility function** for the redshift at which photons were last scattered, $e^{-\tau} d\tau/dz$, is sharply peaked, and is well fitted by a Gaussian of mean redshift 1070 and standard deviation in redshift 80. As illustrated in figure 8, these properties are in practice insensitive to the cosmological parameters. Thus, when we look at the sky, we can expect to see in all directions photons that originate from a **last-scattering surface** at $z \simeq 1100$.

3 Inflation – I

Topics to be covered:

- Initial condition problems
- Dynamics of scalar fields
- Noether's theorem

3.1 Initial condition problems

The expanding universe of the big-bang model is surprising in many ways: (1) What caused the expansion? (2) Why is the expansion so close to flat – i.e. $\Omega \sim 1$ today? (3) Why is the universe close to isotropic (the same in all directions)? (4) Why does it contain structure? Some of these problems may seem larger than others, but when examined in detail all point to something missing in our description of the early stages of cosmological history.

QUANTUM GRAVITY LIMIT In principle, $T \to \infty$ as $R \to 0$, but there comes a point at which this extrapolation of classical physics breaks down. This is where the thermal energy of typical particles is such that their de Broglie wavelength is smaller than their Schwarzschild radius: quantum black holes clearly cause difficulties with the usual concept of background spacetime. Equating $2\pi\hbar/(mc)$ to $2Gm/c^2$ yields a characteristic mass for quantum gravity known as the **Planck mass**. This mass, and the corresponding length $\hbar/(m_{\rm P}c)$ and time $\ell_{\rm P}/c$ form the system of **Planck units**:

$$\begin{aligned} m_{\rm P} &\equiv \sqrt{\frac{\hbar c}{G}} \simeq 10^{19}\text{GeV} \\ \ell_{\rm P} &\equiv \sqrt{\frac{\hbar G}{c^3}} \simeq 10^{-35}\text{m} \\ t_{\rm P} &\equiv \sqrt{\frac{\hbar G}{c^5}} \simeq 10^{-43}\text{s}. \end{aligned} \tag{80}$$

The Planck time therefore sets the origin of time for the classical phase of the big bang. It is incorrect to extend the classical solution to $R = 0$ and conclude that the universe began in a singularity of

infinite density. A common question about the big bang is 'what happened at $t < 0$?', but in fact it is not even possible to get to zero time without adding new physical laws.

NATURAL UNITS To simplify the appearance of equations, it is common practice in high-energy physics to adopt **natural units**, where we take

$$k = \hbar = c = \mu_0 = \epsilon_0 = 1. \tag{81}$$

This convention makes the meaning of equations clearer by reducing the algebraic clutter, and is also useful in the construction of intuitive arguments for the order of magnitude of quantities of interest. Hereafter, natural units will frequently be adopted, although it will occasionally be convenient to re-insert explicit powers of $\hbar$ *etc.*

The adoption of natural units corresponds to fixing the units of charge, mass, length and time relative to each other. This leaves one free unit, usually taken to be energy. Natural units are thus one step short of the Planck system, in which $G = 1$ also, so that all units are fixed and all physical quantities are dimensionless. In natural units, the following dimensional equalities hold:

$$\begin{aligned} [E] = [T] &= [m] \\ [L] &= [m]^{-1} \end{aligned} \tag{82}$$

Hence, the dimensions of energy density are

$$[u] = [m]^4, \tag{83}$$

with units often quoted in GeV^4. It is however often of interest to express things in Planck units: energy as a multiple of m_{P}, energy density as a multiple of m_{P}^4 *etc.* The gravitational constant itself is then

$$G = m_{\mathrm{P}}^{-2}. \tag{84}$$

FLATNESS PROBLEM Now to quantify the first of the many puzzles concerning initial conditions. From the Friedmann equation, we can write the density parameter as a function of era:

$$\Omega(a) = \frac{8\pi G\rho(a)}{H^2(a)} = \frac{\Omega_v + \Omega_m a^{-3} + \Omega_r a^{-4}}{\Omega_v + \Omega_m a^{-3} + \Omega_r a^{-4} - (\Omega - 1)a^{-2}} \tag{85}$$

(and corresponding expressions for the $\Omega(a)$ corresponding to any one component just by picking the appropriate term on the top line). This tells us that, if the total Ω is unity today, it was always unity (a geometrical statement: if $k = 0$, it can't make a continuous transition to $k = \pm 1$). But if $\Omega \neq 1$, how does $\Omega(a)$ evolve? It should be clear that $\Omega(a) \to 1$ at very large and very small a, provided Ω_v is nonzero in the former case, and provided Ω_m or Ω_r is nonzero in the latter case (without vacuum energy, $\Omega = 1$ is unstable). In short, the $\Omega = 1$ state is an **attractor**, looking in either direction in time. It has long been clear that this presents a puzzle with regard to the initial conditions. These will be radiation dominated, so we have

$$\Omega(a_{\rm init}) \simeq 1 + \frac{(\Omega - 1)}{\Omega_r}\, a_{\rm init}^2. \tag{86}$$

If we are willing to consider a Planck-scale origin with $a_{\rm init} \sim 10^{-32}$, then clearly conditions at that time must be flat to perhaps 60 powers of 10. A more democratic initial condition might be thought to have $\Omega(a_{\rm init}) - 1$ of order unity, so some mechanism to make it very small (or zero) is clearly required. This 'how could the universe have known?' argument is a general basis for a prejudice that $\Omega = 1$ holds exactly today.

HORIZON PROBLEM We have already mentioned the puzzle that it has apparently been impossible to establish causal contact throughout the present observable universe. Consider the integral for the horizon length:

$$r_{\rm H} = \int \frac{c\,dt}{R(t)}. \tag{87}$$

The standard radiation-dominated $R \propto t^{1/2}$ law makes this integral converge near $t = 0$. To solve the horizon problem and allow causal contact over the whole of the region observed at last scattering requires a universe that expands 'faster than light' near $t = 0$: $R \propto t^\alpha$, with $\alpha > 1$. It is tempting to assert that the observed homogeneity *proves* that such causal contact must once have occurred, but this means that the equation of state at early times must have been different. Indeed, if we look at Friedmann's equation in its second form,

$$\ddot{R} = -4\pi G R(\rho + 3p/c^2)/3, \tag{88}$$

and realize that $R \propto t^\alpha$, with $\alpha > 1$ implies an accelerating expansion, we see that what is needed is negative pressure:

$$\rho c^2 + 3p < 0. \tag{89}$$

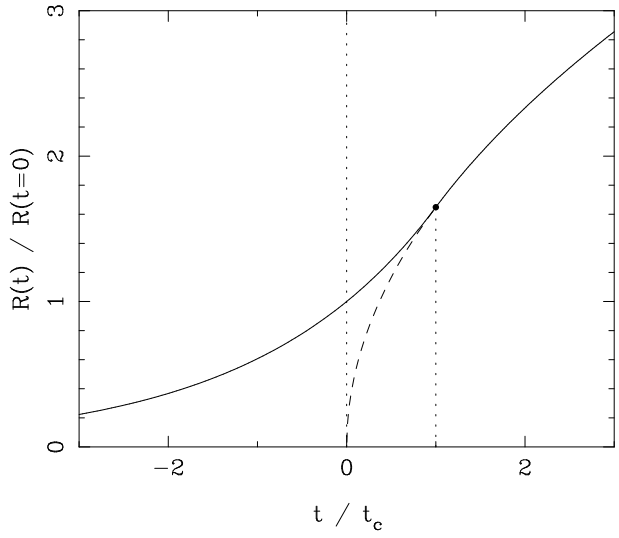

Figure 9. Illustrating the true history of the scale factor in the simplest possible inflationary model. Here, the universe stays in an exponential de Sitter phase for an indefinite time until its equation of state abruptly changes from vacuum dominated to radiation dominated at time t_c. This must occur in such a way as to match R and R, leading to the solid curve, where the plotted point indicates the join. For $0 < t < t_c$, the dashed curve indicates the time dependence we would infer if vacuum energy was ignored. This reaches $R = 0$ at $t = 0$: the classical 'big bang'. The inflationary solution clearly removes this feature, placing any singularity at large negative time. The universe is much older than we would expect from observations at $t > t_c$, which is one way of seeing how the horizon problem can be evaded.

DE SITTER SPACE The familiar example of negative pressure is vacuum energy, and this is therefore a hint that the universe may have been vacuum-dominated at early times. The Friedmann equation in the $k = 0$ vacuum-dominated case has the **de Sitter solution**:

$$R \propto \exp Ht, \tag{90}$$

where $H = \sqrt{8\pi G\rho_{\rm vac}/3}$. This is the basic idea of the **inflationary universe**: vacuum repulsion can cause the universe to expand at an ever-increasing rate. This launches the Hubble expansion,

and solves the horizon problem by stretching a small causally-connected patch to a size large enough to cover the whole presently-observable universe.

This is illustrated by in figure 9, where we assume that the universe can be made to change its equation of state abruptly from vacuum dominated to radiation dominated at some time t_c. Before t_c, we have $R \propto \exp Ht$; after t_c, $R \propto t^{1/2}$. We have to match R and $\dot{R}$ at the join; it is then easy to show that $t_c = 1/2H$. In principle, the question 'what happened before the big bang?' is now answered: there was no big bang. There might have still been a singularity at large negative time, but one could imagine the de Sitter phase being of indefinite duration. In a sense, then, an inflationary start to the expansion would in reality be a very slow one – as compared to the common popular description of 'an extraordinarily rapid phase of expansion'.

This idea of a non-singular origin to the universe was first proposed by the Soviet cosmologist E.B. Gliner, in 1969. He suggested no mechanism by which the vacuum energy could change its level, however. Before trying to plug this critical gap, we can note that an early phase of vacuum-dominated expansion can also solve the flatness problem. Consider the Friedmann equation,

$$\dot{R}^2 = \frac{8\pi G\rho R^2}{3} - kc^2. \tag{91}$$

In a vacuum-dominated phase, ρR^2 increases as the universe expands. This term can therefore always be made to dominate over the curvature term, making a universe that is close to being flat (the curvature scale has increased exponentially). In more detail, the Friedmann equation in the vacuum-dominated case has three solutions:

$$R \propto \begin{cases} \sinh Ht & (k=-1) \\ \cosh Ht & (k=+1) \\ \exp Ht & (k=0), \end{cases} \tag{92}$$

where $H = \sqrt{8\pi G\rho_{\rm vac}/3}$. Note that H is not the Hubble parameter at an arbitrary time (unless $k = 0$), but it becomes so exponentially fast as the hyperbolic trigonometric functions tend to the exponential. If we assume that the initial conditions are not fine tuned (*i.e.* $\Omega = O(1)$ initially), then maintaining the expansion for a factor f produces

$$\Omega = 1 + O(f^{-2}). \tag{93}$$

This can solve the flatness problem, provided f is large enough. To obtain Ω of order unity today requires $|\Omega - 1| \lesssim 10^{-52}$ at the GUT epoch, and so

$$\ln f \gtrsim 60 \tag{94}$$

e-foldings of expansion are needed; it will be proved below that this is also exactly the number needed to solve the horizon problem. It then seems almost inevitable that the process should go to completion and yield $\Omega = 1$ to measurable accuracy today. This is one of the most robust predictions of inflation (although, as we have seen, the expectation of flatness is fairly general).

HOW MUCH INFLATION DO WE NEED? To be quantitative, we have to decide when inflation is to happen. The earliest possible time is at the Planck era, $t \simeq 10^{-43}$ s, at which point the causal scale was $ct \simeq 10^{-35}$ m. What comoving scale is this? The redshift is roughly (ignoring changes in $g_{\rm eff}$) the Planck energy (10^{19} GeV) divided by the CMB energy ($kT \simeq 10^{-3.6}$ eV), or

$$z_{\rm P} \simeq 10^{31.6}. \tag{95}$$

This expands the Planck length to 0.4 mm today. This is far short of the present horizon ($\sim 6000\,h^{-1}\,$Mpc), by a factor of nearly 10^{30}, or e^{69}. It is more common to assume that inflation happened at a safer distance from quantum gravity, at about the GUT energy of 10^{15} GeV. The GUT-scale horizon needs to be stretched by 'only' a factor e^{60} in order to be compatible with observed homogeneity. This tells us a minimum duration for the inflationary era:

$$\Delta t_{\rm inflation} > 60\,H^{-1}_{\rm inflation}. \tag{96}$$

The GUT energy corresponds to a time of about 10^{-35} s in the conventional radiation-dominated model, and we have seen that this switchover time should be of order $H^{-1}_{\rm inflation}$. Therefore, the whole inflationary episode need last no longer than about 10^{-33} s.

3.2 Dynamics of scalar fields

Since 1981, these ideas have been set on a more specific foundation using models for a variable vacuum energy that come from particle physics. There are many variants, but the simplest concentrate on **scalar fields**. These are fields like the electromagnetic field, but differing in a number of respects. First, the field has only one degree of freedom: just a number that varies with position, not a vector like the EM field. The wave equation obeyed by such a field in flat space is the **Klein–Gordon equation**:

$$\frac{1}{c^2}\ddot{\phi} - \nabla^2\phi + (m^2c^2/\hbar^2)\phi = 0, \tag{97}$$

which is just the standard wave equation if $m = 0$. This is easy to derive just by substituting the de Broglie relations $\mathbf{p} = -i\hbar\boldsymbol{\nabla}$ and $E = i\hbar\partial/\partial t$ into $E^2 = p^2c^2 + m^2c^4$. To apply this to cosmology, we neglect the spatial derivatives, since we imagine some initial domain in which we have a **homogeneous scalar field**. This synchronizes the subsequent dynamics of $\phi(t)$ throughout the observable universe (*i.e.* the patch that we inflate). The differential equation is now

$$\ddot{\phi} = -\frac{d}{d\phi}V(\phi); \quad V(\phi) = (m^2c^4/\hbar^2)\phi^2/2. \tag{98}$$

This is just a harmonic oscillator equation, and we can see that the field will oscillate in the potential, with 'kinetic energy' $T = \dot{\phi}^2/2$. This behaviour is rather different to the familiar oscillations of the electromagnetic field: if the field is homogeneous, it does not oscillate. This is because the familiar energy density in electromagnetism ($\epsilon_0E^2/2 + B^2/2\mu_0$) is entirely kinetic energy in this analogy (to see this, write the fields in terms of the potentials: $\mathbf{B} = \boldsymbol{\nabla}\wedge\mathbf{A}$ and $\mathbf{E} = -\boldsymbol{\nabla}\phi - \dot{\mathbf{A}}$. We don't see coherent oscillations in electromagnetism because the photon has no mass.

We will show below that, not only does $V(\phi)$ play the role of a potential energy in the equation of motion, it acts as a physical energy density in space. This potential energy density is equivalent to a vacuum density: its gravitational properties are repulsive and can cause an inflationary phase of exponential expansion. In this simple model, the universe is started in a potential-dominated state, and inflates until the field falls enough that the kinetic energy becomes important. In practical models, this stage will be associated with **reheating**: although weakly interacting, the field does couple to other particles, and its oscillations can generate other particles – thus transforming the scalar-field energy into energy of a normal radiation-dominated universe.

LAGRANGIANS AND FIELDS To understand what scalar fields can do for cosmology, it is necessary to use some elements of the more powerful Lagrangian description of the dynamics. We will try to keep this fairly informal. Consider first a classical system of particles: the **Lagrangian** L is defined as the difference of the kinetic and potential energies, $L = T - V$, for some set of particles with coordinates $q_i(t)$. **Euler's equation** gives an equation of motion for each particle

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right) = \frac{\partial L}{\partial q_i}. \tag{99}$$

As a sanity check, consider a single particle in a potential in 1D: $L = m\dot{x}^2/2 - V(x)$. $\partial L/\partial \dot{x} = m\dot{x}$, so we get $m\ddot{x} = -\partial V/\partial x$, as desired. The advantage of the Lagrangian formalism, of course, is that it is not necessary to use Cartesian coordinates. In passing, we note that the formalism also supplies a general definition of momentum:

$$p_i \equiv \frac{\partial L}{\partial \dot{q}_i}, \tag{100}$$

which again is clearly sensible for Cartesian coordinates.

A field may be regarded as a dynamical system, but with an infinite number of degrees of freedom. How do we handle this? A hint is provided by electromagnetism, where we are familiar with writing the total energy in terms of a density which, as we are dealing with generalized mechanics, we may formally call the Hamiltonian density:

$$H = \int \mathcal{H}\, dV = \int \left(\frac{\epsilon_0 E^2}{2} + \frac{B^2}{2\mu_0}\right)\, dV. \tag{101}$$

This suggests that we write the Lagrangian in terms of a **Lagrangian density** $\mathcal{L}$: $L = \int \mathcal{L}\, dV$. This quantity is of such central importance in quantum field theory that it is usually referred to (incorrectly) simply as 'the Lagrangian'. The equation of motion that corresponds to Euler's equation is now the **Euler–Lagrange equation**

$$\frac{\partial}{\partial x^\mu}\left[\frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)}\right] = \frac{\partial \mathcal{L}}{\partial \phi}, \tag{102}$$

where we use the shorthand $\partial_\mu \phi \equiv \partial\phi/\partial x^\mu$. Note the downstairs index for consistency: in special relativity, $x^\mu = (ct, \mathbf{x})$, $x^\mu = (ct, -\mathbf{x}) = g_{\mu\nu}x^\nu$. The Lagrangian $\mathcal{L}$ and the field equations are

therefore generally equivalent, although the Lagrangian arguably seems more fundamental: we can obtain the field equations given the Lagrangian, but inverting the process is less straightforward.

For quantum mechanics, we want a Lagrangian that will yield the Klein–Gordon equation. If ϕ is a single real scalar field, then the required Lagrangian is

$$\mathcal{L} = \tfrac{1}{2}\partial^\mu\phi\partial_\mu\phi - V(\phi); \quad V(\phi) = \tfrac{1}{2}\mu^2\phi^2. \tag{103}$$

Again, we will be content with checking that this does the right thing in a simple case: the homogeneous model, where $\mathcal{L} = \dot{\phi}^2/2 - V(\phi)$. This is now just like the earlier example, and gives $\ddot{\phi} = -\partial V/\partial\phi$, as required.

NOETHER'S THEOREM The final ingredient we need before applying scalar fields to cosmology is to understand that they can be treated as a fluid with thermodynamic properties like pressure. these properties are derived by a profoundly important general argument that relates the existence of global symmetries to conservation laws in physics. In classical mechanics, conservation of energy and momentum arise by considering Euler's equation

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right) - \frac{\partial L}{\partial q_i} = 0, \tag{104}$$

where $L = \sum_i T_i - V_i$ is a sum over the difference in kinetic and potential energies for the particles in a system. If L is independent of all the position coordinates q_i, then we obtain conservation of momentum (or angular momentum, if q is an angular coordinate): $p_i \equiv \partial L/\partial\dot{q}_i =$ constant for each particle. More realistically, the potential will depend on the q_i, but homogeneity of space says that the Lagrangian as a whole will be unchanged by a **global translation**: $q_i \to q_i + dq$, where dq is some constant. Using Euler's equation, this gives conservation of total momentum:

$$dL = \sum_i \frac{\partial L}{\partial q_i}\, dq \quad \Rightarrow \quad \frac{d}{dt}\sum_i p_i = 0. \tag{105}$$

If L has no explicit dependence on t, then

$$\frac{dL}{dt} = \sum_i \left(\frac{\partial L}{\partial q_i}\dot{q}_i + \frac{\partial L}{\partial \dot{q}_i}\ddot{q}_i\right) = \sum_i (\dot{p}_i\dot{q}_i + p_i\ddot{q}_i), \tag{106}$$

which leads us to define the **Hamiltonian** as a further constant of the motion

$$H \equiv \sum_i p_i \dot{q}_i - L = \text{constant}. \tag{107}$$

Something rather similar happens in the case of quantum (or classical) field theory: the existence of a global symmetry leads directly to a conservation law. The difference between discrete dynamics and field dynamics, where the Lagrangian is a *density*, is that the result is expressed as a **conserved current** rather than a simple constant of the motion. Suppose the Lagrangian has no explicit dependence on spacetime (*i.e.* it depends on x^μ only implicitly through the fields and their 4-derivatives). As above, we write

$$\frac{d\mathcal{L}}{dx^\mu} = \frac{\partial \mathcal{L}}{\partial \phi}\,\frac{\partial \phi}{\partial x^\mu} + \frac{\partial \mathcal{L}}{\partial(\partial_\nu \phi)}\,\frac{\partial(\partial_\nu \phi)}{\partial x^\mu}, \tag{108}$$

Using the Euler–Lagrange equation to replace $\partial\mathcal{L}/\partial\phi$ and collecting terms results in

$$\frac{d}{dx^\nu}\left[\frac{\partial \mathcal{L}}{\partial(\partial_\nu \phi)}\,\frac{\partial \phi}{\partial x^\mu} - \mathcal{L} g^{\mu\nu}\right] \equiv \frac{d}{dx^\nu} T^{\mu\nu} = 0. \tag{109}$$

This is a conservation law, as we can see by analogy with a simple case like the conservation of charge. There, we would write

$$\partial_\mu J^\mu = \dot{\rho} + \boldsymbol{\nabla} \cdot \mathbf{j} = 0, \tag{110}$$

where ρ is the charge density, $\mathbf{j}$ is the current density, and J^μ is the 4-current. We have effectively four such equations (one for each value of ν) so there must be four conserved quantities: clearly energy and the four components of momentum. Conservation of 4-momentum is expressed by $T^{\mu\nu}$, which is the 4-current of 4-momentum. For a simple fluid, it is just

$$T^{\mu\nu} = \text{diag}(\rho c^2, p, p, p), \tag{111}$$

so now we can read off the density and pressure generated by a scalar field. Note immediately the important consequence for cosmology: a potential term $-V(\phi)$ in the Lagrangian produces $T^{\mu\nu} = V(\phi) g^{\mu\nu}$. This is the $p = -\rho$ equation of state characteristic of the cosmological constant. If we now follow the evolution of ϕ, the cosmological 'constant' changes and we have the basis for models of inflationary cosmology.

4 Inflation – II

Topics to be covered:

- Models for inflation
- Slow roll dynamics
- Ending inflation

4.1 Equation of motion

Most of the main features of inflation can be illustrated using the simplest case of a single real scalar field, with Lagrangian

$$\mathcal{L} = \tfrac{1}{2}\partial_\mu\phi\,\partial^\mu\phi - V(\phi) = \tfrac{1}{2}(\dot{\phi}^2 - \nabla^2\phi) - V(\phi). \tag{112}$$

It turns out that we can get inflation with even the simple mass potential $V(\phi) = m^2\,\phi^2/2$, but it is easy to keep things general. Noether's theorem gives the energy–momentum tensor for the field as

$$T^{\mu\nu} = \partial^\mu\phi\partial^\nu\phi - g^{\mu\nu}\mathcal{L}. \tag{113}$$

From this, we can read off the energy density and pressure:

$$\begin{aligned} \rho &= T^{00} = \tfrac{1}{2}\dot{\phi}^2 + V(\phi) + \tfrac{1}{2}(\nabla\phi)^2 \\ p &= T^{11} = \tfrac{1}{2}\dot{\phi}^2 - V(\phi) - \tfrac{1}{6}(\nabla\phi)^2. \end{aligned} \tag{114}$$

If the field is constant both spatially and temporally, the equation of state is then $p = -\rho$, as required if the scalar field is to act as a cosmological constant; note that derivatives of the field spoil this identification.

We now want to revisit the equation of motion for the scalar field, but with the critical difference that we place the field in the expanding universe. Everything so far has been special relativity, so we don't have quite enough formalism to derive the full equation of motion, but it is

$$\ddot{\phi} + 3H\dot{\phi} - \nabla^2\phi + dV/d\phi = 0. \tag{115}$$

This is a wave equation similar to the one in flat space. The **Hubble drag** term $3H\dot{\phi}$ is the main new feature: loosely, it reflects the fact that the redshifting effects of expansion will drain energy from the field oscillations.

This is not hard to prove in the homogeneous case, which is the main one of interest for inflationary applications. This is because $\nabla\phi = \nabla_{\rm comoving}\,\phi/R$. Since R increases exponentially, these perturbations are damped away: assuming V is large enough for inflation to start in the first place, inhomogeneities rapidly become negligible. In the homogeneous limit, we can simply appeal to energy conservation:

$$\frac{d\ln\rho}{d\ln a} = -3(1+w) = -3\dot{\phi}^2/(\dot{\phi}^2/2 + V), \tag{116}$$

following which the relations $H = d\ln a/dt$ and $\dot{V} = \dot{\phi}V'$ can be used to change variables to t, and the damped oscillator equation for ϕ follows.

4.2 The slow-roll approximation

The solution of the equation of motion becomes tractable if we both ignore spatial inhomogeneities in ϕ and make the **slow-rolling approximation** that the $\ddot{\phi}$ term is negligible. The physical motivation here is to say that we are most interested in behaviour close to de Sitter space, so that the potential dominates the energy density. This requires

$$\dot{\phi}^2/2 \ll |V(\phi)|; \tag{117}$$

differentiating this gives $\ddot{\phi} \ll |dV/d\phi|$, as required. We therefore have a simple slow-rolling equation for homogeneous fields:

$$3H\dot{\phi} = -dV/d\phi. \tag{118}$$

In combination with Friedmann's equation in the natural-unit form

$$H^2 = \frac{8\pi}{3m_{\rm P}^2}(\dot{\phi}^2/2 + V) \simeq \frac{8\pi}{3m_{\rm P}^2}V, \tag{119}$$

This gives a powerful but simple apparatus for deducing the expansion history of any inflationary model.

The conditions for inflation can be cast into useful dimensionless forms. The basic condition $V \gg \dot{\phi}^2$ can now be rewritten using the slow-roll relation as

$$\epsilon \equiv \frac{m_{\rm P}^2}{16\pi} (V'/V)^2 \ll 1. \tag{120}$$

Also, we can differentiate this expression to obtain the criterion $V'' \ll V'/m_{\rm P}$, or $m_{\rm P}V''/V \ll V'/V \sim \sqrt{\epsilon}/m_{\rm P}$. This gives a requirement for the second derivative of V to be small, which we can write as

$$\eta \equiv \frac{m_{\rm P}^2}{8\pi} (V''/V) \ll 1 \tag{121}$$

These two criteria make perfect intuitive sense: the potential must be flat in the sense of having small derivatives if the field is to roll slowly enough for inflation to be possible.

4.3 Inflationary models

The curse and joy of inflationary modelling is that nothing is known about the **inflaton** field ϕ, nor about its potential. We therefore consider simple classes of possible example models, with varying degrees of physical motivation.

If we think about a single field, models can be divided into two basic classes, as illustrated in figure 10. The simplest are **large-field inflation** models, in which the field is strongly displaced from the origin. There is nothing to prevent the scalar field from reaching the minimum of the potential – but it can take a long time to do so, and the universe meanwhile inflates by a large factor. In this case, inflation is realized by means of 'inertial confinement'. The opposite is when the potential is something like the Higgs potential, where the gradient vanishes at the origin: this is a model of **small-field inflation**. In principle, the field can stay at $\phi = 0$ forever if it is placed exactly there. One would say that the universe then inhabited a state of **false vacuum**, as opposed to the true vacuum at $V = 0$ (but it is important to be clear that there is no fundamental reason why the minimum should be at zero density exactly; we will return to this point).

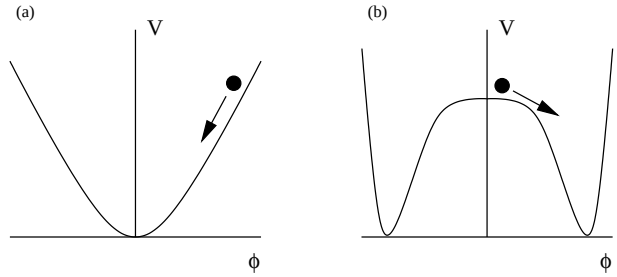

Figure 10. The two main classes of single-field inflation models: (a) large-field inflation; (b) small-field inflation. The former is motivated by a mass-like potential, the latter by something more like the Higgs potential.

The first inflation model (Guth 1981) was of the small-field type, but large-field models have tended to be considered more plausible, for two reasons. The first is to do with initial conditions. If inflation starts from anywhere near to thermal equilibrium at a temperature $T_{\rm GUT}$, we expect thermal fluctuations in ϕ; the potential should generally differ from its minimum by an amount $V \sim T_{\rm GUT}^4$. How then is the special case needed to trap the potential near $\phi = 0$ to arise? We have returned to the sort of fine-tuned initial conditions from which inflation was designed to save us. The other issue with simple small-field models relates to the issue of how inflation ends. This can be viewed as a form of phase transition, which is continuous or second order in the case of large-field models. For small-field models, however, the transition to the true vacuum can come about by quantum tunnelling, so that the transition is effectively discontinuous and first order. As we will discuss further below, this can lead to a universe that is insufficiently homogeneous to be consistent with observations.

CHAOTIC INFLATION MODELS Most attention is therefore currently paid to the large-field models where the field finds itself some way from its potential minimum. This idea is also termed **chaotic inflation**: there could be primordial chaos, within which conditions might vary. Some parts may attain the vacuum-dominated conditions needed for inflation, in which case they will expand hugely, leaving a universe inside a single bubble – which could be the one we inhabit. In principle this bubble has an edge, but if inflation persists for sufficiently long, the distance to this nastiness is so much greater than the current particle horizon that its existence has no testable consequences.

A wide range of inflation models of this kind is possible, but it will suffice here to discuss two simple special cases:

(1) **Polynomial inflation.** If the potential is taken to be $V \propto \phi^\alpha$, then the scale-factor behaviour can be very close to exponential. This becomes less true as α increases, but investigations are usually limited to ϕ^2 and ϕ^4 potentials on the grounds that higher powers are nonrenormalizable.

(2) **Power-law inflation.** On the other hand, $a(t) \propto t^p$ would suffice, provided $p > 1$. The potential required to produce this behaviour is

$$V(\phi) \propto \exp\left(\sqrt{\frac{16\pi}{p\, m_{\rm P}^2}}\,\phi\right). \tag{122}$$

This is an exact solution, not a slow-roll approximation.

HYBRID INFLATION One way in which the symmetric nature of the initial condition for small-field inflation can be made more plausible is to go beyond the space of single-field inflation. The most popular model in this generalized class is **hybrid inflation**, in which there are two fields, with potential

$$V(\phi,\psi) = \frac{1}{4\lambda}(M^2 - \lambda\psi^2)^2 + U(\phi) + \tfrac{1}{2}g^2\psi^2\phi^2. \tag{123}$$

We can think of this as being primarily $V(\psi)$, but with the form of V controlled by the second field, ϕ. For $\phi = 0$, we have the standard symmetry-breaking potential; but for large ϕ, $\phi > M/g$, the dependence on ψ becomes parabolic. Evolution in this parabolic trough at large ϕ can thus naturally lower ψ close to $\psi = 0$. If this happens, we have inflation driven by ϕ as the inflaton, with $V(\phi) = U(\phi) + \lambda M^2/4$. This extra constant in the potential raises H, so the Hubble damping term is particularly high, keeping the field from rolling away from $\psi = 0$ until near to $\phi = 0$.

Hybrid inflation therefore has the ability to make some of the features of the simplest inflation models seem more plausible, while introducing sufficient extra complexity that one can try to test

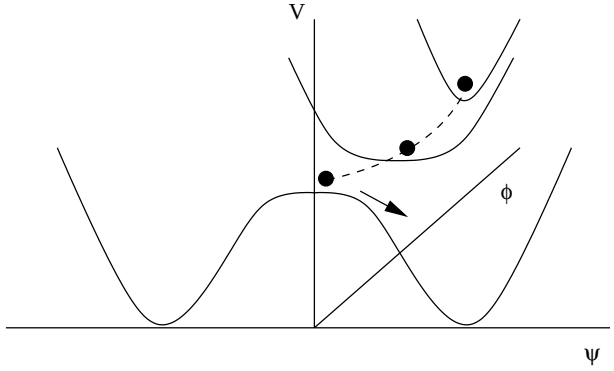

Figure 11. A sketch of the potential in hybrid inflation. For $\phi = 0$, $V(\psi)$ has the symmetry-breaking form of the potential for small-field inflation, but for large ϕ there is a simple quadratic minimum in $V(\psi)$. Evolution in this potential can drive conditions towards $\psi = 0$ while ϕ is large, preparing the way for something similar to small-field inflation.

the robustness of the predictions of the simple models. The form of the Lagrangian is also claimed to have some fundamental motivation (although this has been said of many Lagrangians). As a result, hybrid inflation is rather popular with inflationary theorists.

CRITERIA FOR INFLATION Successful inflation in any of these models requires > 60 e-foldings of the expansion. The implications of this are easily calculated using the slow-roll equation, which gives the number of e-foldings between ϕ_1 and ϕ_2 as

$$N = \int H\, dt = -\frac{8\pi}{m_{\rm P}^2} \int_{\phi_1}^{\phi_2} \frac{V}{V'}\, d\phi \tag{124}$$

For a potential that resembles a smooth polynomial, $V' \sim V/\phi$, and so we typically get $N \sim (\phi_{\rm start}/m_{\rm P})^2$, assuming that inflation terminates at a value of ϕ rather smaller than at the start. The criterion for successful inflation is thus that the initial value of the field exceeds the Planck scale:

$$\phi_{\rm start} \gg m_{\rm P}. \tag{125}$$

This is the real origin of the term 'large-field': it means that ϕ has to be large in comparison to the Planck scale. By the same argument, it is easily seen that this is also the criterion needed to make the slow-roll parameters ϵ and $\eta \ll 1$. To summarize, any model in which the potential is sufficiently flat that slow-roll inflation can commence will probably achieve the critical 60 e-foldings.

It is interesting to review this conclusion for some of the specific inflation models listed above. Consider a mass-like potential $V = m^2\phi^2$. If inflation starts near the Planck scale, the fluctuations in V are presumably $\sim m_{\rm P}^4$ and these will drive $\phi_{\rm start}$ to $\phi_{\rm start} \gg m_{\rm P}$ provided $m \ll m_{\rm P}$; similarly, for $V = \lambda\phi^4$, the condition is weak coupling: $\lambda \ll 1$. Any field with a rather flat potential will thus tend to inflate, just because typical fluctuations leave it a long way from home in the form of the potential minimum.

This requirement for weak coupling and/or small mass scales near the Planck epoch is suspicious, since quantum corrections will tend to re-introduce the Planck scale. In this sense, especially with the appearance of the Planck scale as the minimum required field value, it is not clear that the aim of realizing inflation in a classical way distinct from quantum gravity has been fulfilled.

4.4 Ending inflation

BUBBLE NUCLEATION AND THE GRACEFUL EXIT In small-field inflation, as in with Guth's initial idea, the potential is trapped at $\phi = 0$, and eventually undergoes a first-order phase transition. This model suffers from the problem that it predicts residual inhomogeneities after inflation is over that are far too large. This is easily seen: because the transition is first-order, it proceeds by **bubble nucleation**, where the vacuum tunnels between false and true vacua. However, the region occupied by these bubbles will grow as a causal process, whereas outside the bubbles the exponential expansion of inflation continues. This means that it is very difficult for the bubbles to percolate and eliminate the false vacuum everywhere, as is needed for an end to inflation. Instead, inflation continues indefinitely, with the bubbles of true vacuum having only a small filling factor at any time. This **graceful exit problem** motivated variants in which the phase transition is second order, and proceeds continuously by the field rolling slowly but freely down the potential.

REHEATING As we have seen, slow-rolling behaviour requires the field derivatives to be negligible; but the relative importance of time derivatives increases as V approaches zero (if the minimum is indeed at zero energy). Even if the potential does not steepen, sooner or later we will have $\epsilon \simeq 1$ or $|\eta| \simeq 1$ and the inflationary phase will cease. Instead of rolling slowly 'downhill', the field will oscillate

about the bottom of the potential, with the oscillations becoming damped by the $3H\dot{\phi}$ friction term. Eventually, we will be left with a stationary field that either continues to inflate without end, if $V(\phi = 0) > 0$, or which simply has zero density.

However, this conclusion is incomplete, because we have so far neglected the couplings of the scalar field to matter fields. Such couplings will cause the rapid oscillatory phase to produce particles, leading to **reheating**. Thus, even if the minimum of $V(\phi)$ is at $V = 0$, the universe is left containing roughly the same energy density as it started with, but now in the form of normal matter and radiation – which starts the usual FRW phase, albeit with the desired special 'initial' conditions.

As well as being of interest for completing the picture of inflation, it is essential to realize that these closing stages of inflation are the *only* ones of observational relevance. Inflation might well continue for a huge number of e-foldings, all but the last few satisfying $\epsilon, \eta \ll 1$. However, the scales that left the de Sitter horizon at these early times are now vastly greater than our observable horizon, c/H_0, which exceeds the de Sitter horizon by only a finite factor – about e^{60} for GUT-scale inflation, as we saw earlier. Realizing that the observational regime corresponds only to the terminal phases of inflation is both depressing and stimulating: depressing, because ϕ may well not move very much during the last phases – our observations relate only to a small piece of the potential, and we cannot hope to recover its form without substantial *a priori* knowledge; stimulating, because observations even on very large scales must relate to a period where the simple concepts of exponential inflation and scale-invariant density fluctuations were coming close to breaking down. This opens the possibility of testing inflation theories in a way that would not be possible with data relating to only the simpler early phases. These tests take the form of tilt and gravitational waves in the final perturbation spectrum, to be discussed further below.

5 Fluctuations from inflation

Topics to be covered:

- Description of inhomogeneity
- Mechanisms for fluctuation generation
- Tilt and tensor modes
- Eternal inflation

5.1 The perturbed universe

We now need to consider the greatest achievement of inflation, which was not anticipated when the theory was first put forward: it provides a concrete mechanism for generating the seeds of structure in the universe. In essence, the idea is that the inevitable small quantum fluctuations in the inflaton field ϕ are transformed into residual classical fluctuations in density when inflation is over. The details of this process can be technical, and could easily fill a lecture course. The following treatment is therefore simplified as far as possible, while still making contact with the full results.

QUANTIFYING INHOMOGENEITY The first issue we have to deal with is how to quantify departures from uniform density. Frequently, an intuitive Newtonian approach can be used, and we will adopt this wherever possible. But we should begin with a quick overview of the relativistic approach to this problem, to emphasise some of the big issues that are ignored in the Newtonian method.

Because relativistic physics equations are written in a covariant form in which all quantities are independent of coordinates, relativity does not distinguish between *active* changes of coordinate (e.g. a Lorentz boost) or *passive* changes (a mathematical change of variable, normally termed a gauge transformation). This generality is a problem, as we can see by asking how some scalar quantity S (which might be density, temperature etc.) changes under a gauge transformation $x^\mu \to x'^\mu = x^\mu + \epsilon^\mu$. A gauge transformation induces the usual Lorentz transformation coefficients dx'^μ/dx^ν (which have no effect on a scalar), but also involves a translation that relabels spacetime points. We therefore have $S'(x^\mu + \epsilon^\mu) = S(x^\mu)$, or

$$S'(x^\mu) = S(x^\mu) - \epsilon^\alpha \partial S/\partial x^\alpha. \tag{126}$$

Consider applying this to the case of a uniform universe; here ρ only depends on time, so that

$$\rho' = \rho - \epsilon^0 \dot{\rho}. \tag{127}$$

An effective density perturbation is thus produced by a local alteration in the time coordinate: when we look at a universe with a fluctuating density, should we really think of a uniform model in which time is wrinkled? This ambiguity may seem absurd, and in the laboratory it could be resolved empirically by constructing the coordinate system directly – in principle by using light signals. This shows that the cosmological horizon plays an important role in this topic: perturbations with wavelength $\lambda \lesssim ct$ inhabit a regime in which gauge ambiguities can be resolved directly via common sense. The real difficulties lie in the super-horizon modes with $\lambda \gtrsim ct$. Within inflationary models, however, these difficulties can be overcome, since the true horizon is $\gg ct$.

The most direct general way of solving these difficulties is to construct perturbation variables that are explicitly independent of gauge. A comprehensive technical discussion of this method is given in chapter 7 of Mukhanov's book, and we summarize the essential elements here, largely without proof.

Firstly, metric perturbations can be split into three classes: **scalar perturbations**, which are described by scalar functions of spacetime coordinate, and which correspond to growing density perturbations; **vector perturbations**, which correspond to vorticity perturbations, and **tensor perturbations**, which correspond to gravitational waves. Here, we shall concentrate mainly on scalar perturbations.

A key result is that scalar perturbations can be described by just two gauge-invariant 'potentials' (functions of spacetime coordinates). Since these are gauge-invariant, we may as well write the perturbed metric in a particular gauge that makes things look as simple as possible. This is the **longitudinal gauge** in which the time and space parts of the RW metric are perturbed separately:

$$d\tau^2 = (1 + 2\Psi)dt^2 - (1 - 2\Phi)\gamma_{ij}\, dx^i\, dx^j. \tag{128}$$

Health warning: there are different conventions, and the symbols for the potentials are sometimes swapped, or signs flipped.

A second key result is that inserting the longitudinal metric into the Einstein equations shows that Ψ and Φ are identical in the case of fluid-like perturbations where off-diagonal elements of the energy–momentum tensor vanish. In this case, the longitudinal gauge becomes identical to the

Newtonian gauge, in which perturbations are described by a single scalar field, which is the gravitational potential:

$$d\tau^2 = (1+2\Phi)dt^2 - (1-2\Phi)\gamma_{ij}\,dx^i\,dx^j, \tag{129}$$

and this should be quite familiar. If we consider small scales, so that the spatial metric γ_{ij} becomes that of flat space, then this form matches, for example, the Schwarzschild metric with $\Phi = -GM/r$, in the limit $\Phi/c^2 \ll 1$.

The conclusion is thus that the gravitational potential can for many purposes give an effectively gauge-invariant measure of cosmological perturbations. The advantage of this fact is that the gravitational potential is a familiar object, which we can manipulate and use our Newtonian intuition. This is still not guaranteed to give correct results on scales greater than the horizon, however, so a fully relativistic approach is to be preferred. But with the length restrictions of this course, it is hard to go beyond the Newtonian approach. The main results of the full theory can at least be understood and made plausible in this way.

Informally, the potential Φ is a measure of space-time curvature which solves the gauge issue and has meaning on super-horizon scales. A key property, which is perhaps intuitively reasonable, is that Φ is constant in time for perturbations with wavelengths much larger than the horizon. Conversely, interesting effects can happen inside the horizon, which imprints characteristic scale-dependent features on the cosmological inhomogeneities. A full justification of the constancy of Φ using a complete relativistic treatment would take too much space in the present course, and we will generally discuss perturbations using a Newtonian approach. This does yield the correct conclusion regarding the constancy of Φ, but we should be clear that this is at best a consistency check, since we will use a treatment of gravity that is restricted to static fields.

FLUCTUATION POWER SPECTRA From the Newtonian point of view, potential fluctuations are directly related to those in density via Poisson's equation:

$$\nabla^2\Phi/a^2 = 4\pi G(1+3w)\,\bar{\rho}\,\delta, \tag{130}$$

where we have defined a dimensionless fluctuation amplitude

$$\delta \equiv \frac{\rho - \bar{\rho}}{\bar{\rho}}. \tag{131}$$

the factor of a^2 is there so we can use comoving length units in ∇^2 and the factor $(1+3w)$ accounts for the relativistic active mass density $\rho + 3p$.

We are very often interested in asking how these fluctuations depend on scale, which amounts to making a Fourier expansion:

$$\delta(\mathbf{x}) = \sum \delta_k e^{-i\mathbf{k}\cdot\mathbf{x}}, \tag{132}$$

where $\mathbf{k}$ is the comoving wavevector. What are the allowed modes? If the field were periodic within some box of side L, we would have the usual harmonic boundary conditions

$$k_x = n\,\frac{2\pi}{L}, \qquad n = 1, 2\cdots, \tag{133}$$

and the inverse Fourier relation would be

$$\delta_k(\mathbf{k}) = \left(\frac{1}{L}\right)^3 \int \delta(\mathbf{x}) \exp\bigl(i\mathbf{k}\cdot\mathbf{x}\bigr)\, d^3x. \tag{134}$$

Working in Fourier space in this way is powerful because it immediately gives a way of solving Poisson's equation and relating fluctuations in density and potential. For a single mode, $\nabla^2 \to -k^2$, and so

$$\Phi_k = -4\pi G(1+3w)a^2\,\bar\rho\,\delta_k/k^2. \tag{135}$$

The fluctuating density field can be described by its statistical properties. The mean is zero by construction; the variance is obtained by taking the volume average of δ^2:

$$\langle\delta^2\rangle = \sum |\delta_k|^2. \tag{136}$$

To see this result, write the lhs instead as $\langle\delta\delta^*\rangle$ (makes no difference for a real field), and appreciate that all cross terms integrate to zero via the boundary conditions. For obvious reasons, the quantity

$$P(k) \equiv |\delta_k|^2 \tag{137}$$

is called the **power spectrum**. Note that, in an isotropic universe, we assume that P will be independent of direction of the wavevector in the limit of a large box: the fluctuating density field is statistically **isotropic**. In applying this apparatus, we would not want the (arbitrary) box size to appear. This happens naturally: as the box becomes big, the modes are finely spaced and a sum over modes is replaced by an integral over k space times the usual density of states, $(L/2\pi)^3$:

$$\langle \delta^2 \rangle = \sum |\delta_k|^2 \to \frac{L^3}{(2\pi)^3} \int P(k)\, d^3k = \int \Delta^2(k)\, d\ln k. \tag{138}$$

In the last step, we have defined the combination

$$\Delta^2(k) \equiv \frac{L^3}{(2\pi)^3}\, 4\pi k^3\, P(k), \tag{139}$$

which absorbs the box size into the definition of a dimensionless power spectrum, which gives the contribution to the variance from each logarithmic range of wavenumber (or wavelength). Despite the attraction of a dimensionless quantity, one still frequently sees plots of $P(k)$ – and often in a dimensionally fudged form in which $L = 1$ is assumed, and P given units of volume.

6 Structure formation – I

6.1 Newtonian equations of motion

We have decided that perturbations will in most cases effectively be described by the Newtonian potential, Φ. Now we need to develop an equation of motion for Φ, or equivalently for the density fluctuation $\rho \equiv (1+\delta)\bar{\rho}$. In the Newtonian approach, we treat dynamics of cosmological matter exactly as we would in the laboratory, by finding the equations of motion induced by either pressure or gravity. We begin by casting the problem in comoving units:

$$\begin{aligned} \mathbf{x}(t) &= a(t)\mathbf{r}(t) \\ \delta\mathbf{v}(t) &= a(t)\mathbf{u}(t), \end{aligned} \tag{140}$$

so that $\mathbf{x}$ has units of proper length, i.e. it is an **Eulerian coordinate**. First note that the comoving peculiar velocity $\mathbf{u}$ is just the time derivative of the comoving coordinate $\mathbf{r}$:

$$\dot{\mathbf{x}} = \dot{a}\mathbf{r} + a\dot{\mathbf{r}} = H\mathbf{x} + a\dot{\mathbf{r}}, \tag{141}$$

where the rhs must be equal to the Hubble flow $H\mathbf{x}$, plus the peculiar velocity $\delta\mathbf{v} = a\mathbf{u}$.

The equation of motion follows from writing the Eulerian equation of motion as $\ddot{\mathbf{x}} = \mathbf{g}_0 + \mathbf{g}$, where $\mathbf{g} = -\boldsymbol{\nabla}\Phi/a$ is the peculiar acceleration, and $\mathbf{g}_0$ is the acceleration that acts on a particle in a homogeneous universe (neglecting pressure forces to start with, for simplicity). Differentiating $\mathbf{x} = a\mathbf{r}$ twice gives

$$\ddot{\mathbf{x}} = a\dot{\mathbf{u}} + 2\dot{a}\mathbf{u} + \frac{\ddot{a}}{a}\,\mathbf{x} = \mathbf{g}_0 + \mathbf{g}. \tag{142}$$

The unperturbed equation corresponds to zero peculiar velocity and zero peculiar acceleration: $(\ddot{a}/a)\,\mathbf{x} = \mathbf{g}_0$; subtracting this gives the perturbed equation of motion

$$\dot{\mathbf{u}} + 2(\dot{a}/a)\mathbf{u} = \mathbf{g}/a = -\boldsymbol{\nabla}\Phi/a^2. \tag{143}$$

This equation of motion for the peculiar velocity shows that $\mathbf{u}$ is affected by gravitational acceleration and by the **Hubble drag** term, $2(\dot{a}/a)\mathbf{u}$. This arises because the peculiar velocity falls with time as a particle attempts to catch up with successively more distant (and therefore more rapidly receding) neighbours. In the absence of gravity, we get $\delta v \propto 1/a$: momentum redshifts away, just as with photon energy.

The peculiar velocity is directly related to the evolution of the density field, through conservation of mass. This is described by the usual continuity equation $\dot{\rho} = -\boldsymbol{\nabla}\cdot(\rho\mathbf{v}$, where $\rho = \bar{\rho}(1+\delta)$ and proper length units are assumed. If we use comoving length units, the mean density is constant and this is easily transformed to

$$\dot{\delta} = -\boldsymbol{\nabla}\cdot[(\mathbf{1}+\delta)\mathbf{u}]. \tag{144}$$

This simplifies further if we restrict ourselves to a the linear approximation where $\delta \ll 1$, and neglect terms that are second order in the perturbation, yielding the linearized continuity equation

$$\dot{\delta} = -\boldsymbol{\nabla}\cdot\mathbf{u}. \tag{145}$$

The solutions of these equations can be decomposed into modes either parallel to $\mathbf{g}$ or independent of $\mathbf{g}$ (these are the homogeneous and inhomogeneous solutions to the equation of motion). The homogeneous case corresponds to no peculiar gravity – i.e. zero density perturbation. This is consistent with the linearized continuity equation, $\boldsymbol{\nabla}\cdot\mathbf{u} = -\dot{\delta}$, which says that it is possible

to have **vorticity modes** with $\boldsymbol{\nabla} \cdot \mathbf{u} = 0$ for which $\dot{\delta}$ vanishes, so there is no growth of structure in this case. The proper velocities of these vorticity modes decay as $v = au \propto a^{-1}$, as with the kinematic analysis for a single particle.

GROWING MODE For the growing mode, it is most convenient to eliminate $\mathbf{u}$ by taking the divergence of the equation of motion for $\mathbf{u}$, and the time derivative of the continuity equation. This requires a knowledge of $\boldsymbol{\nabla} \cdot \mathbf{g}$, which comes via Poisson's equation: $\boldsymbol{\nabla} \cdot \mathbf{g} = 4\pi G a \rho_0 \delta$. The resulting 2nd-order equation for δ is

$$\ddot{\delta} + 2\frac{\dot{a}}{a}\dot{\delta} = 4\pi G \rho_0 \, \delta. \tag{146}$$

This is easily solved for the $\Omega_m = 1$ case, where $4\pi G\rho_0 = 3H^2/2 = 2/3t^2$, and a power-law solution works:

$$\delta(t) \propto t^{2/3} \quad \text{or} \quad t^{-1}. \tag{147}$$

The first solution, with $\delta(t) \propto a(t)$ is the growing mode, corresponding to the gravitational instability of density perturbations. Given some small initial seed fluctuations, this is the simplest way of creating a universe with any desired degree of inhomogeneity.

RADIATION-DOMINATED UNIVERSE The analysis so far does not apply when the universe was radiation dominated ($c_s = c/\sqrt{3}$). For this period of the early Universe it is therefore common to resort to general relativity perturbation theory or use special relativity fluid mechanics and Newtonian gravity with a relativistic source term (see e.g. Section 15.2 of Peacock 1999). In the interests of brevity and completeness we simply quote the result of this analysis, which is

$$\delta(t) \propto t \quad \text{or} \quad t^{-1}; \tag{148}$$

thus the growing mode during radiation domination ($a \propto t^{1/2}$) has $\delta(t) \propto a(t)^2$.

The results for matter domination and radiation domination can be combined to say that gravitational potential perturbations are independent of time (at least while $\Omega = 1$). Poisson's equation tells us that $-k^2\Phi/a^2 \propto \rho\,\delta$; since $\rho \propto a^{-3}$ for matter domination or a^{-4} for radiation, that gives $\Phi \propto \delta/a$ or δ/a^2 respectively, so that

$$\Phi = \text{constant} \tag{149}$$

in either case. In other words, the metric fluctuations resulting from potential perturbations are frozen, at least for perturbations with wavelengths greater than the horizon size. This simple result at least has an intuitive plausibility for the radiation-dominated case, despite the lack of a derivation.

Lastly, it will also be useful to express the growing mode in one other way. We have growing modes proportional to $t^{2/3}$ or t in the matter and radiation eras, during which the scale factor changes as $t^{2/3}$ or $t^{1/2}$ respectively. Now consider the particle horizon: its proper size is $\sim ct$, so the comoving value scales as $t^{1/3}$ or $t^{1/2}$ respectively (divide by $a[t]$). Thus, in either era,

$$\delta(t) \propto D_{\rm H}^2(t). \tag{150}$$

MODELS WITH NON-CRITICAL DENSITY We have solved the growth equation for the matter-dominated $\Omega = 1$ case. It is possible to cope with other special cases (e.g. matter + curvature) with some effort. In the general case (especially with a general vacuum having $w \neq -1$), it is necessary to integrate the differential equation numerically. At high z, we always have the matter-dominated $\delta \propto a$, and this serves as an initial condition. In general, we can write

$$\delta(a) \propto a\, f[\Omega_m(a)], \tag{151}$$

where the factor f expresses a deviation from the simple growth law. For flat models with $\Omega_m + \Omega_v = 1$, a useful approximation is $f \simeq \Omega_m^{0.23}$, which is less marked than $f \simeq \Omega_m^{0.65}$ in the $\Lambda = 0$ case. This reflects the more rapid variation of Ω_v with redshift; if the cosmological constant is important dynamically, this only became so very recently, and the universe spent more of its history in a nearly Einstein–de Sitter state by comparison with an open universe of the same Ω_m. Interestingly, this difference is erased if we look at the growth *rate* , in which case we have the almost universal formula

$$f_g \equiv \frac{d\ln\delta}{d\ln a} = \Omega_m(a)^\gamma, \tag{152}$$

where γ is close to 0.6, independently of whether there is significant vacuum energy.

6.2 Pressure and the shape of the matter power spectrum

So far, we have mainly considered the collisionless component. For the photon-baryon gas, all that changes is that the peculiar acceleration gains a term from the pressure gradients:

$$\mathbf{g} = -\boldsymbol{\nabla}\Phi/a - \boldsymbol{\nabla}p/(a\rho). \tag{153}$$

The pressure fluctuations are related to the density perturbations via the sound speed

$$c_s^2 \equiv \frac{\partial p}{\partial \rho}. \tag{154}$$

Now think of a plane-wave disturbance $\delta \propto e^{-i\mathbf{k}\cdot\mathbf{r}}$, where $\mathbf{k}$ and $\mathbf{r}$ are in comoving units. All time dependence is carried by the amplitude of the wave. The linearized equation of motion for δ then gains an extra term on the rhs from the pressure gradient:

$$\ddot{\delta} + 2\frac{\dot{a}}{a}\dot{\delta} = \delta\bigl(4\pi G\rho_0 - c_s^2 k^2/a^2\bigr). \tag{155}$$

This shows that there is a critical proper wavelength, the **Jeans length**, at which we switch from the possibility of gravity-driven growth for long-wavelength modes to standing sound waves at short wavelengths. This critical length is

$$\lambda_{\rm J}^{\rm proper} = \frac{2\pi}{k_{\rm J}^{\rm proper}} = c_s\sqrt{\frac{\pi}{G\rho}}. \tag{156}$$

Prior to matter-radiation equality, the speed of sound for a radiation-dominated fluid with $p = u/3$ is $c_s = c/\sqrt{3}$, so this Jeans length is close to the horizon size.

A perturbation of given comoving wavelength will start out larger than the horizon, but the horizon grows with time, and so the perturbation 'enters the horizon' (not very good terminology, but standard). After that time, pressure forces dominate over gravity, and the perturbation oscillates as a standing sound wave, which turns out to have a constant amplitude in δ during the radiation era; but zero-pressure growth would result in δ increasing with time, so the amplitude of small-scale fluctuations falls relative to the growing mode. Since the growing mode corresponds to constant

potential, this means that the potential decays as $1/\delta_{\rm grow} \propto 1/a^2$. All this can be solved exactly for the radiation-dominated era, and the effective damping of the initial potential fluctuation is

$$\frac{\Phi}{\Phi_i} = 3(\sin x - x\cos x)/x^3; \quad x \equiv kc_s\eta, \tag{157}$$

where η is conformal time, $d\eta \equiv dt/a(t)$, which is thus equal to the comoving particle horizon size. We will see later that the imprint of these **acoustic oscillations** is visible in the microwave background.

At these early times, the dark matter is a minority constituent of the universe, but it suffers an interesting and critical effect from the above behaviour of the coupled baryon-photon fluid (which is glued together by Thomson scattering). The small-scale damping of the waves results in the radiation becoming smooth, which breaks the usual adiabatic relation in which the matter density and the photon number density have equal perturbations. Now recall the growth equation for matter perturbations, neglecting pressure:

$$\ddot{\delta} + 2\frac{\dot{a}}{a}\dot{\delta} = 4\pi G\rho_0\,\delta. \tag{158}$$

The rhs contains the combination $\rho_0\delta$; this is just the fluctuation in density, which drives the gravitational growth. But the growing mode with Φ constant only arises if ρ_0 is the total density, which also sets the timescale for expansion. If the majority constituent of the universe (the radiation, at early times) is uniform, then the rhs becomes $\propto \rho_m\delta$, which is $\ll \rho_0\delta$. Thus the growth switches off.

Figure 12 shows a schematic of the resulting growth history for matter density fluctuations. For scales greater than the horizon, perturbations in matter and radiation can grow together, so fluctuations at early times grow at the same rate, independent of wavenumber. But this growth ceases once the perturbations 'enter the horizon' – i.e. when the horizon grows sufficiently to exceed the perturbation wavelength. At this point, growth ceases. For fluids (baryons) it is the radiation pressure that prevents the perturbations from collapsing further. For collisionless matter the rapid radiation driven expansion prevents the perturbation from growing again until matter radiation equality.

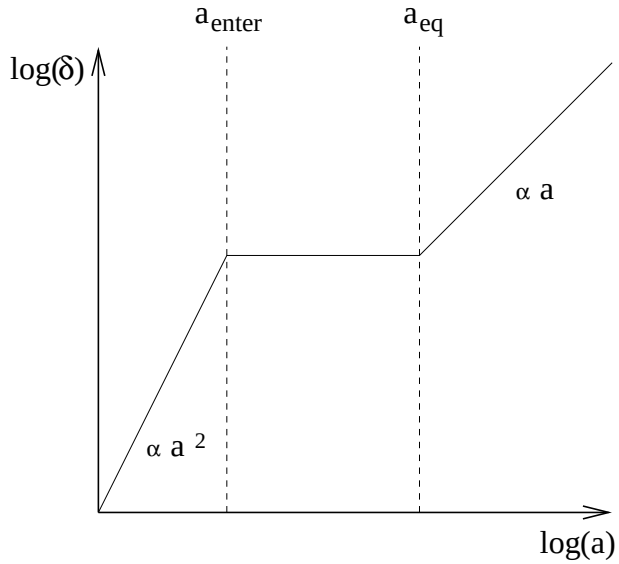

Figure 12. A schematic of the suppression of fluctuation growth during the radiation dominated phase when the density perturbation enters the horizon at $a_{\rm enter} < a_{\rm eq}$.

This effect (called the Mészáros effect) is critical in shaping the late-time power spectrum (as we will show) as the universe preserves a 'snapshot' of the amplitude of the mode at horizon crossing. Before this process operates, inflation predicts an approximately scale invariant initial Zeldovich spectrum where $P_i(k) \propto k$. How does the Mészáros effect modify the shape of this initial power spectrum?

Figure 13 shows that the smallest physical scales (largest k scales) will be affected first and experience the strongest suppression to their amplitude. The largest physical scale fluctuations (smallest k scales) will be unaffected as they will enter the horizon after matter-radiation equality. We can therefore see that there will be a turnover in the power spectrum at a characteristic scale given by the horizon size at matter-radiation equality.

From Figure 12 we can see that when a fluctuation enters the horizon before matter-radiation equality its growth is suppressed by $f = (a_{\rm enter}/a_{\rm eq})^2$. A fluctuation k enters the horizon when

$D_{\rm H} \simeq 1/k$. As $D_{\rm H} = c/aH(a)$ and $H(a) \propto a^{-2}$ during radiation domination we see that the fluctuations are suppressed by a factor $f \propto k^{-2}$ and that the power spectrum on large k scales follows a k^{-3} power law.

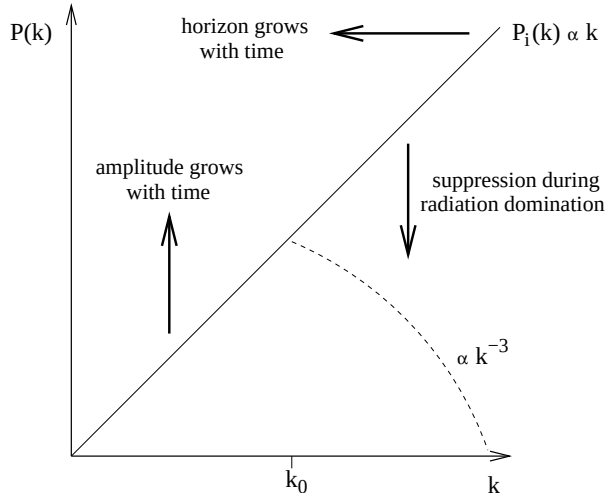

Figure 13. Schematic of the how the Mészáros effect modifies the initial power spectrum. Note log scale.

6.3 Transfer functions and characteristic scales

The above discussion can be summed up in the from of the linear **transfer function** for density perturbations, where we factor out the long-wavelength growth law from a term that expresses how growth is modulated as a function of wavenumber:

$$\delta(a) \propto g(a) T_k. \tag{159}$$

While curvature is negligible, we have seen that $g(a)$ is proportional to the square of conformal time for adiabatic perturbations. In principle, there is a transfer function for each constituent of the

universe, and these evolve with time. As we have discussed, however, the different matter ingredients tend to come together at late times, and the overall transfer function tends to something that is the same for all matter components and which does not change with time for low redshifts. This late-time transfer function is therefore an important tool for cosmologists who want to predict observed properties of density fields in the current universe.

We have discussed the main effects that contribute to the form of the transfer function, but a full calculation is a technical challenge. In detail, we have a mixture of matter (both collisionless dark particles and baryonic plasma) and relativistic particles (collisionless neutrinos and collisional photons), which does not behave as a simple fluid. Particular problems are caused by the change in the photon component from being a fluid tightly coupled to the baryons by Thomson scattering, to being collisionless after recombination. Accurate results require a solution of the Boltzmann equation to follow the evolution of the full phase-space distribution. This was first computed accurately by Bond & Szalay (1983), and is today routinely available via public-domain codes such as CMBFAST.

Some illustrative results are shown in figure 14. Leaving aside the isocurvature models, all adiabatic cases have $T \to 1$ on large scales – i.e. there is growth at the universal rate (which is such that the amplitude of potential perturbations is constant until the vacuum starts to be important at $z \lesssim 1$). The different shapes of the functions can be understood intuitively in terms of a few special length scales, as follows:

(1) Horizon length at matter-radiation equality. The main bend visible in all transfer functions is due to the Mészáros effect (discussed above), which arises because the universe is radiation dominated at early times.

$$T_k \simeq \begin{cases} 1 & kD_{\rm H}(z_{\rm eq}) \ll 1 \\ [kD_{\rm H}(z_{\rm eq})]^{-2} & kD_{\rm H}(z_{\rm eq}) \gg 1. \end{cases} \tag{160}$$

This process continues until the universe becomes matter dominated. We therefore expect a characteristic 'break' in the fluctuation spectrum around the comoving horizon length at this time, which we have seen is $D_{\rm H}(z_{\rm eq}) = 16\,(\Omega_m h^2)^{-1}{\rm Mpc}$. Since distances in cosmology always scale as h^{-1}, this means that $\Omega_m h$ should be observable.

(2) Free-streaming length. This relatively gentle filtering away of the initial fluctuations is all that applies to a universe dominated by Cold Dark Matter, in which random velocities are negligible. A CDM universe thus contains fluctuations in the dark matter on all scales, and structure formation proceeds via hierarchical process in which nonlinear structures grow via mergers. Examples of CDM would be thermal relic WIMPs with masses of order 100 GeV, but a

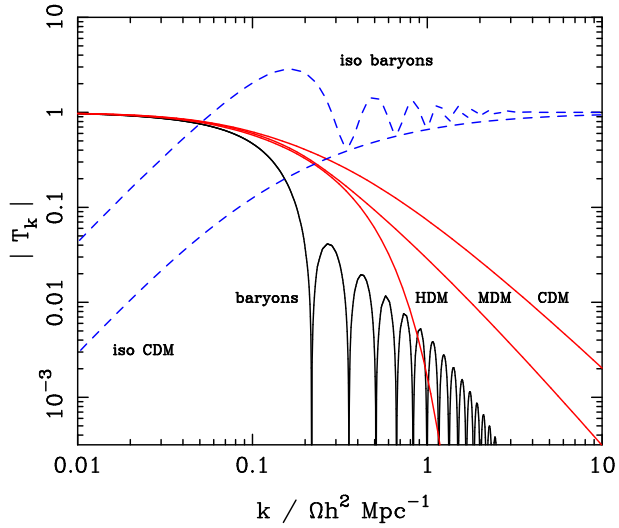

Figure 14. A plot of transfer functions for various adiabatic models, in which $T_k \to 1$ at small k. A number of possible matter contents are illustrated: pure baryons; pure CDM; pure HDM. For dark-matter models, the characteristic wavenumber scales proportional to $\Omega_m h^2$, marking the break scale corresponding to the horizon length at matter-radiation equality. The scaling for baryonic models does not obey this exactly; the plotted case corresponds to $\Omega_m = 1$, $h = 0.5$.

more interesting case arises when thermal relics have lower masses. For collisionless dark matter, perturbations can be erased simply by free streaming: random particle velocities cause blobs to disperse. At early times ($kT > mc^2$), the particles will travel at c, and so any perturbation that has entered the horizon will be damped. This process switches off when the particles become non-relativistic, so that perturbations are erased up to proper lengthscales of $\simeq ct(kT = mc^2)$. This translates to a comoving horizon scale ($2ct/a$ during the radiation era) at $kT = mc^2$ of

$$L_{\rm free-stream} = 112\,(m/{\rm eV})^{-1}\,{\rm Mpc} \qquad (161)$$

(in detail, the appropriate figure for neutrinos will be smaller by $(4/11)^{1/3}$ since they have a smaller temperature than the photons). A light neutrino-like relic that decouples while it is relativistic satisfies

$$\Omega_\nu h^2 = m/94.1\,\mathrm{eV} \tag{162}$$

Thus, the damping scale for HDM (Hot Dark Matter) is of order the bend scale. The existence of galaxies at $z \simeq 6$ tells us that the coherence scale must have been below about 100 kpc, so the DM mass must exceed about 1 keV.

A more interesting (and probably practically relevant) case is when the dark matter is a mixture of hot and cold components. The free-streaming length for the hot component can therefore be very large, but within range of observations. The dispersal of HDM fluctuations reduces the CDM growth rate on all scales below $L_{\mathrm{free-stream}}$ – or, relative to small scales, there is an enhancement in large-scale power.

(3) Acoustic horizon length. The horizon at matter-radiation equality also enters in the properties of the baryon component. Since the sound speed is of order c, the largest scales that can undergo a single acoustic oscillation are of order the horizon at this time. The transfer function for a pure baryon universe shows large modulations, reflecting the number of oscillations that have been completed before the universe becomes matter dominated and the pressure support drops. The lack of such large modulations in real data is one of the most generic reasons for believing in collisionless dark matter. Acoustic oscillations persist even when baryons are subdominant, however, and can be detectable as lower-level modulations in the transfer function. At matter-radiation equality, the dark matter has a smoothly declining transfer function – but the baryons have an oscillating transfer function, so the spatial distribution of these two components is different. Once the sound speed drops, gravity will pull these components together, and their transfer functions will tend to become identical. But since baryons are about 5% of the total matter content, the resulting final 'compromise' transfer function has acoustic oscillations imprinted into it at the few per cent level. We will say more about this later.

SPECTRUM NORMALIZATION We now have a full recipe for specifying the matter power spectrum: Historically, this is done in a slightly awkward way. First suppose we wanted to consider smoothing the density field by convolution with some **window**. One simple case is to imagine averaging within a sphere of radius R. For the effect on the power spectrum, we need the Fourier transform of this filter:

$$\sigma^2(R) = \int \Delta^2(k)\,|W_k|^2\,d\ln k; \quad W_k = \frac{3}{(kR)^3}(\sin kR - kR\cos kR). \tag{163}$$

Unlike the power spectrum, $\sigma(R)$ is monotonic, and the value at any scale is sufficient to fix the normalization. The traditional choice is to specify σ_8, corresponding to $R = 8\,h^{-1}\,\mathrm{Mpc}$. As a final complication, this measure is normally taken to apply to the rms in the filtered *linear-theory* density field. The best current estimate is $\sigma_8 \simeq 0.8$, so clearly nonlinear corrections matter in interpreting this number. The virtue of this convention is that it is then easy to calculate the spectrum normalization at any early time.

6.4 Relic fluctuations from inflation

OVERVIEW It was realized very quickly after the invention of inflation that the theory might also solve the other big puzzle with the initial condition of the universe. When we study gravitational instability, we will see that the present-day structure requires that the universe at even the Planck era would have had to possess a finite degree of inhomogeneity. Inflation suggests an audacious explanation for this structure, which is that it is an amplified form of the quantum fluctuations that are inevitable when the universe is sufficiently small. The present standard theory of this process was worked out by a number of researchers and generally agreed at a historic 1982 Nuffield conference in Cambridge.

The essence of the idea can be seen in figure 15. This reminds us that de Sitter space contains an **event horizon**, in that the comoving distance that particles can travel between a time t_0 and $t = \infty$ is finite,

$$r_{\rm EH} = \int_{t_0}^{\infty} \frac{c\,dt}{R(t)}; \tag{164}$$

this is not to be confused with the particle horizon, where the integral would be between 0 and t_0. With $R \propto \exp(Ht)$, the proper radius of the horizon is given by $R_0 r_{\rm EH} = c/H$. The exponential expansion literally makes distant regions of space move faster than light, so that points separated by $> c/H$ can never communicate with each other. If we imagine expanding the inflaton, ϕ, using comoving Fourier modes, then there are two interesting limits for the mode wavelength:

(1) 'Inside the horizon': $a/k \ll c/H$. Here the de Sitter expansion is negligible, just as we neglect the modern vacuum energy in the Solar system. The fluctuations in ϕ can be calculated exactly as in flat-space quantum field theory.

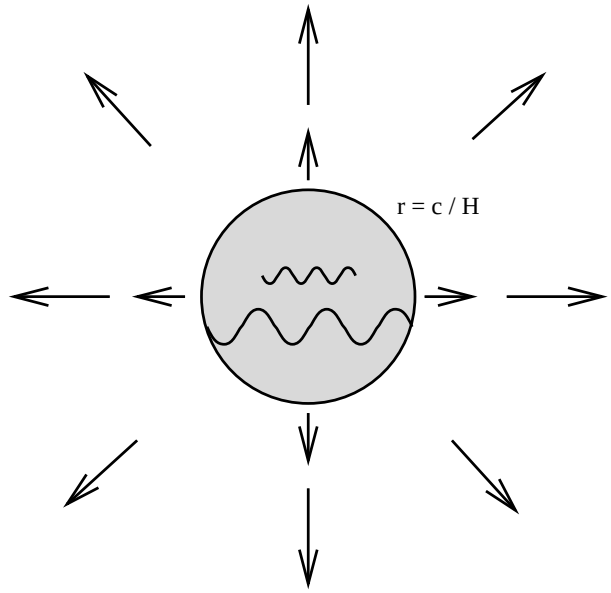

Figure 15. The event horizon in de Sitter space. Particles outside the sphere at $r = c/H$ can never receive light signals from the origin, nor can an observer at the origin receive information from outside the sphere. The exponential expansion quickly accelerates any freely falling observers to the point where their recession from the origin is effectively superluminal. The wave trains represent the generation of fluctuations in this spacetime. Waves with $\lambda \ll c/H$ effectively occupy flat space, and so undergo the normal quantum fluctuations for a vacuum state. As these modes (of fixed comoving wavelength) are expanded to sizes $\gg c/H$, causality forces the quantum fluctuation to become frozen as a classical amplitude that can seed large-scale structure.

(2) 'Outside the horizon': $a/k \gg c/H$. Now the mode has a wavelength that exceeds the scale over which causal influences can operate. Therefore, it must now act as a 'frozen' quantity, which has the character of a classical disturbance. This field fluctuation can act as the seed for subsequent density fluctuations.

Before going any further, we can immediately note that a natural prediction will be a spectrum of perturbations that are nearly *scale invariant*. This means that the metric fluctuations of spacetime

receive equal levels of distortion from each decade of perturbation wavelength, and may be quantified in terms of the dimensionless power spectrum, Δ^2_Φ, of the Newtonian gravitational potential, Φ ($c = 1$):

$$\Delta^2_\Phi \equiv \frac{d\,\sigma^2(\Phi)}{d\,\ln k} = \text{constant} \equiv \delta^2_{\rm H}. \tag{165}$$

The origin of the term 'scale-invariant' is clear: since potential fluctuations modify spacetime, this is equivalent to saying that spacetime must be a fractal: it has the same level of deviation from the exact RW form on each level of resolution. It is common to denote the level of metric fluctuations by $\delta_{\rm H}$ – the **horizon-scale amplitude** (which we know to be about 10^{-5}). The justification for this name is that the potential perturbation is of the same order as the density fluctuation on the scale of the horizon at any given time. We can see this from Poisson's equation in Fourier space:

$$\Phi_k = -\frac{a^2}{k^2}\,4\pi G\,\bar\rho\,\delta_k = -(3/2)\,\frac{a^2}{k^2}\,\Omega_m H^2\,\delta_k \tag{166}$$

(where we have taken a $w = 0$ pressureless equation of state). This says that $\Phi_k/c^2 \sim \delta_k$ when the reciprocal of the physical wavenumber is c/H, i.e. is of order the horizon size.

The intuitive argument for scale invariance is that de Sitter space is invariant under time translation: there is no natural origin of time under exponential expansion. At a given time, the only length scale in the model is the horizon size c/H, so it is inevitable that the fluctuations that exist on this scale are the same at all times. By our causality argument, these metric fluctuations must be copied unchanged to larger scales as the universe exponentiates, so that the appearance of the universe is independent of the scale at which it is viewed.

If we accept this rough argument, then the implied density power spectrum is interesting. because of the relation between potential and density, it must be

$$\Delta^2(k) \propto k^4, \tag{167}$$

So the density field is very strongly inhomogeneous on small scales. Another way of putting this is in terms of a standard power-law notation for the non-dimensionless spectrum:

$$P(k) \propto k^n; \quad n = 1. \tag{168}$$

To get a feeling for what this means, consider the case of a matter distribution built up by the random placement of particles. It is not hard to show that this corresponds to **white noise**: a power spectrum that is independent of scale – i.e. $n = 0$. Recall the inverse Fourier relation:

$$\delta_k(\mathbf{k}) = \left(\frac{1}{L}\right)^3 \int \delta(\mathbf{x}) \exp\bigl(i\mathbf{k}\cdot\mathbf{x}\bigr)\, d^3x. \tag{169}$$

Here, the density field is a sum of spikes at the locations of particles. Because the placement is random, the contribution of each spike is a complex number of phase uniformly distributed between 0 and 2π, independent of k. Conversely, the $n = 1$ 'scale-invariant' spectrum thus represents a density field that is super-uniform on large scales, but with enhanced small-scale fluctuations.

This $n = 1$ spectrum was considered a generic possibility long before inflation, and is also known as the **Zeldovich spectrum**. It is possible to alter this prediction of scale invariance only if the expansion is non-exponential; but we have seen that such deviations must exist towards the end of inflation. As we will see, it is natural for n to deviate from unity by a few %, and this is one of the predictions of inflation.

A MORE DETAILED TREATMENT We now need to give an outline of the exact treatment of inflationary fluctuations, which will allow us to calculate both the scale dependence of the spectrum and the absolute level of fluctuations. This can be a pretty technical subject, but it is possible to take a simple approach and still give a flavour of the main results and how they arise.

To anticipate the final answer, the inflationary prediction is of a horizon-scale amplitude

$$\delta_{\rm H} = \frac{H^2}{2\pi\,\dot\phi} \tag{170}$$

which can be understood as follows. Imagine that the main effect of fluctuations is to make different parts of the universe have fields that are perturbed by an amount $\delta\phi$. In other words, we are dealing with various copies of the same rolling behaviour $\phi(t)$, but viewed at different times

$$\delta t = \frac{\delta\phi}{\dot\phi}. \tag{171}$$

These universes will then finish inflation at different times, leading to a spread in energy densities (figure 16). The horizon-scale density amplitude is given by the different amounts that the universes have expanded following the end of inflation:

$$\delta_{\rm H} \simeq H\ \delta t = \frac{H}{\dot\phi}\,\delta\phi = \frac{H}{\dot\phi} \times \frac{H}{2\pi} = \frac{H^2}{2\pi\,\dot\phi}. \tag{172}$$

The $\delta_{\rm H} \simeq H\ \delta t$ argument relies on $R(t) \propto \exp(Ht)$ and that $\delta_{\rm H}$ is of order the fractional change in R. We will not attempt here to do better than justify the order of magnitude.

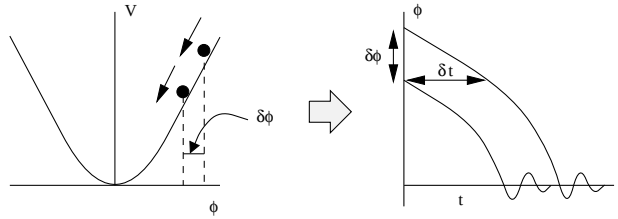

Figure 16. This plot shows how fluctuations in the scalar field transform themselves into density fluctuations at the end of inflation. Different points of the universe inflate from points on the potential perturbed by a fluctuation $\delta\phi$, like two balls rolling from different starting points. Inflation finishes at times separated by δt in time for these two points, inducing a density fluctuation $\delta = H\delta t$.

The last step uses the crucial input of quantum field theory, which says that the rms $\delta\phi$ is given by $H/2\pi$, and we now sketch the derivation of this result, What we need to do is consider the equation of motion obeyed by perturbations in the inflaton field. The basic equation of motion is

$$\ddot\phi + 3H\dot\phi - \nabla^2\phi + V'(\phi) = 0, \tag{173}$$

and we seek the corresponding equation for the perturbation $\delta\phi$ obtained by starting inflation with slightly different values of ϕ in different places. Suppose this perturbation takes the form of a comoving plane-wave perturbation of comoving wavenumber k and amplitude A: $\delta\phi = A\exp(i\mathbf{k}\cdot\mathbf{x} - ikt/a)$. If the slow-roll conditions are also assumed, so that V' may be treated as

a constant, then the perturbed field $\delta\phi$ obeys the first-order perturbation of the equation of motion for the main field:

$$[\ddot{\delta\phi}] + 3H[\dot{\delta\phi}] + (k/a)^2[\delta\phi] = 0, \tag{174}$$

which is a standard wave equation for a massless field evolving in an expanding universe.

Having seen that the inflaton perturbation behaves in this way, it is not much work to obtain the quantum fluctuations that result in the field at late times (*i.e.* on scales much larger than the de Sitter horizon). First consider the fluctuations in flat space on scales well inside the horizon. In principle, this requires quantum field theory, but the vacuum fluctuations in ϕ can be derived by a simple argument using the uncertainty principle. First of all, note that the sub-horizon equation of motion is just that for a simple harmonic oscillator: $[\ddot{\delta\phi}] + \omega^2[\delta\phi] = 0$, where $\omega = k/a$. For an oscillator of mass m and position coordinate q, the rms uncertainty in q in the ground state is

$$q_{\rm rms} = \left(\frac{\hbar}{2m\omega}\right)^{1/2}. \tag{175}$$

This can be derived immediately from the uncertainty principle, which says that the minimum uncertainty is

$$\langle(\delta p)^2\rangle\langle(\delta q)^2\rangle = \hbar^2/4. \tag{176}$$

For a classical oscillation with $q(t) \propto e^{i\omega t}$, the momentum is $p(t) = m\dot{q} = i\omega m q(t)$. Quantum uncertainty can be thought of as saying that we lack a knowledge of the amplitude of the oscillator, but in any case the amplitudes in momentum and coordinate must be related by $p_{\rm rms} = m\omega q_{\rm rms}$. The uncertainty principle therefore says

$$m^2\omega^2 q_{\rm rms}^4 = \hbar^2/4, \tag{177}$$

which yields the required result.

For scalar field fluctuations, our 'coordinate' q is just the field $\delta\phi$, the oscillator frequency is $\omega = k/a$, and we now revert to $\hbar = 1$. What is the analogue of the mass of the oscillator in this case? Recall that a Lagrangian, L, has a momentum $p = \partial L/\partial\dot{q}$ corresponding to each coordinate. For the present application, the kinetic part of the Lagrangian density is

$$\mathcal{L}_{\rm kinetic} = a^3\dot{\phi}^2/2, \tag{178}$$

and the 'momentum' conjugate to ϕ is $p = a^3\dot{\phi}$. In the current case, p is a momentum *density*, since $\mathcal{L}$ is a Lagrangian density; we should therefore multiply p by a comoving volume V, so the analogue of the SHO mass is $m = a^3V$.

The uncertainty principle therefore gives us the variance of the zero-point fluctuations in $\delta\phi$ as

$$\langle(\delta\phi)^2\rangle = \left(2\,(a^3V)\,(k/a)\,\right)^{-1}, \tag{179}$$

so we adopt an rms field amplitude from quantum fluctuations of

$$\delta\phi = (a^3V)^{-1/2}\,(2k/a)^{-1/2}\,e^{-ikt/a}. \tag{180}$$

This is the correct expression that results from a full treatment in quantum field theory.

With this boundary condition, it straightforward to check by substitution that the following expression satisfies the evolution equation:

$$\delta\phi = (a^3V)^{-1/2}\,(2k/a)^{-1/2}\,e^{ik/aH}\,(1 + iaH/k) \tag{181}$$

(remember that H is a constant, so that $(d/dt)[aH] = H\dot{a} = aH^2$ *etc.*). At early times, when the horizon is much larger than the wavelength, $aH/k \ll 1$, and so this expression is the flat-space result, except that the time dependence looks a little odd, being $\exp(ik/aH)$. However, since $(d/dt)[k/aH] = -k/a$, we see that the oscillatory term has a leading dependence on t of the desired $-kt/a$ form.

At the opposite extreme, $aH/k \gg 1$, the squared fluctuation amplitude becomes frozen out at the value

$$\langle 0|\,|\phi_k|^2\,|0\rangle = \frac{H^2}{2k^3V}, \tag{182}$$

where we have emphasised that this is the vacuum expectation value. The fluctuations in ϕ depend on k in such a way that the fluctuations per decade are constant:

$$\frac{d\,(\delta\phi)^2}{d\,\ln k} = \frac{4\pi k^3V}{(2\pi)^3}\,\langle 0|\,|\phi_k|^2\,|0\rangle = \left(\frac{H}{2\pi}\right)^2 \tag{183}$$

(the factor $V/(2\pi)^3$ comes from the density of states in the Fourier transform, and cancels the $1/V$ in the field variance; $4\pi k^2\,dk = 4\pi k^3\,d\ln k$ comes from the k-space volume element).

This completes the argument. The initial quantum zero-point fluctuations in the field have been transcribed to a constant classical fluctuation that can eventually manifest itself as large-scale structure. The rms value of fluctuations in ϕ can be used as above to deduce the power spectrum of mass fluctuations well after inflation is over. In terms of the variance per $\ln k$ in potential perturbations, the answer is

$$\begin{aligned} \delta_{\rm H}^2 \equiv \Delta_\Phi^2(k) &= \frac{H^4}{(2\pi\dot\phi)^2} \\ H^2 &= \frac{8\pi}{3}\,\frac{V}{m_{\rm P}^2} \\ 3H\dot\phi &= -V', \end{aligned} \tag{184}$$

where we have also written once again the slow-roll condition and the corresponding relation between H and V, since manipulation of these three equations is often required in derivations.

TENSOR PERTURBATIONS Later in the course, we will compare the predictions of this inflationary apparatus with observations of the fluctuating density field of the contemporary universe. It should be emphasised again just what an audacious idea this is: that all the structure around us was seeded by quantum fluctuations while the universe was of subnuclear scale. It would be nice if we could verify this radical assumption, and there is one basic test: if the idea of quantum fluctuations is correct, it should apply to every field that was present in the early universe. In particular, it should apply to the gravitational field. This corresponds to metric perturbations in the form of a tensor $h^{\mu\nu}$, whose coefficients have some typical amplitude h (not the Hubble parameter). This **spatial strain** is what is measured by gravity-wave telescopes such as LIGO: the separation between a pair of freely-suspended masses changes by a fractional amount of order h as the wave passes. These experiments can be fabulously precise, with a current sensitivity of around $h = 10^{-21}$.

What value of h does inflation predict? For scalar perturbations, small-scale quantum fluctuations lead to an amplitude $\delta\phi = H/2\pi$ on horizon exit, which transforms to a metric fluctuation $\delta_{\rm H} = H\delta\phi/\dot\phi$. Tensor modes behave similarly – except that h must be dimensionless, whereas ϕ

has dimensions of mass. On dimensional grounds, then, the formula for the tensor fluctuations is plausible:

$$h_{\rm rms} \sim H/m_{\rm P}. \tag{185}$$

But unlike fluctuations in the inflaton, the tensor fluctuation do not affect the progress of inflation: once generated, they play no further part in events and survive to the present day. Detection of these primordial tensor perturbations would not only give confidence in the basic inflationary picture, but would measure rather directly the energy scale of inflation.

INFLATON COUPLING The calculation of density inhomogeneities sets an important limit on the inflation potential, which is that it is very weakly self-coupled. To see what this means, write a power-law potential as $V(\phi) = \lambda m_{\rm P}^4 (\phi/m_{\rm P})^\alpha$, for which inflation will only proceed while $\phi \gg m_{\rm P}$. To verify this, calculate the inflationary parameter

$$\epsilon = \frac{m_{\rm P}^2}{16\pi}(V'/V)^2 = \frac{m_{\rm P}^2 \alpha^2}{16\pi\phi^2}, \tag{186}$$

so $\epsilon \ll 1$ requires $\phi \gg m_{\rm P}$. Now consider the amplitude of inflationary fluctuations:

$$\delta_{\rm H}^2 = \frac{H^4}{(2\pi\dot\phi)^2} = \frac{128\pi}{3}\left(\frac{V^3}{m_{\rm P}^6\, V'^2}\right) = \frac{128\pi}{3\alpha^2}\,\lambda(\phi/m_{\rm P})^{\alpha+2}. \tag{187}$$

The observed $\delta_{\rm H} \sim 10^{-5}$, plus the constraint $\phi \gg m_{\rm P}$, therefore requires

$$\lambda \ll 10^{-12}. \tag{188}$$

For a mass-like potential $V = m^2\phi^2$, $m^2 = \lambda m_{\rm P}$ in the above notation, so our constraint is

$$m \ll 10^{-6} m_{\rm P}. \tag{189}$$

Thus the inflaton has to be a light particle and/or the self-coupling λ has to be small. This is one way to see that the only known scalar field (the Higgs field) cannot be the inflaton: at large ϕ the Higgs potential is dominated by a $\lambda\phi^4$ dependence with λ of order unity.

These constraints appear to suggest a defect in inflation, in that we should be able to use the theory to *derive* $\delta_{\rm H} \sim 10^{-5}$. The amplitude of $\delta_{\rm H}$ is one of the most important numbers in cosmology, and it is vital to know whether there is a simple explanation for its magnitude. One way to view this is to express the horizon-scale amplitude as

$$\delta_{\rm H} \sim \frac{V^{1/2}}{m_{\rm P}^2\,\epsilon^{1/2}}. \tag{190}$$

We have argued that inflation will end with ϵ of order unity; if the potential were to have the characteristic value $V \sim E_{\rm GUT}^4$ then this would give the simple result

$$\delta_{\rm H} \sim \left(\frac{m_{\rm GUT}}{m_{\rm P}}\right)^2. \tag{191}$$

TILT Finally, deviations from exact exponential expansion must exist at the end of inflation, and the corresponding change in the fluctuation power spectrum is a potential test of inflation. Define the **tilt** of the fluctuation spectrum as follows:

$$\text{tilt} \equiv 1 - n \equiv -\frac{d\ln\delta_{\rm H}^2}{d\,\ln k}. \tag{192}$$

We then want to express the tilt in terms of parameters of the inflationary potential, ϵ and η. These are of order unity when inflation terminates; ϵ and η must therefore be evaluated when the observed universe left the horizon, recalling that we only observe the last 60-odd e-foldings of inflation. The way to introduce scale dependence is to write the condition for a mode of given comoving wavenumber to cross the de Sitter horizon,

$$a/k = H^{-1}. \tag{193}$$

Since H is nearly constant during the inflationary evolution, we can replace $d/d\,\ln k$ by $d\,\ln a$, and use the slow-roll condition to obtain

$$\frac{d}{d\,\ln k} = a\,\frac{d}{da} = \frac{\dot\phi}{H}\,\frac{d}{d\phi} = -\frac{m_{\rm P}^2}{8\pi}\,\frac{V'}{V}\,\frac{d}{d\phi}. \tag{194}$$

We can now work out the tilt, since the horizon-scale amplitude is

$$\delta_{\rm H}^2 = \frac{H^4}{(2\pi\dot{\phi})^2} = \frac{128\pi}{3}\left(\frac{V^3}{m_{\rm P}^6\,V'^2}\right), \tag{195}$$

and derivatives of V can be expressed in terms of the dimensionless parameters ϵ and η. The tilt of the density perturbation spectrum is thus predicted to be

$$1 - n = 6\epsilon - 2\eta \tag{196}$$

For most models in which the potential is a smooth polynomial-like function, $|\eta| \simeq |\epsilon|$. Since ϵ has the larger coefficient and is positive by definition, the simplest inflation models tend to predict that the spectrum of scalar perturbations should be slightly tilted, in the sense that n is slightly less than unity.

It is interesting to put flesh on the bones of this general expression and evaluate the tilt for some specific inflationary models. This is easy in the case of power-law inflation with $a \propto t^p$ because the inflation parameters are constant: $\epsilon = \eta/2 = 1/p$, so that the tilt here is always

$$1 - n = 2/p \tag{197}$$

In general, however, the inflation derivatives have to be evaluated explicitly on the largest scales, 60 e-foldings prior to the end of inflation, so that we need to solve

$$60 = \int H\,dt = \frac{8\pi}{m_{\rm P}^2}\int_{\phi_{\rm end}}^{\phi}\frac{V}{V'}\,d\phi. \tag{198}$$

A better motivated choice than power-law inflation would be a power-law potential $V(\phi) \propto \phi^\alpha$; many chaotic inflation models concentrate on $\alpha = 2$ (mass-like term) or $\alpha = 4$ (highest renormalizable power). Here, $\epsilon = m_{\rm P}^2\alpha^2/(16\pi\phi^2)$, $\eta = \epsilon \times 2(\alpha-1)/\alpha$, and

$$60 = \frac{8\pi}{m_{\rm P}^2}\int_{\phi_{\rm end}}^{\phi}\frac{\phi}{\alpha}\,d\phi = \frac{4\pi}{m_{\rm P}^2\alpha}(\phi^2 - \phi_{\rm end}^2). \tag{199}$$

It is easy to see that $\phi_{\rm end} \ll \phi$ and that $\epsilon = \alpha/240$, leading finally to

$$1 - n = (2+\alpha)/120. \tag{200}$$

The predictions of simple chaotic inflation are thus very close to scale invariance in practice: $n = 0.97$ for $\alpha = 2$ and $n = 0.95$ for $\alpha = 4$. However, such a tilt has a significant effect over the several decades in k from CMB anisotropy measurements to small-scale galaxy clustering. These results are in some sense the default inflationary predictions: exact scale invariance would be surprising, as would large amounts of tilt. Either observation would indicate that the potential must have a more complicated structure, or that the inflationary framework is not correct.

6.5 Stochastic eternal inflation

These fluctuations in the scalar field can affect the progress of inflation itself. They can be thought of as adding a random-walk element to the classical rolling of the scalar field down the trough defined by $V(\phi)$. In cases where ϕ is too close to the origin for inflation to persist for sufficiently long, it is possible for the quantum fluctuations to push ϕ further out – creating further inflation in a self-sustaining process. This is the concept of **stochastic inflation**.

Consider the scalar field at a given point in the inflationary universe. Each e-folding of the expansion produces new classical fluctuations, which add incoherently to those previously present. If the field is sufficiently far from the origin in a polynomial potential, these fluctuations produce a random walk of $\phi(t)$ that overwhelms the classical trajectory in which ϕ tries to roll down the potential, as follows. The classical amplitude from quantum fluctuations is $\delta\phi = H/2\pi$, and a new disturbance of the same rms will be added for every $\Delta t = 1/H$. The slow-rolling equation says that the trajectory is $\dot{\phi} = -V'/3H$; we also have $H^2 = 8\pi V/3m_{\rm P}^2$, so that the classical change in ϕ is $\Delta\phi = -m_{\rm P}^2 V'/8\pi V$ in a time $\Delta t = 1/H$. Consider $V = \lambda|\phi|^n/(nm_{\rm P}^{n-4})$, for which these two changes in ϕ will be equal at $\phi \sim \phi^* = m_{\rm P}/\lambda^{1/(n+2)}$. For smaller ϕ, the quantum fluctuations will have a negligible effect on the classical trajectory; for larger ϕ, the equation of motion will become stochastic. The resulting random walk will send some parts of the universe to ever larger values of ϕ, so inflation never entirely ends. This **eternal inflation** is the basis for the concept of the **inflationary multiverse**: different widely-separated parts of the universe will inflate by different amounts, producing in effect separate universes with distinct formation histories.

7 Structure formation – II

The equations of motion are nonlinear, and we have only solved them in the limit of linear perturbations. We now discuss evolution beyond the linear regime, first considering the full numerical solution of the equations of motion, and then a key analytic approximation by which the 'exact' results can be understood.

N-BODY MODELS The exact evolution of the density field is usually performed by means of an **N-body simulation**, in which the density field is represented by the sum of a set of fictitious discrete particles. We need to solve the equations of motion for each particle, as it moves in the gravitational field due to all the other particles. Using comoving units for length and velocity ($\mathbf{v} = a\mathbf{u}$), we have previously seen the equation of motion

$$\frac{d}{dt}\mathbf{u} = -2\frac{\dot{a}}{a}\mathbf{u} - \frac{1}{a^2}\boldsymbol{\nabla}\Phi, \tag{201}$$

where Φ is the Newtonian gravitational potential due to density perturbations. The time derivative is already in the required form of the convective time derivative observed by a particle, rather than the partial $\partial/\partial t$.

In outline, this is straightforward to solve, given some initial positions and velocities. Defining some timestep dt, particles are moved according to $d\mathbf{x} = \mathbf{u}\ dt$, and their velocities updated according to $d\mathbf{u} = \dot{\mathbf{u}}\ dt$, with $\dot{\mathbf{u}}$ given by the equation of motion (in practice, more sophisticated time integration schemes are used). The hard part is finding the gravitational force, since this involves summation over $(N-1)$ other particles each time we need a force for one particle. All the craft in the field involves finding clever ways in which all the forces can be evaluated in less than the raw $O(N^2)$ computations per timestep. We will have to omit the details of this, unfortunately, but one obvious way of proceeding is to solve Poisson's equation on a mesh using a Fast Fourier Transform. This can convert the $O(N^2)$ time scaling to $O(N \ln N)$, which is a qualitative difference given that N can be as large as 10^{10}.

These non-linear effects boost the amplitude of the power spectrum at small physical scales (large k scales) as can be seen in Figure 17 . For cosmological observations we need to understand these non-linear effects to high precision. This is one of the issues facing modern day cosmology and non-linear effects can only be calculated through large scale suites of HPC N-body simulations.

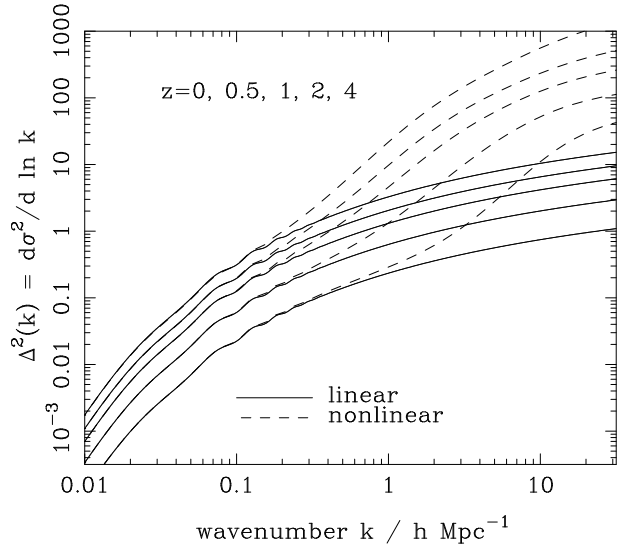

Figure 17. ΛCDM power spectrum normalised by $\sigma_8 = 0.9$. The linear power spectrum is show solid and the non-linear power spectrum is shown dashed using the fitting formula from Smith et al. 2003.

THE SPHERICAL MODEL N-body models can yield evolved density fields that are nearly exact solutions to the equations of motion, but working out what the results mean is then more a question of data analysis than of deep insight. Where possible, it is important to have analytic models that guide the interpretation of the numerical results. The most important model of this sort is the spherical density perturbation, which can be analysed immediately using the tools developed for the Friedmann models, since Birkhoff's theorem tells us that such a perturbation behaves in exactly the same way as part of a closed universe. The equations of motion are the same as for the scale factor, and we can therefore write down the **cycloid solution** immediately. For a matter-dominated universe, the relation between the proper radius of the sphere and time is

$$\begin{aligned} r &= A(1-\cos\theta) \\ t &= B(\theta-\sin\theta). \end{aligned} \tag{202}$$

It is easy to eliminate θ to obtain $\ddot{r} = -GM/r^2$, and the relation $A^3 = GMB^2$ (use e.g. $\dot{r} = (dr/d\theta)/((dt/d\theta)$, which gives $\dot{r} = [A/B]\sin\theta/[1-\cos\theta]$). Expanding these relations up to order θ^5 gives $r(t)$ for small t:

$$r \simeq \frac{A}{2}\left(\frac{6t}{B}\right)^{2/3}\left[1-\frac{1}{20}\left(\frac{6t}{B}\right)^{2/3}\right], \tag{203}$$

and we can identify the density perturbation within the sphere:

$$\delta \simeq \frac{3}{20}\left(\frac{6t}{B}\right)^{2/3}. \tag{204}$$

This all agrees with what we knew already: at early times the sphere expands with the $a \propto t^{2/3}$ Hubble flow and density perturbations grow proportional to a.

We can now see how linear theory breaks down as the perturbation evolves. There are three interesting epochs in the final stages of its development, which we can read directly from the above solutions. Here, to keep things simple, we compare only with linear theory for an $\Omega = 1$ background.

(1) **Turnround**. The sphere breaks away from the general expansion and reaches a maximum radius at $\theta = \pi$, $t = \pi B$. At this point, the true density enhancement with respect to the background is just $[A(6t/B)^{2/3}/2]^3/r^3 = 9\pi^2/16 \simeq 5.55$.

(2) **Collapse**. If only gravity operates, then the sphere will collapse to a singularity at $\theta = 2\pi$.

(3) **Virialization**. Clearly, collapse to a point is highly idealized. Consider the time at which the sphere has collapsed by a factor 2 from maximum expansion ($\theta = 3\pi/2$). At this point, it has kinetic energy K related to potential energy V by $V = -2K$. This is the condition for equilibrium, according to the **virial theorem**. Conventionally, it is assumed that this stable virialized radius is eventually achieved only at the collapse time, at which point the density contrast is $\rho/\bar{\rho} = (6\pi)^2/2 \simeq 178$ and $\delta_{\rm lin} \simeq 1.686$.

These calculations are the basis for a common 'rule of thumb', whereby one assumes that linear theory applies until $\delta_{\rm lin}$ is equal to some δ_c a little greater than unity, at which point virialization is deemed to have occurred. Although the above only applies for $\Omega = 1$, analogous results can be

worked out from the full $\delta_{\rm lin}(z,\Omega)$ and $t(z,\Omega)$ relations. These indicate that $\delta_{\rm lin} \simeq 1$ is a good criterion for collapse for any value of Ω likely to be of practical relevance. The density contrast at virialization tends to be higher in low-density universes, where the faster expansion means that, by the time a perturbation has turned round and collapsed to its final radius, a larger density contrast has been produced. For real non-spherical systems, it is not clear that this effect is meaningful, and in practice a fixed density contrast of around 200 is used to define the **virial radius** that marks the boundary of an object.

The spherical model thus gives some insight into why the end result of a nonlinear N-body calculation is dominated by nearly spherical 'haloes' of dark matter. These systems can be of varying mass, depending on the size of the region that exceeds the threshold density contrast, spanning the range from single galaxies to the most massive clusters of galaxies. Galaxy formation is thus simplified to the task of understanding the fate of gas within these haloes, although it is complicated by the fact that halo growth proceeds via merging. Several galaxies may form at high redshift within separate low-mass haloes, but these then merge into a single group with a common halo – and **subhaloes** within it, which are clumps of dark-matter substructure that mark the residue of the haloes in which the galaxies formed, and in which they continue to reside.

HALOES AND CLUSTERING The existence of dark-matter haloes gives an immediate insight into the nonlinear properties of the dark-matter density field, via an approximation known as the **halo model**. Start by distributing point seeds throughout the universe with number density n, in which case the power spectrum of the resulting density field is just shot noise:

$$\Delta^2(k) = \frac{4\pi}{n} \left(\frac{k}{2\pi}\right)^3. \tag{205}$$

The density field for a distribution of clumps is produced by convolution of the initial field of delta-functions, so the power spectrum is simply modified by the squared Fourier transform of the clump density profile:

$$\Delta^2(k) = \frac{4\pi}{n} \left(\frac{k}{2\pi}\right)^3 |W_k|^2, \tag{206}$$

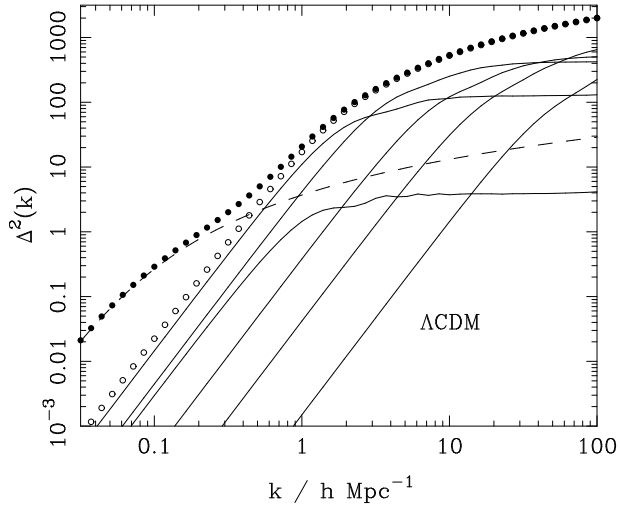

Figure 18. The decomposition of the mass power spectrum according to the halo model, for the flat $\Omega_m = 0.3$, $\Gamma = 0.2$, $\sigma_8 = 0.8$ case. The dashed line shows linear theory, and the open circles show the predicted 1-halo contribution. Adding in linear theory to produce the correct large-scale clustering yields the solid points. The full lines show the contribution of different mass ranges to the 1-halo term: bins of width a factor 10 in width, starting at $10^{10} - 10^{11} h^{-1} M_\odot$ and ending at $10^{15} - 10^{16} h^{-1} M_\odot$. The more massive haloes have larger virial radii and hence filter the power spectrum on progressively larger scales. The majority of the quasilinear power is contributed by the haloes near the peak in the mass function at $10^{14} - 10^{15} h^{-1} M_\odot$.

where the 'window function' cuts off the power at $k \sim 1/r_v$, depending on the virial radius (which we have seen is an increasing function of mass). Integrating over the halo mass function, we get

$$\Delta^2_{\rm halo}(k) = 4\pi \left(\frac{k}{2\pi}\right)^3 \frac{\int M^2 \, |W_k(M)|^2 \, f(M) \, dM}{\left[\int M \, f(M) \, dM\right]^2}. \qquad (207)$$

The normalization term $\int M \, f(M) \, dM$ just gives the total background density, ρ_b, since $f(M) \, dM$ is the number density of haloes in the mass range dM.

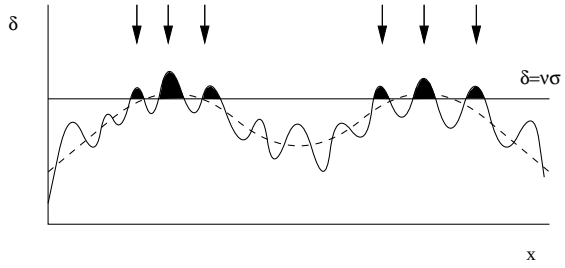

Figure 19. The high-peak bias model. If we decompose a density field into a fluctuating component on galaxy scales, together with a long-wavelength 'swell' (shown dashed), then those regions of density that lie above a threshold in density of ν times the rms will be strongly clustered. If proto-objects are presumed to form at the sites of these high peaks (shaded, and indicated by arrows), then this is a population with Lagrangian bias – i.e. a non-uniform spatial distribution even prior to dynamical evolution of the density field.

So far, we have ignored any spatial correlations in the halo positions. A simple guess for amending this is to add the linear power spectrum to the power generated by the halo structure:

$$\Delta^2_{\rm tot} = \Delta^2_{\rm halo} + \Delta^2_{\rm linear}. \tag{208}$$

The justification for this is that the extra small-scale power introduced by nonlinear evolution is associated with the internal structure of the haloes. In practice, this model works extremely well, giving a good description of the power spectrum on all scales (see figure 19). This is a novel way of looking at the features in the nonlinear spectrum, particularly the steep rise between $k \simeq 0.5\,h\,{\rm Mpc}^{-1}$ and $k \simeq 5\,h\,{\rm Mpc}^{-1}$, and the flattening on smaller scales.

GALAXY BIAS This model can be extended readily by allowing for the fact that different classes of galaxy will be associated with different mass haloes. Most simply, haloes below a certain mass will contain insufficient mass to constitute even a single galaxy; beyond this, the efficiency of galaxy formation will depend in a complex an nonlinear way on halo enviroment. All this can be encoded in the **halo occupation number**, $N(M)$. This mass-dependent weighting is especially important, since the amplitude of apparent density fluctuations depends on halo mass, with the most massive haloes having the highest amplitude.

We now consider the central mechanism of biased clustering, in which a rare high density fluctuation, corresponding to a massive object, collapses sooner if it lies in a region of large-scale overdensity. This 'helping hand' from the long-wavelength modes means that overdense regions contain an enhanced abundance of massive objects with respect to the mean, so that these systems display enhanced clustering. The basic mechanism can be immediately understood via the diagram in figure 19; it was first clearly analysed by Kaiser (1984) in the context of rich clusters of galaxies.

The key ingredient of this analysis is the mass function of dark-matter haloes. The universe fragments into virialized systems such that $f(M)\,dM$ is the number density of haloes in the mass range dM. The simplest analyses of the mass function rest on the concept of a density threshold: collapse to a virialized object is deemed to have occurred where linear-theory δ averaged over a sphere containing mass M reaches some critical value δ_c. Generally, we shall assume the value $\delta_c = 1.686$ appropriate for spherical collapse in an Einstein–de Sitter universe. Now imagine that this situation is perturbed, by adding some constant shift ϵ to the density perturbations over some large region. The effect of this is to perturb the threshold: fluctuations now only need to reach $\delta = \delta_c - \epsilon$ in order to achieve collapse. Of course, the field ϵ can hardly be imposed by hand; instead, we make the **peak-background split**, in which δ is mentally decomposed into a small-scale and a large-scale component – which we identify with ϵ. The scale above which the large-scale

Without going into details, it should be obvious that the abundance of high-mass haloes is highly sensitive to any change of the threshold: the probability of exceeding the $\delta > \delta_c$ threshold is dominated by a Gaussian factor $\exp(-\nu^2/2)$, where $\nu \equiv \delta_c/\sigma(M)$ and σ is the fractional rms in the density field filtered on a mass scale M; thus high-mass objects correspond to large values of ν, and the density field is dominated by typical haloes with $\nu \sim 1$. The result is a mass-dependent **linear bias parameter**, where

$$\delta_g = b(M)\delta, \tag{209}$$

and $b(M)$ increases $\propto \nu^2$ at large mass. Once we have averaged over a distribution of galaxy occupation numbers, the overall bias parameter can vary from less than unity for dwarf galaxies to 2 or more for massive ellipticals. Thus, in probing primordial density fluctuations using galaxy surveys, the bias is often treated as an unknown free parameter.

8 CMB anisotropies – I

So far, we have concentrated on describing perturbations in the matter density, and will go on to discuss ways in which these may be observed. But first, we should put in place the corresponding machinery for the fluctuations in the radiation density. These can be observed directly in terms of fluctuations in the temperature of the CMB, which relate to the density fluctuation field at $z \simeq 1100$. We therefore have the chance to observe both current cosmic structure and its early seeds. By putting the two together and requiring consistency, the cosmological model can be pinned down with amazing precision.

8.1 Anisotropy mechanisms

Fluctuations in the 2D temperature perturbation field are treated similarly to density fluctuations, except that the field is expanded in spherical harmonics, so modes of different scales are labelled by multipole number, ℓ:

$$\frac{\delta T}{T}(\hat{\mathbf{q}}) = \sum a_\ell^m Y_{\ell m}(\hat{\mathbf{q}}), \tag{210}$$

where $\hat{\mathbf{q}}$ is a unit vector that specifies direction on the sky. The spherical harmonics satisfy the orthonormality relation $\int Y_{\ell m} Y^*_{\ell' m'}\, d^2q = \delta_{\ell\ell'}\delta_{mm'}$, so the variance in temperature averaged over the sky is

$$\left\langle \left(\frac{\delta T}{T}\right)^2 \right\rangle = \frac{1}{4\pi}\sum_{\ell,m} |a_\ell^m|^2 = \frac{1}{4\pi}\sum_\ell (2\ell+1)\, C_\ell \tag{211}$$

The spherical harmonics are familiar as the eigenfunctions of the angular part of ∇^2, and there are $2\ell+1$ modes of given ℓ, hence the notation for the angular power spectrum, C_ℓ. For $\ell \gg 1$, the spherical harmonics become equivalent to Fourier modes, in which the angular wavenumber is ℓ; therefore one can associate a 'wavelength' $2\pi/\ell$ with each mode.

Once again, it is common to define a 'power per octave' measure for the temperature fluctuations:

$$\mathcal{T}^2(\ell) = \ell(\ell+1)C_\ell/2\pi \tag{212}$$

(although shouldn't $\ell(\ell+1)$ be $\ell(\ell+1/2)$? – see later). Note that $\mathcal{T}^2(\ell)$ is a power per $\ln\ell$; the modern trend is often to plot CMB fluctuations with a linear scale for ℓ – in which case one should really use $\mathcal{T}^2(\ell)/\ell$.

We now list the mechanisms that cause **primary anisotropies** in the CMB (as opposed to **secondary anisotropies**, which are generated by scattering along the line of sight). There are three basic primary effects, illustrated in figure 20, which are important on respectively large, intermediate and small angular scales:

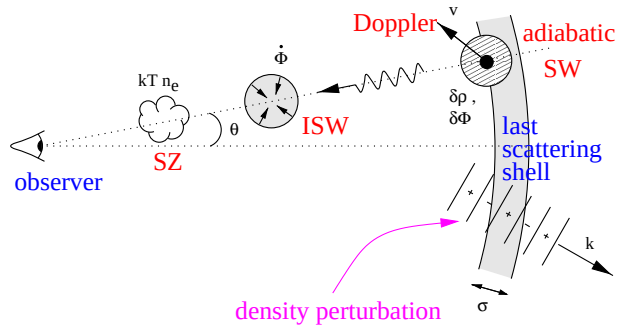

Figure 20. Illustrating the physical mechanisms that cause CMB anisotropies. The shaded arc on the right represents the last-scattering shell; an inhomogeneity on this shell affects the CMB through its potential, adiabatic and Doppler perturbations. Further perturbations are added along the line of sight by time-varying potentials (Rees–Sciama effect) and by electron scattering from hot gas (Sunyaev–Zeldovich effect). The density field at last scattering can be Fourier analysed into modes of wavevector $\mathbf{k}$. These spatial perturbation modes have a contribution that is in general damped by averaging over the shell of last scattering. Short-wavelength modes are more heavily affected (i) because more of them fit inside the scattering shell, and (ii) because their wavevectors point more nearly radially for a given projected wavelength.

(1) Gravitational (Sachs–Wolfe) perturbations. Photons from high-density regions at last scattering have to climb out of potential wells, and are thus redshifted:

$$\frac{\delta T}{T} = \frac{1}{3}(\Phi/c^2). \tag{213}$$

The factor 1/3 is a surprise, which arises because Φ has two effects: (i) it redshifts the photons we see, so that an overdensity *cools* the background as the photons climb out, $\delta T/T = \Phi/c^2$; (ii) it causes time dilation at the last-scattering surface, so that we seem to be looking at a younger (and hence *hotter*) universe where there is an overdensity. The time dilation is $\delta t/t = \Phi/c^2$; since the time dependence of the scale factor is $a \propto t^{2/3}$ and $T \propto 1/a$, this produces the counterterm $\delta T/T = -(2/3)\Phi/c^2$.

(2) Intrinsic (adiabatic) perturbations. In high-density regions, the coupling of matter and radiation can compress the radiation also, giving a higher temperature:

$$\frac{\delta T}{T} = \frac{\delta(z_{\rm LS})}{3}, \tag{214}$$

(3) Velocity (Doppler) perturbations. The plasma has a non-zero velocity at recombination, which leads to Doppler shifts in frequency and hence brightness temperature:

$$\frac{\delta T}{T} = \frac{\delta \mathbf{v} \cdot \hat{\mathbf{r}}}{c}. \tag{215}$$

To the above list should be added 'tensor modes': anisotropies due to a background of primordial gravitational waves, potentially generated during an inflationary era (see below).

There are in addition effects generated along the line of sight. One important effect is the integrated Sachs-Wolfe effect (**ISW effect**), which arises when the potential perturbations evolve:

$$\frac{\delta T}{T} = \frac{1}{c^2}\int (\dot{\Psi} + \dot{\Phi})\, dt. \tag{216}$$

In the usual $\Psi = \Phi$ limit, this is twice as large as one might have expected from Newtonian intuition. This factor 2 thus has an origin that is similar to the factor 2 for relativistic light deflection (where

the one-line argument is that the gravitational potential modifies both the time and space parts of the metric, and each contribute equally to the effective change in the coordinate speed of light). But the ISW effect is a little more subtle, and we shall just accept the result as intuitively plausible. As we have seen, the potential Φ stays constant in the linear regime during the matter-dominated era, as long as $\Omega_m \simeq 1$, so the source term for the ISW effect vanishes for much of the universe's history. The ISW effect then becomes only important quite near to the last scattering redshift (because radiation is still important) and at low z (because of Λ).

Other foreground effects are to do with the development of nonlinear structure, and are mainly on small scales (principally the Sunyaev–Zeldovich effect from IGM Comptonization). The exception is the effect of reionization; to a good approximation, this merely damps the fluctuations on all scales:

$$\frac{\delta T}{T} \quad \rightarrow \quad \frac{\delta T}{T} \exp -\tau, \tag{217}$$

where the optical depth must exceed $\tau \simeq 0.04$, based on the highest-redshift quasars and the BBN baryon density. As we will see later, CMB polarization data have detected a signature consistent with $\tau = 0.1 \pm 0.03$, implying reionization at $z \simeq 10$.

8.2 Power spectrum

We now need to see how the angular power spectrum of the CMB arises from the implementation of these effects. The physical separation we have made is useful for insight, although it is not exactly how things are calculated in practice. We have not been able to spend time going into the detailed formalism used on CMB anisotropies, and the details will have to be omitted here – although the actual equations to be integrated are not enormously complicated. For the present purpose, we will make a few comments about why the exact approach is complicated, and then retreat to a simpler approximate treatment.

The natural approach is to start in Fourier space and consider a density fluctuation of given wavevector $\mathbf{k}$; if we can work out how this appears as an induced temperature fluctuation on the CMB sky, then the problem can be solved by superposition. The wavevector $\mathbf{k}$ sets a natural polar axis, and the temperature anisotropy corresponds to knowing the photon phase-space distribution at our location in space (i.e. the distribution of the photons in energy and as a function of angle with respect to $\mathbf{k}$). Evolving this function is hard principally because of the coupling between radiation and matter, which is by Thomson scattering. Scattering a beam of photons that come from a given direction will tend to push the electron in the opposite direction, so a nett force requires an

anisotropic the photon distribution. In fact, it is clear that the force must be proportional to the dipole moment of the distribution function, and this is obviously a problem: it couples the evolution of the number of photons travelling at a given direction with a knowledge of the whole distribution. Mathematically, we have an integro-differential equation.

In practice, rather than trying to solve numerically for the photon distribution function (normally denoted by Θ), we can carry out a multipole transform to work with Θ_ℓ. The integro-differential equation then becomes a set of equations that couple different ℓ values. These have to be solved as a large set of equations (we will see that the CMB power spectrum contains signal at least to $\ell \gtrsim 1000$), and when this is done we still have to integrate over k space. It took many years to solve this numerical challenge, and even then the computations were very slow. But a key event in cosmology was the 1996 release of CMBFAST, a public Boltzmann code that allowed computation of the CMB angular power spectrum sufficiently rapidly that a large range of models could be investigated by non-specialists.

TIGHT-COUPLING PROJECTION APPROACH An alternative approximate method is to imagine that the temperature anisotropies exist as a 3D spatial field. The last-scattering surface can be envisaged as a slice through this field, so the angular properties are really just a question of understanding the projection that is involved. This works reasonably well in the **tight coupling** limit where photons and baryons are a single fluid – but this is of course breaking down at last scattering, where the photon mean free path is becoming large.

The projection is easily performed in the **flat-sky approximation**, where we ignore the curvature of the celestial sphere. The angular wavenumber is then just $\ell = KD_{\rm H}$, where $D_{\rm H}$ is the distance to the last-scattering surface and K is a 2D transverse physical wavenumber ($K^2 = k_x^2 + k_y^2$). The relation between 3D and 2D power spectra is easily derived: we just add up the power along the unused axis, k_z:

$$P_{\rm 2D}(k_x, k_y) = \sum_{k_z} P_{\rm 3D}(k_x, k_y, k_z) = \frac{L}{2\pi} \int_{-\infty}^{\infty} P_{\rm 3D}(k)\, dk_z. \tag{218}$$

In terms of dimensionless power, this is

$$\Delta^2_{\rm 2D}(K) = \left(\frac{L}{2\pi}\right)^2 2\pi K^2\, P_{\rm 2D}(K) = K^2 \int_0^{\infty} \Delta^2_{\rm 3D}(k)\, dk_z/k^3, \tag{219}$$

where $k^2 = K^2 + k_z^2$. The 2D spectrum is thus a smeared version of the 3D one, but the relation is pleasingly simple for a scale-invariant spectrum in which $\Delta^2_{\rm 3D}(k)$ is a constant:

$$\Delta^2_{\rm 2D}(K) = \Delta^2_{\rm 3D}. \tag{220}$$

The important application of this is to the Sachs-Wolfe effect, where the 3D dimensionless spectrum of interest is that of the potential, $\Delta^2_\Phi = \delta^2_{\rm H}$. This shows that the angular spectrum of the CMB should have a flat portion at low ℓ that measures directly the metric fluctuations.

This is the signature that formed the first detection of CMB anisotropies – by COBE in 1992; we will see below that this corresponds to

$$\delta_{\rm H} \simeq 3 \times 10^{-5}. \tag{221}$$

This immediately determines the large-scale matter power spectrum in the universe today. We know from Poisson's equation that the relation between potential and density power spectra at scale factor a is

$$\Delta^2_\Phi = (4\pi G\rho_m a^2/k^2)^2 \Delta^2(a) \equiv \delta^2_{\rm H}. \tag{222}$$

Converting to the present, $\Delta^2 = a^{-2}\Delta^2(a) f(\Omega_m)^2$, and we get

$$\Delta^2 = (4/9)\delta^2_{\rm H} \left(\frac{ck}{H_0}\right)^4 \Omega_m^{-2} f(\Omega_m)^2 \tag{223}$$

(where $f(\Omega_m)$, $\simeq \Omega_m^{0.23}$ for a flat universe, is the growth suppression factor). This expression is modified on small scales by the transfer function, but it shows how mass fluctuations today can be deduce from CMB anisotropies. As an aside, a more informal argument in the opposite direction is to say that we can estimate the depth of potential wells today:

$$v^2 \sim \frac{GM}{r} \quad \Rightarrow \quad \frac{\Phi}{c^2} \sim \frac{v^2}{c^2}, \tag{224}$$

so the potential well of the richest clusters with velocity dispersion $\sim 1000\,{\rm km\,s^{-1}}$ is of order 10^{-5} deep. It is therefore no surprise to see this level of fluctuation on the CMB sky.

Finally, it is also possible with some effort to calculate the full spherical-harmonic spectrum from the 3D spatial spectrum. For a scale-invariant spectrum, the result is

$$C_\ell = \frac{6}{\ell(\ell+1)}\, C_2, \tag{225}$$

which is why the broad-band measure of the 'power per log ℓ' is defined as

$$\mathcal{T}^2(\ell) = \frac{\ell(\ell+1)}{2\pi}\, C_\ell. \tag{226}$$

Finally, a word about units. The temperature fluctuation $\Delta T/T$ is dimensionless, but anisotropy experiments generally measure ΔT directly, independent of the mean temperature. It is therefore common practice to quote $\mathcal{T}^2$ in units of $(\mu\mathrm{K})^2$.

CHARACTERISTIC SCALES We now want to look at the smaller-scale features of the CMB. The current data are contrasted with some CDM models in figure 21. The key feature that is picked out is the dominant peak at $\ell \simeq 220$, together with harmonics of this scale at higher ℓ. How can these features be understood?

The main point to appreciate is that the gravitational effects are the ones that dominate on large angular scales. This is easily seen by contrasting the temperature perturbations from the gravitational and adiabatic perturbations:

$$\frac{\delta T}{T} \simeq \frac{1}{3}\frac{\Phi}{c^2} \quad \text{(gravity)}; \qquad \frac{\delta T}{T} \simeq \frac{1}{3}\frac{\delta\rho}{\rho} \quad \text{(adiabatic)}. \tag{227}$$

Poisson's equation says $\nabla^2\Phi = -k^2\Phi = 4\pi G\rho(\delta\rho/\rho)$, so there is a critical (proper) wavenumber where these two effects are equal: $k_{\mathrm{crit}}^2 \sim G\rho/c^2$. The age of the universe is always $t \sim (G\rho)^{-1/2}$, so this says that

$$k_{\mathrm{crit}} \sim (ct)^{-1}. \tag{228}$$

In other words, perturbations with wavelengths above the horizon size at last scattering generate $\delta T/T$ via gravitational redshift, but on smaller scales it is adiabatic perturbations that matter.

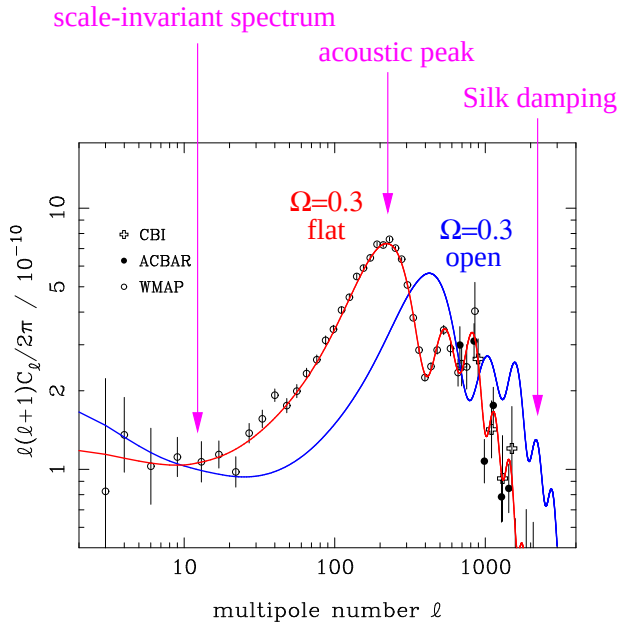

Figure 21. Angular power spectra $\mathcal{T}^2(\ell) = \ell(\ell+1)C_\ell/2\pi$ for the CMB, plotted against angular wavenumber ℓ in radians^{-1}. For references to the experimental data, see Spergel et al. (2006). The two lines show model predictions for adiabatic scale-invariant CDM fluctuations, calculated using the CMBFAST package (Seljak & Zaldarriaga 1996). These have $(n, \Omega_m, \Omega_b, h) = (1, 0.3, 0.05, 0.65)$ and have respectively $\Omega_v = 1-\Omega_m$ ('flat') and $\Omega_v = 0$ ('open'). The main effect is that open models shift the peaks to the right, as discussed in the text.

The significance of the main **acoustic peak** is therefore that it picks out the (sound) horizon at last scattering. The redshift of last scattering is almost independent of cosmological parameters at $z_{\rm LS} \simeq 1100$, as we have seen. If we assume that the universe is matter dominated at last scattering, the horizon size is

$$D_{\rm H}^{\rm LS} = 184\,(\Omega_m h^2)^{-1/2}{\rm Mpc}. \qquad (229)$$

The angle this subtends is given by dividing by the current size of the horizon (strictly, the comoving angular-diameter distance to $z_{\rm LS}$). Again, for a matter-dominated model with $\Lambda = 0$, this is

$$D_{\rm H} = 6000\,\Omega_m^{-1}\,h^{-1}{\rm Mpc} \quad \Rightarrow \quad \theta_{\rm H} = D_{\rm H}^{\rm LS}/D_{\rm H} = 1.8\,\Omega_m^{0.5}\,{\rm degrees}. \tag{230}$$

Figure 21 shows that heavily open universes thus yield a main CMB peak at scales much smaller than the observed $\ell \simeq 220$, and these can be ruled out. Indeed, open models were disfavoured for this reason long before any useful data existed near the peak, simply because of strict upper limits at $\ell \simeq 1500$ (Bond & Efstathiou 1984). In contrast, a flat vacuum-dominated universe has $D_{\rm H} \simeq 6000\,\Omega_m^{-0.4}\,h^{-1}$Mpc, so the peak is predicted at $\ell \simeq 2\pi/(184/6000) \simeq 200$ almost independent of parameters. These expression lie behind the common statement that the CMB data require a flat universe – although it turns out that large degrees of spatial curvature *and* Λ can also match the CMB well.

The second dominant scale is imposed by the fact that the last-scattering surface is fuzzy – with a width in redshift of about $\delta z = 80$. This imposes a radial smearing over scales $\sigma_r = 7(\Omega_m h^2)^{-1/2}$ Mpc. This subtends an angle

$$\theta_r \simeq 4\,{\rm arcmin}, \tag{231}$$

for flat models. This is partly responsible for the fall in power at high ℓ (Silk damping also contributes). Finally, a characteristic scale in many density power spectra is set by the horizon at $z_{\rm eq}$. This is $16(\Omega h^2)^{-1}$ Mpc and subtends a similar angle to θ_r.

REIONIZATION As mentioned previously, it is plausible that energy output from young stars and AGN at high redshift can reionize the intergalactic medium. Certainly, we know empirically from the lack of Gunn–Peterson neutral hydrogen absorption in quasars that such **reheating** did occur, and at a redshift in excess of 6. The consequences for the microwave background of this reionization depend on the Thomson-scattering optical depth:

$$\tau = \int \sigma_{\rm T}\,n_e\,d\ell_{\rm prop} = \int \sigma_{\rm T}\,n_e\,\frac{c}{H_0}\,\frac{dz}{(1+z)\sqrt{1-\Omega_m+\Omega_m(1+z)^3}} \tag{232}$$

(for a flat model). If we re-express the electron number density in terms of the baryon density parameter as

$$n_e = \Omega_b\,\frac{3H_0^2}{8\pi G\mu m_p}\,(1+z)^3, \tag{233}$$

where the parameter μ is approximately 1.143 for a gas of 25% helium by mass, and do the integral over redshift, we get

$$\tau = 0.04h\,\frac{\Omega_b}{\Omega_m}\left[\sqrt{1+\Omega_m z(3+3z+z^2)}-1\right] \simeq 0.04h\,\frac{\Omega_b}{\Omega_m^{1/2}} z^{3/2}. \tag{234}$$

Predictions from CDM galaxy formation models tend to predict a reheating redshift between 10 and 15, thus τ between 0.1 and 0.2 for standard parameters. The main effect of this scattering is to damp the CMB fluctuations by a factor $\exp(-\tau)$, but this does not apply to the largest-scale angular fluctuations. To see this, think backwards: where could a set of photons scattered at z have come from? If they are scattered by an angle of order unity, they can be separated at the last-scattering surface by at most the distance from z to $z_{\rm LS}$ – which is almost exactly the horizon size at z. The geometry of this situation is shown in figure 22, which illustrates that the critical angle is the angle subtended today by the horizon size at the redshift of the secondary scattering; for a flat model, this is approximately $z^{-1/2}$ radians, so modes with $\ell < z^{1/2}$ are unaffected. This turns out to be a critical factor in changing the apparent shape of the CMB power spectrum.

9 CMB anisotropies – II

Having given an outline of the physical mechanisms that contribute to the CMB anisotropies, we now examine how the CMB is used in conjunction with other probes to pin down the cosmological model.

The information we gain from the CMB is dominated by the main acoustic peak at $\ell = 220$, and it is interesting to ask what this tells us. We have argued that the location of this feature marks the angle subtended by the acoustic horizon at last scattering, which has been given as $D_{\rm H}^{\rm LS} = 184\,(\Omega_m h^2)^{-1/2}$Mpc. Using the current size of the horizon, the angle subtended in a flat model is

$$D_{\rm H} = 6000\,\Omega_m^{-0.4}\,h^{-1}{\rm Mpc} \quad \Rightarrow \quad \theta_{\rm H} = D_{\rm H}^{\rm LS}/D_{\rm H} \propto \Omega_m^{-0.1}, \tag{235}$$

so there is very little dependence of peak location on cosmological parameters. This contrast between little dependence on density for flat models and a large density dependence for models with no cosmological constant is often used to argue that the CMB proves flatness; but this ignores the case where both curvature and Λ are important, and independent constraints on the density are needed before this possibility can be ruled out.

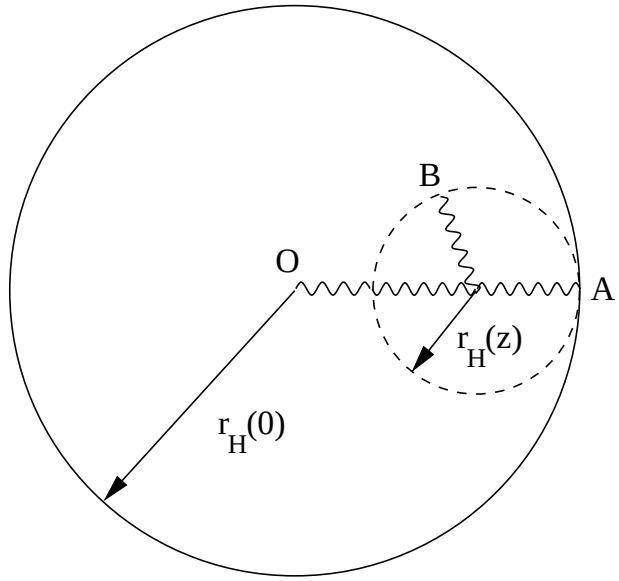

Figure 22. Illustrating the effect of secondary scattering of CMB photons owing to reionization at redshift z. An observer O sees radiation apparently from A, which in fact originates at point B somewhere on the last scattering shell around z (shown dotted). To a good approximation, the distance to last scattering is the horizon distance; thus CMB fluctuations of angular scale beyond $r_{\rm H}(z)/r_{\rm H}(0)$ cannot be erased by this secondary scattering.

However, this argument is incomplete in detail because the earlier expression for $D_{\rm H}(z_{\rm LS})$ assumes that the universe is completely matter dominated at last scattering, and this is not perfectly true. The comoving sound horizon size at last scattering is defined by

$$D_{\rm S}(z_{\rm LS}) \equiv \frac{1}{H_0\Omega_m^{1/2}} \int_0^{a_{\rm LS}} \frac{c_{\rm S}}{(a+a_{\rm eq})^{1/2}}\, da \tag{236}$$

where vacuum energy is neglected at these high redshifts; the expansion factor $a \equiv (1+z)^{-1}$ and $a_{\rm LS}, a_{\rm eq}$ are the values at last scattering and matter-radiation equality respectively. In practice, $z_{\rm LS} \simeq 1100$ independent of the matter and baryon densities, and $c_{\rm S}$ is fixed by Ω_b. Thus the main

effect is that $a_{\rm eq}$ depends on Ω_m. Dividing by $D_{\rm H}(z=0)$ therefore gives the angle subtended today by the light horizon as

$$\theta_{\rm H} \simeq \frac{\Omega_m^{-0.1}}{\sqrt{1+z_{\rm LS}}} \left[\sqrt{1+\frac{a_{\rm eq}}{a_{\rm LS}}} - \sqrt{\frac{a_{\rm eq}}{a_{\rm LS}}} \right], \tag{237}$$

where $z_{\rm LS} = 1100$ and $a_{\rm eq} = (23900\,\omega_m)^{-1}$. This remarkably simple result captures most of the parameter dependence of CMB peak locations within flat ΛCDM models. Differentiating this equation near a fiducial $\omega_m = 0.13$ gives

$$\left.\frac{\partial \ln \theta_{\rm H}}{\partial \ln \Omega_m}\right|_{\omega_m} = -0.1; \qquad \left.\frac{\partial \ln \theta_{\rm H}}{\partial \ln \omega_m}\right|_{\Omega_m} = \tfrac{1}{2}\left(1+\frac{a_{\rm LS}}{a_{\rm eq}}\right)^{-1/2} = +0.25, \tag{238}$$

Thus for moderate variations from a 'fiducial' flat model, the CMB peak multipole number scales approximately as $\ell_{\rm peak} \propto \Omega_m^{-0.15} h^{-0.5}$, i.e. the condition for constant CMB peak location is well approximated as

$$\Omega_m h^{3.3} = \text{constant}, \tag{239}$$

provided spatial curvature vanishes. It is now clear how LSS data combines with the CMB: $\Omega_m h$ is the main combination probed by the matter power spectrum so this approximate degeneracy is strongly broken using the combined data.

9.1 Degeneracy breaking with detailed CMB data

Although the main horizon-scale peak in the power spectrum dominates the appearance of the CMB, giving degenerate information about cosmological parameters, the fine detail of the pattern is also important. As the quality of the CMB measurements improve, more information can be extracted, and the parameter degeneracies are increasingly broken by the CMB alone. Regarding the structure around the peak, two physical effects are important in giving this extra information:

(1) **Early ISW**. We have seen that the transition from radiation domination to matter domination occurs only just before last scattering. Although we have proved that potential fluctuations Φ stay constant during the radiation and matter eras (while vacuum and curvature

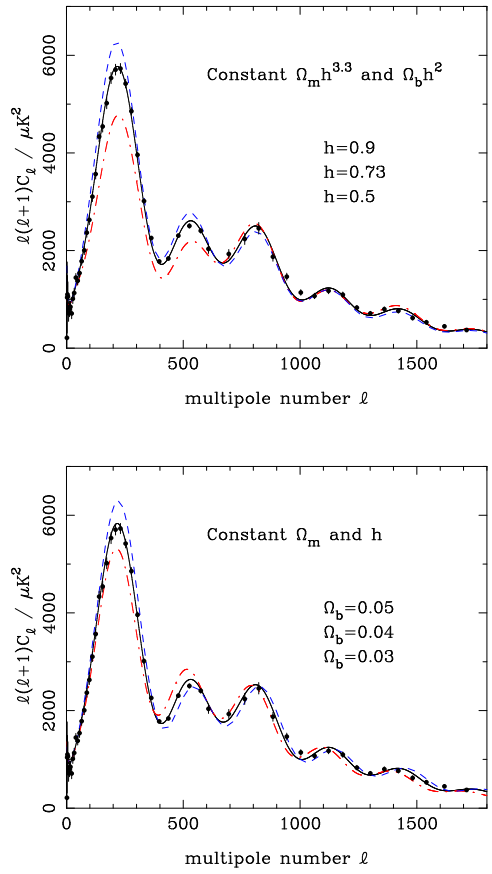

Figure 23. The location of the principal peak in the CMB power spectrum is largely determined by the combination $\Omega_m h^{3.3}$, representing the scaling of the angular size of the horizon at last scattering. The two other main characteristics are the rise to the peak, and the fall to the second and subsequent maxima. The former (the height of the peak above the Sachs-Wolfe plateau) is influenced by the early-time ISW effect. This is illustrated in the first panel, where we fix $\Omega_b h^2$ and hence the sound speed. For fixed peak location, higher h gives lower matter density, and hence a higher peak from the early-time ISW effect (all models are normalized at $\ell = 20$). The second panel shows the influence of varying the baryon density at constant matter density, where we see that a higher baryon fraction increases the amplitude of the acoustic oscillations. Thus, if we assume flatness, the three observables of the peak location and drops to either side suffice to determine Ω_m, Ω_b and h.

are negligible), this is not true at the junction, and there is a small change in Φ during the radiation–matter transition (by a factor 9/10: see chapter 7 of Mukhanov's book). This introduces an additional ISW effect, which boosts the amplitude of the peak, especially for models with low $\Omega_m h^2$, which brings $z_{\rm eq}$ right down to $z_{\rm LS}$ (see the first panel of figure 23).

(2) **Baryon loading**. If we keep the overall matter density fixed but alter the baryon fraction, the sound speed at last scattering changes. This has the effect of making a change in the amplitude of the acoustic oscillations beyond the first peak: the drop to the second peak is more pronounced if the baryon fraction is high (see the second panel of figure 23).

Overall, the kind of precision data now delivered by WMAP and SPT (shown in Figure 24) allows these effects to be measured, and the degeneracy between Ω_m, Ω_b and h broken without external data.

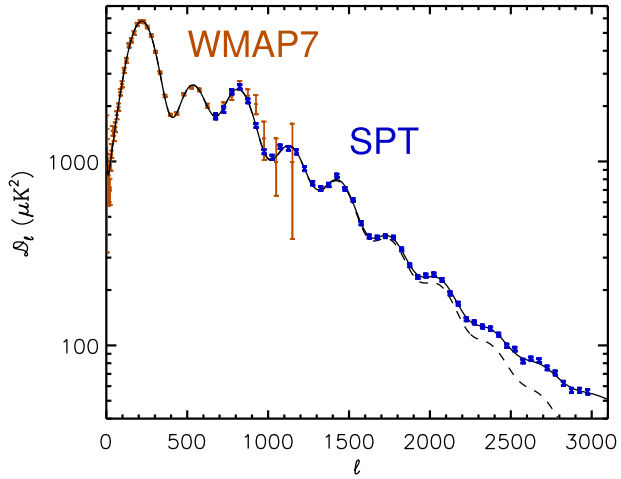

Figure 24. The CMB power spectrum as measured on large scales by WMAP and on smaller scales by the South Pole Telescope. Here, $\mathcal{D}_\ell/T_0^2$ means the same as our $\mathcal{T}^2$, i.e. $\ell(\ell+1)C_\ell/2\pi$. Amazingly, as many as nine acoustic peaks are visible, without any real need for a guiding model. Note that for $\ell \gtrsim 2000$ the primordial CMB signal gains a background contribution from point sources (dusty starburst galaxies and the SZ effect from the gas in clusters and groups of galaxies)

9.2 Tensor modes

All of our discussion to date applies to models in which scalar modes dominate. But we know that gravity-wave metric perturbations in the form of a traceless symmetric tensor $h^{\mu\nu}$ are also possible, and that inflation predicts that a background of such waves is generated, with amplitude

$$h_{\rm rms} \sim H_{\rm inflation}/m_{\rm P}. \tag{240}$$

These tensor metric distortions are observable via the large-scale CMB anisotropies, where the tensor modes produce a spectrum with the same scale dependence as the Sachs–Wolfe gravitational redshift from scalar metric perturbations. In the scalar case, we have $\delta T/T \sim \phi/3c^2$, *i.e.* of order the Newtonian metric perturbation; similarly, the tensor effect is

$$\left(\frac{\delta T}{T}\right)_{\rm GW} \sim h_{\rm rms}. \tag{241}$$

Could the large-scale CMB anisotropies actually be tensor modes? This would be tremendously exciting, since it would be a direct window into the inflationary era. The Hubble parameter in inflation is $H^2 = 8\pi G\rho/3 \sim V(\phi)/m_{\rm P}^2$, so that

$$\left(\frac{\delta T}{T}\right)_{\rm GW} \sim h_{\rm rms} \sim H/m_{\rm P} \sim V^{1/2}/m_{\rm P}^2. \tag{242}$$

A measurement of the tensor modes in the CMB would therefore tell us directly the energy scale of inflation: $E_{\rm inflation} \sim V^{1/4}$. This is more direct than the scalar signature, which was

$$\delta_{\rm H} \sim \frac{V^{1/2}}{m_{\rm P}^2\,\epsilon^{1/2}}, \tag{243}$$

where ϵ is the principal slow-roll parameter (dimensionless version of the gradient-squared of the potential).

From these relations, we can see that the **tensor-to-scalar ratio** in the large-scale CMB power spectra just depends on ϵ:

$$r \equiv \mathcal{T}_{\rm T}^2/\mathcal{T}_{\rm S}^2 = 16\epsilon \tag{244}$$

(putting in the factor of 16 from an exact analysis). We have argued that ϵ cannot be too small if inflation is to end, so significant tensor contributions to the CMB anisotropy are a clear prediction. As a concrete example, consider power-law inflation with $a \propto t^p$, where we showed that $\epsilon = \eta/2 = 1/p$. In this case,

$$r = 8(1 - n_s), \tag{245}$$

so the larger the tilt, the more important the tensors. We will see below that there is fairly strong evidence for a non-zero tilt with $n_s \simeq 0.96$, so the simplest expectation would be a tensor contribution of $r \simeq 0.3$. Of course, this only applies for a large-field model like power-law inflation; it is quite possible to have a small-field model with $|\eta| \gg |\epsilon|$, in which case there can be tilt without tensors.

An order unity tensor contribution would imply metric distortions at the level of 10^{-5}, which might sound easy to detect directly. The reason this is not so is that the small-scale tensor fluctuations are reduced today: their energy density (which is $\propto h^2$) redshifts away as a^{-4} once they enter the horizon. This redshifting produces a break in the spectrum of waves, reminiscent of the matter transfer spectrum, so that the tensor contribution to the CMB declines for $\ell \gtrsim 100$. This redshifting means that the present-day metric distortions are more like 10^{-27} on relevant scales (kHz gravity waves) than the canonical 10^{-5}. Even so, direct detection of these relic gravity waves can be contemplated, but this will be challenging in the extreme. At the current rate of progress in technology, the necessary sensitivity may be achieved around 2050; but the signal may be higher than in the simple models, so one should be open to the possibility of detecting this ultimate probe of the early universe.

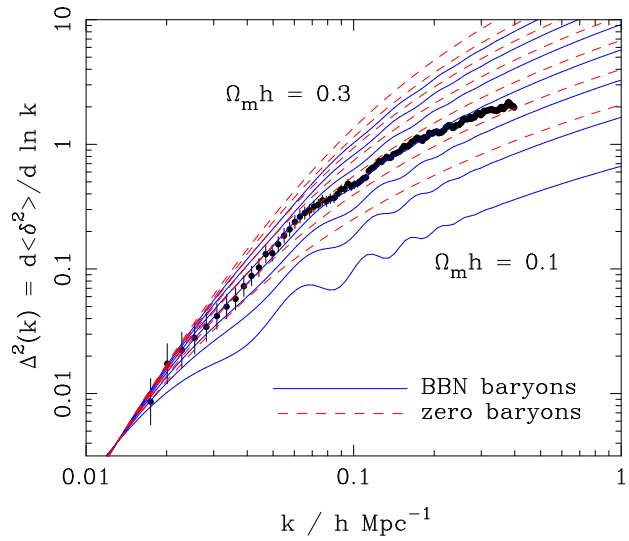

Figure 25. The galaxy power spectrum from the 2dF Galaxy Redshift Survey, shown as the contribution to the fractional density variance per $\ln k$ against wavenumber (spatial wavelength is $\lambda = 2\pi/k$). The data are contrasted with CDM models having scale-invariant primordial fluctuations ($n_s = 1$) and $\Omega_m h = 0.1$, 0.15, 0.2, 0.25, 0.3. The dotted lines show pure CDM models, whereas the solid lines show the effect of baryons at the nucleosynthesis level (assuming $\Omega_b = 0.04$ and $h = 0.7$).

10 Combined constraints on the cosmological model

We have shown that, given perfect data, the CMB anisotropy power spectrum alone is able to determine the main cosmological parameters. But current data are still some way from being ideal – and this capability weakens when we expand the model to include ingredients that are as yet undetected, but which have a reasonable theoretical motivation. However, additional information from large-scale structure cures these problems effectively.

THE GALAXY POWER SPECTRUM A key aim in observational cosmology has long been to use the expected feature at the $z_{\rm eq}$ horizon scale to measure the density of the universe. Data on galaxy clustering is now sufficiently good that this can be done quite accurately. The measured

power spectrum from the 2dF Galaxy Redshift Survey is contrasted with CDM models (for which $|\delta_k|^2 \propto k^n T_k^2$) in figure 25. The curvature of the spectrum is clearly measured, leading to the constraint

$$\Omega_m h = 0.168 \pm 0.016. \tag{246}$$

For $h = 0.7 \pm 10\%$, as indicated by absolute external measurements, this gives $\Omega_m = 0.24 \pm 0.03$. The 2dFGRS results also give a detection of the expected baryon features, leading to a measurement of the baryon fraction:

$$\Omega_b/\Omega_m = 0.185 \pm 0.046 \tag{247}$$

(see Cole et al. astro-ph/0501174). Although this is not as accurate a measurement of the baryon fraction as we obtain from the CMB, it is a more direct piece of evidence that collisionless dark matter is needed; with only baryonic matter, the galaxy power spectrum would be expected to display the same order-unity oscillations that we see in the CMB power spectrum.

In order to reach these conclusions, however, it is necessary to make an assumption about the primordial spectrum, which was taken to be scale-invariant with $n = 1$. Values $n < 1$ would correspond to a larger inferred density, and LSS data cannot break this degeneracy with tilt. The best way to constrain n is to combine with data on CMB anisotropies; as we have discussed, these probe larger scales and give a robust measure of n, which indeed turns out to be very close to unity. Similarly, LSS data do not make any statement about the curvature of the universe. Again, this can be measured from the CMB given the use of LSS data to limit possible combinations of matter content and h, and so break the geometrical degeneracy.

COMBINED CONSTRAINTS FROM CMB+LSS Following the superb WMAP results (first announced in Spergel et al. 2006; astro-ph/0603449, with arxiv:1212.5226 giving the final results), the statistics of the temperature field are measured sufficiently precisely that many of the parameter degeneracies we have worried about are broken, at least weakly. This comes partly from the polarization measurements, and also via the ISW effect. In general, what we have to do is explore a multidimensional parameter space, which can easily be 11-dimensional, as shown in Table 1.

This is frequently reduced to 7 free parameters (ignoring tensors and the neutrino mass fraction, and assuming $w = -1$): a scalar CDM universe. In this case, the interesting parameter to focus on is the curvature. The **likelihood** of the data given the model parameters is regarded as a probability density for the parameters, and we **marginalize** by integrating this distribution

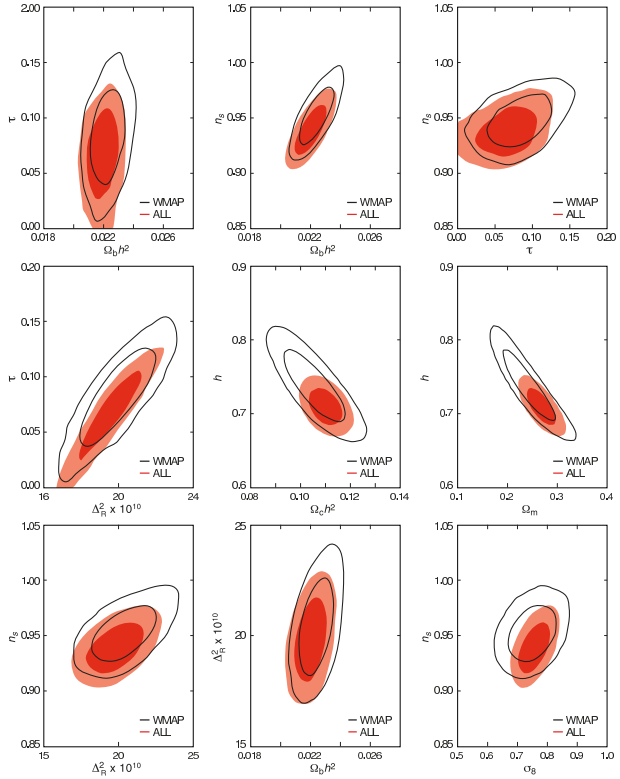

Figure 26. The basic WMAP3 confidence contours on the key cosmological parameters for flat scalar-only models (from Spergel et al. 2006).

over the uninteresting parameters. This leaves a probability distribution for the curvature, which is sharply peaked about zero:

$$\Omega_k = -0.0059 \pm 0.0040. \tag{248}$$

This is normally taken as sufficient empirical justification (in addition to inflationary prejudice) to assuming exact flatness when trying to set constraints on more exotic ingredients (tensors; $w \neq -1$).

Table 1. Cosmological parameters.

Parameter	Meaning
ω_{dm}	Physical density of dark matter
ω_b	Physical density of baryons
ω_v	Physical density of vacuum
w	Equation of state of vacuum
ω_k	Curvature 'density'
n_s	Scalar spectral index
r	Tensor-to-scalar ratio
n_t	Tensor spectral index
σ_8	Spectrum normalization
τ	Optical depth from reionization
f_ν	Neutrino mass fraction

Table 2. Constraints on the basic 6-parameter model (flat; no tensors) from WMAP+SPT in combination with large-scale structure (from Lowell et al. 2012 arxiv:1210.7231).

Parameter	WMAP + SPT + LSS
σ_8	0.827 ± 0.015
τ	0.076 ± 0.012
n_s	0.952 ± 0.008
ω_b	0.0220 ± 0.003
ω_m	0.117 ± 0.002
h	0.691 ± 0.009
$\Rightarrow \Omega_m$	0.291 ± 0.011

But so far these are not required, and there is a very well specified 6-parameter standard model, as shown in figure 27 and Table 2.

The impressive thing here is the specification of a relatively low optical depth due to reionization, leading to evidence in favour of $n_s < 1$; exact scale-invariance would need a larger optical depth, and thus stronger large-scale polarization than observed. The detection of tilt (a roughly 6σ rejection of the $n_s = 1$ model) has to be considered an impressive success for inflation, given that such deviations from scale invariance were a clear prediction. So should we consider

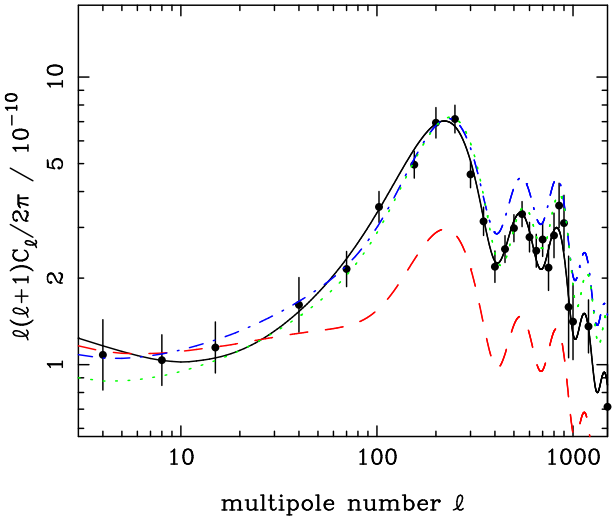

Figure 27. The tensor degeneracy. Adding a large tensor component to an $n_s = 1$ scalar model (solid line) greatly lowers the peak (dashed line), once COBE normalization is imposed. Tilting to $n_s = 1.3$ cures this (dot-dashed line), but the 2nd and subsequent harmonics are too high. Raising the baryon density by a factor 1.5 (dotted line) leaves us approximately back where we started.

inflation to be proved? Perhaps not yet, but one is certainly encouraged to look more closely at the tensor signal.

LIMITS ON THE TENSOR FRACTION The possibility of a large tensor component yields additional degeneracies, as shown in figure 27. An $n_s = 1$ model with a large tensor component can be made to resemble a zero-tensor model with large blue tilt ($n_s > 1$) and high baryon content. this is only weakly broken with current data, as shown in figure 27. This illustrates that we cannot be sure about the 'detection' of tilt: the data can be well matched with $n_s = 1$, but then a substantial tensor fraction is needed.

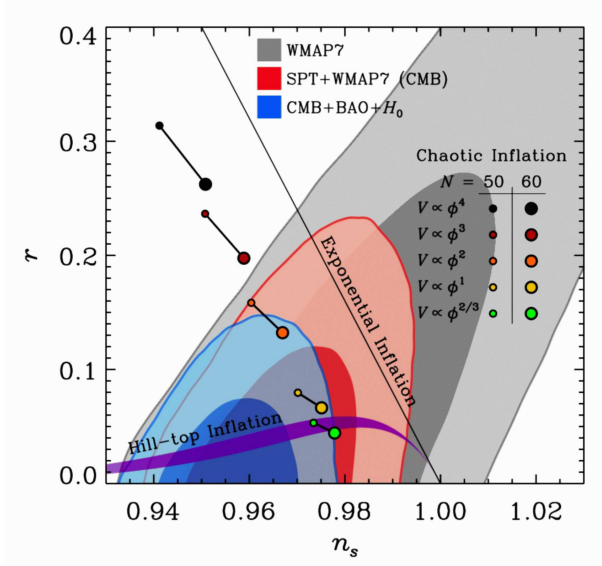

Figure 28. The marginalized confidence contours on the inflationary $r - n$ plane (from Lowell et al. 2012 arxiv:1210.7231), using CMB data from WMAP+SPT plus constraints from large-scale structure. Not only is the simple scale-invariant zero-tensor model hugely disfavoured, but all simple power-law inflationary models are in trouble (prior to 2012, ϕ^4 inflation was ruled out, but ϕ^2 inflation matched the data well).

So far, the tensor contribution to the large-angle anisotropy power spectrum is limited to a fraction $r \lesssim 0.15$. To do much better, we need to detect the characteristic 'B-mode' polarization. The B modes are excited only by tensors, so all future large-scale polarization experiments will be searching for this signature; it will not be easy, even if the foregrounds are gentle. Planck will only be able to detect tensors if $r \gtrsim 0.1$, although the ultimate limit from cosmic variance is more like $r \simeq 10^{-5}$. This sounds like there is a lot of future scope, but it should be recalled that the energy scale of inflation scales as the tensor $C_\ell^{1/4}$. Therefore, we will need a degree of luck with the energy scale if there is to be a detection.

11 The puzzle of dark energy

11.1 Cosmological effects of the vacuum

One of the most radical conclusions of recent cosmological research has been the necessity for a non-zero vacuum density. This was detected on the assumption that Einstein's **cosmological constant**, Λ, might contribute to the energy budget of the universe. But if this ingredient is a reality, it raises many questions about the physical origin of the vacuum energy; as we will see, a variety of models may lead to something similar in effect to Λ, and the general term **dark energy** is used to describe these.

The properties of dark energy can be probed by the same means that we used to deduce its existence in the first place: via its effect on the expansion history of the universe. The vacuum density is included in the Friedmann equation, independent of the equation of state

$$\dot{R}^2 - \frac{8\pi G}{3}\,\rho\, R^2 = -kc^2. \tag{249}$$

At the outset, then we should be very clear that the deduced existence of dark energy depends on the correctness of the Friedmann equation, and this is not guaranteed. Possibly we have the wrong theory of gravity, and we have to replace the Friedmann equation by something else. Alternative models do exist, particularly in the context of extra dimensions, and these must be borne in mind. Nevertheless, as a practical framework, it makes sense to stick with the Friedmann equation and see if we can get consistent results. If this programme fails, we may be led in the direction of more radical change.

To insert vacuum energy into the Friedmann equation, we need the equation of state

$$w \equiv p/\rho\, c^2 \tag{250}$$

If this is constant, adiabatic expansion of the vacuum gives

$$\frac{8\pi G\rho}{3H_0^2} = \Omega_v a^{-3(w+1)}. \tag{251}$$

More generally, we can allow w to vary; in this case, we should regard $-3(w+1)$ as $d\ln\rho/d\ln a$, so that

$$\frac{8\pi G\rho}{3H_0^2} = \Omega_v \exp\left(\int -3(w(a)+1)\; d\ln a\right). \tag{252}$$

In general, we therefore need

$$H^2(a) = H_0^2 \left[\Omega_v e^{\int -3(w(a)+1)\, d\ln a} + \Omega_m a^{-3} + \Omega_r a^{-4} - (\Omega - 1) a^{-2} \right]. \tag{253}$$

Some complete dynamical model is needed to calculate $w(a)$. Given the lack of a unique model, a common empirical parameterization is

$$w(a) = w_0 + w_a(1 - a). \tag{254}$$

Frequently it is sufficient to stick with constant w; most experiments are sensitive to w at a particular redshift of order unity, and w at this redshift can be estimated with little dependence on whether we allow dw/dz to be non-zero.

If w is negative at all, this leads to models that become progressively more vacuum-dominated as time goes by. When this process is complete, the scale factor should vary as a power of time. The case $w < -1$ is particularly interesting, sometimes known as **phantom dark energy**. Here the vacuum energy density will eventually diverge, which has two consequences: this singularity happens in a finite time, rather than asymptotically; as it does so, vacuum repulsion will overcome the normal electromagnetic binding force of matter, so that all objects will be torn apart in the **big rip**. Integrating the Friedmann equation forward, ignoring the current matter density, the time to this event is

$$t_{\rm rip} - t_0 \simeq \frac{2}{3} H_0^{-1} |1 + w|^{-1} (1 - \Omega_m)^{-1/2}. \tag{255}$$

OBSERVABLE EFFECTS OF THE VACUUM The comoving distance-redshift relation is one of the chief diagnostics of w. The general definition is

$$D \equiv R_0 r = \int_0^z \frac{c}{H(z)}\, dz. \tag{256}$$

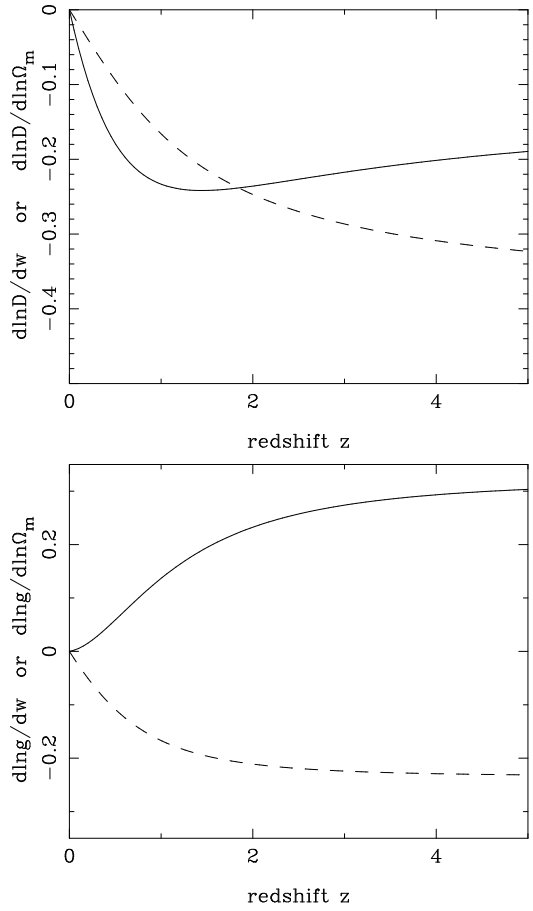

Figure 29. Perturbation around $\Omega_m = 0.25$ of distance-redshift and growth-redshift relations. Solid line shows the effect of increase in w; dashed line the effect of increase in Ω_m

Perturbing this about a fiducial $\Omega_m = 0.25$ $w = -1$ model shows a **sensitivity multiplier** of about 5 – i.e. a measurement of w to 10% requires D to 2%. Also, there is a near-perfect degeneracy with Ω_m, so this parameter must be known very well before the effect of varying w becomes detectable.

The other main diagnostic of w is its effect on the growth of density perturbations. These are also sensitive to the vacuum, as may be seen from the growth equation:

$$\ddot{\delta} + 2\frac{\dot{a}}{a}\dot{\delta} = 4\pi G\rho_0\delta. \tag{257}$$

The vacuum energy manifests itself in the factor of H in the 'Hubble drag' term $2(\dot{a}/a)\dot{\delta}$. For flat models with $w = -1$, we have seen that the growing mode for density perturbations is approximately as $g(a) \propto a\Omega(a)^{0.23}$. If w is made more negative, this makes the growth law closer to the Einstein–de Sitter $g(a) \propto a$ (for very large negative w, the vacuum was unimportant until very recently). Therefore, increasing w (making it less negative) has an effect in the same sense as *decreasing* Ω_m. As shown in figure 29, the degeneracy between variations in Ω_m and w thus has the opposite sign to the degeneracy in $D(z)$. Ideally, one would therefore try to observe both effects.

11.2 Observing the properties of dark energy

What are the best ways to measure w? We have seen that the two main signatures are alterations to the distance-redshift relation and the perturbation growth rate. It is possible to use both of these effects in the framework we have been discussing: observing the perturbed universe in both the CMB and large-scale structure.

In the CMB, the main observable is the angle subtended by the horizon at last scattering

$$\theta_{\rm H} = D(z_{\rm LS})/D(z=0). \tag{258}$$

This has the approximate scaling with cosmological parameters (for a flat universe)

$$\theta_{\rm H} \propto (\Omega_m h^{3.3})^{0.15}\Omega_m^{\alpha-0.4}; \quad \alpha(w) = -2w/(1-3.8w). \tag{259}$$

The latter term comes from a convenient approximation for the current horizon size:

$$D_0 = 2\frac{c}{H_0}\Omega_m^{-\alpha(w)}. \tag{260}$$

At first sight, this looks bad: the single observable of the horizon angle depends on three parameters (four, if we permit curvature). Thus, even in a flat model, we can only pin down w if we know both Ω_m and h.

However, if we have more detail on the CMB than just the main peak location, then we have seen that the $\Omega_m - h$ degeneracy is weakly broken, and that this situation improves with information from large-scale structure, which yields an estimate of $\Omega_m h$. In effect, we have two constraints on the $\Omega_m - h$ plane that are consistent if $w = -1$, but this is not the case for other values of w. In this way, the current combined constraints from CMB plus alternative probes (LSS and the Supernova Hubble diagram) yield an impressive accuracy:

$$w = -0.926^{+0.054}_{-0.053}, \tag{261}$$

for a spatially flat model – see Spergel et al. (2006). The confidence contours are plotted in detail in figure 30, and it is clear that so far there is very good consistency with a simple cosmological constant. But as we will see, plenty of models exist in which some deviation is predicted. The next goal of the global cosmology community is therefore to push the errors on w down substantially – to about 1%. There is no guarantee that this will yield any signal, but certainly it will cut down the range of viable models for dark energy.

One of the future tools for improving the accuracy in w will be large-scale structure. We have seen how this helps pin down the parameter degeneracies inherent in a CMB-only analysis, but it also contains unique information from the acoustic horizon. Earlier, we approximated this without considering how the speed of sound would depend on the baryon density; a good approximation to the exact result is

$$D_a \simeq 60\,(\Omega_m h^2)^{-0.25}(\Omega_b h^2)^{-0.08}\ \mathrm{Mpc}. \tag{262}$$

This forms a standard measuring rod, as seen in the 'baryon wiggles' in the galaxy power spectrum. In future galaxy surveys, the measurement of this signature as a function of redshift will be a further useful geometrical probe.

11.3 Quintessence

The simplest physical model for dynamical vacuum energy is a scalar field. We know from inflationary models that this can yield something close in properties to a cosmological constant, and so we can immediately borrow the whole apparatus for modelling vacuum energy at late times. This idea of scalar fields as a dynamical substitute for Λ was first explored by Ratra & Peebles (1988). Of course, this means yet another scalar field that is introduced without much or any motivation from

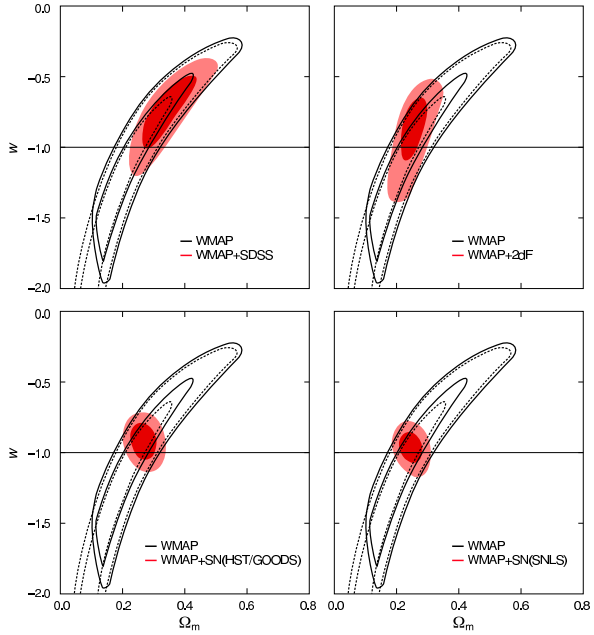

Figure 30. The marginalized WMAP3 confidence contours on the plane of dark-energy equation of state (w) vs Ω_m (from Spergel et al. 2006). A flat universe is assumed, although this is not critical to the conclusions.

fundamental physics. This hypothetical field is given the fanciful name 'quintessence', implying a new addition to the ancient Greek list of elements (fire, air, earth, water).

The Lagrangian density for a scalar field is as usual of the form of a kinetic minus a potential term:

$$\mathcal{L} = \tfrac{1}{2}\partial_\mu\phi\,\partial^\mu\phi - V(\phi). \tag{263}$$

In familiar examples of quantum fields, the potential would be

$$V(\phi) = \tfrac{1}{2}\,m^2\,\phi^2, \tag{264}$$

where m is the mass of the field. However, as before we keep the potential function general at this stage.

Suppose the Lagrangian has no explicit dependence on spacetime (*i.e.* it depends on x^μ only implicitly through the fields and their 4-derivatives). Noether's theorem then gives the energy–momentum tensor for the field as

$$T^{\mu\nu} = \partial^\mu\phi\partial^\nu\phi - g^{\mu\nu}\mathcal{L}. \tag{265}$$

From this, we can read off the energy density and pressure:

$$\begin{aligned} \rho &= \tfrac{1}{2}\dot{\phi}^2 + V(\phi) + \tfrac{1}{2}(\nabla\phi)^2 \\ p &= \tfrac{1}{2}\dot{\phi}^2 - V(\phi) - \tfrac{1}{6}(\nabla\phi)^2. \end{aligned} \tag{266}$$

If the field is constant both spatially and temporally, the equation of state is then $p = -\rho$, as required if the scalar field is to act as a cosmological constant; note that derivatives of the field spoil this identification.

For a homogeneous field, we have the equation of motion

$$\ddot{\phi} + 3H\dot{\phi} + dV/d\phi = 0, \tag{267}$$

which is most easily derived via energy conservation:

$$\frac{d\ln\rho}{d\ln a} = -3(1+w) = -3\dot{\phi}^2/(\dot{\phi}^2/2 + V), \tag{268}$$

following which the relations $H = d\ln a/dt$ and $\dot{V} = \dot{\phi}V'$ can be used to change variables to t, and the damped oscillator equation for ϕ follows. The solution of the equation of motion becomes tractable if we make the **slow-rolling approximation** that $|\ddot{\phi}|$ is negligible in comparison with $|3H\dot{\phi}|$ and $|dV/d\phi|$, so that

$$3H\dot{\phi} = -dV/d\phi. \tag{269}$$

From this, we know that a sufficiently flat potential can provide a dynamical vacuum that is arbitrarily close to a cosmological constant in its equation of state. However, there are good reasons why we might want to imagine the slow-roll conditions being violated in the case of dark energy.

COSMIC COINCIDENCE AND QUINTESSENCE Accepting the reality of vacuum energy raises a difficult question. If the universe contains a constant vacuum density and normal matter with $\rho \propto a^{-3}$, there is a unique epoch at which these two contributions cross over, and we seem to be living near to that time. This coincidence calls for some explanation.

We already have one coincidence, in that we live relatively close in time to the era of matter-radiation equality ($z \sim 10^3$, as opposed to $z \sim 10^{28}$ for the GUT era). This is relatively simple to understand: structure formation cannot begin until after $z_{\rm eq}$, and so we would expect observers to appear before the universe has expanded much beyond this point. The vacuum coincidence problem could therefore be solved if the vacuum density was some dynamical entity that was triggered to become Λ-like by the change in expansion history at $z_{\rm eq}$. Zlatev, Wang & Steinhardt (1999) suggested how this might happen. We have seen that the density and pressure for a quintessence field will be

$$\begin{aligned} \rho_\phi &= \dot\phi^2/2 + V \\ p_\phi &= \dot\phi^2/2 - V. \end{aligned} \tag{270}$$

This gives us two extreme equations of state: (i) vacuum-dominated, with $V \gg \dot\phi^2/2$, so that $p = -\rho$; (ii) kinetic-dominated, with $V \ll \dot\phi^2/2$, so that $p = \rho$. In the first case, we know that ρ does not alter as the universe expands, so the vacuum rapidly tends to dominate over normal matter. In the second case, the equation of state is the unusual $\Gamma = 2$, so we get the rapid behaviour $\rho \propto a^{-6}$. If a quintessence-dominated universe starts off with a large kinetic term relative to the potential, it may seem that things should always evolve in the direction of being potential-dominated. However, this ignores the detailed dynamics of the situation: for a suitable choice of potential, it is possible to have a **tracker field**, in which the kinetic and potential terms remain in a constant proportion, so that we can have $\rho \propto a^{-\alpha}$, where α can be anything we choose.

Putting this condition in the equation of motion shows that the potential is required to be exponential in form. The Friedmann equation with $\rho \propto a^{-\alpha}$ requires $a \propto t^{2/\alpha}$, so we have $\rho \propto t^{-2}$ as usual. But now both V and $\dot\phi^2$ must scale in the same way as ρ, so that $\dot\phi \propto 1/t$. Both the $\ddot\phi$ and $3H\dot\phi$ terms are therefore proportional to V, so an exponential potential solves the equation of motion. More importantly, we can generalize to the case where the universe contains scalar field and ordinary matter. Suppose the latter obeys $\rho_m \propto a^{-\alpha}$; it is then possible to have the scalar-field density obeying the same $\rho \propto a^{-\alpha}$ law, provided

$$V(\phi) \propto \exp[-\lambda\phi/M], \tag{271}$$

where $M = m_{\rm P}/\sqrt{8\pi}$. The scalar-field density is $\rho_\phi = (\alpha/\lambda^2)\rho_{\rm total}$. The impressive thing about this solution is that the quintessence density stays a fixed fraction of the total, whatever the overall equation of state: it automatically scales as a^{-4} at early times, switching to a^{-3} after matter-radiation equality.

This is not quite what we need, but it shows how the effect of the overall equation of state can affect the rolling field. Because of the $3H\dot{\phi}$ term in the equation of motion, ϕ 'knows' whether or not the universe is matter dominated. This suggests that a more complicated potential than the exponential may allow the arrival of matter domination to trigger the desired Λ-like behaviour. Zlatev, Wang & Steinhardt suggested two potentials which might achieve this:

$$V(\phi) = M^{4+\beta}\phi^{-\beta} \quad \text{or} \quad \mathrm{V}(\phi) = \mathrm{M}^4[\exp(\mathrm{m_P}/\phi) - 1]. \tag{272}$$

They show that these can yield an evolution in $w(t)$ so that it switches from $w \simeq 1/3$ in the radiation era to $w \simeq -1$ today.

However, a degree of fine-tuning is still required, in that the trick only works for $M \sim 1$ meV, so there is no natural reason for tracking to cease at matter-radiation equality. The idea of tracker fields thus does not remove completely the puzzle concerning the level of present-day vacuum energy. But such models are at least testable: because the Λ-like behaviour only switches on quite recently, it is hard to complete the transition, and the prediction is of something around $w \simeq -0.8$ today. As we have seen, this can be firmly ruled out with current data. These ideas about the dynamical vacuum are therefore already interesting testable science.

11.4 Modifying gravity

An alternative point of view on dark energy, which is receiving increasing interest in the research literature is to suggest that dark energy may not be a genuine physical entity at all. All our current knowledge about it comes from the Friedmann equation:

$$H^2(a) = H_0^2 \left[\Omega_m a^{-3} + \Omega_r a^{-4} - (\Omega - 1)a^{-2} + \Omega_v\right]. \tag{273}$$

In other words, the expansion history of the universe cannot be satisfied without adding a constant to the rhs. But this could mean that the standard Friedmann equation was wrong all along and that the presence of the constant indicates the need for changes to the theory of gravity.

This possibility is frequently termed 'violation of general relativity', but one should be clear at the outset that this is a misnomer: general relativity means assuming the existence of a metric and writing physics equations in covariant form, most simply by using relativistic invariants. Einstein's field equations are the simplest set consistent with this requirement, but are easy to generalise. This is most easily seen by using the Lagrangian formalism and writing the Einstein-Hilbert action:

$$S \propto \int (R + 2\Lambda)\, \sqrt{-g}\, d^4x^\mu, \tag{274}$$

where R is the Ricci scalar. Einstein's field equations arise from requiring a stationary action, and it is now obvious how to generate a more complex theory: replace $R + \Lambda$ by some other scalar; a popular choice is $f(R)$. This substitution has to be done with care, however, since there exist stringent constraints on deviations from Einstein gravity in the Solar System. The value of R is proportional to the matter density, which is about 10^6 times larger in interplanetary space than on cosmological scales. Thus what is required empirically is $f(R) \simeq R$ when R is large, but $f(R) \to$ constant as $R \to 0$. One might legitimately ask whether it is plausible that nature should carefully make sure that modifications of gravity are locally undetectable in this way.

A further popular way in which gravity might have non-standard properties is if the universe has more than the normal 3+1 spacetime dimensions. This was first introduced in the **Kaluza-Klein** picture, in which our universe is a lower-dimensional hypersurface in a higher-dimensional system. Such models were first discussed in the 1920s, and the device chosen to hide the extra dimensions was that they were **compactified** and have the topology of a very small cylinder in the hidden direction(s). A more recent development has been the **brane world** model, in which the extra dimension is not assumed to be small. There is then a larger space, termed the **bulk**, which lies away from the (mem)brane on which our universe is located. If Einstein gravity applies to the

joint space of bulk and brane, and matter is confined only to the brane, then it has been shown that the apparent Friedmann equation on the Brane is of the form

$$H^2(a) \propto \rho^2 + \mathcal{C}/a^4. \tag{275}$$

This quadratic dependence on density is startling and inconsistent with nucleosynthesis, whatever the value of the **dark radiation** term parameterized by $\mathcal{C}$. More realistic brane models allow a bulk cosmological constant, so that the metric is **warped** and no longer of the form $d\tau^2 = g_{\mu\nu}dx^\mu dx^\nu - dw^2$. These generalized brane models are known as **Randall-Sundrum models**.

This may all seem pointless if the aim is simply to come up with an alternative model that gives an expansion history $a(t)$ that is just like the standard case with matter plus dark energy. But more recent work has emphasised that it is possible to tell the difference by looking at the growth of structure. Informally speaking, we are exploring the possibility that gravity may have a different strength on the 10-Gpc scale of the entire visible universe than it does on small scales. Here, 'small scales' can mean as large as the kpc scales of galaxies, since the central parts of these can be explained dynamically using standard gravity and no dark matter. We stress that the aim here is to dispose of dark energy, not dark matter: that is the subject of a more radical programme known as **MOND**, or Modified Newtonian Dynamics. There is then the possibility that the behaviour on the intermediate 10-Mpc scales of large-scale structure may be a diagnostic of modified gravity. An empirical parameterization has been developed to deal with this:

$$f_g \equiv \frac{d\ln\delta}{d\ln a} \simeq \Omega_m(a)^\gamma. \tag{276}$$

The standard model is well fitted by $\gamma \simeq 0.55$, but many of the modified models discussed above require values of γ that differ from this by of order 0.1. The parameter γ thus forms one natural target for observers, to be added to w as an empirical description of fundamental cosmology. To complete the set, we note that gravitational lensing adds a specific degree of freedom in that it is able to probe the sum of the two metric potentials, $\Psi + \Phi$:

$$\eta = \Phi/\Psi. \tag{277}$$

A large number of future cosmological surveys are thus gearing up to measure these parameters and search for deviations from $(w, \gamma, \eta) = (-1, 0.55, 1)$. Whether or not one expects this search to succeed, is is undeniably good for science that cosmology is able to test the correctness of Einstein gravity, rather than simply assuming it.

11.5 The anthropic landscape

Whether or not one finds the 'essence' approach to dark energy compelling, there remains one big problem. All the models are constructed using Lagrangians with a particular zero level. All quintessence potentials have the field rolling down towards $V = 0$, and k-essence models lack a potential altogether. They are therefore subject to the classical dilemma of the cosmological constant: adding a pure constant to the Lagrangian has no affect on field dynamics, but mimics a cosmological constant. With so many possible contributions to this vacuum energy from the zero-point energies of different fields (if nothing else), it seems contrived to force $V(\phi)$ to asymptote to zero without a reason.

To review why zero is a problematic value for the vacuum density, recall what we mean by the vacuum: $|0\rangle$, or zero occupation number for each wave mode inside a given box. But standard quantum mechanics assigns a zero-point energy of $\hbar\omega/2$ to each mode. Integrating $\hbar\omega/2c^2$ per mode over k-space (with a degeneracy of 2 for polarization) gives a total density of

$$\rho_{\rm vac} = \frac{\hbar}{2\pi^2 c^5} \int \omega^3 \, d\omega, \tag{278}$$

which diverges horribly. Is it possible that the upper limit of the integral should be finite? This would be the case if space were a lattice, which is perhaps conceivable on some unobservably small scale. However, even with a cutoff at the hardly microscopic level of $\lambda \sim 1$ mm, $\rho_{\rm vac}$ already exceeds the critical density of the universe ($\sim 10^{-26}{\rm kg\,m^{-3}}$). We can express things in terms of an energy scale E_v by writing the dimensional scaling

$$\rho_v = \frac{\hbar}{c} \left(\frac{E_v}{\hbar c} \right)^4, \tag{279}$$

or simply $\rho_v = E_v^4$ in natural units. if we adopt the values $\Omega_v = 0.75$ and $h = 0.73$ for the key cosmological parameters, then $E_v = 2.39$ meV is known to a tolerance of about 1%. What is a natural choice for E_v? A case can be made for E_v lying at the Planck scale, since quantum gravity effects must destroy the flat-space assumptions of quantum field theory. This would give a vacuum density 120 power of 10 larger than observed. But this is over-dramatising the problem: one should focus on E_v rather than E_v^4. Also, the solution may lurk at much smaller energies. In unbroken supersymmetry, there would be an exact cancellation of the zero point energy of bosonic and fermionic oscillators, and the scale of supersymmetry breaking could be as low as 10 TeV. So the vacuum problem is perhaps that the energy scale of the vacuum is 'only' 15 powers of 10 smaller than seems reasonable – a lot fewer than 120 powers of 10, but still enough to cause a problem.

It should however be clear that this prediction is hard to make fixed, partly because of our ignorance of the field content of the universe, and because these zero-point contributions can be supplemented by classical contributions from $V(\phi)$ of any number of scalar fields. This problem has been sharpened by recent developments in string theory, known under the heading of the **landscape**. For the present purpose, this can be regarded as requiring the introduction of a large number of additional scalar fields, each with an associated potential. If we assume that a vacuum state is defined by these fields sitting at the minimum of their various potentials, then the effective cosmological constant can vary. It has been estimated that there are about 10^{500} distinct minima, which divides the natural vacuum density of $E_{\rm P}^4$ into what is almost a continuous range from the point of view of observations – so we can have almost any effective value of Λ we like.

This leads us in the direction of anthropic arguments, which are able to limit Λ to some extent: if the universe had become vacuum-dominated at $z > 1000$, gravitational instability would have been impossible – so that galaxies, stars and observers would not have been possible (Weinberg 1989). Indeed, Weinberg made the astonishingly prescient prediction on this basis that a non-zero vacuum density would be detected at Ω_v of order unity, since there was no reason for it to be much smaller.

MANY UNIVERSES At first sight, this argument seems quite appealing, but it rapidly leads us into deep waters. How can we talk about changing Λ? It has the value that it has. We are implicitly invoking an **ensemble picture** in which there are many universes with differing properties. This is a big step (although exciting, if this turns out to be the only way to explain the vacuum level we see). In fact, the idea of an ensemble emerges inevitably from the framework of inflationary cosmology, since the fluctuations in the scalar field can affect the progress of inflation itself. We have used this idea to look at the changes in when inflation ends – but fluctuations can affect the field at all stages of its evolution. They can be thought of as adding a random-walk element to the classical rolling of the scalar field down the trough defined by $V(\phi)$. In cases where ϕ is too close to the origin for inflation to persist for sufficiently long, it is possible for the quantum fluctuations to push ϕ further out – creating further inflation in a self-sustaining process. This is the concept of **stochastic eternal inflation** due to Linde. Sufficiently far from the origin, the random walk effect of fluctuations becomes more marked and can overwhelm the classical downhill rolling. This means that some regions of space can inflate for an indefinite time, and a single inflating universe automatically breaks up into different bubbles with their own histories. Some random subset of these eventually random-walk close enough to the origin that the classical end of inflation can occur, thus creating a set of 'universes' each of which can potentially host observers.

With this as a starting point, the question now becomes whether we can arrange for the different members of this ensemble to have different values of Λ. This is easily achieved. Let there be some quintessence field with a very flat potential, so that it is capable of simulating Λ

effectively. Quantum fluctuations during inflation can also displace this field, so that each member of the **multiverse** would have a different Λ.

THE DISTRIBUTION OF Λ We are now almost in a position to calculate a probability distribution for Λ. First, we have to set some ground rules: what will vary and what will be held fixed? We should try to change as little as possible, so we assume that all universes have the same values for

(1) The Baryon fraction $f_b = \rho_b/\rho_m$.

(2) The entropy per particle $S = (T/2.73)^3/\Omega_m h^2$

(3) The horizon-scale inhomogeneity $\delta_{\rm H} \simeq 10^{-5}$.

It is far from clear that these minimal assumptions are correct. For example, in the string theory **landscape**, there is no unique form for low-energy particle physics, but instead a large number of possibilities in which numbers such as the fine-structure constant, neutrino masses etc. are different. From the point of view of understanding Λ, we need there to be at least 10^{100} possible states so that at least some have Λ smaller than the natural $m_{\rm p}^4$ density by a sufficient factor. The landscape hypothesis provides this variation in Λ, but does not support the idea that particle physics is otherwise invariant. Still, it makes sense to start with the simplest forms of anthropic variation: if this can be ruled out, it might be taken as evidence in favour of the fuller landscape picture.

We then take a Bayesian viewpoint to the distribution of Λ given the existence of observers:

$$P(\Lambda \mid \text{Observer}) \propto P_{\rm prior}(\Lambda) P(\text{Observer} \mid \Lambda), \tag{280}$$

where we need both the prior distribution of Λ between different members of the ensemble and how the chance of getting an observer is modified by Λ. The latter factor should be proportional to the number of stars, which is generally take to be proportional to the fraction of the baryons that are incorporated into nonlinear structures. We can estimate this using the Press-Schechter apparatus to get the collapse fraction into systems of a galaxy-scale mass. The exact definition of this is not very important, since the CDM power spectrum is very flat on small scales: any mass at all close to $10^{12}\,M_\odot$ gives similar answers.

The more difficult part is the prior distribution of Λ, and a common argument is to say that it has a uniform distribution – which seems reasonable enough if we are to allow it to have either sign, but know that we will be interested in a very small range near zero. This is the startling proposition

of the anthropic model: the vacuum density takes large ranges, and in almost all realizations, the values are comparable in magnitude to the natural scale $m_{\rm P}^4$; such models are stupendously inimical to life.

We therefore have the simple model

$$dP(\rho_v) \propto f_c \, d\rho_v, \tag{281}$$

where f_c is the collapse fraction into galaxy-scale objects. For large values of Λ, growth ceases at high redshift, and f_c is exponentially suppressed. But things are less clear-cut if $\Lambda < 0$. Here the universe eventually recollapses, and the high density means that the collapse fraction always tends to unity. So why do we not observe $\Lambda < 0$? The answer is that we have to cut off the calculation at late stages of recollapse: once the universe becomes too hot, star-formation may be affected and in any case there is little time for life to form.

With this proviso, figure 31 shows the posterior distribution of Λ conditional on the existence of observers in the multiverse. Provided we consider recollapse only to a maximum temperature of about 10 K, the observed figure is matched well by the anthropic prediction: with this cutoff, most observers will see a positive Λ, and something of order 10% of observers will see Λ as big as we do, or smaller.

So is the anthropic explanation the correct one? Many people find the hypothesis too radical: why postulate an infinity of universes in order to explain a detail of one of them? Certainly, if an alternative explanation for the 'why now' problem existed in the form of e.g. a naturally successful quintessence model, one might tend to prefer that. But so far, there is no such alternative. The longer this situation persists, the more we will be forced to accept that the universe we see can only be understood by making proper allowance for our role as observers.

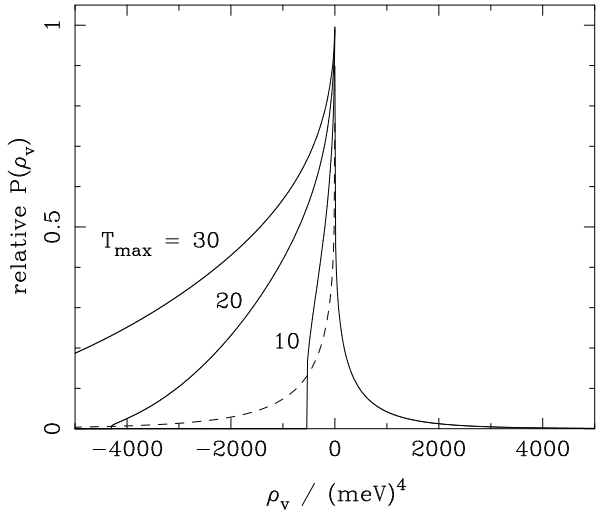

Figure 31. The collapse fraction as a function of the vacuum density, which is assumed to give the relative weighting of different models. The dashed line for negative density corresponds to the expanding phase only, whereas the solid lines for negative density include the recollapse phase, up to maximum temperatures of 10 K, 20 K, 30 K.