COSMOLOGY AND PARTICLE PHYSICS

J.A. Peacock
Institute for Astronomy, University of Edinburgh, Royal Observatory, Edinburgh EH9 3HJ, UK

Abstract
These lectures cover some of the basics of modern cosmology, assuming relatively little prior knowledge of the subject. They are organised into three main sections: (1) Models of the expanding universe (the Robertson-Walker metric, dynamics and the equation of state, the hot big bang, initial conditions and inflation); (2) Dark matter (astrophysical mass measurements, particle candidates for dark matter, constraints on dark matter from galaxy haloes, dark matter and cosmological perturbations); (3) Structure formation (statistics of cosmological density fields, generation of fluctuations via inflation, observations of large-scale structure, fluctuations in the microwave background).

1 THE ISOTROPIC UNIVERSE

1.1 The Robertson–Walker metric

Cosmological investigation began by considering the simplest possible mass distribution: one whose properties are **homogeneous** (constant density) and **isotropic** (the same in all directions). From this symmetry, the only allowed velocity field on a local scale is expansion (or contraction) with velocity proportional to distance:

$$\boxed{\mathbf{v} = H\mathbf{r}.} \tag{1}$$

Having chosen a model mass distribution, the next step is to solve the field equations to find the corresponding metric. Since our model is a particularly symmetric one, it is perhaps not too surprising that many of the features of the metric can be deduced from symmetry alone – and indeed will apply even if Einstein's equations are replaced by something more complicated. These general arguments were put forward independently by H.P. Robertson and A.G. Walker in 1936.

Cosmological time The first point to note is that something suspiciously like a universal time exists in an isotropic universe. Consider a set of observers in different locations, all of whom are at rest with respect to the matter in their vicinity (these characters are usually termed **fundamental observers**). We can envisage them as each sitting on a different galaxy, and so receding from each other with the general expansion. We can define a global time coordinate 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. The coordinate is useful globally rather than locally because the clocks can be synchronized by the exchange of light signals between observers, who agree to set their clocks to a standard time when *e.g.* the universal homogeneous density reaches some given value. Using this time coordinate plus isotropy, we already have enough information to conclude that the metric must take the following form:

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

Here, we have used the equivalence principle to say that the proper time interval between two distant events would look locally like special relativity to a fundamental observer on the spot: for them, $c^2 d\tau^2 = c^2 dt^2 - dx^2 - dy^2 - dz^2$. Since we use the same time coordinate as they do, our only difficulty is in the spatial part of the metric: relating their dx *etc.* to spatial coordinates centred on us.

Because of spherical symmetry, the spatial part of the metric can be decomposed into a radial and a transverse part (in spherical polars, $d\psi^2 = d\theta^2 + \sin^2\theta\, d\phi^2$). Distances have been decomposed into a product of a time-dependent **scale factor** $R(t)$ and a time-independent **comoving coordinate** r. The functions f and g are arbitrary; however, we can choose our radial coordinate such that either $f = 1$ or $g = r^2$, to make things look as much like Euclidean space as possible. Furthermore, we can determine the form of the remaining function from symmetry arguments.

To get some feeling for the general answer, it should help to think first about a simpler case: the metric on the surface of a sphere. A balloon being inflated is a common popular analogy for the expanding universe, and it will serve as a two-dimensional example of a space of constant curvature. If we call the polar angle in spherical polars r instead of the more usual θ, then the element of length on the surface of a sphere of radius R is

$$d\sigma^2 = R^2 \left(dr^2 + \sin^2 r\, d\phi^2 \right). \tag{3}$$

It is possible to convert this to the metric for a 2-space of constant **negative curvature** by the device of considering an imaginary radius of curvature, $R \to iR$. If we simultaneously let $r \to ir$, we obtain

$$d\sigma^2 = R^2 \left(dr^2 + \sinh^2 r\, d\phi^2 \right). \tag{4}$$

These two forms can be combined by defining a new radial coordinate that makes the transverse part of the metric look Euclidean:

$$d\sigma^2 = R^2 \left(\frac{dr^2}{1 - kr^2} + r^2\, d\phi^2 \right), \tag{5}$$

where $k = +1$ for positive curvature and $k = -1$ for negative curvature.

An isotropic universe has the same form for the comoving spatial part of its metric as the surface of a sphere. This is no accident, since it it possible to define the equivalent of a sphere in higher numbers of dimensions, and the form of the metric is always the same. For example, a **3-sphere** embedded in four-dimensional Euclidean space would be defined as the coordinate relation $x^2 + y^2 + z^2 + w^2 = R^2$. Now define the equivalent of spherical polars and write $w = R\cos\alpha$, $z = R\sin\alpha\cos\beta$, $y = R\sin\alpha\sin\beta\cos\gamma$, $x = R\sin\alpha\sin\beta\sin\gamma$, where α, β and γ are three arbitrary angles. Differentiating with respect to the angles gives a four-dimensional vector (dx, dy, dz, dw), and it is a straightforward exercise to show that the squared length of this vector is

$$|(dx, dy, dz, dw)|^2 = R^2 \left[d\alpha^2 + \sin^2\alpha \left(d\beta^2 + \sin^2\beta\, d\gamma^2 \right) \right], \tag{6}$$

which is the Robertson–Walker metric for the case of positive spatial curvature. This $k = +1$ metric describes a **closed universe**, in which a traveller who sets off along a trajectory of fixed β and γ will eventually return to their starting point (when $\alpha = 2\pi$). In this respect, the positively curved 3D universe is identical to the case of the surface of a sphere: it is finite, but unbounded. By contrast, the $k = -1$ metric describes an **open universe** of infinite extent; as before, changing to negative spatial curvature replaces $\sin\alpha$ with $\sinh\alpha$, and α can be made as

large as we please without returning to the starting point. The $k = 0$ model describes a **flat universe**, which is also infinite in extent. This can be thought of as a limit of either of the $k = \pm 1$ cases, where the curvature scale R tends to infinity.

Notation and conventions The Robertson–Walker metric (which we shall often write in the shorthand **RW metric**) may be written in a number of different ways. The most compact forms are those where the comoving coordinates are *dimensionless*. Define the very useful function

$$\boxed{S_k(r) = \begin{cases} \sin r & (k = 1) \\ \sinh r & (k = -1) \\ r & (k = 0), \end{cases}} \tag{7}$$

and its cosine-like analogue, $C_k(r) \equiv \sqrt{1 - kS_k^2(r)}$. The metric can now be written in the preferred form that we shall use throughout:

$$\boxed{c^2 d\tau^2 = c^2 dt^2 - R^2(t) \left[dr^2 + S_k^2(r)\, d\psi^2 \right].} \tag{8}$$

The most common alternative is to use a different definition of comoving distance, $S_k(r) \to r$, so that the metric becomes

$$c^2 d\tau^2 = c^2 dt^2 - R^2(t) \left(\frac{dr^2}{1 - kr^2} + r^2 d\psi^2 \right). \tag{9}$$

There should of course be two different symbols for the different comoving radii, but each is often called r in the literature, so we have to learn to live with this ambiguity; the presence of terms like $S_k(r)$ or $1 - kr^2$ will usually indicate which convention is being used. Alternatively, one can make the scale factor dimensionless, defining

$$\boxed{a(t) \equiv \frac{R(t)}{R_0},} \tag{10}$$

so that $a = 1$ at the present.

The redshift At small separations, where things are Euclidean, the proper separation of two fundamental observers is just $R(t)\, dr$, so that we obtain Hubble's law with

$$\boxed{H = \frac{\dot{R}}{R}.} \tag{11}$$

At large separations where spatial curvature becomes important, the concept of radial velocity becomes a little more slippery – but in any case how could one measure it directly in practice? At small separations, the recessional velocity gives the Doppler shift

$$\frac{\nu_{\rm emit}}{\nu_{\rm obs}} \equiv 1 + z \simeq 1 + \frac{v}{c}. \tag{12}$$

This defines the **redshift** z in terms of the shift of spectral lines. What is the equivalent of this relation at larger distances? Since photons travel on null geodesics of zero proper time, we see directly from the metric that

$$r = \int \frac{c\,dt}{R(t)}. \tag{13}$$

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. Photons that are emitted at later times will be received at later times, but these changes in $t_{\rm emit}$ and $t_{\rm obs}$ cannot alter the integral, since r is a comoving quantity. This requires the condition $dt_{\rm emit}/dt_{\rm obs} = R(t_{\rm emit})/R(t_{\rm obs})$, which means that events on distant galaxies time-dilate according to how much the universe has expanded since the photons we see now were emitted. Clearly (think of events separated by one period), this dilation also applies to frequency, and we therefore get

$$\boxed{\frac{\nu_{\rm emit}}{\nu_{\rm obs}} \equiv 1+z = \frac{R(t_{\rm obs})}{R(t_{\rm emit})}.} \tag{14}$$

In terms of the normalized scale factor $a(t)$ we have simply $a(t) = (1+z)^{-1}$. Photon wavelengths therefore stretch with the universe, as is intuitively reasonable.

1.2 Dynamics of the expansion

Expansion and geometry The equation of motion for the scale factor can be obtained in a quasi-Newtonian fashion. Consider a sphere about some arbitrary point, and let the radius be $R(t)r$, where r is arbitrary. The motion of a point at the edge of the sphere will, in Newtonian gravity, be influenced only by the interior mass. We can therefore write down immediately a differential equation (**Friedmann's equation**) that expresses conservation of energy: $(\dot{R}r)^2/2 - GM/(Rr) =$ constant. 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**. General relativity becomes even more vital in giving us the constant of integration in Friedmann's equation:

$$\boxed{\dot{R}^2 - \frac{8\pi G}{3}\rho R^2 = -kc^2.} \tag{15}$$

Note that this equation covers all contributions to ρ, *i.e.* those from matter, radiation and vacuum; it is independent of the equation of state.

For a given rate of expansion, there is thus a **critical density** that will yield $k = 0$, making the comoving part of the metric look Euclidean:

$$\boxed{\rho_c = \frac{3H^2}{8\pi G}.} \tag{16}$$

A universe with density above this critical value will be **spatially closed**, whereas a lower-density universe will be **spatially open**.

It is sometimes convenient to work with the time derivative of the Friedmann equation, because acceleration arguments in dynamics can often be more transparent than energy ones.

Differentiating with respect to time requires a knowledge of $\dot{\rho}$, but this can be eliminated by means of conservation of energy: $d[\rho c^2 R^3] = -p d[R^3]$. We then obtain

$$\ddot{R} = -4\pi G R(\rho c^2 + 3p)/3. \tag{17}$$

Both this equation and the Friedmann equation in fact arise as independent equations from different components of Einstein's equations for the RW metric.

Density parameters etc. The 'flat' universe with $k = 0$ arises for a particular **critical density**. We are therefore led to define a **density parameter** as the ratio of density to critical density:

$$\boxed{\Omega \equiv \frac{\rho}{\rho_c} = \frac{8\pi G\rho}{3H^2}.} \tag{18}$$

Since ρ and H change with time, this defines an epoch-dependent density parameter. The current value of the parameter should strictly be denoted by Ω_0. Because this is such a common symbol, we shall keep the formulae uncluttered by normally dropping the subscript; the density parameter at other epochs will be denoted by $\Omega(z)$. The critical density therefore just depends on the rate at which the universe is expanding. If we now also define a dimensionless (current) Hubble parameter as

$$\boxed{h \equiv \frac{H_0}{100\,\mathrm{km\,s^{-1}Mpc^{-1}}},} \tag{19}$$

then the current density of the universe may be expressed as

$$\begin{aligned} \rho_0 &= 1.88 \times 10^{-26} \Omega h^2 \ \mathrm{kg\,m^{-3}} \\ &= 2.78 \times 10^{11} \Omega h^2 \ M_\odot \,\mathrm{Mpc}^{-3}. \end{aligned} \tag{20}$$

A powerful approximate model for the energy content of the universe is to divide it into pressureless matter ($\rho \propto R^{-3}$), radiation ($\rho \propto R^{-4}$) and vacuum energy (ρ constant). 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; the third relation applies for Einstein's **cosmological constant**. In terms of observables, this means that the density is written as

$$\frac{8\pi G\rho}{3} = H_0^2(\Omega_v + \Omega_m a^{-3} + \Omega_r a^{-4}) \tag{21}$$

(introducing the normalized scale factor $a = R/R_0$). For some purposes, this separation is unnecessary, since the Friedmann equation treats all contributions to the density parameter equally:

$$\boxed{\frac{kc^2}{H^2R^2} = \Omega_m(a) + \Omega_r(a) + \Omega_v(a) - 1.} \tag{22}$$

Thus, a flat $k = 0$ universe requires $\sum \Omega_i = 1$ at all times, whatever the form of the contributions to the density, even if the equation of state cannot be decomposed in this simple way.

Lastly, it is often necessary to know the present value of the scale factor, which may be read directly from the Friedmann equation:

$$\boxed{R_0 = \frac{c}{H_0}\left[(\Omega - 1)/k\right]^{-1/2}.} \tag{23}$$

The Hubble constant thus sets the **curvature length**, which becomes infinitely large as Ω approaches unity from either direction.

Solutions to the Friedmann equation The Friedmann equation may be solved most simply in 'parametric' form, by recasting it in terms of the conformal time $d\eta = c\,dt/R$ (denoting derivatives with respect to η by primes):

$$R'^2 = \frac{8\pi G}{3c^2}\rho R^4 - kR^2. \tag{24}$$

Because $H_0^2 R_0^2 = kc^2/(\Omega - 1)$, the Friedmann equation becomes

$$a'^2 = \frac{k}{(\Omega - 1)}\left[\Omega_r + \Omega_m a - (\Omega - 1)a^2 + \Omega_v a^4\right], \tag{25}$$

which is straightforward to integrate provided $\Omega_v = 0$.

To the observer, the evolution of the scale factor is most directly characterised by the change with redshift of the Hubble parameter and the density parameter; the evolution of $H(z)$ and $\Omega(z)$ is given immediately by the Friedmann equation in the form $H^2 = 8\pi G\rho/3 - kc^2/R^2$. Inserting the above dependence of ρ on a gives

$$\boxed{H^2(a) = H_0^2\left[\Omega_v + \Omega_m a^{-3} + \Omega_r a^{-4} - (\Omega - 1)a^{-2}\right].} \tag{26}$$

This is a crucial equation, which can be used to obtain the relation between redshift and comoving distance. The radial equation of motion for a photon is $R\,dr = c\,dt = c\,dR/\dot{R} = c\,dR/(RH)$. With $R = R_0/(1+z)$, this gives

$$\boxed{\begin{aligned} R_0 dr &= \frac{c}{H(z)}\,dz \\ &= \frac{c}{H_0}\left[(1-\Omega)(1+z)^2 + \Omega_v + \Omega_m(1+z)^3 + \Omega_r(1+z)^4\right]^{-1/2} dz. \end{aligned}} \tag{27}$$

This relation is arguably the single most important equation in cosmology, since it shows how to relate comoving distance to the observables of redshift, Hubble constant and density parameters.

Lastly, using the expression for $H(z)$ with $\Omega(a) - 1 = kc^2/(H^2R^2)$ gives the redshift dependence of the total density parameter:

$$\boxed{\Omega(z) - 1 = \frac{\Omega - 1}{1 - \Omega + \Omega_v a^2 + \Omega_m a^{-1} + \Omega_r a^{-2}}.} \tag{28}$$

This last equation is very important. It tells us that, at high redshift, all model universes apart from those with only vacuum energy will tend to look like the $\Omega = 1$ model. If $\Omega \neq 1$, then in

the distant past $\Omega(z)$ must have differed from unity by a tiny amount: the density and rate of expansion needed to have been finely balanced for the universe to expand to the present. This tuning of the initial conditions is called the **flatness problem** and is one of the motivations for the applications of quantum theory to the early universe.

Matter-dominated universe From the observed temperature of the microwave background (2.73 K) and the assumption of three species of neutrino at a slightly lower temperature (see below), we deduce that the total relativistic density parameter is $\Omega_r h^2 \simeq 4.2 \times 10^{-5}$, so at present it should be a good approximation to ignore radiation. However, the different redshift dependences of matter and radiation densities mean that this assumption fails at early times: $\rho_m/\rho_r \propto (1+z)^{-1}$. One of the critical epochs in cosmology is therefore the point at which these contributions were equal: the redshift of **matter–radiation equality**

$$\boxed{1 + z_{\rm eq} \simeq 23\,900\ \Omega h^2.} \tag{29}$$

At redshifts higher than this, the universal dynamics were dominated by the relativistic-particle content. By a coincidence discussed below, this epoch is close to another important event in cosmological history: **recombination**. Once the temperature falls below $\simeq 10^4$ K, ionized material can form neutral hydrogen. Observational astronomy is only possible from this point on, since Thomson scattering from electrons in ionized material prevents photon propagation. In practice, this limits the maximum redshift of observational interest to about 1000; unless Ω is very low or vacuum energy is important, a matter-dominated model is therefore a good approximation to reality.

Models with vacuum energy The solution of the Friedmann equation becomes more complicated if we allow a significant contribution from vacuum energy – *i.e.* a non-zero cosmological constant. Detailed discussions of the problem are given by Felten & Isaacman (1986) and Carroll, Press & Turner (1992); the most important features are outlined below.

The Friedmann equation itself is independent of the equation of state, and just says $H^2R^2 = kc^2/(\Omega - 1)$, whatever the form of the contributions to Ω. In terms of the cosmological constant itself, we have

$$\Omega_v = \frac{8\pi G\rho_v}{3H^2} = \frac{\Lambda c^2}{3H^2}. \tag{30}$$

de Sitter space Before going on to the general case, it is worth looking at the endpoint of an outwards perturbation of Einstein's static model, first studied by de Sitter and named after him. This universe is completely dominated by vacuum energy, and is clearly the limit of the unstable expansion, since the density of matter redshifts to zero while the vacuum energy remains constant. Consider again the Friedmann equation in its general form $\dot{R}^2 - 8\pi G\rho R^2/3 = -kc^2$: since the density is constant and R will increase without limit, the two terms on the lhs must eventually become almost exactly equal and the curvature term on the rhs will be negligible. Thus, even if $k \neq 0$, the universe will have a density that differs only infinitesimally from the critical, so that we can solve the equation by setting $k = 0$, in which case

$$R \propto \exp Ht, \quad H = \sqrt{\frac{8\pi G\rho_v}{3}} = \sqrt{\frac{\Lambda c^2}{3}}. \tag{31}$$

An interesting interpretation of this behaviour was promoted in the early days of cosmology by Eddington: the cosmological constant is what *caused* the expansion.

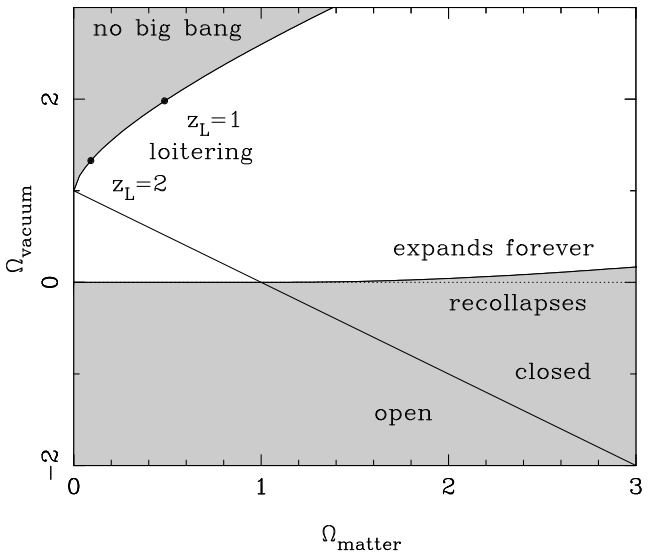

Fig. 1: This plot shows the different possibilities for the cosmological expansion as a function of matter density and vacuum energy. Models with total $\Omega > 1$ are always spatially closed (open for $\Omega < 1$), although closed models can still expand to infinity if $\Omega_v \neq 0$. If the cosmological 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.

Bouncing and loitering models Returning to the general case of models with a mixture of energy in the vacuum and normal components, we have to distinguish three cases. For models that start from a big bang (in which case radiation dominates completely at the earliest times), the universe will either recollapse or expand forever. The latter outcome becomes more likely for low densities of matter and radiation, but high vacuum density. It is however also possible to have models in which there is no big bang: the universe was collapsing in the distant past, but was slowed by the repulsion of a positive Λ term and underwent a 'bounce' to reach its present state of expansion. Working out the conditions for these different events is a matter of integrating the Friedmann equation. For the addition of Λ, this can only in general be done numerically. However, we can find the conditions for the different behaviours described above analytically, at least if we simplify things by ignoring radiation. The equation in the form of the time-dependent Hubble parameter looks like

$$\frac{H^2}{H_0^2} = \Omega_v(1 - a^{-2}) + \Omega_m(a^{-3} - a^{-2}) + a^{-2}, \tag{32}$$

and we are interested in the conditions under which the lhs vanishes, defining a turning point in the expansion. Setting the rhs to zero yields a cubic equation, and it is possible to give the conditions under which this has a solution (see Felten & Isaacman 1986). The main results of this analysis are summed up in figure 1. Since the radiation density is very small today, the main task of relativistic cosmology is to work out where on the $\Omega_{\rm matter} - \Omega_{\rm vacuum}$ plane the real universe lies. The existence of high-redshift objects rules out the bounce models, so that the idea of a hot big bang cannot be evaded.

Flat universe The most important model in cosmological research is that with $k = 0 \Rightarrow \Omega_{\rm total} = 1$; when dominated by matter, this is often termed the **Einstein–de Sitter** model.

Paradoxically, this importance arises because it is an unstable state: as we have seen earlier, the universe will evolve away from $\Omega = 1$, given a slight perturbation. For the universe to have expanded by so many **e-foldings** (factors of e expansion) and yet still have $\Omega \sim 1$ implies that it was very close to being spatially flat at early times.

It now makes more sense to work throughout in terms of the normalized scale factor $a(t)$, so that the Friedmann equation for a matter–radiation mix is

$$\dot{a}^2 = H_0^2 \left(\Omega_m a^{-1} + \Omega_r a^{-2} \right), \tag{33}$$

which may be integrated to give the time as a function of scale factor:

$$H_0 t = \frac{2}{3\Omega_m^2} \left[\sqrt{\Omega_r + \Omega_m a} \, (\Omega_m a - 2\Omega_r) + 2\Omega_r^{3/2} \right]; \tag{34}$$

this goes to $\frac{2}{3} a^{3/2}$ for a matter-only model, and to $\frac{1}{2} a^2$ for radiation only.

One further way of presenting the model's dependence on time is via the density. Following the above, it is easy to show that

$$\boxed{\begin{array}{ll} t = \sqrt{\dfrac{1}{6\pi G \rho}} & \text{(matter domination)} \\ t = \sqrt{\dfrac{3}{32\pi G \rho}} & \text{(radiation domination).} \end{array}} \tag{35}$$

Because Ω_r is so small, the deviations from a matter-only model are unimportant for $z \lesssim 1000$, and so the distance–redshift relation for the $k = 0$ matter plus radiation model is effectively just that of the $\Omega_m = 1$ Einstein–de Sitter model. An alternative $k = 0$ model of greater observational interest has a significant cosmological constant, so that $\Omega_m + \Omega_v = 1$ (radiation being neglected for simplicity). This may seem contrived, but once $k = 0$ has been established, it cannot change: individual contributions to Ω must adjust to keep in balance. The advantage of this model is that it is the only way of retaining the theoretical attractiveness of $k = 0$ while changing the age of the universe from the relation $H_0 t_0 = 2/3$, which characterises the Einstein–de Sitter model. Since much observational evidence indicates that $H_0 t_0 \simeq 1$, this model has received a good deal of interest in recent years. To keep things simple we shall neglect radiation, so that the Friedmann equation is

$$\dot{a}^2 = H_0^2 [\Omega_m a^{-1} + (1 - \Omega_m) a^2], \tag{36}$$

and the $t(a)$ relation is

$$H_0 t(a) = \int_0^a \frac{x \, dx}{\sqrt{\Omega_m x + (1 - \Omega_m) x^4}}. \tag{37}$$

The x^4 on the bottom looks like trouble, but it can be rendered tractable by the substitution $y = \sqrt{x^3 |\Omega_m - 1| / \Omega_m}$, which turns the integral into

$$H_0 t(a) = \frac{2}{3} \, \frac{S_k^{-1}(\sqrt{a^3 |\Omega_m - 1| / \Omega_m})}{\sqrt{|\Omega_m - 1|}}. \tag{38}$$

Here, k in S_k is used to mean sin if $\Omega_m > 1$, otherwise sinh; these are still $k = 0$ models. This $t(a)$ relation is compared to models without vacuum energy in figure 2. Since there is nothing special about the current era, we can clearly also rewrite this expression as

$$H(a)\, t(a) = \frac{2}{3} \, \frac{S_k^{-1}(\sqrt{|\Omega_m(a) - 1| / \Omega_m(a)})}{\sqrt{|\Omega_m(a) - 1|}} \simeq \frac{2}{3} \, \Omega_m(a)^{-0.3}, \tag{39}$$

where we include a simple approximation that is accurate to a few % over the region of interest ($\Omega_m \gtrsim 0.1$). In the general case of significant Λ but $k \neq 0$, this expression still gives a very good approximation to the exact result, provided Ω_m is replaced by $0.7\Omega_m - 0.3\Omega_v + 0.3$ (Carroll, Press & Turner 1992).

Horizons For photons, the radial equation of motion is just $c\,dt = R\,dr$. How far can a photon get in a given time? The answer is clearly

$$\Delta r = \int_{t_0}^{t_1} \frac{c\,dt}{R(t)} = \Delta\eta, \tag{40}$$

i.e. just the interval of conformal time. What happens as $t_0 \to 0$ in this expression? We can replace dt by $dR/\dot{R}$, which the Friedmann equation says is $\propto dR/\sqrt{\rho R^2}$ at early times. Thus, this integral converges if $\rho R^2 \to \infty$ as $t_0 \to 0$, otherwise it diverges. Provided the equation of state is such that ρ changes faster than R^{-2}, light signals can only propagate a finite distance between the big bang and the present; there is then said to be a **particle horizon**. Such a horizon therefore exists in conventional big bang models, which are dominated by radiation at early times.

1.3 Observations in cosmology

We can now assemble some essential formulae for interpreting cosmological observations. Since we will mainly be considering the post-recombination epoch, these apply for a matter-dominated model only. Our observables are redshift, z, and angular difference between two points on the sky, $d\psi$. We write the metric in the form

$$c^2 d\tau^2 = c^2 dt^2 - R^2(t)\left[dr^2 + S_k^2(r)\,d\psi^2\right], \tag{41}$$

so that the *comoving* volume element is

$$\boxed{dV = 4\pi[R_0 S_k(r)]^2\, R_0 dr.} \tag{42}$$

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

$$\boxed{d\ell = d\psi\ R_0 S_k(r)/(1+z).} \tag{43}$$

Probably the most important relation for observational cosmology is that between monochromatic flux density and luminosity. Start by assuming isotropic emission, so that the photons emitted by the source pass with a uniform flux density through any sphere surrounding the source. We can now make a shift of origin, and consider the RW metric as being centred on the source; however, because of homogeneity, the comoving distance between the source and the observer is the same as we would calculate when we place the origin at our location. The photons from the source are therefore passing through a sphere, on which we sit, of proper surface area $4\pi[R_0 S_k(r)]^2$. But redshift still affects the flux density in four further ways: photon energies and arrival rates are redshifted, reducing the flux density by a factor $(1+z)^2$; opposing this, the bandwidth $d\nu$ is reduced by a factor $1+z$, so the energy flux per unit bandwidth goes down by one power of $1+z$; finally, the observed photons at frequency ν_0 were emitted at

frequency $\nu_0(1+z)$, so the flux density is the luminosity at this frequency, divided by the total area, divided by $1+z$:

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

A word about units: L_ν in this equation would be measured in units of $\mathrm{W\,Hz^{-1}}$. Recognizing that emission is often not isotropic, it is common to consider instead the luminosity emitted into unit solid angle – in which case there would be no factor of 4π, and the units of L_ν would be $\mathrm{W\,Hz^{-1}\,sr^{-1}}$.

The flux density received by a given observer can be expressed by definition as the product of the **specific intensity** I_ν (the flux density received from unit solid angle of the sky) and the solid angle subtended by the source: $S_\nu = I_\nu\, d\Omega$. Combining the angular size and flux-density relations thus gives the relativistic version of surface-brightness conservation. This 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). We can integrate over ν_0 to obtain the corresponding total or **bolometric** formulae, which are needed *e.g.* for spectral-line emission:

$$S_{\rm tot} = \frac{L_{\rm tot}}{4\pi R_0^2 S_k^2(r)(1+z)^2}; \tag{46}$$

$$I_{\rm tot} = \frac{B_{\rm tot}}{(1+z)^4}. \tag{47}$$

The form of the above relations lead to the following definitions for particular kinds of distances:

$$\boxed{\begin{aligned} \textbf{angular}-\textbf{diameter distance}: &\quad D_{\rm A} = (1+z)^{-1} R_0 S_k(r) \\ \textbf{luminosity distance}: &\quad D_{\rm L} = (1+z)\; R_0 S_k(r). \end{aligned}} \tag{48}$$

The last element needed for the analysis of observations is a relation between redshift and age for the object being studied. This brings in our earlier relation between time and comoving radius (consider a null geodesic traversed by a photon that arrives at the present):

$$c\, dt = R_0\, dr/(1+z). \tag{49}$$

Distance–redshift relation The general relation between comoving distance and redshift was given earlier as

$$\begin{aligned} R_0 dr &= \frac{c}{H(z)}\, dz \\ &= \frac{c}{H_0}\, \left[(1-\Omega)(1+z)^2 + \Omega_v + \Omega_m(1+z)^3 + \Omega_r(1+z)^4\right]^{-1/2}\, dz. \end{aligned} \tag{50}$$

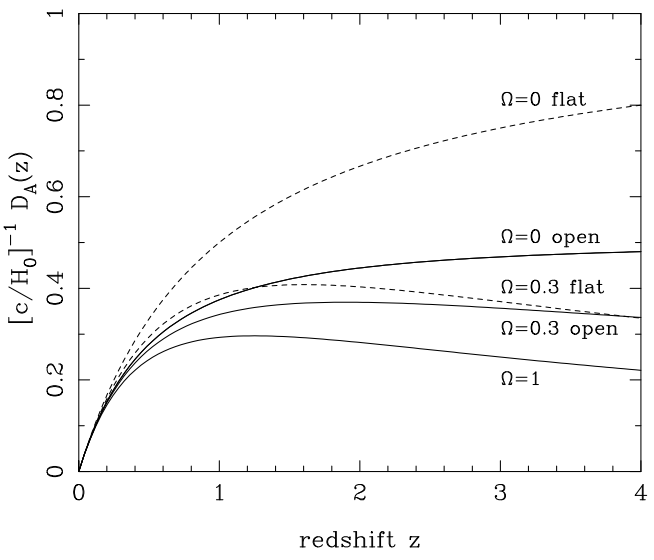

Fig. 2: A plot of dimensionless angular-diameter distance versus redshift for various cosmologies. Solid lines show models with zero vacuum energy; dashed lines show flat models with $\Omega_m + \Omega_v = 1$. In both cases, results for $\Omega_m = 1, 0.3, 0$ are shown; higher density results in lower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

For a matter-dominated Friedmann model, this means that the distance of an object from which we receive photons today is

$$R_0 r = \frac{c}{H_0} \int_0^z \frac{dz'}{(1+z')\sqrt{1+\Omega z'}}. \tag{51}$$

Integrals of this form often arise when manipulating Friedmann models; they can usually be tackled by the substitution $u^2 = k(\Omega - 1)/[(\Omega(1+z)]$. This substitution produces **Mattig's formula** (1958), which is one of the single most useful equations in cosmology as far as observers are concerned:

$$\boxed{R_0 S_k(r) = \frac{2c}{H_0} \frac{\Omega z + (\Omega - 2)[\sqrt{1+\Omega z} - 1]}{\Omega^2 (1+z)}.} \tag{52}$$

2 THE HOT BIG BANG

Adiabatic expansion What was the state of matter in the early phases of the big bang? Since the present-day expansion will cause the density to decline in the future, conditions in the past must have corresponded to high density – and thus to high temperature. We can deal with this quantitatively by looking at the thermodynamics of the fluids that make up a uniform cosmological model.

The expansion is clearly **adiathermal**, since the symmetry means that there can be no net heat flow through any surface. If the expansion is also reversible, then we can go one step further, because entropy change is defined in terms of the heat that flows during a reversible change. If no heat flows during a reversible change, then entropy must be conserved, and the expansion will be **adiabatic**. This can only be an approximation, since there will exist irreversible microscopic

processes. In practice, however, it will be shown below that the effects of these processes are overwhelmed by the entropy of thermal background radiation in the universe. It will therefore be an excellent approximation to treat the universe as if the matter content were a simple dissipationless fluid undergoing a reversible expansion. This means that, for a ratio of specific heats Γ, we get the usual adiabatic behaviour

$$T \propto R^{-3(\Gamma-1)}. \tag{53}$$

For radiation, $\Gamma = 4/3$ and we get just $T \propto 1/R$. A simple model for the energy content of the universe is to distinguish pressureless 'dust-like' matter (in the sense that $p \ll \rho c^2$) from relativistic 'radiation-like' matter (photons plus neutrinos). If these are assumed not to interact, then the energy densities scale as

$$\begin{aligned} \rho_m &\propto R^{-3} \\ \rho_r &\propto R^{-4} \end{aligned} \tag{54}$$

The universe must therefore have been **radiation dominated** at some time in the past, where the densities of matter and radiation cross over. To anticipate, we know that the current radiation density corresponds to thermal radiation with $T \simeq 2.73\mathrm{K}$. We shall shortly show that one expects to find, in addition to this **cosmic microwave background**CMB (CMB), a background in neutrinos that has 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

$$\boxed{1 + z_{\mathrm{eq}} = 23\,900\,\Omega h^2\,(T/2.73\mathrm{K})^{-4}.} \tag{55}$$

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.

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 $\hbar/(m_{\mathrm{P}}c)$ and time ℓ_{P}/c form the system of **Planck units**:

$$\begin{aligned} m_{\mathrm{P}} &\equiv \sqrt{\frac{\hbar c}{G}} \simeq 10^{19}\mathrm{GeV} \\ \ell_{\mathrm{P}} &\equiv \sqrt{\frac{\hbar G}{c^3}} \simeq 10^{-35}\mathrm{m} \\ t_{\mathrm{P}} &\equiv \sqrt{\frac{\hbar G}{c^5}} \simeq 10^{-43}\mathrm{s}. \end{aligned} \tag{56}$$

The Planck time therefore sets the origin of time for the classical phase of the big bang.

Collisionless equilibrium backgrounds We need the thermodynamics of a possibly relativistic perfect gas. We consider some box of volume $V = L^3$, and say that we will analyse the quantum

mechanics of particles in the box by taking the system to be periodic on scale L. Quantum fields in the box are expanded in plane waves, with allowed wavenumbers $k_x = n\,2\pi/L$ *etc.*; these **harmonic boundary conditions** for the allowed eigenstates in the box lead to the density of states in k space:

$$dN = g\,\frac{V}{(2\pi)^3}\,d^3k \tag{57}$$

(where g is a degeneracy factor for spin *etc.*). This expression is nice because it is **extensive** ($N \propto V$) and hence the number density n is independent of V. The equilibrium **occupation number** for a quantum state of energy ϵ is given generally by

$$\langle f \rangle = \left[e^{(\epsilon-\mu)/kT} \pm 1\right]^{-1} \tag{58}$$

(+ 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. More formally, the Helmholtz free energy $F = E - TS$ is minimized in equilibrium for a system at constant temperature and volume. Since $dF = -SdT - PdV + \mu dN$, $dF/dN = 0 \Rightarrow \mu = 0$. Thus, in terms of momentum space, the thermal equilibrium **background number density** of particles is

$$\boxed{n = g\,\frac{1}{(2\pi\hbar)^3}\int_0^\infty \frac{4\pi\,p^2 dp}{e^{\epsilon(p)/kT} \pm 1},} \tag{59}$$

where $\epsilon = \sqrt{m^2c^4 + p^2c^2}$ 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

$$n = \left(\frac{kT}{c}\right)^3 \frac{4\pi g}{(2\pi\hbar)^3}\int_0^\infty \frac{y^2 dy}{e^y \pm 1}. \tag{60}$$

(2) Non-relativistic limit. Here we can neglect the ± 1 in the occupation number, in which case

$$n = e^{-mc^2/kT}(2mkT)^{3/2}\frac{4\pi g}{(2\pi\hbar)^3}\int_0^\infty e^{-y^2}y^2 dy. \tag{61}$$

This shows us that the background 'switches on' at about $kT \sim mc^2$; at this energy, photons and other species in equilibrium will have sufficient energy to create particle-antiparticle pairs, which is how such an equilibrium background would be created. The point at which $kT \sim mc^2$ for some particle is known as a **threshold**.

Similar reasoning 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:

$$\boxed{u = \rho\,c^2 = g\,\frac{1}{(2\pi\hbar)^3}\int_0^\infty \frac{4\pi\;p^2\;dp}{e^{\epsilon(p)/kT} \pm 1}\;\epsilon(p).} \tag{62}$$

In the same way, we can get the pressure from kinetic theory: $P = n\langle pv\rangle/3 = n\langle p^2c^2/\epsilon\rangle/3$, where v is the particle velocity, and is related to its momentum and energy by $p = (\epsilon/c^2)v$. The pressure is therefore given by the following integral:

$$\boxed{P = g\,\frac{1}{(2\pi\hbar)^3}\int_0^\infty \frac{4\pi\ p^2\,dp}{e^{\epsilon/kT}\pm 1}\,\frac{p^2c^2}{3\epsilon}.} \tag{63}$$

Clearly, in the ultrarelativistic limit with $\epsilon \simeq pc$, the pressure obeys $P = \rho c^2/3$. In the nonrelativistic limit, the pressure is just $P = nkT$ (see below), whereas the density is dominated by the rest mass: $\rho = mn$, and therefore $P \ll \rho c^2/3$. At a threshold, the equation of state thus departs slightly from $P = \rho c^2/3$, even if the universe is radiation dominated on either side of the critical temperature.

Entropy of the background One quantity that is of considerable importance is the **entropy** of a thermal background. This may be derived in several ways. The most direct is to note that both 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,

$$dE = TdS - PdV \quad \Rightarrow \quad \left(\frac{E}{V}dV + \frac{\partial E}{\partial T}dT\right) = \left(T\frac{S}{V}dV + T\frac{\partial S}{\partial T}dT\right) - PdV. \tag{64}$$

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

$$\boxed{S = \frac{E + PV}{T}} \tag{65}$$

Using the above integral for the pressure, the entropy is

$$S = \frac{4\pi g V}{(2\pi\hbar)^3}\int_0^\infty \frac{p^2dp}{e^{\epsilon/kT}\pm 1}\left(\frac{\epsilon}{T} + \frac{p^2c^2}{3\epsilon T}\right), \tag{66}$$

which becomes $S = 3.602Nk$ (bosons) or $4.202Nk$ (fermions) in the ultrarelativistic limit and $S = (mc^2/kT)Nk$ in the non-relativistic limit. So, for radiation, the entropy is just proportional to the number of particles. For this reason, the ratio of photon to baryon number densities $n_\gamma/n_{\rm B}$ is sometimes called the **entropy per baryon**.

Formulae for ultrarelativistic backgrounds We now summarize the most useful results from this discussion, which are the thermodynamic quantities for massless particles. These formulae are required time and time again in calculations of conditions in the early universe. Consider first bosons, such as the microwave background. Evaluating the dimensionless integrals encountered earlier (only possible numerically in the case of n) gives energy, number and entropy densities:

$$\boxed{\begin{aligned} u &= g\,\frac{\pi^2}{30}\,kT\,\left(\frac{kT}{\hbar c}\right)^3 = 3P \\ \frac{s}{k} &= g\,\frac{2\pi^2}{45}\,\left(\frac{kT}{\hbar c}\right)^3 = 3.602n \end{aligned}} \tag{67}$$

(remember that $g = 2$ for photons).

It is also expected that there will be a fermionic relic background of neutrinos left over from the big bang. Assume for now that the neutrinos are massless. In this case, the thermodynamic properties can be obtained from those of black-body radiation by the following trick:

$$\frac{1}{e^x+1} = \frac{1}{e^x-1} - \frac{2}{e^{2x}-1}. \tag{68}$$

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 then get the corresponding fermionic results. The entropy requires just a little more care. Although we have said that entropy density is proportional to number density, in fact the entropy density for an ultrarelativistic gas was shown above to be $s = (4/3)u/T$, and so the fermionic factor is the same as for energy density:

$$\boxed{\begin{aligned} n_{\rm F} &= \frac{3}{4}\frac{g_{\rm F}}{g_{\rm B}} n_{\rm B} \\ u_{\rm F} &= \frac{7}{8}\frac{g_{\rm F}}{g_{\rm B}} u_{\rm B}. \\ s_{\rm F} &= \frac{7}{8}\frac{g_{\rm F}}{g_{\rm B}} s_{\rm B}. \end{aligned}} \tag{69}$$

Using these rules, it is usually possible to forget about the precise nature of the relativistic background in the universe and count bosonic degrees of freedom, given the effective degeneracy factor for u or s:

$$g_* \equiv \sum_{\rm bosons} g_i + \frac{7}{8}\sum_{\rm fermions} g_j, \tag{70}$$

although this definition needs to be modified if some species have different temperatures (see below).

Neutrino decoupling At the later stages of the big bang, energies are such that only light particles survive in equilibrium: γ, ν and the three leptons e, μ, τ. If neutrinos could be maintained in equilibrium, the lepton-antilepton pairs would annihilate as the temperature fell still further ($T_\tau = 10^{13.3}$ K, $T_\mu = 10^{12.1}$ K, $T_e = 10^{9.7}$ K), and the end result would be that the products of these annihilations would be shared among the only massless particles. However, in practice the weak reactions that maintain the neutrinos in thermal equilibrium 'switch off' at $T \simeq 10^{10}$ K. This **decoupling** is discussed in more detail below, but is a general cosmological phenomenon, which arises whenever the interaction timescales exceed the local Hubble time, leaving behind abundances of particles frozen at the values they had when last in thermal equilibrium. Two-body reaction rates scale proportional to density, times a cross-section that is often a declining function of energy, so that the interaction time changes at least as fast as R^{-3}. In contrast, the Hubble time changes no faster than R^{-2} (in the radiation era), so that there is inevitably a crossover. For neutrinos, this point occurs at a redshift of $\sim 10^{10}$, whereas the photons of the microwave background typically last interacted with matter at $z \simeq 1000$.

The effect of the electron-positron annihilation is therefore to enhance the numbers of photons relative to neutrinos. It is easy to see what quantitative effect this has: although we may talk loosely about the energy of $e^\pm$ annihilation going into photons, what is actually conserved 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

$$s(\gamma + e^+ + e^-) = \frac{11}{4} s(\gamma). \tag{71}$$

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. Equivalently, given the observed photon temperature today, we infer the existence of a neutrino background with a temperature

$$\boxed{T_\nu = \left(\frac{4}{11}\right)^{1/3} T_\gamma = 1.95\,\mathrm{K},} \tag{72}$$

for $T_\gamma = 2.73$ K. Although it is hard to see how such low energy neutrinos could ever be detected directly, their gravitation is certainly not negligible: they contribute an energy density that is a factor $(7/8) \times (4/11)^{4/3}$ times that of the photons (the fact that neutrinos have $g = 1$ whereas photons have $g = 2$ is cancelled by the fact that neutrinos and antineutrinos are distinguishable particles). For three neutrino species, this enhances the energy density in relativistic particles by a factor 1.68.

Massive neutrinos Although for many years the conventional wisdom was that neutrinos were massless, this assumption began to be increasingly challenged around the end of the 1970s. Theoretical progress in understanding the origin of masses in particle physics meant that it was no longer natural for the neutrino to be completely devoid of mass. Also, experimental evidence (Reines *et al.* 1980), which in fact turned out to be erroneous, seemed to imply a non-zero mass of $m \sim 10$ eV for the electron neutrino. The consequences of this for cosmology could be quite profound, as relic neutrinos are expected to be very abundant. The above section showed that $n(\nu + \overline{\nu}) = (3/4)n(\gamma;\, T = 1.95\,\mathrm{K})$. That yields a total of 113 relic neutrinos in every cm^3 for each species.

The consequences of giving these particles a mass are easily worked out provided the mass is small enough. If this is the case, then the neutrinos were ultrarelativistic at decoupling and their statistics were those of massless particles. As the universe expands to $kT < m_\nu c^2$, the total number of neutrinos is preserved. Furthermore, their momentum redshifts as $p \propto 1/R$, so that the momentum-space distribution today will just be a redshifted version of the ultrarelativistic form. As discussed more fully below, the momentum-space distribution stays exactly what would have been expected for thermal-equilibrium neutrinos, even though they have long since decoupled. However, this illusion is broken once the temperature falls below $kT < m_\nu c^2$, because the effect of the rest-mass energy on the equilibrium occupation number causes the nonrelativistic momentum distribution to differ from the relativistic one. We therefore obtain the present-day mass density in neutrinos just by multiplying the zero-mass number density by m_ν, and the consequences for the cosmological density are easily worked out to be

$$\boxed{\Omega h^2 = \frac{\sum m_i}{93.5\,\mathrm{eV}}.} \tag{73}$$

For a low Hubble parameter $h \simeq 0.5$, an average mass of only 8 eV will suffice to close the universe. In contrast, the current laboratory limits to the neutrino masses are

$$\begin{aligned} \nu_e &\lesssim 15\,\mathrm{eV} \\ \nu_\mu &\lesssim 0.17\,\mathrm{MeV} \\ \nu_\tau &\lesssim 24\,\mathrm{MeV}. \end{aligned} \tag{74}$$

3 RELICS OF THE BIG BANG

The massive neutrino is the simplest example of a relic of the big bang: a particle that once existed in equilibrium, but which has decoupled and thus preserves a 'snapshot' of the properties of the universe at the time the particle was last in thermal equilibrium. The aim of this section is to give a little more detail on the processes that determine the final abundance of these relics.

3.1 Freeze-out

So far, we have used a simple argument, which decrees that the relic abundance becomes fixed when expansion and interaction timescales are equal. To do better than this, it is necessary to look at the differential equation that governs the abundance of particle species in the expanding universe. This is the **Boltzmann equation**, which considers the **phase-space density**: the joint probability density for finding a particle in a given volume element and in a given range of momentum, denoted by $f(\mathbf{x}, \mathbf{p})$. The general form of this equation is

$$\boxed{\frac{\partial f}{\partial t} + (\dot{\mathbf{x}} \cdot \boldsymbol{\nabla}_{\mathbf{x}})\, f + (\dot{\mathbf{p}} \cdot \boldsymbol{\nabla}_{\mathbf{p}})\, f = \dot{f}_c.} \tag{75}$$

The lhs is just the fluid-dynamical convective derivative of the phase-space density, generalized to 6D space. The rhs is the collisional term, and the equation therefore just says that groups of particles maintain their phase-space density as they stream through phase space, unless modified by collisions (**Liouville's theorem**). The truth of this theorem is easily seen informally in one spatial dimension: a small square element $dx\, dv_x$ becomes sheared to a parallelogram of unchanged area, and so the phase-space density is unaltered.

The Boltzmann equation has been written with respect to a fixed system of laboratory coordinates, but it is quite easily adapted to the expanding universe. We should now interpret the particle velocities as being relative to a set of uniformly expanding observers. A particle that sets off from $r = 0$ with some velocity will effectively slow down as it tries to overtake distant receding observers. After time t, the particle will have travelled $x = vt$, and so encountered an observer with velocity $dv = Hx$. According to this observer, the particle's momentum is now reduced by $dp = mdv = mHvt = Hpt$. There is therefore the appearance of a **Hubble drag** force:

$$\frac{\dot{p}}{p} = -H. \tag{76}$$

In the presence of density fluctuations, this needs to be supplemented by gravitational forces, which as usual manifest themselves through the affine connection. In a sense, most of cosmological theory comes down to solving the Boltzmann equation for photons plus neutrinos plus collisionless dark matter, coupled to the matter fluid via gravity in all cases and also by Thomson scattering in the case of photons. Since the interesting processes are operating at early times when the density fluctuations are small, this is an exercise in first-order relativistic perturbation theory. The technical difficulties in detail mean this will have to be omitted here (see Peebles 1980; Efstathiou 1990); when discussing perturbations, the main results can usually be understood in terms of a fluid approximation, and this approach is pursued below.

Things are much easier in the case of homogeneous backgrounds, where spatial derivatives can be neglected. The Boltzmann equation then has the simple form

$$\frac{\partial f}{\partial t} - Hp\frac{\partial f}{\partial p} = \dot{f}_c. \tag{77}$$

The collision term is usually dominated by particle–antiparticle annihilations (assuming for the moment that the numbers of each are identical, so that there is no significant asymmetry):

$$\dot{f}_c = -\int \langle \sigma v \rangle \, f \, \bar{f} \, d^3\bar{p}, \tag{78}$$

where $\langle \sigma v \rangle$ is the velocity-averaged product of the cross-section and the velocity. We can take $\langle \sigma v \rangle$ outside the integral even if it is not constant provided it is evaluated at some suitable average energy. Integrating over momentum then gives the moment equation for the number density,

$$\dot{n} + 3Hn = -\langle \sigma v \rangle n^2 + S, \tag{79}$$

where S is a source term added to represent the production of particles from thermal processes – effectively pair creation. This term is 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

$$S = \langle \sigma v \rangle n_T^2. \tag{80}$$

If we define comoving number densities $N \equiv a^3 n$, the rate equation can be rewritten in the simple form

$$\boxed{\frac{d\ln N}{d\ln a} = -\frac{\Gamma}{H}\left[1 - \left(\frac{N_T}{N}\right)^2\right],} \tag{81}$$

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 $z \simeq 1000$. The CMB nevertheless still appears to be equilibrium black-body radiation because the number density of photons has fallen by the right amount to compensate for the redshifting of photon energy. This sounds like an incredible coincidence, but is in fact quite inevitable when looked at from the quantum-mechanical point of view. This says that the occupation number of a given mode, $= (\exp \hbar\omega/kT - 1)^{-1}$ for thermal radiation, is an adiabatic invariant that does not change as the universe expands – only the frequency alters, and thus the apparent temperature.

Now consider the opposite case, where the thermal solution would be nonrelativistic, with

$$N_T \propto T^{-3/2} \exp(-mc^2/kT). \tag{82}$$

If the background is to stay at the equilibrium value, the lhs of the rate equation must therefore be $\gg -1$. 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,

$$N(a \to \infty) = N(\Gamma/H = 1). \tag{83}$$

Exact numerical solutions of the rate equation almost always turn out very close to this simple rule (see chapter 5 of Kolb & Turner 1990).

3.2 Recombination and last scattering

One of the critical epochs 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.

There is a problem: highly excited atoms can be produced by a series of small transitions, but to reach the ground state requires the production of photons at least as energetic as the $2P \to 1S$ spacing (Lyman α, with $\lambda = 1216$Å). Multiple absorption of these photons will cause reionization once they become abundant, so it would now appear that recombination can never occur at all (unlike a finite HII region, where the Lyα photons can escape; see *e.g.* Osterbrock 1974). There is a way out, however, using **two-photon emission**. The $2S \to 1S$ 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. This gives the mechanism we need for transferring the ionization energy into photons with $\lambda > \lambda_{\mathrm{Ly}\alpha}$.

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

$$\frac{d(nx)}{dt} = -R\,(nx)^2\,\frac{\Lambda_{2\gamma}}{\Lambda_{2\gamma} + \Lambda_{\mathrm{U}}(T)}, \tag{84}$$

where n is the number density of protons, x is the fractional ionization, R is the recombination coefficient ($R \simeq 3 \times 10^{-17} T^{-1/2}\,\mathrm{m^3 s^{-1}}$), $\Lambda_{2\gamma}$ is the two-photon decay rate, and $\Lambda_{\mathrm{U}}(T)$ 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. A fuller discussion (see chapter 6 of Peebles 1993) would include a number of other processes: depopulation of the ground state by inverse 2-photon absorption; redshifting of Ly alpha photons due to the universal expansion, which can prevent them being re-absorbed. However, at the redshifts of practical interest (1000 to 10), the simplified equation captures the main effect.

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 $\Lambda_{\mathrm{U}}(T)$ are functions of temperature, and thus of redshift only, so that any parameter dependence is carried just by n^2, which scales $\propto (\Omega_{\mathrm{B}} h^2)^2$, where Ω_{B} is the baryonic density parameter. Similarly, the lhs depends on $\Omega_{\mathrm{B}} h^2$ through n; the other parameter dependence comes if we convert time derivatives to derivatives with respect to redshift:

$$\frac{dt}{dz} \simeq -3.09 \times 10^{17} (\Omega h^2)^{-1/2}\, z^{-5/2}\ \mathrm{s}, \tag{85}$$

for a matter-dominated model at large redshift (Ω is the total density parameter). Putting these together, the fractional ionization must scale as

$$\boxed{x(z) \propto \frac{(\Omega h^2)^{1/2}}{\Omega_{\rm B} h^2}.} \tag{86}$$

The last-scattering shell Putting in all the relevant processes, Jones & Wyse (1985) found the fractional ionization x near $z = 1000$ to be well approximated by

$$x(z) = 2.4 \times 10^{-3} \frac{(\Omega h^2)^{1/2}}{\Omega_B h^2} \left(\frac{z}{1000} \right)^{12.75}. \tag{87}$$

The scaling with Ω and h has a marvelous consequence. If we work out the optical depth to Thomson scattering, $\tau = \int n_e x \sigma_{\rm T} dr_{\rm prop}$, we find just

$$\boxed{\tau(z) = 0.37 \left(\frac{z}{1000} \right)^{14.25},} \tag{88}$$

independent of cosmological parameters. The rate equation causes $x(z)$ to scale in just the right way that the optical depth is a completely robust quantity. Because τ changes rapidly with redshift, the distribution 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 1065 and standard deviation in redshift 80. Thus, when we look at the sky, we can expect to see in all directions photons that originate from a **last-scattering surface** at $z \simeq 1065$. This independence of parameters is not quite exact in detail, however, and very accurate work needs to solve the evolution equations exactly (*e.g.* appendix C of Hu & Sugiyama 1995).

The microwave background In a famous piece of serendipity, the redshifted radiation from the last-scattering photosphere was detected as a 2.73 K microwave background by Penzias & Wilson (1965). Since the initial detection of the microwave background at $\lambda = 7.3$ cm, measurements of the spectrum have been made over an enormous range of wavelengths, from the depths of the Rayleigh–Jeans regime at 74 cm to well into the Wien tail at 0.5 mm. The most accurate measurements come from **COBE** – the NASA cosmic background explorer satellite. Early data showed the spectrum to be very close to a pure Planck function (Mather *et al.* 1990), and the final result verifies the lack of any distortion with breathtaking precision. The COBE temperature measurement and 95% confidence range of

$$\boxed{T = 2.728 \pm 0.004\,{\rm K}} \tag{89}$$

improves significantly on the ground-based experiments. The lack of distortion in the shape of the spectrum is astonishing, and limits the chemical potential to $|\mu| < 9 \times 10^{-5}$ (Fixsen *et al.* 1996). These results also allow the limit $y \lesssim 1.5 \times 10^{-5}$ to be set on the Compton-scattering distortion parameter. These limits are so stringent that many competing cosmological models can be eliminated.

3.3 Primordial nucleosynthesis

At sufficiently early times, the temperature of the universe reaches the point where nuclear reactions can occur ($T \sim 10^9$K). The abundance of light elements that results from these early reactions is fixed by an argument that can be outlined quite simply. In equilibrium, the numbers of neutrons and protons should vary as

$$\frac{N_n}{N_p} = e^{-\Delta mc^2/kT} \simeq e^{-1.5(10^{10}\,\mathrm{K}/T)}. \tag{90}$$

The reason that neutrons exist today is that the timescale for the weak interactions needed to keep this equilibrium set up eventually becomes longer than the expansion timescale. The reactions thus rapidly cease, and the neutron–proton ratio undergoes **freeze-out** at some characteristic value. In practice this occurs at $N_n/N_p \simeq 1/6$. If most of the neutrons ended up in ^{4}He, we would then expect 25% He by mass – which is very nearly what we see. One of the critical calculations in cosmology is therefore to calculate this freeze-out process in detail. As the following outline of the analysis will show, the result is due to a complex interplay of processes and the fact that there is a significant primordial abundance of anything other than hydrogen is a consequence of a number of coincidences.

Neutron freeze-out The only nuclear reactions that matter initially are the weak interactions that convert between protons and neutrons:

$$\begin{aligned} p + e &\leftrightarrow n + \nu \\ p + \bar{\nu} &\leftrightarrow n + e^+. \end{aligned} \tag{91}$$

The main systematics of the relic abundances of the light elements can be understood by looking at the neutron to proton ratio and how it evolves.

The number density of neutrons obeys the kinetic equation

$$\frac{d\,n_n}{dt} = (\Lambda_{pe} + \Lambda_{p\nu})\,n_p - (\Lambda_{ne} + \Lambda_{n\nu})\,n_n, \tag{92}$$

where the rate coefficients Λ_i refer to the four possible processes given above. To these two-body processes should also be added the spontaneous decay of the neutron, which has the e-folding lifetime of

$$\boxed{\tau_n = 887 \pm 2\ \mathrm{s}} \tag{93}$$

(according to the Particle Data Group). We neglect this for now, as it turns out that neutron freeze-out happens at slightly earlier times.

The quantum-field calculation of the rate coefficients is not difficult, because of the simple form of the Fermi Lagrangian for the weak interaction. Recall that this is proportional to the Fermi constant (G_{F}) times the fields of the various particles that participate, and that each field is to be thought of as a sum of creation and annihilation operators. This means that the matrix elements for all processes where various particles change their occupation numbers by one are the same, and the rate coefficients differ only by virtue of integration over states for the particles involved.

For example, the rate for neutron decay is

$$\tau_n^{-1} \propto G_{\mathrm{F}}^2 \left(\frac{1}{[2\pi\hbar]^3}\right)^2 2 \int d^3p_e\, d^3p_\nu\, \delta(\epsilon_e + \epsilon_\nu - Q), \tag{94}$$

where Q is the neutron–proton energy difference. This expression can be more or less written down at sight. The delta function expresses conservation of energy and as usual arises automatically from integration over space, rather than having to be put in by hand. The factor of 2 before the integral allows for electron helicity, but there is no need to be concerned with the overall constant of proportionality. By performing the integral, this expression can be reduced to

$$\tau_n^{-1} \propto \frac{G_{\rm F}^2}{2\pi^4}\, 1.636\, m_e^5 \quad \text{(natural units)}. \tag{95}$$

The reason why the overall constant of proportionality is not needed is that all other related weak processes have rates of the same form. The only difference is that, unlike the free decay in a vacuum just analysed, we need to include the probabilities that the initial and final states are occupied. For example, in $n + \nu \to p + e$, we need a factor n_ν for the initial neutrino state and $1 - n_e$ for the final electron state (effectively to allow for the fermionic equivalent of stimulated plus spontaneous emission; if the particles involved were bosons, this would become $1 + n$). The rate for this process is therefore

$$\Lambda_{n\nu} = [1.636 m_e^5 \tau_n]^{-1} \int_0^\infty \frac{p_e\, \epsilon_e\, p_\nu^2\, dp_\nu}{[1 + \exp(p_\nu/kT)]\,[1 + \exp(-\epsilon_e/kT)]}, \tag{96}$$

where $\epsilon_e = p_\nu + Q$. Very similar integrals can be written down for the other processes involved.

Since $Q \sim m_e c^2$, it is clear that at high temperatures $kT \gg m_e c^2$, all the rate coefficients will be of the same form; both p_e and ϵ_e can be replaced by p_ν in top and bottom of the integral, leaving a single rate coefficient to be determined by numerical integration:

$$\boxed{\Lambda = 13.893\ \tau_n^{-1} \left(\frac{kT}{m_e c^2}\right)^5 = \left(\frac{10^{10.135}\,\mathrm{K}}{T}\right)^{-5}\ \mathrm{s}^{-1}.} \tag{97}$$

Since the number density of the thermal background of neutrinos and electrons is proportional to T^3, this says that the effective cross-section for these weak interactions scales proportional to [energy]^2. The radiation-dominated era has

$$t = \sqrt{\frac{3}{32\pi G\rho}} = \left(\frac{10^{10.125}\,\mathrm{K}}{T}\right)^2\ \mathrm{s}, \tag{98}$$

allowing for three massless neutrinos. The T^{-2} dependence of the expansion timescale is much slower than the interaction timescale, which changes as T^{-5}, so there is a quite sudden transition between thermal equilibrium and freeze-out, suggesting that weak interactions switch off, freezing the neutron abundance at a temperature of $T \simeq 10^{10.142}\,\mathrm{K}$. This implies an equilibrium neutron–proton ratio of

$$\frac{N_n}{N_p} = e^{-Q/kT} = e^{-10^{10.176\,\mathrm{K}}/T} \simeq 0.34. \tag{99}$$

This obviously cannot be a precisely correct result, because the freeze-out condition was calculated assuming a temperature well above the electron mass threshold, whereas it appears that freeze-out actually occurs at about this critical temperature. The rate Λ is in fact a little larger at the threshold than the high-temperature extrapolation would suggest, so the neutron abundance is in practice lower.

Neutrino freeze-out There is a yet more serious complication, because another important process that occurs around the same time is neutrino decoupling. The weak reaction that keeps neutrino numbers in equilibrium at this late time is $\nu + \bar{\nu} \leftrightarrow e^+ + e^-$. The rate for this process can be found in exactly the same way as above. At high energies, the result is identical, save only that the squared matrix element involved is smaller by a factor of about 5 because of the axial coupling of nucleons: the neutrino rate scales as $G_{\rm F}^2$, as opposed to $G_{\rm F}^2(1 + 2g_{\rm A}^2)$ for nucleons. This leads to the neutrinos decoupling at a slightly higher temperature:

$$\boxed{T(\nu \text{ decoupling}) \simeq 10^{10.5} \text{ K}.} \tag{100}$$

This is uncomfortably close to the electron mass threshold, but just sufficiently higher that it is not a bad approximation to say that all e^+e^- annihilations go into photons rather than neutrinos. During the time at which nucleons are decoupling, the neutrino and photon temperatures are therefore becoming different, and a detailed calculation must account for this. The resulting freeze-out temperature is very close to 10^{10} K, at which point the neutron-to-proton ratio is about 1:3.

Unfortunately, we are still not finished, because neutrons are not stable. It does not matter what abundance of them freezes out: unless they can be locked away in nuclei before $t = 887$ s, the relic abundance will decay freely to zero. The freeze-out point occurs at an age of a few seconds, so there are only a few e-foldings of expansion available in which to salvage some neutrons. So far, a remarkable sequence of coincidences has been assembled, in that the freeze-out of neutrinos and nucleons happens at about the same time as $e^+ - e^-$ annihilation, which is also a time of order τ_n. It may seem implausible that we can add one more – *i.e.* that nuclear reactions will become important at about the same time – but this is just what does happen.

Construction of nucleons This coincidence is not surprising, since the deuteron binding energy of 2.225 MeV is only 4.3 times larger than $m_e c^2$ and only 1.7 times larger than the neutron–proton mass difference. At higher temperatures, the strong interaction $n + p = {\rm D} + \gamma$ is fast enough to produce deuterium, but thermal equilibrium favours a small deuterium fraction – *i.e.* typical photons are energetic enough to disrupt deuterium nuclei very easily. Furthermore, because of the large photon-to-baryon ratio, the photons can keep the deuterium from forming until the temperature has dropped well below the binding energy of the deuteron. The equilibrium abundance of deuterium is set in much the same way as the ionization abundance of hydrogen, and so obeys an equation that is identical (apart from spin-degeneracy factors) to the Saha equation for hydrogen ionization:

$$\frac{n_{\rm D}}{n_p n_n} = \frac{3}{4} \frac{(2\pi\hbar)^3}{(2\pi k T m_p m_n / m_{\rm D})^{3/2}} \exp(\chi/kT), \tag{101}$$

where χ is the binding energy. As with hydrogen ionization, this defines an abrupt transition between the situation where deuterium is rare and where it dominates the equilibrium. The terms outside the exponential keep the deuterium density low until $kT \ll \chi$: the $n_{\rm D} = n_n$ crossover occurs at

$$\boxed{T_{\rm deuteron} \simeq 10^{8.9} \text{ K},} \tag{102}$$

or a time of about 3 minutes. An exact integration of the weak-interaction kinetic equation for

the neutron abundance at $n_{\rm D} = n_n$ (including free neutron decay, which is significant) gives

$$\boxed{\frac{n_n}{n_p} \simeq 0.163\ (\Omega_{\rm B} h^2)^{0.04}\ (N_\nu/3)^{0.2}} \tag{103}$$

(see *e.g.* chapter 4 of Kolb & Turner 1990; chapter 6 of Peebles 1993). The dependences on the baryon density and on the number of neutrino species are easily understood. A high baryon density means that the Saha equation gives a higher deuterium abundance, increasing the temperature at which nuclei finally form and giving a higher neutron abundance because fewer of them have decayed. The effect of extra neutrino species is to increase the overall rate of expansion, so that neutron freeze-out happens earlier, again raising the abundance.

The primordial helium abundance The argument so far has produced a universe consisting of just hydrogen and deuterium, but this is not realistic, as one would expect ^{4}He to be preferred on thermodynamic grounds, owing to its greater binding energy per nucleon (7 MeV, as opposed to 1.1 MeV for deuterium). In practice, the production of helium must await the synthesis of significant quantities of deuterium, which we have seen happens at a temperature roughly one-third that at which helium would be expected to dominate. What the thermodynamic argument does show, however, is that it is expected that the deuterium will be rapidly converted to helium once significant nucleosynthesis begins. This argument is what allows us to expect that the helium abundance can be calculated from the final n/p ratio. If all neutrons are incorporated into ^{4}He, then the number density of hydrogen is set by the remaining protons: $n_{\rm H} = n_p - n_n$. The mass fraction of helium, Y, is unity minus the hydrogen fraction, so that

$$\boxed{Y = 1 - \frac{n_p - n_n}{n_p + n_n} = 2\left(1 + \frac{n_p}{n_n}\right)^{-1}.} \tag{104}$$

For the earlier n/p ratio of 0.163, this gives $Y = 0.28$.

The 'observed' value of Y is in the region of $Y = 0.22$ to 0.23 (*e.g.* Pagel 1994), and there exists something of a difference of opinion on whether this is marvelously close agreement, or evidence for something seriously wrong with the standard model.

The number of particle generations Increasing the number of neutrino species widens the gap between theory and observation by $\Delta Y \simeq 0.01$ for each additional neutrino species. It is therefore clear that strong limits can be set on the number of unobserved species, and thus on the number of possible additional families in particle physics. For many years, these nucleosynthesis limits were stronger than those that existed from particle physics experiments. This changed in 1990, with a critical series of experiments carried out in **LEP**, which was the first experiment to produce Z^0 particles in large numbers. The particles are not seen directly, but their presence is inferred by detecting a peak in the energy-dependent cross-sections for producing pairs of leptons (l) or hadrons (h). The interpretation is that the peak is a 'resonance' due to the production of a Z^0 as an intermediate state, and that the energy of the peak measures the Z^0 mass:

$$e^+ + e^- \rightarrow Z^0 \rightarrow \begin{cases} l, \bar{l} \\ h, \bar{h}. \end{cases} \tag{105}$$

The width of the peak measures the Z^0 lifetime, through the uncertainty principle, and this gives a means of counting the numbers of neutrino species. The Z^0 can decay to pairs of neutrinos so

long as their rest mass sums to less than 91.2 GeV; more species increase the decay rate, and increase the Z^0 width, which measures the total decay rate.

Since 1990, these arguments have required N to be very close to 3 (see the Opal consortium, 1990); it is a matter of detailed argument over the helium data as to whether $N = 4$ was ruled out from cosmology prior to this. In any case, it is worth noting that these two routes do not measure exactly the same thing: both are sensitive only to relativistic particles, with upper mass scales of about 1 MeV and 100 GeV in the cosmological and accelerator cases respectively. If LEP had measured $N = 5$, that would have indicated extra species of rather massive neutrinos. The fact that both limits in fact agree is therefore good evidence for the correctness of the standard model, containing only three families.

Other light-element abundances The same thermodynamic arguments that say that helium should be favoured at temperatures around 0.1 MeV say that more massive nuclei would be preferred in equilibrium at lower temperatures. However, by the time helium synthesis is accomplished, the density and temperature are too low for significant synthesis of heavier nuclei to proceed: the lower density means that reactions tend to freeze out, even for a constant cross-section, and the need for penetration of the nuclear Coulomb barrier means that cross-sections decline rapidly as the temperature decreases.

Apart from helium, the main nuclear residue of the big bang is therefore those deuterium nuclei that escape being mopped up into helium, plus a trace of ^{3}He, which is produced en route to ^{4}He: $\mathrm{D} + p \to {}^3\mathrm{He}$, followed by ${}^3\mathrm{He} + n \to {}^4\mathrm{He}$ (the alternative route, of first $\mathrm{D} + n$, then p also happens, but the intermediate tritium is not so strongly bound). There also exist extremely small fractions of other elements: ^{7}Li ($\sim 10^{-9}$ by mass) and ^{7}Be ($\sim 10^{-11}$). Unlike helium, the critical feature of these abundances is that they are rather sensitive to density. One of the major achievements of big bang cosmology is that it can account simultaneously for the abundances of H, ^{2}D, ^{3}He, ^{4}He and ^{7}Li – but *only for a low-density universe*. A proper understanding of the abundances really requires a numerical solution of the coupled rate equations for all the nuclear reactions, putting in the temperature-dependent cross-sections. This careful piece of numerical physics was first carried out impressively soon after the discovery of the microwave background, by Wagoner, Fowler & Hoyle (1967). At least the sense of the answer can be understood intuitively, however. We have seen that helium formation occurs at very nearly a fixed temperature, depending only weakly on density or neutrino species. The residual deuterium will therefore freeze out at about this temperature, leaving a number density fixed at whatever sets the reaction rate low enough to survive for a Hubble time. Since this density is a fixed quantity, the *proportion* of the baryonic density that survives as deuterium (or ^{3}He) should thus decline roughly as 1/(density).

This provides a relatively sensitive means of weighing the baryonic content of the universe. A key event in the development of cosmology was thus the determination of the D/H ratio in the interstellar medium, carried out by the COPERNICUS UV satellite in the early 1970s (Rogerson & York 1973). This gave $\mathrm{D/H} \simeq 2 \times 10^{-5}$, providing the first evidence for a low baryonic density, as follows. Figure 3 shows how the abundances of light elements vary with the cosmological density, according to detailed calculations. The baryonic density in these calculations is traditionally quoted in the field as the reciprocal of the entropy per baryon:

$$\boxed{\eta \equiv (n_p + n_n)/n_\gamma = 2.74 \times 10^{-8} (T/2.73\,\mathrm{K})^{-3} \Omega_\mathrm{B} h^2.} \qquad (106)$$

Figure 3 shows that this deuterium abundance favours a low density, $\Omega_\mathrm{B} h^2 \simeq 0.02$, and data on other elements give answers close to this. The constraint obtained from a comparison between

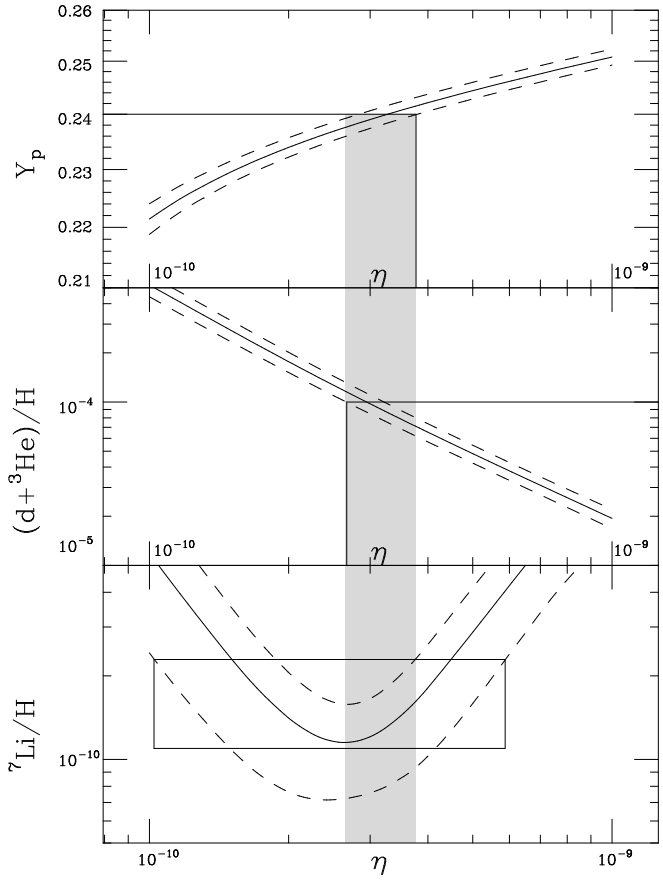

Fig. 3: The predicted primordial abundances of the light elements, as a function of the baryon-to-photon ratio η (Smith, Kawano & Malaney 1993). For a microwave-background temperature of 2.73 K, this is related to the baryonic density parameter via $\Omega_{\rm B}h^2 = \eta/2.74\times10^{-8}$. Concordance with the data is found only for $\eta \simeq 3\times10^{-10}$, shown by the shaded strip.

nucleosynthesis predictions and observational data is rather tight:

$$\boxed{0.010 \lesssim \Omega_{\rm B}h^2 \lesssim 0.015} \tag{107}$$

(*e.g.* Walker *et al.* 1991; Smith, Kawano & Malaney 1993). This comparison is of course non-trivial, because we can only observe abundances in present-day stars and gas, rather than in primordial material. Nevertheless, making the best allowances possible for production or destruction of light elements in the course of stellar evolution, the conclusions obtained from different species are in remarkable agreement: baryons cannot close the universe. If $\Omega = 1$, the dark matter must be non-baryonic.

3.4 Baryogenesis

In discussing nucleosynthesis, we have taken it for granted that photons outnumber baryons in the universe by a large factor. Since baryons and antibaryons will annihilate at low temperatures, this is reasonable; but in that case why are there any baryons at all? A thermal background at high enough temperatures will contain equal numbers of protons and antiprotons, and this symmetry will be maintained as the particles annihilate. As usual, there would be some low

frozen-out abundance of both species at late times, but the universe would display a **matter-antimatter symmetry**. A recurring theme in cosmological debate has been to ask what happened to the antimatter. It appears 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? One attractive idea is that the matter-antimatter asymmetry may have been set in the GUT era, before the universe cooled below the critical temperature of $\sim 10^{15}$ GeV.

The main idea to be exploited is that GUTs erase the distinction between baryons and leptons, treating them as different states of the same underlying unity. This raises the conceptual possibility of reactions that can generate a net baryon asymmetry, so long as the temperature is high enough that the GUT symmetry is not broken. The particles that will be involved are the gauge bosons of the GUT, since these are the ones that mediate the baryon $\leftrightarrow$ lepton exchange reactions – *e.g.* the X & Y bosons of $SU(5)$. In particular, decay processes such as

$$X^{4/3} \rightarrow \begin{cases} e^+ + \bar{d} \\ 2u \end{cases} \tag{108}$$

are a promising direct source of baryon-number violation.

These mechanisms provide the first of three general **Sakharov conditions** for baryosynthesis (published in 1967, well before the invention of GUTs):

(1) $\Delta B \neq 0$ reactions; (2) CP violation; (3) non-equilibrium conditions.

The second condition requires an asymmetry between particles and antiparticles. Recall what the symmetries C and P mean: a given observed reaction is possible (and will proceed with the same rate) if particles are replaced by antiparticles (C) or viewed in a mirror (P). Although P is violated by the weak interaction to the extent that only left-handed neutrinos are produced, the combined symmetry CP is obeyed in almost all cases. Consider the effect on the above X-boson decay if CP were to hold exactly: if a fraction f of decays produce $e^+\bar{d}$, then decays of $\bar{X}$ will produce the opposite baryon number via e^-d the same fraction of the time, in which case no net asymmetry can be created. What we need is for the two fractions to be different, and such a process is observed in the laboratory in the form of the neutral kaons: both K_0 and $\overline{K}_0$ decay to either two or three pions, but with branching ratios that differ at the 10^{-3} level.

The third condition is necessary in order to prevent reverse reactions from erasing any baryon asymmetry as soon as it has been created. As in many cases in the expanding universe, the crucial physical results are contained in the ability of reactions to freeze out.

The challenge of baryosynthesis is to predict the observed asymmetry (in the form of a baryon-to-photon ratio $n_{\rm B}/n_\gamma \sim 10^{-9}$. In principle, this can be done once the GUT model is given, and a connection can be made between laboratory measurements of CP violation and the baryon content of the universe. The simplest model for baryosynthesis would consist of a single Majorana particle, whose decays favour the production of baryons over antibaryons:

$$\begin{aligned} \Gamma(X \rightarrow \Delta B = +1) &= \tfrac{1}{2}(1+\epsilon)\,\Gamma \\ \Gamma(X \rightarrow \Delta B = -1) &= \tfrac{1}{2}(1-\epsilon)\,\Gamma. \end{aligned} \tag{109}$$

Here, $\Gamma = 1/\tau$ is the decay rate, and ϵ parameterizes the CP violation. The kinetic equations governing the effect of decays on the X number density and on baryon number B can be written down immediately following our earlier discussion of the Boltzmann equation:

$$\begin{aligned} \dot{n}_X + 3Hn_X &= -\Gamma\,(n_X - n_X^T) \\ \dot{n}_{\rm B} + 3Hn_{\rm B} &= \epsilon\Gamma\,(n_X - n_X^T). \end{aligned} \tag{110}$$

The terms involving the thermal-equilibrium X density, n_X^T, allow for inverse processes, and are deduced by asking what source term makes the lhs vanish in equilibrium, as before. What they say is that it is impossible for the X boson to decay when it is relativistic; we have to wait until $kT < mc^2$, so that n_X^T is suppressed. Note that the second equation explicitly makes clear the third Sakharov criterion for a violation of equilibrium. These equations are incomplete, as they neglect two-body processes that would contribute to the changing X abundance, such as $X + \bar{X}$ annihilation. The simplified form will apply after the X abundance has frozen out. In terms of comoving densities $N \equiv a^3 n$, the equation for baryon number is just

$$\dot{N}_{\rm B} = -\epsilon \dot{N}_X \quad \Rightarrow \quad N_{\rm B} = \epsilon (N_X^{\rm init} - N_X), \tag{111}$$

so that the final baryon comoving density tends to $\epsilon N_X^{\rm init}$ as the X's decay away. The baryon-to-entropy ratio produced by this process is then

$$\frac{N_{\rm B}}{s} \simeq \frac{\epsilon}{g_*} \exp\left(-\frac{m_X c^2}{kT_f}\right), \tag{112}$$

where the entropy density is defined here as g_* times the photon density, and the last term allows for the freeze-out suppression of the X density relative to massless backgrounds. Since the required ratio is $\sim 10^{-9}$ and $g_* \sim 100$ at early times, we need $\epsilon \gtrsim 10^{-7}$. Note that the freeze-out point cannot be very late: even for $\epsilon = 1$, $kT \gtrsim m_X c^2/16$ is needed.

Other mechanisms Baryosynthesis via GUT decay as above is the simplest mechanism, but there are other possibilities. First note that the above picture may well be inconsistent with an inflationary origin for the universe. Inflation generally involves a GUT-scale phase transition that leaves the universe reheated to a temperature somewhat below that of the GUT scale. Any pre-existing baryon asymmetry would be rendered irrelevant by the inflationary expansion, and things would not be hot enough afterwards for GUT processes to operate.

It is possible that baryosynthesis may proceed at still lower temperatures, since baryon non-conserving processes may even occur as part of the electroweak phase transition at $T \sim 200\,{\rm GeV}$. This is a surprise, since the electroweak Lagrangian contains no terms that would violate baryon number: leptons and hadrons are explicitly contained in different multiplets. This constraint may possibly be evaded by quantum tunnelling, but the exact extent to which baryon number violation may be realized in practice within the standard model is still a matter of debate (see *e.g.* section 6.8 of Kolb & Turner 1990; Grigoriev *et al.* 1992; Moore 1996; Ellis 1997).

4 INFLATIONARY COSMOLOGY

The standard isotropic cosmology is a very successful framework for interpreting observations, but prior to the early 1980s there were certain questions that had to be avoided. The initial conditions of the big bang appear to be odd in a number of ways; these puzzles are encapsulated in a set of classical 'problems', as follows.

The horizon problem Standard cosmology contains a particle horizon of comoving radius

$$r_{\rm H} = \int_0^t \frac{c\,dt}{R(t)}, \tag{113}$$

which converges because $R \propto t^{1/2}$ in the early radiation-dominated phase. At late times, the integral is largely determined by the matter-dominated phase, for which

$$D_{\rm H} = R_0\, r_{\rm H} \simeq \frac{6000}{\sqrt{\Omega z}}\, h^{-1}\,{\rm Mpc}. \tag{114}$$

The horizon at last scattering ($z \sim 1000$) was thus only ~ 100 Mpc in size, subtending an angle of about 1 degree. Why then are the large number of causally disconnected regions we see on the microwave sky all at the same temperature?

The flatness problem The $\Omega = 1$ universe is unstable:

$$[1 - 1/\Omega(z)] = f(z)\,[1 - 1/\Omega], \tag{115}$$

where $f(z) = (1+z)^{-1}$ in the matter-dominated era and $f(z) \propto (1+z)^{-2}$ for radiation domination, so that $f(z) \simeq (1+z_{\rm eq})/(1+z)^2$ at early times. To get $\Omega \simeq 1$ today requires a **fine tuning of Ω** in the past, which becomes more and more precisely constrained as we increase the redshift at which the initial conditions are presumed to have been imposed. Ignoring annihilation effects, $1 + z = T_{\rm init}/2.7$ K and $1 + z_{\rm eq} \simeq 10^4$, so that the required fine tuning is

$$|\Omega(t_{\rm init}) - 1| \lesssim 10^{-22}\,(E_{\rm init}/{\rm GeV})^2. \tag{116}$$

At the Planck epoch, which is the natural initial time, this requires a deviation of only 1 part in 10^{60}.

The expansion problem Even the most obvious fact of the cosmological expansion is unexplained. Although general relativity forbids a static universe, this is not enough to understand the expansion. As shown above, the gravitational dynamics of the cosmological scale factor $R(t)$ are just those of a cannonball travelling vertically in the Earth's gravity. Suppose we see a cannonball rising at a given time $t = t_0$: it may be true to say that it has $r = r_0$ and $v = v_0$ at this time because at a time Δt earlier it had $r = r - v_0\Delta t$ and $v = v_0 - g\Delta t$, but this is hardly a satisfying explanation for the motion of a cannonball that was in fact fired by a cannon. Nevertheless, this is the only level of explanation that classical cosmology offers: the universe expands now because it did so in the past. Although it is not usually included in the list, one might thus with justice add an 'expansion problem' as perhaps the most fundamental in the catalogue of classical cosmological problems.

For many years, it was assumed that any solution to these difficulties would have to await a theory of quantum gravity. The classical singularity can be approached no closer than the Planck time of $\sim 10^{-43}$ s, and so the initial conditions for the classical evolution following this time must have emerged from behind the presently impenetrable barrier of the quantum gravity epoch. There remains a significant possibility that this policy of blaming everything on quantum gravity may be correct, but the great development of cosmology in the 1980s was the realization that the explanation of the initial-condition puzzles might involve physics at lower energies: 'only' 10^{15} GeV. Although this idea, now known as inflation, cannot be considered to be firmly established, the ability to treat gravity classically puts the discussion on a much less speculative foundation. What has emerged is a general picture of the early universe that has compelling simplicity, which moreover may be subject to observational verification. What follows is an outline of the main features of inflation; for more details see *e.g.* chapter 8 of Kolb & Turner (1990); Brandenberger (1990); Liddle & Lyth (1993).

Equation of state for inflation The list of problems with conventional cosmology provides a strong hint that the equation of state of the universe may have been very different at very early times. 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 $\alpha > 1$. If such a phase had existed, the integral for the comoving horizon would have diverged, and there would be no difficulty in understanding the overall homogeneity of the

universe – this could then be established by causal processes. Indeed, it is tempting to assert that the observed homogeneity *proves* that such causal contact must once have occurred. This phase of accelerated expansion is the most general feature of what has become known as the **inflationary universe**.

What condition does this place on the equation of state? In the integral for $r_{\rm H}$, we can replace dt by $dR/\dot{R}$, which the Friedmann equation says is $\propto dR/\sqrt{\rho R^2}$ at early times. Thus, the horizon diverges provided the equation of state is such that ρR^2 vanishes or is finite as $R \to 0$. For a perfect fluid with $p \equiv (\Gamma - 1)\epsilon$ as the relation between pressure and energy density, we have the adiabatic dependence $p \propto R^{-3\Gamma}$, and the same dependence for ρ if the rest-mass density is negligible. A period of inflation therefore needs

$$\boxed{\Gamma < 2/3 \quad \Rightarrow \quad \rho c^2 + 3p < 0.} \tag{117}$$

An alternative way of seeing that this criterion is sensible is that the 'active mass density' $\rho + 3p/c^2$ then vanishes. Since this quantity forms the rhs of Poisson's equation generalized to relativistic fluids, it is no surprise that the vanishing of $\rho + 3p/c^2$ allows a coasting solution with $R \propto t$.

Such a criterion can also solve the flatness problem. Consider the Friedmann equation,

$$\dot{R}^2 = \frac{8\pi G \rho R^2}{3} - kc^2. \tag{118}$$

As we have seen, the density term on the rhs must exceed the curvature term by a factor of at least 10^{60} at the Planck time, and yet a more natural initial condition might be to have the matter and curvature terms being of comparable order of magnitude. However, an inflationary phase in which ρR^2 increases as the universe expands can clearly make the curvature term relatively as small as required, provided inflation persists for sufficiently long.

de Sitter space and inflation We have seen that inflation will require an equation of state with negative pressure, and the only familiar example of this is the $p = -\rho c^2$ relation that applies for vacuum energy; in other words, we are led to consider inflation as happening in a universe dominated by a cosmological constant. As usual, any initial expansion will redshift away matter and radiation contributions to the density, leading to increasing dominance by the vacuum term. If the radiation and vacuum densities are initially of comparable magnitude, we quickly reach a state where the vacuum term dominates. The Friedmann equation in the vacuum-dominated case has three solutions:

$$R \propto \begin{cases} \sinh Ht & (k=-1) \\ \cosh Ht & (k=+1) \\ \exp Ht & (k=0), \end{cases} \tag{119}$$

where $H = \sqrt{\Lambda c^2/3} = \sqrt{8\pi G \rho_{\rm vac}/3}$; all solutions evolve towards the exponential $k = 0$ solution, known as **de Sitter space**. 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.

Because de Sitter space clearly has H^2 and ρ in the right ratio for $\Omega = 1$ (obvious, since $k = 0$), the density parameter in all models tends to unity as the Hubble parameter tends to H. 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

$$\Omega = 1 + O(f^{-2}). \tag{120}$$

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

$$\boxed{\ln f \gtrsim 60} \tag{121}$$

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. There is only a rather small range of e-foldings (60 ± 2, say) around the critical value for which Ω today can be of order unity without it being equal to unity to within the tolerance set by density fluctuations ($\pm 10^{-5}$), and it would constitute an unattractive fine-tuning to require that the expansion hit this narrow window exactly.

This gives the first of two strong **predictions of inflation**: that the universe must be spatially flat

$$\boxed{\text{inflation} \quad \Rightarrow \quad k = 0.} \tag{122}$$

Note that this need not mean the Einstein–de Sitter model; the alternative possibility is that a vacuum contribution is significant in addition to matter, so that $\Omega_m + \Omega_v = 1$. Astrophysical difficulties in finding evidence for $\Omega_m = 1$ are thus one of the major motivations, through inflation, for taking the idea of a large cosmological constant seriously.

4.1 Inflation field dynamics

The general concept of inflation rests on being able to achieve a negative-pressure equation of state. This can be realized in a natural way by quantum fields in the early universe.

Quantum fields at high temperatures The critical fact we shall need from quantum field theory is that quantum fields can produce an energy density that mimics a cosmological constant. The discussion will be restricted to the case of a scalar field ϕ (complex in general, but often illustrated using the case of a single real field). The restriction to scalar fields is not simply for reasons of simplicity, but because the scalar sector of particle physics is relatively unexplored. While vector fields such as electromagnetism are well understood, it is expected in many theories of unification that additional scalar fields such as the Higgs field will exist. We now need to look at what these can do for cosmology.

The Lagrangian density for a scalar field is as usual of the form of a kinetic minus a potential term:

$$\mathcal{L} = \tfrac{1}{2} \partial_\mu \phi \, \partial^\mu \phi - V(\phi). \tag{123}$$

In familiar examples of quantum fields, the potential would be

$$V(\phi) = \tfrac{1}{2} m^2 \, \phi^2, \tag{124}$$

where m is the mass of the field in natural units. However, it will be better to keep the potential function general at this stage. As usual, Noether's theorem gives the energy–momentum tensor for the field as

$$T^{\mu\nu} = \partial^\mu \phi \partial^\nu \phi - g^{\mu\nu} \mathcal{L}. \tag{125}$$

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

$$\begin{aligned} \rho &= \tfrac{1}{2}\dot{\phi}^2 + V(\phi) + \tfrac{1}{2}(\nabla\phi)^2 \\ p &= \tfrac{1}{2}\dot{\phi}^2 - V(\phi) - \tfrac{1}{6}(\nabla\phi)^2. \end{aligned} \tag{126}$$

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.

If ϕ is a (complex) Higgs field, then the symmetry-breaking Mexican hat potential might be assumed:

$$V(\phi) = -\mu^2|\phi|^2 + \lambda|\phi|^4. \tag{127}$$

At the classical level, such potentials determine where $|\phi|$ will be found in equilibrium: at the potential minimum. In quantum terms, this goes over to the **vacuum expectation value** $\langle 0|\phi|0\rangle$. However, these potentials do not include the inevitable fluctuations that will arise in thermal equilibrium. We know how to treat these in classical systems: at non-zero temperature a system of fixed volume will minimize not its potential energy, but the **Helmholtz free energy** $F = V - TS$, S being the entropy. The calculation of the entropy is technically complex, since it involves allowance for quantum interactions with a thermal bath of background particles. However, the main result can be justified, as follows. The effect of the thermal interaction must be to add an interaction term to the Lagrangian $\mathcal{L}_{\rm int}(\phi, \psi)$, where ψ is a thermally fluctuating field that corresponds to the heat bath. In general, we would expect $\mathcal{L}_{\rm int}$ to have a quadratic dependence on $|\phi|$ around the origin: $\mathcal{L}_{\rm int} \propto |\phi|^2$ (otherwise we would need to explain why the second derivative either vanishes or diverges); the coefficient of proportionality will be an effective mass^2 that depends on the thermal fluctuations in ψ. On dimensional grounds, this coefficient must be proportional to T^2, although a more detailed analysis would be required to obtain the constant of proportionality.

There is thus a temperature-dependent **effective potential** that we have to minimize:

$$V_{\rm eff}(\phi, T) = V(\phi, 0) + aT^2|\phi|^2. \tag{128}$$

The effect of this on the symmetry-breaking potential depends on the form of the zero-temperature $V(\phi)$. If the function is taken to be the simple Higgs form $V = -\mu^2 + \lambda\phi^4$, then the temperature-dependent part simply modifies the effective value of μ^2: $\mu^2_{\rm eff} = \mu^2 - aT^2$. At very high temperatures, the potential will be parabolic, with a minimum at $|\phi| = 0$; below the critical temperature, $T_c = \mu/\sqrt{a}$, the ground state is at $|\phi| = [\mu^2_{\rm eff}/(2\lambda)]^{1/2}$ and the symmetry is broken. At any given time, there is only a single minimum, and so this is a second-order phase transition.

It is easy enough to envisage more complicated behaviour, as illustrated in figure 4. This plots the potential

$$V_{\rm eff}(\phi, T) = \lambda|\phi|^4 - b|\phi|^3 + aT^2|\phi|^2, \tag{129}$$

which displays two critical temperatures. At very high temperatures, the potential will have a parabolic minimum at $|\phi| = 0$; at T_1, a second minimum appears in $V_{\rm eff}$ at $|\phi| \neq 0$, and this will be the global minimum for some $T_2 < T_1$. For $T < T_2$, the state at $|\phi| = 0$ is known as the **false vacuum**, whereas the global minimum is known as the **true vacuum**. For this particular form of potential, the second minimum around $\phi = 0$ always exists, so that there is a potential barrier preventing a transition to the false vacuum. This can be overcome by adding a small $-\mu^2|\phi|^2$ component to the potential, so that there will be a third critical temperature at which the curvature around the origin changes sign, leaving only one minimum in the potential.

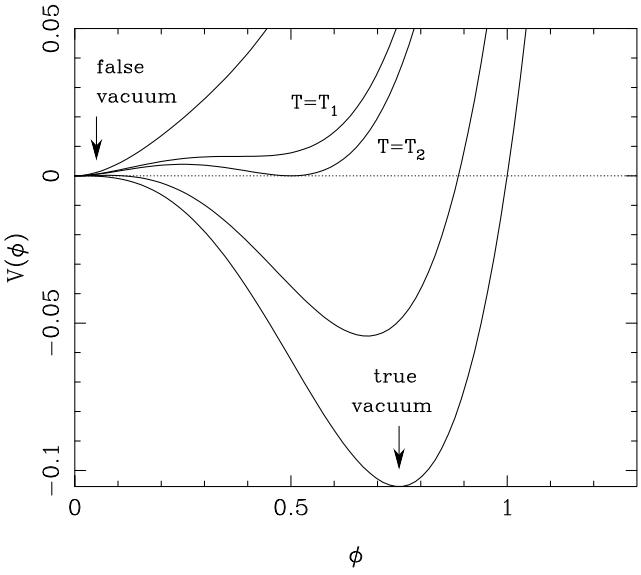

Fig. 4: The temperature-dependent effective potential $V = T^2|\phi|^2 - |\phi|^3 + |\phi|^4$, illustrated at several temperatures: $T^2 = 0.5$, 9/32, 1/4, 0.1, 0. For $T > T_1 = (9/32)^{1/2} \simeq 0.53$, only the false vacuum is available; for $T < T_2 = 1/2$ the true vacuum is energetically favoured and the potential approaches the zero-temperature form.

Alternatively, once the barrier is small enough, quantum tunnelling can take place and free ϕ to move. The universe is no longer trapped in the false vacuum and can make a first-order phase transition to the true vacuum state.

The crucial point to note for cosmology is that there is an energy-density difference between the two vacuum states:

$$\Delta V = \frac{\mu^4}{2\lambda} \tag{130}$$

If we say that the zero of energy is such that $V = 0$ in the true vacuum, this implies that the false-vacuum symmetric state displays an effective cosmological constant. On dimensional grounds, this must be an energy density $\sim m^4$ in natural units, where m is the energy at which the phase transition occurs. For GUTs, $m \simeq 10^{15}$ GeV; in laboratory units, this implies

$$\rho_{\rm vac} = \frac{(10^{15}{\rm GeV})^4}{\hbar^3 c^5} \simeq 10^{80}\ {\rm kg\,m^{-3}}. \tag{131}$$

The inevitability of such a colossal vacuum energy in models with GUT-scale symmetry breaking was the major motivation for the concept of inflation as originally envisaged by Guth (1981). At first sight, the overall package looks highly appealing, since the phase transition from false to true vacuum both terminates inflation and also reheats the universe to the GUT temperature, allowing the possibility that GUT-based reactions that violate baryon-number conservation can generate the observed matter/antimatter asymmetry. Because the transition is first-order, the original inflation model is known as **first-order inflation**.

However, while a workable inflationary cosmology will very probably deploy the three basic elements of vacuum-driven expansion, fluctuation generation and reheating, it has become clear that such a model must be more complex than Guth's initial proposal. To explain where the problems arise, we need to look in more detail at the functioning of the inflation mechanism.

Dynamics of the inflation field Treating the field classically (*i.e.* considering the expectation value $\langle\phi\rangle$, we get from energy–momentum conservation ($T^{\mu\nu}_{;\nu} = 0$) the equation of motion

$$\boxed{\ddot{\phi} + 3H\dot{\phi} - \nabla^2\phi + dV/d\phi = 0.} \tag{132}$$

This can also be derived more easily by the direct route of writing down the action $S = \int \mathcal{L}\sqrt{-g}\ d^4x$ and applying the Euler–Lagrange equation that arises from a stationary action ($\sqrt{-g} = R^3(t)$ for an FRW model, which is the origin of the Hubble drag term $3H\dot{\phi}$).

The solution of the equation of motion becomes tractable if we both ignore spatial inhomogeneities in ϕ and make the **slow-rolling approximation** that $|\ddot{\phi}|$ is negligible in comparison with $|3H\dot{\phi}|$ and $|dV/d\phi|$. Both these steps are required in order that inflation can happen; we have shown above that the vacuum equation of state only holds if in some sense ϕ changes slowly both spatially and temporally. Suppose there are characteristic temporal and spatial scales T and X for the scalar field; the conditions for inflation are that the negative-pressure equation of state from $V(\phi)$ must dominate the normal-pressure effects of time and space derivatives:

$$V \gg \phi^2/T^2, \quad V \gg \phi^2/X^2, \tag{133}$$

hence $|dV/d\phi| \sim V/\phi$ must be $\gg \phi/T^2 \sim \ddot{\phi}$. The $\ddot{\phi}$ term can therefore be neglected in the equation of motion, which then takes the slow-rolling form for homogeneous fields:

$$\boxed{3H\dot{\phi} = -dV/d\phi.} \tag{134}$$

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

$$\boxed{\epsilon \equiv \frac{m_{\rm P}^2}{16\pi}\,(V'/V)^2 \ll 1.} \tag{135}$$

Also, we can differentiate this expression to obtain the criterion $V'' \ll V'/m_{\rm P}$. Using slow-roll once more gives $3H\dot{\phi}/m_{\rm P}$ for the rhs, which is in turn $\ll 3H\sqrt{V}/m_{\rm P}$ because $\dot{\phi}^2 \ll V$, giving finally

$$\boxed{\eta \equiv \frac{m_{\rm P}^2}{8\pi}\,(V''/V) \ll 1} \tag{136}$$

(recall that for de Sitter space $H = \sqrt{8\pi G V(\phi)/3} \sim \sqrt{V}/m_{\rm P}$ in natural units). 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.

Similar arguments can be made for the spatial parts. However, they are less critical: what matters is the value of $\nabla\phi = \nabla_{\rm comoving}\,\phi/R$. 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. This 'stretching' of field gradients as we increase the cosmological horizon beyond the value predicted in classical cosmology also solves a related problem that was historically important in motivating the invention of inflation – the **monopole problem**. Monopoles are point-like topological defects that would be expected to arise in any

phase transition at around the GUT scale ($t \sim 10^{-35}$ s). If they form at approximately one per horizon volume at this time, then it follows that the present universe would contain $\Omega \gg 1$ in monopoles. This unpleasant conclusion is avoided if the horizon can be made much larger than the classical one at the end of inflation; the GUT fields have then been aligned over a vast scale, so that topological-defect formation becomes extremely rare.

Ending inflation Although spatial derivatives of the scalar field can thus be neglected, the same is not always true for time derivatives. Although they may be negligible initially, the relative importance of time derivatives increases as ϕ rolls down the potential and V approaches zero (leaving aside the subtle question of how we know that 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 (see figure 5). 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. This would be a most boring universe to inhabit, but fortunately there is a more realistic way in which inflation can end. We have neglected so far 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. If inflation terminated by reheating to the GUT temperature, then the expansion factor required to reach the present epoch is

$$a_{\rm GUT}^{-1} \simeq E_{\rm GUT}/E_\gamma. \tag{137}$$

The comoving horizon size at the end of inflation was therefore

$$d_{\rm H}(t_{\rm GUT}) \simeq a_{\rm GUT}^{-1}\,[c/H_{\rm GUT}] \simeq [E_{\rm P}/E_\gamma]\,E_{\rm GUT}^{-1}, \tag{138}$$

where the last expression in natural units uses $H \simeq \sqrt{V}/E_{\rm P} \simeq E_{\rm GUT}^2/E_{\rm P}$. For a GUT energy of 10^{15}GeV, this is about 10 m. This is a sobering illustration of the magnitude of the horizon problem; if we relied on causal processes at the GUT era to produce homogeneity, then the universe would only be smooth in patches a few comoving metres across. To solve the problem, we need enough e-foldings of inflation to have stretched this GUT-scale horizon to the present horizon size

$$\boxed{N_{\rm obs} = \ln\left[\frac{3000h^{-1}\,{\rm Mpc}}{(E_{\rm P}/E_\gamma)E_{\rm GUT}^{-1}}\right] \simeq 60.} \tag{139}$$

By construction, this is enough to solve the horizon problem, and it is also the number of e-foldings needed to solve the flatness problem. This is no coincidence, since we saw earlier that the criterion in this case was

$$N \gtrsim \tfrac{1}{2}\,\ln\left(\frac{a_{\rm eq}}{a_{\rm GUT}^2}\right). \tag{140}$$

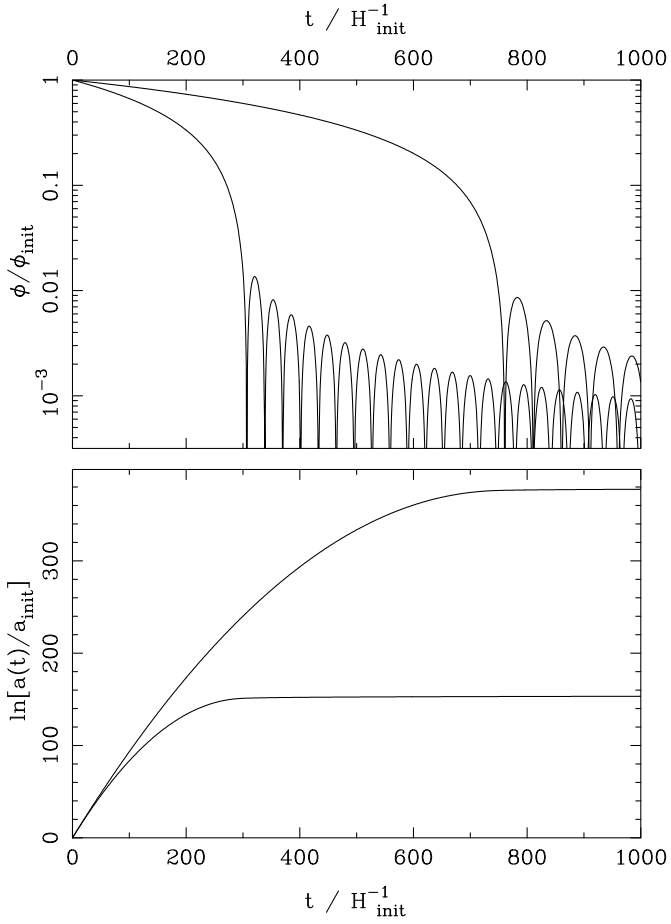

Fig. 5: A plot of the exact solution for the scalar field in a model with a $V \propto \phi^2$ potential. The top panel shows how the absolute value of ϕ falls smoothly with time during the inflationary phase, and then starts to oscillate when inflation ends. The bottom panel shows the evolution of the scale factor. We see the initial exponential behaviour flattening as the vacuum energy ceases to dominate. The two models shown have starting points of $W_i \equiv V_i/(\phi_i^2 H_i^2) = 0.002$ and 0.005; the former (upper lines in each panel) gives about 380 e-foldings of inflation, the latter (lower lines) only 150. According to the $\epsilon = \eta = 1$ criterion, inflation in these models ends at respectively $t = 730$ and 240. The observationally relevant part of inflation is the last 60 e-foldings, and the behaviour of the scale factor is significantly non-exponential in this regime.

Now, $a_{\rm eq} = \rho_\gamma/\rho$, and $\rho = 3H^2\Omega/(8\pi G)$. In natural units, this translates to $\rho \sim E_{\rm P}^2(c/H_0)^{-2}$, or $a_{\rm eq}^{-1} \sim E_{\rm P}^2(c/H_0)^{-2}/E_\gamma^4$. The expression for N is then identical to that in the case of the horizon problem: the same number of e-folds will always solve both.

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

$$N = \int H\, dt = -\frac{8\pi}{m_{\rm P}^2} \int_{\phi_1}^{\phi_2} \frac{V}{V'}\, d\phi \tag{141}$$

For any potential that is relatively smooth, $V' \sim V/\phi$, and so we 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:

$$\boxed{\phi_{\rm start} \gg m_{\rm P}.} \tag{142}$$

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. Counterexamples can of course be constructed, but they have to be somewhat special cases.

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 $\sim m_{\rm P}^4$ 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. In a sense, inflation is realized by means of 'inertial confinement': there is nothing to prevent the scalar field from reaching the minimum of the potential – but it takes a long time to do so, and the universe has meanwhile inflated by a large factor.

4.2 Relic fluctuations from inflation

We have seen that de Sitter space contains a true event horizon, of proper size c/H. This suggests that there will be thermal fluctuations present, as with a black hole, for which the **Hawking temperature** is $kT_{\rm H} = \hbar c/(4\pi r_s)$. This analogy is close, but imperfect, and the characteristic temperature of de Sitter space is a factor 2 higher:

$$kT_{\rm deSitter} = \frac{\hbar H}{2\pi}. \tag{143}$$

This existence of thermal fluctuations is one piece of intuitive motivation for expecting fluctuations in the quantum fields that are present in de Sitter space, but is not so useful in detail. In practice, we need a more basic calculation, which is to see how the zero-point fluctuations in small-scale quantum modes freeze out as classical density fluctuations once the modes have been inflated to super-horizon scales.

The details of this calculation are given below. However, 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 rms fluctuations, σ,

in Newtonian gravitational potential, Φ ($c = 1$):

$$\boxed{\delta_{\rm H}^2 \equiv \Delta_\Phi^2 \equiv \frac{d\,\sigma^2(\Phi)}{d\,\ln k} = \text{constant}.} \tag{144}$$

The notation $\delta_{\rm H}$ arises because the potential perturbation is of the same order as the density fluctuation on the scale of the horizon at any given time.

It is commonly argued that the prediction of scale invariance arises because 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. After inflation ceases, the resulting fluctuations (at constant amplitude on the scale of the horizon) give us the Zeldovich spectrum**Zeldovich** or scale-invariant spectrum**scale-invariant** spectrum. The problem with this argument is that it ignores the issue of how the perturbations evolve while they are outside the horizon; we have only really calculated the amplitude for the last generation of fluctuations – *i.e.* those that are on the scale of the horizon at the time inflation ends. Fluctuations generated at earlier times will be inflated outside the de Sitter horizon, and will re-enter the FRW horizon at some time after inflation has ceased.

The evolution during this period is a topic where some care is needed, since the description of these large-scale perturbations is sensitive to the gauge freedom in general relativity. A technical discussion is given in *e.g.* Mukhanov, Feldman & Brandenberger (1992); for the present, we shall rely on simply motivating the inflationary result, which is that potential perturbations re-enter the horizon with the same amplitude they had on leaving. This may be made reasonable in two ways. Perturbations outside the horizon are immune to causal effects, so it is hard to see how any large-scale non-flatness in spacetime could 'know' whether it was supposed to grow or decline.

We therefore argue that the inflationary process produces a universe that is fractal-like in the sense that scale-invariant fluctuations correspond to a metric that has the same 'wrinkliness' per log length-scale. It then suffices to calculate that amplitude on one scale – *i.e.* the perturbations that are just leaving the horizon at the end of inflation, so that super-horizon evolution is not an issue. It is possible to alter this prediction of scale invariance only if the expansion is non-exponential; we have seen that such deviations plausibly do exist towards the end of inflation, so it is clear that exact scale invariance is not to be expected.

To anticipate the detailed treatment, the inflationary prediction is of a horizon-scale amplitude

$$\boxed{\delta_{\rm H} = \frac{H^2}{2\pi\,\dot{\phi}}} \tag{145}$$

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

$$\delta t = \frac{\delta\phi}{\dot{\phi}}. \tag{146}$$

These universes will then finish inflation at different times, leading to a spread in energy densities (figure 6). The horizon-scale density amplitude is given by the different amounts that the

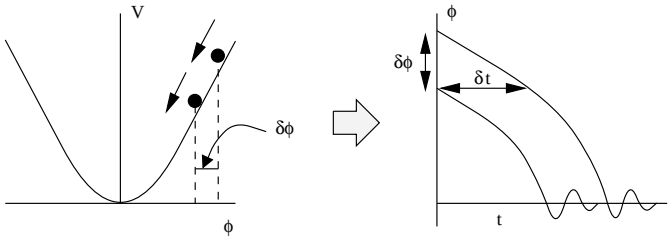

Fig. 6: This plot shows how fluctuations in the scalar field transform themselves into density fluctuations at the end of inflation. Different points of the universe inflate from points on the potential perturbed by a fluctuation $\delta\phi$, like two balls rolling from different starting points. Inflation finishes at times separated by δt in time for these two points, inducing a density fluctuation $\delta = H\delta t$.

universes have expanded following the end of inflation:

$$\delta_{\rm H} \simeq H\ \delta t = \frac{H^2}{2\pi\,\dot\phi}, \tag{147}$$

where the last step uses the crucial input of quantum field theory, which says that the rms $\delta\phi$ is given by $H/2\pi$.

The fluctuation spectrum We now need to go over this vital result in rather more detail (see Liddle & Lyth 1993 for a particularly clear treatment). First, consider the equation of motion obeyed by perturbations in the inflaton field. The basic equation of motion is

$$\ddot\phi + 3H\dot\phi - \nabla^2\phi + V'(\phi) = 0, \tag{148}$$

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:

$$[\ddot{\delta\phi}] + 3H[\dot{\delta\phi}] + (k/a)^2[\delta\phi] = 0, \tag{149}$$

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: the field would be expanded as

$$\phi_k = \omega_k a_k + \omega_k^* a_k^\dagger, \tag{150}$$

and the field variance would be

$$\langle 0|\ |\phi_k|^2\ |0\rangle = |\omega_k|^2. \tag{151}$$

To solve the general problem, we only need to find how the amplitude ω_k changes as the universe expands. The idea is to start from the situation where we are well inside the horizon ($k/a \gg H$), in which case flat-space quantum theory will apply, and end at the point of interest outside the horizon (where $k/a \ll H$).

Returning now to the calculation, we want to know how the mode amplitude changes as the wavelength passes through the horizon. Initially, we have the standard result from flat-space quantum field theory, which can be rewritten in comoving units as

$$\omega_k = a^{-3/2}\,(2k/a)^{-1/2}\,e^{-ikt/a}. \tag{152}$$

The powers of the scale factor, $a(t)$, just allow for expanding the field in comoving wavenumbers k. The field amplitude contains a normalizing factor of $V^{-1/2}$, V being a proper volume; hence the $a^{-3/2}$ factor, if we use comoving $V = 1$. Another way of looking at this is that the proper number density of inflatons goes as a^{-3} as the universe expands. With this boundary condition, it straightforward to check by substitution that the following expression satisfies the evolution equation:

$$\boxed{\omega_k = a^{-3/2}\,(2k/a)^{-1/2}\,e^{-ik/aH}\,(1 + iaH/k)} \tag{153}$$

(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 ω_k 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. In the limit of very early times, the period of oscillation is $\ll H^{-1}$, so a is effectively a constant from the point of view of the epoch where quantum fluctuations dominate.

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

$$\langle 0|\,|\phi_k|^2\,|0\rangle = \frac{H^2}{2k^3}. \tag{154}$$

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 fluctuations in ϕ depend on k in such a way that the fluctuations per decade are constant:

$$\frac{d\,(\delta\phi)^2}{d\,\ln k} = \frac{4\pi k^3}{(2\pi)^3}\,\langle 0|\,|\phi_k|^2\,|0\rangle = \left(\frac{H}{2\pi}\right)^2 \tag{155}$$

(the factor $(2\pi)^{-3}$ comes from the Fourier transform; $4\pi k^2\,dk = 4\pi k^3\,d\ln k$ comes from the k-space volume element). This completes the argument. 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

$$\boxed{\begin{aligned} &\delta_{\rm H}^2 \equiv \Delta_\Phi^2(k) = \frac{H^4}{(2\pi\dot{\phi})^2} \\ &H^2 = \frac{8\pi}{3}\,\frac{V}{m_{\rm P}^2} \\ &3H\dot{\phi} = -V', \end{aligned}} \tag{156}$$

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

This result calls for a number of comments. First, if H and $\dot{\phi}$ are both constant then the predicted spectrum is exactly scale invariant, with some characteristic inhomogeneity on

the scale of the horizon. As we have seen, exact de Sitter space with constant H will not be strictly correct for most inflationary potentials; nevertheless, in most cases the main points of the analysis still go through. The fluctuations in ϕ start as normal flat-space fluctuations (and so not specific to de Sitter space), which change their character as they are advected beyond the horizon and become frozen-out classical fluctuations. All that matters is that the Hubble parameter is roughly constant for the few e-foldings that are required for this transition to happen. If H does change with time, the number to use is the value at the time that a mode of given k crosses the horizon. Even if H were to be precisely constant, there remains the dependence on $\dot{\phi}$, which again will change as different scales cross the horizon. This means that different inflationary models display different characteristic deviations from a nearly scale-invariant spectrum, and this is discussed in more detail below.

Two other characteristics of the perturbations are more general: they will be Gaussian and adiabatic in nature. A Gaussian density field is one for which the joint probability distribution of the density at any given number of points is a multivariate Gaussian. The easiest way for this to arise in practice is for the density field to be constructed as a superposition of Fourier modes with independent random phases; the Gaussian property then follows from the central limit theorem. It is easy to see in the case of inflation that this requirement will be satisfied: the quantum commutation relations only apply to modes of the same k, so that modes of different wavelength behave independently and have independent zero-point fluctuations.

Gravity waves and tilt The density perturbations left behind as a residue of the quantum fluctuations in the inflaton field during inflation are an important relic of that epoch, but are not the only one. In principle, a further important test of the inflationary model is that it also predicts a background of gravitational waves, whose properties couple with those of the density fluctuations.

It is easy to see in principle how such waves arise. In linear theory, any quantum field is expanded in a similar way into a sum of oscillators with the usual creation and annihilation operators; the above analysis of quantum fluctuations in a scalar field is thus readily adapted to show that analogous fluctuations will be generated in other fields during inflation. In fact, the linearized contribution of a gravity wave, $h_{\mu\nu}$, to the Lagrangian looks like a scalar field $\phi = (m_{\rm P}/4\sqrt{\pi})\, h_{\mu\nu}$, the expected rms gravity-wave amplitude is

$$h_{\rm rms} \sim H/m_{\rm P}. \tag{157}$$

The fluctuations in ϕ are transmuted into density fluctuations, but gravity waves will survive to the present day, albeit redshifted.

This redshifting produces a break in the spectrum of waves. Prior to horizon entry, the gravity waves produce a scale-invariant spectrum of metric distortions, with amplitude $h_{\rm rms}$ per $\ln k$. These 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

$$\left(\frac{\delta T}{T}\right)_{\rm GW} \sim h_{\rm rms} \lesssim \delta_{\rm H} \sim 10^{-5}, \tag{158}$$

where the second step follows because the tensor modes can constitute no more than 100% of the observed CMB anisotropy. The energy density of the waves is $\rho_{\rm GW} \sim m_{\rm P}^2 h^2 k^2$, where $k \sim H(a_{\rm entry})$ is the proper wavenumber of the waves. At horizon entry, we therefore expect

$$\rho_{\rm GW} \sim m_{\rm P}^2\, h_{\rm rms}^2\, H^2(a_{\rm entry}). \tag{159}$$

After horizon entry, the waves redshift away like radiation, as a^{-4}, and generate a present-day energy spectrum per $\ln k$ that is constant for modes that entered the horizon while the universe was radiation dominated (because $a \propto t^{1/2} \Rightarrow H^2 a^4 = \text{constant}$). What is the density parameter of these waves? In natural units, $\Omega = (8\pi/3)\rho/(H^2 m_{\rm P}^2)$, so $\Omega_{\rm GW} \sim h_{\rm rms}^2$ at the time of horizon entry, at which epoch the universe was radiation dominated with $\Omega_r = 1$ to an excellent approximation. Thereafter, the wave density maintains a constant ratio to the radiation density, since both redshift as a^{-4}, giving the present-day density as

$$\boxed{\Omega_{\rm GW} \sim \Omega_r \, (H/m_{\rm P})^2 \sim 10^{-4} V/m_{\rm P}^4.} \tag{160}$$

The gravity-wave spectrum therefore displays a break between constant metric fluctuations on super-horizon scales and constant density fluctuations on small scales. An analogous break also exists in the spectrum of density perturbations in dark matter. If gravity waves make an important contribution to CMB anisotropies, we must have $h_{\rm rms} \sim 10^{-5}$, and so $\Omega_{\rm GW} \sim 10^{-14}$ is expected.

An alternative way of presenting the gravity-wave effect on the CMB anisotropies is via the ratio between the tensor effect of gravity waves and the normal scalar Sachs–Wolfe effect, as first analysed in a prescient paper by Starobinsky (1985). Denote the fractional temperature variance per natural logarithm of angular wavenumber by Δ^2 (constant for a scale-invariant spectrum). The tensor and scalar contributions are respectively

$$\Delta_{\rm T}^2 \sim h_{\rm rms}^2 \sim (H^2/m_{\rm P}^2) \sim V/m_{\rm P}^4. \tag{161}$$

$$\Delta_{\rm S}^2 \sim \delta_{\rm H}^2 \sim \frac{H^2}{\dot\phi} \sim \frac{H^6}{(V')^2} \sim \frac{V^3}{m_{\rm P}^6 \, V'^2}. \tag{162}$$

The ratio of the tensor and scalar contributions to the variance of microwave background anisotropies is therefore proportional to the inflationary parameter ϵ:

$$\frac{\Delta_{\rm T}^2}{\Delta_{\rm S}^2} \simeq 12.4 \, \epsilon, \tag{163}$$

inserting the exact coefficient from Starobinsky (1985). If it could be measured, the gravity-wave contribution to CMB anisotropies would therefore give a measure of ϵ, one of the dimensionless inflation parameters. The less 'de Sitter-like' the inflationary behaviour, the larger the relative gravitational-wave contribution.

Since deviations from exact exponential expansion also manifest themselves as density fluctuations with spectra that deviate from scale invariance, this suggest a potential test of inflation. Define the **tilt** of the fluctuation spectrum as follows:

$$\boxed{\text{tilt} \equiv 1 - n \equiv -\frac{d \ln \delta_{\rm H}^2}{d \, \ln k}.} \tag{164}$$

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,

$$a/k = H^{-1}. \tag{165}$$

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

$$\frac{d}{d\,\ln k} = a\,\frac{d}{da} = \frac{\dot\phi}{H}\,\frac{d}{d\phi} = -\frac{m_{\rm P}^2}{8\pi}\,\frac{V'}{V}\,\frac{d}{d\phi}. \tag{166}$$

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

$$\delta_{\rm H}^2 = \frac{H^4}{(2\pi\dot\phi)^2} = \frac{128\pi}{3}\left(\frac{V^3}{m_{\rm P}^6\,V'^2}\right), \tag{167}$$

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

$$\boxed{1-n = 6\epsilon - 2\eta} \tag{168}$$

In the section below on CMB anisotropies, we discuss whether this relation is observationally testable.

5 EVIDENCE FOR VACUUM ENERGY AT LATE TIMES

The idea of inflation is audacious, but undeniably speculative. However, once we accept the idea that quantum fields can generate an equation of state resembling a cosmological constant, we need not confine this mechanism to GUT-scale energies. There is no known mechanism that requires the minimum of $V(\phi)$ to lie exactly at zero energy, so it is quite plausible that there remains in the universe today some non-zero vacuum energy.

The most direct way of detecting vacuum energy has been the immense recent progress in the use of supernovae as standard candles. Type Ia SNe have been used as standard objects for around two decades, with an rms scatter in luminosity of 40%, and so a distance error of 20%. The big breakthrough came when it was realized that the intrinsic timescale of the SNe correlates with luminosity (brighter SNe last longer). Taking out this effect produces corrected standard candles that are capable of measuring distances to about 5% accuracy. Large search campaigns have made it possible to find of order 100 SNe over the range $0.1 \lesssim z \lesssim 1$, and two teams have used this strategy to make an empirical estimate of the cosmological distance-redshift relation.

The results of the *Supernova cosmology project* (e.g. Perlmutter et al. 1998) and the *High-z supernova search* (e.g. Riess et al. 1998) are highly consistent. Figure 7 shows the Hubble diagram from the latter team. The SNe magnitudes are K-corrected, so that their variation with redshift should be a direct measure of luminosity distance as a function of redshift.

We have seen above that this is written as the following integral, which must usually be evaluated numerically:

$$\begin{aligned} D_{\rm L}(z) = (1+z)R_0 S_k(r) = (1+z)\frac{c}{H_0}|1-\Omega|^{-1/2}\times \\ S_k\left[\int_0^z \frac{|1-\Omega|^{1/2}\,dz'}{\sqrt{(1-\Omega)(1+z')^2+\Omega_v+\Omega_m(1+z')^3}}\right], \end{aligned} \tag{169}$$

where $\Omega = \Omega_m + \Omega_v$, and S_k is sinh if $\Omega < 1$, otherwise sin. It is clear from figure 7 that the empirical distance-redshift relation is very different from the simplest inflationary prediction, which is the $\Omega = 1$ Einstein-de Sitter model; by redshift 0.6, the SNe are fainter than expected in this model by about 0.5 magnitudes. If this model fails, we can try adjusting Ω_m and Ω_v in an attempt to do better. Comparing each such model to the data yields the likelihood

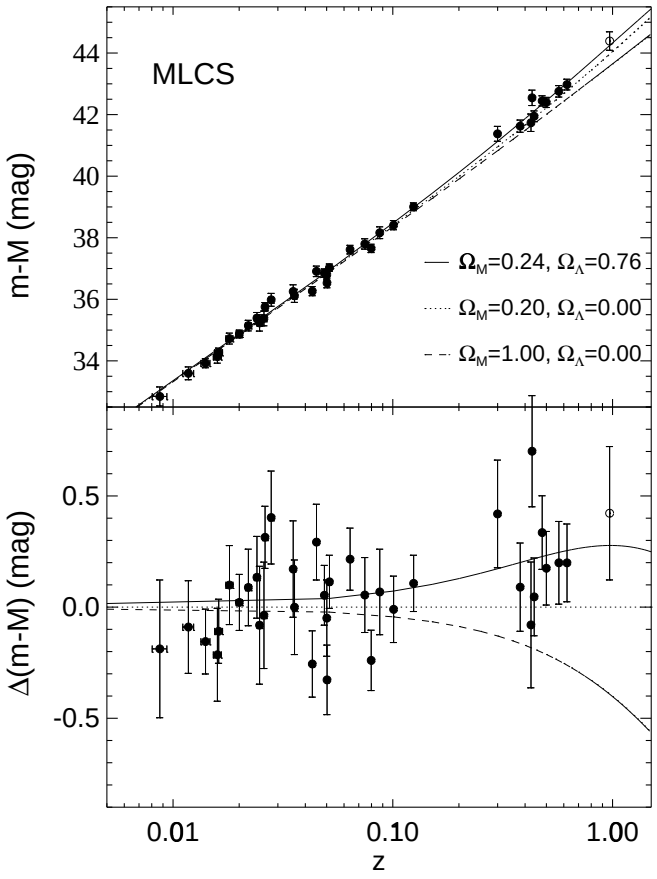

Fig. 7: The Hubble diagram produced by the High-z Supernova search team (Riess et al. 1998). The lower panel shows the data divided by a default model ($\Omega_m = 0.2$, $\Omega_v = 0$). The results lie clearly above this model, favouring a non-zero Λ. The lowest line is the Einstein-de Sitter model, which is in gross disagreement with observation.

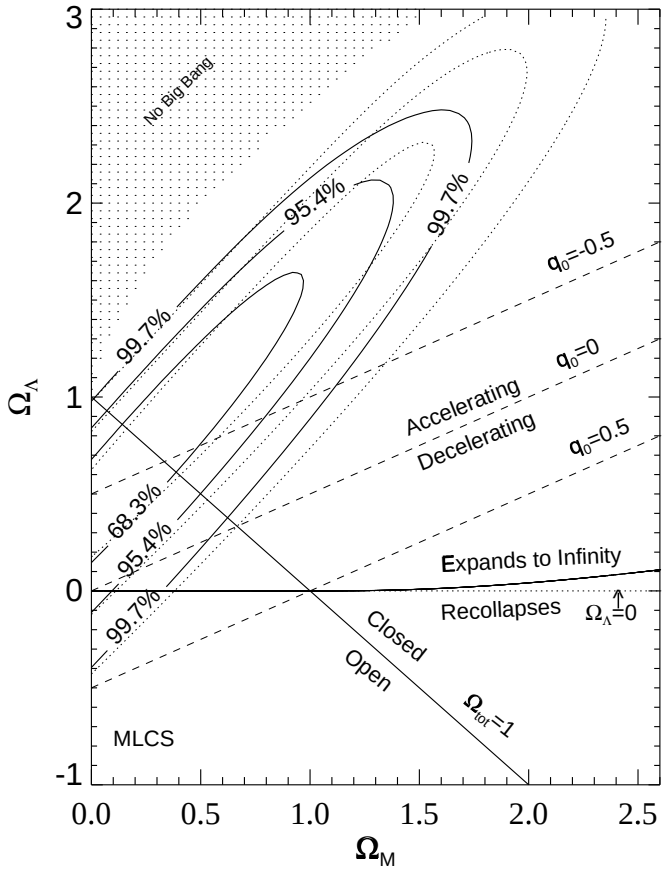

Fig. 8: Confidence contours on the Ω_v–Ω_m plane, according to Riess et al. (1998). Open models of all but the lowest densities are apparently ruled out, and nonzero Λ is strongly preferred. If we restrict ourselves to $k = 0$, then $\Omega_m \simeq 0.3$ is required. The constraints perpendicular to the $k = 0$ line are not very tight, but CMB data can help here in limiting the allowed degree of curvature.

contours shown in figure 8, which can be used in the standard way to set confidence limits on the cosmological parameters. The results very clearly require a low-density universe. For $\Lambda = 0$, a very low density is just barely acceptable, with $\Omega_m \lesssim 0.1$. However, the discussion of the CMB below shows that such a heavily open model is hard to sustain. The preferred model has $\Omega_v \simeq 1$; if we restrict ourselves to the inflationary $k = 0$, then the required parameters are very close to $(\Omega_m, \Omega_v) = (0.3, 0.7)$.

Cosmic coincidence This is an astonishing result – an observational detection of the physical reality of vacuum energy. The error bars continue to shrink, and no convincing systematic error has been suggested that could yield this result spuriously; this is one of the most important achievements of 20th-Century physics.

And yet, 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. One might think of appealing to anthropic ideas, and these can limit Λ to some extent: if the universe became vacuum-dominated at $z > 1000$, gravitational instability as discussed in the next section would have been impossible – so that galaxies, stars and observers would not have been possible. On the other hand, Weinberg (1989)

argues that Λ could have been much larger than its actual value without making observers impossible. Efstathiou (1995) attempted to construct a probability distribution for Λ by taking this to be proportional to the number density of galaxies that result in a given model. However, there is no general agreement on how to set a probability measure for this problem.

It would be more satisfactory if we had some physical mechanism that guaranteed the coincidence, and one possibility has been suggested. 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 $z \sim 10^{80}$ for the GUT era). What is required is a cosmological 'constant' that switches on around the equality era. Zlatev, Wang & Steinhardt (1998) have suggested how this might happen. The idea is to use the vacuum properties of a homogeneous scalar field as the physical origin of the negative-pressure term detected via SNe. This idea of a 'rolling' Λ was first explored by Ratra & Peebles (1988), and there has recently been a tendency towards use of the fanciful term 'quintessence'. In any case, it is important to appreciate that the idea uses exactly the same physical elements that we discussed in the context of inflation: there is some $V(\phi)$, causing the expectation value of ϕ to obey the damped oscillator equation of motion, so the energy density and pressure are

$$\begin{aligned} \rho_\phi &= \dot{\phi}^2/2 + V \\ p_\phi &= \dot{\phi}^2/2 - V. \end{aligned} \tag{170}$$

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. More importantly, we can generalize to the case where the universe contains scalar field and ordinary matter. Suppose the latter dominates, and 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

$$V(\phi) = \frac{2}{\lambda^2}(6/\alpha - 1)\exp[-\lambda\phi]. \tag{171}$$

The scalar-field density is $\rho_\phi = (\alpha/\lambda^2)\rho_{\rm total}$ (see e.g. Liddle & Scherrer 1998). 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 suggest two potentials which might achieve this:

$$V(\phi) = M^{4+\beta}\phi^{-\beta} \quad \text{or} \quad \mathrm{V}(\phi) = \mathrm{M}^4[\exp(\mathrm{m_P}/\phi) - 1]. \tag{172}$$

The evolution in these potentials may be described by $w(t)$, where $w = p/\rho$. We need $w \simeq 1/3$ in the radiation era, changing to $w \simeq -1$ today. The evolution in the inverse exponential potential

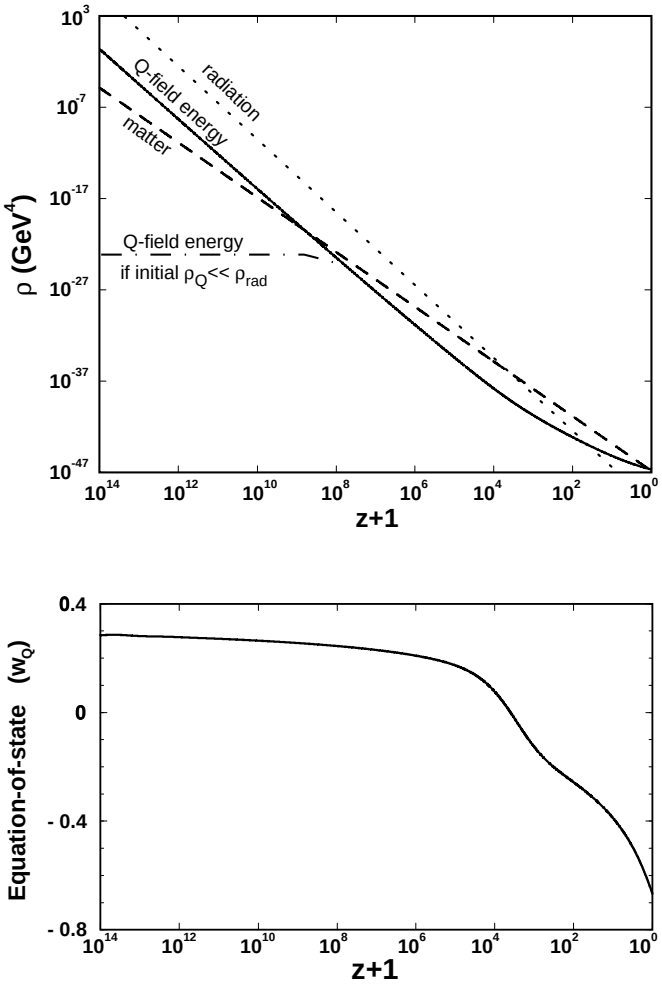

Fig. 9: This figure, taken from Zlatev, Wang & Steinhardt (1998), shows the evolution of the density in the 'quintessence' field (top panel), together with the effective equation of state of the quintessence vacuum (bottom panel), for the case of the inverse exponential potential. This allows vacuum energy to lurk at a few % of the total throughout the radiation era, but switching on a cosmological constant after the universe becomes matter dominated.

is shown in figure 9, demonstrating that the required behaviour can be found. However, a slight fine-tuning is still required, in that the trick only works for $M \sim 1$ meV, so there has to be an energy coincidence with the energy scale of matter-radiation equality.

So, the idea of tracker fields does not remove completely the puzzle concerning the level of present-day vacuum energy. In a sense, relegating the solution to a potential of unexplained form may seem a retrograde step. However, it is at least a testable step: the prediction of figure 9 is that $w \simeq -0.8$ today, so that the quintessence density scales as $\rho \propto a^{-0.6}$. This is a significant difference from the classical $w = -1$ vacuum energy, and it should be detectable as the SNe data improve. The existing data already require approximately $w < -0.5$, so there is the entrancing prospect that the equation of state for the vacuum will soon become the subject of experimental study.

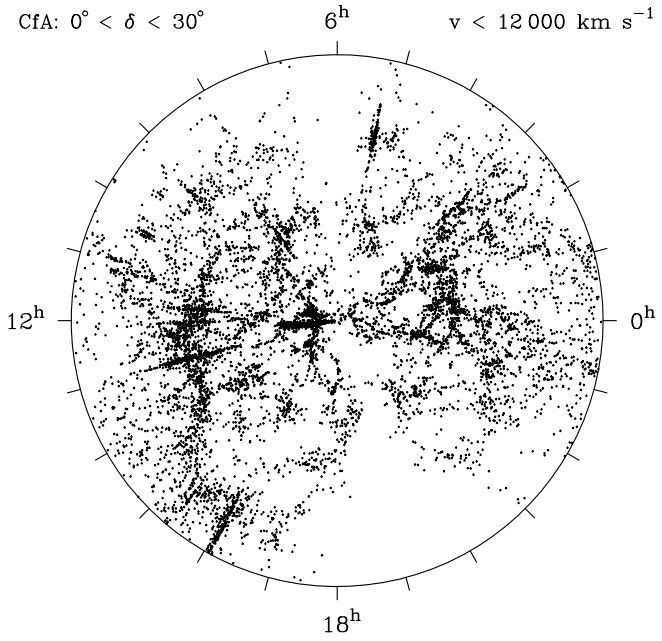

Fig. 10: One of the most dramatic pictures of the large-scale structure in the galaxy distribution is this slice made from the Harvard-Smithsonian Center for Astrophysics redshift survey to $B \simeq 15.5$. The survey coverage is not quite complete; as well as the holes due to the galactic plane around right ascensions $6^{\rm h}$ and $19^{\rm h}$, the rich clusters are somewhat over-represented with respect to a true random sampling of the galaxy population. Nevertheless, this plot emphasizes nicely both the large-scale features such as the 'great wall' on the left, the totally empty void regions, and the radial 'fingers of God' caused by virialized motions in the clusters.

6 DYNAMICS OF STRUCTURE FORMATION

The overall properties of the universe are very close to being homogeneous; and yet telescopes reveal a wealth of detail on scales varying from single galaxies to **large-scale structures** of size exceeding 100 Mpc (see figure 10). The existence of these cosmological structures must be telling us something important about the initial conditions of the big bang, and about the physical processes that have operated subsequently.

The study of cosmological perturbations can be presented as a complicated exercise in linearized general relativity; fortunately, much of the essential physics can be extracted from a Newtonian approach. We start by writing down the fundamental equations governing fluid motion (non-relativistic for now):

$$\begin{aligned} \text{Euler}: &\quad \frac{D\mathbf{v}}{Dt} = -\frac{\boldsymbol{\nabla} p}{\rho} - \boldsymbol{\nabla}\Phi \\ \text{energy}: &\quad \frac{D\rho}{Dt} = -\rho \boldsymbol{\nabla}\cdot\mathbf{v} \\ \text{Poisson}: &\quad \nabla^2\Phi = 4\pi G\rho, \end{aligned} \tag{173}$$

where $D/Dt = \partial/\partial t + \mathbf{v}\cdot\boldsymbol{\nabla}$ is the usual convective derivative. We now produce the **linearized equations of motion** by collecting terms of first order in perturbations about a homogeneous background: $\rho = \rho_0 + \delta\rho$ *etc.* As an example, consider the energy equation:

$$[\partial/\partial t + (\mathbf{v}_0 + \delta\mathbf{v})\cdot\boldsymbol{\nabla}]\,(\rho_0 + \delta\rho) = -(\rho_0 + \delta\rho)\,\boldsymbol{\nabla}\cdot(\mathbf{v}_0 + \delta\mathbf{v}). \tag{174}$$

For no perturbation, the zero-order equation is $(\partial/\partial t + \mathbf{v}_0\cdot\boldsymbol{\nabla})\rho_0 = -\rho_0\boldsymbol{\nabla}\cdot\mathbf{v}_0$; since ρ_0 is homogeneous and $\mathbf{v}_0 = H\mathbf{x}$ is the Hubble expansion, this just says $\dot\rho_0 = -3H\rho_0$. Expanding the

full equation and subtracting the zeroth-order equation gives the equation for the perturbation:

$$(\partial/\partial t + \mathbf{v}_0 \cdot \boldsymbol{\nabla})\,\delta\rho + \delta\mathbf{v} \cdot \boldsymbol{\nabla}(\rho_0 + \delta_\rho) = -(\rho_0 + \delta\rho)\,\boldsymbol{\nabla} \cdot \delta\mathbf{v} - \delta\rho\,\boldsymbol{\nabla} \cdot \mathbf{v}_0. \tag{175}$$

Now, for sufficiently small perturbations, terms containing a product of perturbations such as $\delta\mathbf{v} \cdot \boldsymbol{\nabla}\delta_\rho$ must be negligible in comparison with the first-order terms. Remembering that ρ_0 is homogeneous leaves the linearized equation

$$[\partial/\partial t + \mathbf{v}_0 \cdot \boldsymbol{\nabla}]\,\delta\rho = -\rho_0\,\boldsymbol{\nabla} \cdot \delta\mathbf{v} - \delta\rho\,\boldsymbol{\nabla} \cdot \mathbf{v}_0. \tag{176}$$

It is straightforward to perform the same steps with the other equations; the results look simpler if we define the fractional density perturbation

$$\boxed{\delta \equiv \frac{\delta\rho}{\rho_0}.} \tag{177}$$

As above, when dealing with time derivatives of perturbed quantities, the full convective time derivative D/Dt can always be replaced by $d/dt \equiv \partial/\partial t + \mathbf{v}_0 \cdot \boldsymbol{\nabla}$, which is the time derivative for an observer comoving with the unperturbed expansion of the universe. We then can write

$$\begin{aligned} &\frac{d}{dt}\delta\mathbf{v} = -\frac{\boldsymbol{\nabla}\,\delta p}{\rho_0} - \boldsymbol{\nabla}\,\delta\Phi - (\delta\mathbf{v} \cdot \boldsymbol{\nabla})\mathbf{v}_0 \\ &\frac{d}{dt}\delta = -\boldsymbol{\nabla} \cdot \delta\mathbf{v} \\ &\nabla^2\delta\Phi = 4\pi G\rho_0\delta. \end{aligned} \tag{178}$$

There is now only one complicated term to be dealt with: $(\delta\mathbf{v} \cdot \boldsymbol{\nabla})\mathbf{v}_0$ on the rhs of the perturbed Euler equation. This is best attacked by writing it in components:

$$[(\delta\mathbf{v} \cdot \boldsymbol{\nabla})\mathbf{v}_0]_j = [\delta v]_i \nabla_i [v_0]_j = H\,[\delta v]_j, \tag{179}$$

where the last step follows because $\mathbf{v}_0 = H\,\mathbf{x}_0 \Rightarrow \nabla_i[v_0]_j = H\,\delta_{ij}$. This leaves a set of equations of motion that have no explicit dependence on the global expansion speed v_0; this is only present implicitly through the use of convective time derivatives d/dt.

These equations of motion are written in **Eulerian coordinates**: proper length units are used, and the Hubble expansion is explicitly present through the velocity $\mathbf{v}_0$. The alternative approach is to use the comoving coordinates formed by dividing the Eulerian coordinates by the scale factor $a(t)$:

$$\boxed{\begin{aligned} \mathbf{x}(t) &= a(t)\mathbf{r}(t) \\ \delta\mathbf{v}(t) &= a(t)\mathbf{u}(t). \end{aligned}} \tag{180}$$

The next step is to translate spatial derivatives into comoving coordinates:

$$\boldsymbol{\nabla}_x = \frac{1}{a}\boldsymbol{\nabla}_r. \tag{181}$$

To keep the notation simple, subscripts on $\boldsymbol{\nabla}$ will normally be omitted hereafter, and spatial derivatives will be with respect to comoving coordinates. The linearized equations for conservation of momentum and matter as experienced by fundamental observers moving with the Hubble

flow then take the following simple forms in comoving units:

$$\boxed{\begin{aligned} &\dot{\mathbf{u}} + 2\frac{\dot{a}}{a}\,\mathbf{u} = \frac{\mathbf{g}}{a} - \frac{\boldsymbol{\nabla}\,\delta p}{\rho_0} \\ &\dot{\delta} = -\boldsymbol{\nabla}\cdot\mathbf{u}, \end{aligned}} \tag{182}$$

where dots stand for d/dt. The peculiar gravitational acceleration $\boldsymbol{\nabla}\delta\Phi/a$ is denoted by $\mathbf{g}$.

Before going on, it is useful to give an alternative derivation of these equations, this time working in comoving length units right from the start. First note that the comoving peculiar velocity $\mathbf{u}$ is just the time derivative of the comoving coordinate $\mathbf{r}$:

$$\dot{\mathbf{x}} = \dot{a}\mathbf{r} + a\dot{\mathbf{r}} = H\mathbf{x} + a\dot{\mathbf{r}}, \tag{183}$$

where the rhs must be equal to the Hubble flow $H\mathbf{x}$, plus the peculiar velocity $\delta\mathbf{v} = a\mathbf{u}$. In this equation, dots stand for exact convective time derivatives – *i.e.* time derivatives measured by an observer who follows a particle's trajectory – rather than partial time derivatives $\partial/\partial t$. This allows us to apply the continuity equation immediately in comoving coordinates, since this equation is simply a statement that particles are conserved, independent of the coordinates used. The exact equation is

$$\frac{D}{Dt}\rho_0(1+\delta) = -\rho_0(1+\delta)\,\boldsymbol{\nabla}\cdot\mathbf{u}, \tag{184}$$

and this is easy to linearize because the background density ρ_0 is independent of time when comoving length units are used. This gives the first-order equation $\dot{\delta} = -\boldsymbol{\nabla}\cdot\mathbf{u}$ immediately. The equation of motion follows from writing the Eulerian equation of motion as $\ddot{\mathbf{x}} = \mathbf{g}_0 + \mathbf{g}$, where $\mathbf{g} = \boldsymbol{\nabla}\delta\Phi/a$ is the peculiar acceleration defined earlier, and $\mathbf{g}_0$ is the acceleration that acts on a particle in a homogeneous universe (neglecting pressure forces, for simplicity). Differentiating $\mathbf{x} = a\mathbf{r}$ twice gives

$$\ddot{\mathbf{x}} = a\dot{\mathbf{u}} + 2\dot{a}\mathbf{u} + \frac{\ddot{a}}{a}\,\mathbf{x} = \mathbf{g}_0 + \mathbf{g}. \tag{185}$$

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 $\mathbf{u} + 2(\dot{a}/a)\mathbf{u} = \mathbf{g}$, as before. This derivation is rather more direct than the previous route of working in Eulerian space. Also, it emphasizes that the equation of motion is exact, even though it happens to be linear in the perturbed quantities.

After doing all this, we still have three equations in the four variables δ, $\mathbf{u}$, $\delta\Phi$ and δp. The system needs an equation of state in order to be closed; this may be specified in terms of the sound speed

$$c_s^2 \equiv \frac{\partial p}{\partial \rho}. \tag{186}$$

Now think of a plane-wave disturbance $\delta \propto e^{-i\mathbf{k}\cdot\mathbf{r}}$, where $\mathbf{k}$ is a comoving wavevector; in other words, suppose that the wavelength of a single Fourier mode stretches with the universe. All time dependence is carried by the amplitude of the wave, and so the spatial dependence can be factored out of time derivatives in the above equations (which would not be true with a constant comoving wavenumber k/a). An equation for the amplitude of δ can then be obtained by eliminating $\mathbf{u}$:

$$\boxed{\ddot{\delta} + 2\frac{\dot{a}}{a}\dot{\delta} = \delta(4\pi G\rho_0 - c_s^2 k^2/a^2).} \tag{187}$$

This equation is the one that governs the gravitational amplification of density perturbations.

There is a critical proper wavelength, known as the **Jeans length**, at which we switch from the possibility of exponential growth for long-wavelength modes to standing sound waves at short wavelengths. This critical length is

$$\boxed{\lambda_{\rm J} = c_s\sqrt{\frac{\pi}{G\rho}},} \tag{188}$$

and clearly delineates the scale at which sound waves can cross an object in about the time needed for gravitational free-fall collapse. When considering perturbations in an expanding background, things are more complex. Qualitatively, we expect to have no growth when the 'driving term' on the rhs is negative. However, owing to the expansion, $\lambda_{\rm J}$ will change with time, and so a given perturbation may switch between periods of growth and stasis.

Radiation-dominated universes At early enough times, the universe was radiation dominated ($c_s = c/\sqrt{3}$) and the analysis so far does not apply. It is common to resort to general relativity perturbation theory at this point. However, the fields are still weak, and so it is possible to generate the results we need by using special relativity fluid mechanics and Newtonian gravity with a relativistic source term. For simplicity, assume that accelerations due to pressure gradients are negligible in comparison with gravitational accelerations (*i.e.* restrict the analysis to $\lambda \gg \lambda_{\rm J}$ from the start). The basic equations are then a simplified Euler equation and the full energy and gravitational equations:

$$\begin{aligned}
&\text{Euler}: \quad \frac{D\mathbf{v}}{Dt} = -\boldsymbol{\nabla}\Phi \\
&\text{energy}: \quad \frac{D}{Dt}\left(\rho + p/c^2\right) = \frac{\partial}{\partial t}\left(p/c^2\right) - \left(\rho + p/c^2\right)\boldsymbol{\nabla}\cdot\mathbf{v} \\
&\text{Poisson}: \quad \nabla^2\Phi = 4\pi G(\rho + 3p/c^2).
\end{aligned} \tag{189}$$

For total radiation domination, $p = \rho c^2/3$, and it is easy to linearize these equations as before. The main differences come from factors of 2 and 4/3 due to the non-negligible contribution of the pressure. The result is a continuity equation $\boldsymbol{\nabla}\cdot\mathbf{u} = -(3/4)\dot{\delta}$, and the evolution equation for δ:

$$\boxed{\ddot{\delta} + 2\frac{\dot{a}}{a}\dot{\delta} = \frac{32\pi}{3}G\rho_0\delta,} \tag{190}$$

so the net result of all the relativistic corrections is a driving term on the rhs that is a factor 8/3 higher than in the matter-dominated case.

Solutions for $\delta(t)$ In both matter- and radiation-dominated universes with $\Omega = 1$, we have $\rho_0 \propto 1/t^2$:

$$\begin{aligned}
\text{matter domination } (a \propto t^{2/3}): \quad & 4\pi G\rho_0 = \frac{2}{3t^2} \\
\text{radiation domination } (a \propto t^{1/2}): \quad & 32\pi G\rho_0/3 = \frac{1}{t^2}.
\end{aligned} \tag{191}$$

Every term in the equation for δ is thus the product of derivatives of δ and powers of t, and a power-law solution is obviously possible. If we try $\delta \propto t^n$, then the result is $n = 2/3$ or -1

for matter domination; for radiation domination, this becomes $n = \pm 1$. For the growing mode, these can be combined rather conveniently using the **conformal time** $\eta \equiv \int dt/a$:

$$\boxed{\delta \propto \eta^2.} \tag{192}$$

Recall that η is proportional to the comoving size of the horizon.

The general case It is also interesting to think about the growth of matter perturbations in universes with nonzero vacuum energy, or even possibly some other exotic background with a peculiar equation of state. The differential equation for δ is as before, but $a(t)$ is altered. The way to deal with this is to treat a spherical perturbation as a small universe. Consider the Friedmann equation in the form

$$(\dot{a})^2 = \Omega_0^{\rm tot} H_0^2 a^2 + K, \tag{193}$$

where $K = -kc^2/R_0^2$; this emphasizes that K is a constant of integration. A second constant of integration arises in the expression for time:

$$t = \int_0^a \dot{a}^{-1}\, da + C. \tag{194}$$

This lets us argue as before in the case of decaying modes: if a solution to the Friedmann equation is $a(t, K, C)$, then valid density perturbations are

$$\delta \propto \left(\frac{\partial \ln a}{\partial K}\right)_t \quad \text{or} \quad \left(\frac{\partial \ln a}{\partial C}\right)_t. \tag{195}$$

Since $\partial(\dot{a}^2)/\partial K = 1$, this gives the growing and decaying modes as

$$\boxed{\delta \propto \begin{cases} (\dot{a}/a) \int_0^a (\dot{a})^{-3}\, da & \text{(growing mode)} \\ (\dot{a}/a) & \text{(decaying mode).} \end{cases}} \tag{196}$$

(Heath 1977; see also section 10 of Peebles 1980).

The equation for the growing mode requires numerical integration in general, with $\dot{a}(a)$ given by the Friedmann equation. A very good approximation to the answer is given by Carroll *et al.* (1992):

$$\boxed{\frac{\delta(z=0,\Omega)}{\delta(z=0,\Omega=1)} \simeq \tfrac{5}{2}\Omega_m \left[\Omega_m^{4/7} - \Omega_v + (1+\tfrac{1}{2}\Omega_m)(1+\tfrac{1}{70}\Omega_v)\right]^{-1}.} \tag{197}$$

This fitting formula for the growth suppression in low-density universes is an invaluable practical tool. For flat models with $\Omega_m + \Omega_v = 1$, it says that the growth suppression is less marked than for an open universe – approximately $\Omega^{0.23}$ as against $\Omega^{0.65}$ if $\Lambda = 0$. 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.

Mészáros effect What about the case of collisionless matter in a radiation background? The fluid treatment is not appropriate here, since the two species of particles can interpenetrate.

A particularly interesting limit is for perturbations well inside the horizon: the radiation can then be treated as a smooth, unclustered background that affects only the overall expansion rate. This is analogous to the effect of Λ, but an analytical solution does exist in this case. The perturbation equation is as before

$$\ddot{\delta} + 2\frac{\dot{a}}{a}\dot{\delta} = 4\pi G\rho_m \delta, \tag{198}$$

but now $H^2 = 8\pi G(\rho_m + \rho_r)/3$. If we change variable to $y \equiv \rho_m/\rho_r = a/a_{\rm eq}$, and use the Friedmann equation, then the growth equation becomes

$$\delta'' + \frac{2+3y}{2y(1+y)}\,\delta' - \frac{3}{2y(1+y)}\,\delta = 0 \tag{199}$$

(for $k = 0$, as appropriate for early times). It may be seen by inspection that a growing solution exists with $\delta'' = 0$:

$$\boxed{\delta \propto y + 2/3.} \tag{200}$$

It is also possible to derive the decaying mode. This is simple in the radiation-dominated case ($y \ll 1$): $\delta \propto -\ln y$ is easily seen to be an approximate solution in this limit.

What this says is that, at early times, the dominant energy of radiation drives the universe to expand so fast that the matter has no time to respond, and δ is frozen at a constant value. At late times, the radiation becomes negligible, and the growth increases smoothly to the Einstein–de Sitter $\delta \propto a$ behaviour (Mészáros 1974). The overall behaviour is therefore similar to the effects of pressure on a coupled fluid: for scales greater than the horizon, perturbations in matter and radiation can grow together, but this growth ceases once the perturbations enter the horizon. However, the explanations of these two phenomena are completely different. In the fluid case, the radiation pressure prevents the perturbations from collapsing further; in the collisionless case, the photons have free-streamed away, and the matter perturbation fails to collapse only because radiation domination ensures that the universe expands too quickly for the matter to have time to self-gravitate. Because matter perturbations enter the horizon (at $y = y_{\rm entry}$) with $\dot{\delta} > 0$, δ is not frozen quite at the horizon-entry value, and continues to grow until this initial 'velocity' is redshifted away, giving a total boost factor of roughly $\ln y_{\rm entry}$. This log factor may be seen below in the fitting formulae for the CDM power spectrum.

6.1 The peculiar velocity field

The equations for velocity-field perturbations were developed in section 5.2 as part of the machinery of analysing self-gravitating density fluctuations. There, the velocity field was eliminated, in order to concentrate on the behaviour of density perturbations. However, the peculiar velocity field is of great importance in cosmology, so it is convenient to give a summary that highlights the properties of velocity perturbations.

Consider first a galaxy that moves with some peculiar velocity in an otherwise uniform universe. Even though there is no peculiar gravitational acceleration acting, its velocity will decrease with time as the galaxy attempts to catch up with successively more distant (and therefore more rapidly receding) neighbours. If the proper peculiar velocity is v, then after time dt the galaxy will have moved a proper distance $x = v\,dt$ from its original location. Its near neighbours will now be galaxies with recessional velocities $H\,x = H\,v\,dt$, relative to which the peculiar velocity will have fallen to $v - Hx$. The equation of motion is therefore just

$$\dot{v} = -H\,v = -\frac{\dot{a}}{a}\,v, \tag{201}$$

with the solution $v \propto a^{-1}$: peculiar velocities of nonrelativistic objects suffer redshifting by exactly the same factor as photon momenta. It is often convenient to express the peculiar velocity in terms of its comoving equivalent, $\mathbf{v} \equiv a\,\mathbf{u}$, for which the equation of motion becomes $\dot{u} = -2H\,u$. Thus, in the absence of peculiar accelerations and pressure forces, comoving peculiar velocities redshift away through the **Hubble drag** term $2H\mathbf{u}$.

If we now include the effects of peculiar acceleration, this simply adds the acceleration g on the right-hand side. This gives the equation of motion

$$\dot{\mathbf{u}} + \frac{2\dot{a}}{a}\mathbf{u} = -\frac{\mathbf{g}}{a}, \tag{202}$$

where $\mathbf{g} = \boldsymbol{\nabla}\delta\Phi/a$ is the peculiar gravitational acceleration. Pressure terms have been neglected, so $\lambda \gg \lambda_{\rm J}$. Remember that throughout we are using comoving length units, so that $\nabla_{\rm proper} = \nabla/a$. This equation is the exact equation of motion for a single galaxy, so that the time derivative is $d/dt = \partial/\partial t + \mathbf{u}\cdot\boldsymbol{\nabla}$. In linear theory, the second part of the time derivative can be neglected, and the equation then turns into one that describes the evolution of the linear peculiar velocity field at a fixed point in comoving coordinates.

The solutions for the peculiar velocity field 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 interpretation of these solutions is aided by knowing that the velocity field satisfies the **continuity equation**: $\dot{\rho} = -\boldsymbol{\nabla}\cdot(\rho\mathbf{v})$ in proper units, which obviously takes the same form $\dot{\rho} = -\boldsymbol{\nabla}\cdot(\rho\mathbf{u})$ if lengths and densities are in comoving units. If we express the density as $\rho = \rho_0(1+\delta)$ (where in comoving units ρ_0 is just a number independent of time), the continuity equation takes the form

$$\dot{\delta} = -\boldsymbol{\nabla}\cdot[(1+\delta)\mathbf{u}], \tag{203}$$

which becomes just

$$\boxed{\boldsymbol{\nabla}\cdot\mathbf{u} = -\dot{\delta}} \tag{204}$$

in linear theory when both δ and $\mathbf{u}$ are small. This says that it is possible to have **vorticity modes** with $\boldsymbol{\nabla}\cdot\mathbf{u} = 0$, for which $\dot{\delta}$ vanishes. We have already seen that δ either grows or decays as a power of time, so these modes require zero density perturbation, in which case the associated peculiar gravity also vanishes. These vorticity modes are thus the required homogeneous solutions, and they decay as $v = au \propto a^{-1}$, as with the kinematic analysis for a single particle. For any gravitational-instability theory, in which structure forms via the collapse of small perturbations laid down at very early times, it should therefore be a very good approximation to say that the linear velocity field must be curl-free.

For the growing modes, we want to try looking for a solution $\mathbf{u} = F(t)\mathbf{g}$. Then using continuity plus Gauss's theorem, $\boldsymbol{\nabla}\cdot\mathbf{g} = 4\pi G a\rho\delta$, gives us

$$\delta\mathbf{v} = \frac{2f(\Omega)}{3H\Omega}\mathbf{g}, \tag{205}$$

where the function $f(\Omega) \equiv (a/\delta)d\delta/da$. A very good approximation to this (Peebles 1980) is $g \simeq \Omega^{0.6}$ (a result that is almost independent of Λ; Lahav *et al.* 1991). Alternatively, we can work in Fourier terms. This is easy, as $\mathbf{g}$ and $\mathbf{k}$ are parallel, so that $\boldsymbol{\nabla}\cdot\mathbf{u} = -i\mathbf{k}\cdot\mathbf{u} = -iku$. Thus, directly from the continuity equation,

$$\boxed{\delta\mathbf{v}_{\mathbf{k}} = -\frac{iHf(\Omega)a}{k}\delta_k\hat{\mathbf{k}}.} \tag{206}$$

The $1/k$ factor tells us that cosmological velocities come predominantly from larger-scale perturbations than those that dominate the density field. Deviations from the Hubble flow are therefore in principle a better probe of the inhomogeneity of the universe than large-scale clustering.

6.2 The Boltzmann equation

We now turn to the question of how to treat the matter source, without assuming that it is a fluid. The general approach that should be taken is to consider the phase-space distribution function $f(\mathbf{x}, \mathbf{p})$ – *i.e.* the product of the particle number density and the probability distribution for momentum. The equation that describes the evolution of f is the Boltzmann equation. The general relativistic form of the equation is

$$\left[p^\mu \frac{\partial}{\partial x^\mu} - \Gamma^\mu_{\alpha\beta} p^\alpha p^\beta \frac{\partial}{\partial p^\mu} \right] f = C. \tag{207}$$

This equation is exact for particles affected by gravitational forces and by collisions. The collision term on the rhs, C, has to contain all the appropriate scattering physics (Thomson scattering, in the case of a coupled system of electrons and photons); the gravitational forces are contained implicitly in the connection coefficients. What has to be done is to perturb this equation, using the perturbed metric coefficients, together with their equation of motion derived from the Einstein equations. Although this is easily said, the detailed algebra of the calculation consumes many pages, and it will not be reproduced here (see Peebles 1980; Efstathiou 1990; Bond 1997). The result is a system of coupled differential equations for the distribution functions of the non-fluid components (photons, neutrinos, plus possibly collisionless dark matter), together with the density and pressure of the collisional baryon fluid. Remembering to include gravitational waves, the whole system has to be integrated numerically, starting with a single Fourier mode of wavelength much greater than the horizon scale, and evolving to the present. Finally, the present-day perturbations to observational quantities such as density and radiation specific intensity are constructed by adding together modes of all wavelengths (which evolve independently in the linear approximation).

This, then, is a brief summary of the professional approach to cosmological perturbations. A modern cosmological Boltzmann code, such as that described by Seljak & Zaldarriaga (1996), is a large and sophisticated piece of machinery, which is the final outcome of decades of intellectual effort. Although heroic analytical efforts have been made in an attempt to find alternative methods of calculation (*e.g.* Hu & Sugiyama 1995), results of high precision demand the full approach. For a non-specialist, the best that can be done is to attempt to use simple approximate physical arguments in order to understand the main features of the results; on large scales, this approach is usually quantitatively successful.

6.3 Transfer functions

Real power spectra result from modifications of any primordial power by a variety of processes: growth under self-gravitation; the effects of pressure; dissipative processes. In general, modes of short wavelength have their amplitudes reduced relative to those of long wavelength in this way. The overall effect is encapsulated in the **transfer function**, which gives the ratio of the late-time amplitude of a mode to its initial value:

$$\boxed{T_k \equiv \frac{\delta_k(z=0)}{\delta_k(z)\ D(z)},} \tag{208}$$

where $D(z)$ is the linear growth factor between redshift z and the present. The normalization redshift is arbitrary, so long as it refers to a time before any scale of interest has entered the

horizon. Once we possess the transfer function, it is a most valuable tool. The evolution of linear perturbations back to last scattering obeys the simple growth laws summarized above, and it is easy to see how structure in the universe will have changed during the matter-dominated epoch.

There are in essence two ways in which the power spectrum that exists at early times may differ from that which emerges at the present, both of which correspond to a reduction of small-scale fluctuations:

(1) Jeans mass effects. Prior to matter–radiation equality, we have already seen that perturbations inside the horizon are prevented from growing by radiation pressure. Once $z_{\rm eq}$ is reached, one of two things can happen. If collisionless dark matter dominates, perturbations on all scales can grow. If baryonic gas dominates, the Jeans length remains approximately constant, as follows: The sound speed, $c_{\rm S}^2 = \partial p/\partial\rho$, may be found by thinking about the response of matter and radiation to small adiabatic compressions:

$$\delta p = (4/9)\rho_r c^2(\delta V/V), \qquad \delta\rho = [\rho_m + (4/3)\rho_r](\delta V/V), \tag{209}$$

implying

$$c_{\rm S}^2 = c^2\left(3 + \frac{9}{4}\frac{\rho_m}{\rho_r}\right)^{-1} = c^2\left[3 + \frac{9}{4}\left(\frac{1+z_{\rm rad}}{1+z}\right)\right]^{-1}. \tag{210}$$

Here, $z_{\rm rad}$ is the redshift of equality between matter and photons; $1 + z_{\rm rad} = 1.68(1 + z_{\rm eq})$ because of the neutrino contribution. At $z \ll z_{\rm rad}$, we therefore have $c_{\rm S} \propto \sqrt{1+z}$. Since $\rho = (1+z)^3 3\Omega_{\rm B}H_0^2/(8\pi G)$, the *comoving* Jeans length is constant at

$$\lambda_{\rm J} = \frac{c}{H_0}\left(\frac{32\pi^2}{27\Omega_{\rm B}(1+z_{\rm rad})}\right)^{1/2} = 50\,(\Omega_{\rm B}h^2)^{-1}\ \ {\rm Mpc}. \tag{211}$$

Thus, in either case, one of the critical length scales for the power spectrum will be the horizon distance at $z_{\rm eq}$ ($= 23\,900\Omega h^2$ for $T = 2.73$ K, counting neutrinos as radiation). In the matter-dominated approximation, we get

$$d_{\rm H} = \frac{2c}{H_0}(\Omega z)^{-1/2} \quad \Rightarrow \quad d_{\rm eq} = 39\,(\Omega h^2)^{-1}{\rm Mpc}. \tag{212}$$

The exact distance–redshift relation is

$$R_0 dr = \frac{c}{H_0}\frac{dz}{(1+z)\sqrt{1+\Omega_m z + (1+z)^2\Omega_r}}, \tag{213}$$

from which it follows that the correct answer for the horizon size including radiation is a factor $\sqrt{2}-1$ smaller: $d_{\rm eq} = 16.0\,(\Omega h^2)^{-1}$ Mpc.

It is easy from the above to see the approximate scaling that must be obeyed by transfer functions. Consider the adiabatic case first. Perturbations with $kd_{\rm eq} \ll 1$ always undergo growth as $\delta \propto d_{\rm H}^2$. Perturbations with larger k enter the horizon when $d_{\rm H} \simeq 1/k$; they are then frozen until $z_{\rm eq}$, at which point they can grow again. The missing growth factor is just the square of the change in $d_{\rm H}$ during this period, which is $\propto k^2$. The approximate limits of an adiabatic transfer function would therefore be

$$\boxed{T_k \simeq \begin{cases} 1 & (kd_{\rm eq} \ll 1) \\ [kd_{\rm eq}]^{-2} & (kd_{\rm eq} \gg 1). \end{cases}} \tag{214}$$

For isocurvature perturbations, the situation is the opposite. Consider a perturbation of short wavelength: once it comes well inside the horizon, the photons disperse, and so all the perturbation to the entropy density (which must be conserved) is carried by the matter perturbation. The perturbation thus enters the horizon with the original amplitude δ_i. Thereafter, it grows in the same way as an isothermal perturbation. This means there are two regimes, for perturbations that enter the horizon before and after matter–radiation equality. The former match onto the Mészáros solution, and keep their amplitudes constant until they start to grow after $a_{\rm eq}$. The present-day amplitude for these is $\delta/\delta_i = (3/2)[1/a_{\rm eq}]$. Perturbations that enter after matter–radiation equality start to grow immediately, so that their present amplitude is $\delta/\delta_i \simeq 1/a_{\rm entry}$. Entry occurs when $kd_{\rm H} \simeq 1$, and the horizon evolves as $d_{\rm H} = (2c/H_0)a^{1/2}$ (assuming $\Omega = 1$). Putting these arguments together, the isocurvature transfer function relative to δ_i is

$$\boxed{T_k \simeq \begin{cases} (2/15)\,[kc/H_0]^2 & (kd_{\rm eq} \ll 1) \\ (3/2)\,a_{\rm eq}^{-1} & (kd_{\rm eq} \gg 1) \end{cases}} \tag{215}$$

(a more sophisticated argument is required to obtain the exact factor 2/15 in the long-wavelength limit; see Efstathiou 1990). Since this goes to a constant at high k, it is also common to quote the transfer function relative to this value. This means that $T_k < 1$ at $kd_{\rm eq} \lesssim 1$, and so the isocurvature transfer function is the mirror image of the adiabatic case: one falls where the other rises (see figure 11).

(2) Damping. In addition to having their growth retarded, very small-scale perturbations will be erased entirely, which can happen in one of two ways. For collisionless dark matter, perturbations are 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; for massive particles, this happens long before $z_{\rm eq}$ (resulting in **cold dark matter**;CDM CDM). For massive neutrinos, on the other hand, it happens *at* $z_{\rm eq}$: only perturbations on very large scales survive in the case of **hot dark matter**HDM (HDM). In a purely baryonic universe, the corresponding process is called **Silk damping**: the mean free path of photons due to scattering by the plasma is non-zero, and so radiation can diffuse out of a perturbation, convecting the plasma with it. The typical distance of a random walk in terms of the diffusion coefficient D is $x \simeq \sqrt{Dt}$, which gives a damping length of

$$\lambda_{\rm S} \simeq \sqrt{\lambda d_{\rm H}}, \tag{216}$$

the geometric mean of the horizon size and the mean free path. Since $\lambda = 1/(n\sigma_{\rm T}) = 44.3(1+z)^{-3}(\Omega_{\rm B}h^2)^{-1}$ proper Gpc, we obtain a comoving damping length of

$$\lambda_{\rm S} = 16.3\,(1+z)^{-5/4}(\Omega_{\rm B}^2\Omega h^6)^{-1/4}\ {\rm Gpc}. \tag{217}$$

This becomes close to the Jeans length by the time of last scattering, $1 + z \simeq 1000$.

It is invaluable in practice to have some accurate analytic formulae that fit the numerical results for transfer functions. We give below results for some common models of particular interest (illustrated in figure 11, along with other cases where a fitting formula is impractical). For the models with collisionless dark matter, $\Omega_{\rm B} \ll \Omega$is assumed, so that all lengths scale with the horizon size at matter–radiation equality, leading to the definition

$$q \equiv \frac{k}{\Omega h^2 {\rm Mpc}^{-1}}. \tag{218}$$

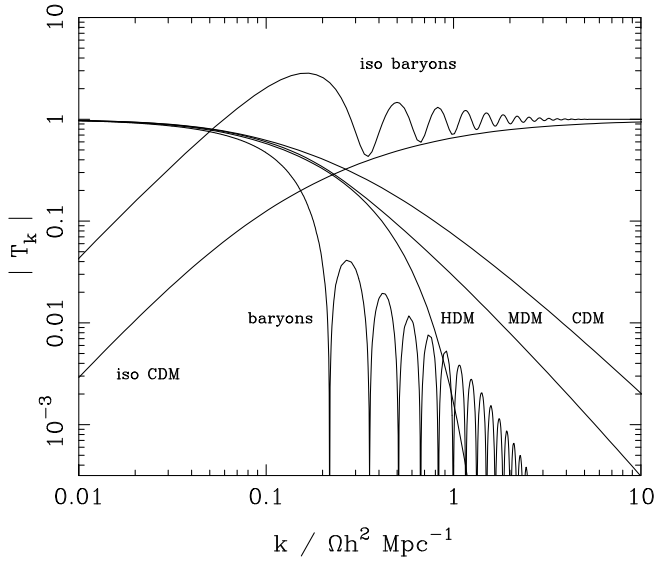

Fig. 11: A plot of transfer functions for various models. For adiabatic models, $T_k \to 1$ at small k, whereas the opposite is true for isocurvature models. A number of possible matter contents are illustrated: pure baryons; pure CDM; pure HDM; MDM (30% HDM, 70% CDM). For dark-matter models, the characteristic wavenumber scales proportional to Ωh^2. The scaling for baryonic models does not obey this exactly; the plotted cases correspond to $\Omega = 1$, $h = 0.5$.

We consider the following cases: (1) Adiabatic CDM; (2) Adiabatic massive neutrinos (1 massive, 2 massless); (3) Isocurvature CDM; these expressions come from Bardeen *et al.* (1986; BBKS). Since the characteristic length-scale in the transfer function depends on the horizon size at matter–radiation equality, the temperature of the CMB enters. In the above formulae, it is assumed to be exactly 2.7 K; for other values, the characteristic wavenumbers scale $\propto T^{-2}$. For these purposes massless neutrinos count as radiation, and three species of these contribute a total density that is 0.68 that of the photons.

$$\begin{aligned}
&(1) \quad T_k = \frac{\ln(1+2.34q)}{2.34q}\left[1+3.89q+(16.1q)^2+(5.46q)^3+(6.71q)^4\right]^{-1/4} \\
&(2) \quad T_k = \exp(-3.9q-2.1q^2) \\
&(3) \quad T_k = (5.6q)^2\left(1+\left[15.0q+(0.9q)^{3/2}+(5.6q)^2\right]^{1.24}\right)^{-1/1.24}
\end{aligned} \tag{219}$$

The case of **mixed dark matter** (MDM:MDM a mixture of massive neutrinos and CDM) is more complex. See Pogosyan & Starobinksy (1995) for a fit in this case.

The above expressions assume pure dark matter, which is unrealistic. At least for CDM models, a non-zero baryonic density lowers the apparent dark-matter density parameter. We can define an apparent shape parameter Γ for the transfer function:

$$\boxed{q \equiv (k/h\ \mathrm{Mpc}^{-1})/\Gamma,} \tag{220}$$

and $\Gamma = \Omega h$ in a model with zero baryon content. This parameter was originally defined by Efstathiou, Bond & White (1992), in terms of a CDM model with $\Omega_\mathrm{B} = 0.03$. Peacock & Dodds (1994) showed that the effect of increasing Ω_B was to preserve the CDM-style spectrum shape,

but to shift to lower values of Γ. This shift was generalized to models with $\Omega \neq 1$ by Sugiyama (1995):

$$\Gamma = \Omega h \, \exp[-\Omega_{\rm B}(1 + \sqrt{2h}/\Omega)]. \tag{221}$$

This formula fails if the baryon content is too large, and the transfer function develops oscillations (see Eisenstein & Hu 1998 for a more accurate approximation in this case).

6.4 The spherical model

An overdense sphere is a very useful nonlinear model, as it behaves in exactly the same way as a closed sub-universe. The density perturbation need not be a uniform sphere: any spherically symmetric perturbation will clearly evolve at a given radius in the same way as a uniform sphere containing the same amount of mass. In what follows, therefore, density refers to the *mean* density inside a given sphere. 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

$$\begin{aligned} r &= A(1 - \cos\theta) \\ t &= B(\theta - \sin\theta), \end{aligned} \tag{222}$$

and $A^3 = GMB^2$, just from $\ddot{r} = -GM/r^2$. Expanding these relations up to order θ^5 gives $r(t)$ for small t:

$$r \simeq \frac{A}{2}\left(\frac{6t}{B}\right)^{2/3}\left[1 - \frac{1}{20}\left(\frac{6t}{B}\right)^{2/3}\right], \tag{223}$$

and we can identify the density perturbation within the sphere:

$$\delta \simeq \frac{3}{20}\left(\frac{6t}{B}\right)^{2/3}. \tag{224}$$

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 $\theta = 2\pi$. This occurs when $\delta_{\rm lin} = (3/20)(12\pi)^{2/3} \simeq 1.69$.

(3) **Virialization**. Consider the time at which the sphere has collapsed by a factor 2 from maximum expansion. 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**. For this reason, many workers take this epoch as indicating the sort of density contrast to be expected as the endpoint of gravitational collapse. This occurs at $\theta = 3\pi/2$, and the corresponding density enhancement is $(9\pi + 6)^2/8 \simeq 147$, with $\delta_{\rm lin} \simeq 1.58$. Some authors prefer to assume that this virialized size is eventually achieved only at collapse, in which case the contrast becomes $(6\pi)^2/2 \simeq 178$.

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; $\delta_{\rm lin} \simeq 1$ is a good criterion for collapse for any value of Ω likely to be of practical relevance. The full density contrast at virialization may be approximated by

$$1 + \delta_{\rm vir} \simeq 178\,\Omega^{-0.7} \tag{225}$$

(although flat Λ-dominated models show less dependence on Ω; Eke *et al.* 1996).

7 COSMOLOGICAL DENSITY FIELDS

The next step is to see how the above theoretical ideas can be confronted with statistical measures of the observed matter distribution, and to summarize what is known about the dimensionless **density perturbation field**

$$\boxed{\delta(\mathbf{x}) \equiv \frac{\rho(\mathbf{x}) - \langle\rho\rangle}{\langle\rho\rangle}.} \tag{226}$$

This quantity need not be assumed to be small. Indeed, some of the most interesting issues arise in understanding the evolution of the density field to large values of δ.

A critical feature of the δ field is that it inhabits a universe that is isotropic and homogeneous in its large-scale properties. This suggests that the statistical properties of δ should also be homogeneous, even though it is a field that describes inhomogeneities. This statement sounds contradictory, and yet it makes perfect sense if there exists an **ensemble of universes**. The concept of an ensemble is used every time we apply probability theory to an event such as tossing a coin: we imagine an infinite sequence of repeated trials, half of which result in heads, half in tails. To say that the probability of heads is 1/2 means that the coin lands heads up in half the members of this ensemble of universes. The analogy of coin tossing in cosmology is that the density at a given point in space will have different values in each member of the ensemble, with some overall variance $\langle\delta^2\rangle$ between members of the ensemble. Statistical homogeneity of the δ field then means that this variance must be independent of position. The actual field found in a given member of the ensemble is a **realization** of the statistical process.

There are two problems with this line of argument: (i) we have no evidence that the ensemble exists; (ii) in any case, we only get to observe one realization, so how is the variance $\langle\delta^2\rangle$ to be measured? The first objection applies to coin tossing, and may be evaded if we understand the physics that generates the statistical process – we only need to *imagine* tossing the coin many times, and we do not actually need to perform the exercise. The best that can be done in answering the second objection is to look at widely separated parts of space, since the δ fields there should be causally unconnected; this is therefore as good as taking measurements from two different member of the ensemble. In other words, if we measure the variance $\langle\delta^2\rangle$ by averaging over a sufficiently large volume, the results would be expected to approach the true ensemble variance, and the averaging operator $\langle\cdots\rangle$ is often used without being specific about which kind of average is intended. Fields that satisfy this property, whereby

$$\text{volume average} \quad \leftrightarrow \quad \text{ensemble average} \tag{227}$$

are termed **ergodic**. Giving a formal proof of ergodicity for a random process is not always easy (Adler 1981); in cosmology it is perhaps best regarded as a common-sense axiom.

7.1 Fourier analysis of density fluctuations

It is often convenient to consider building up a general field by the superposition of many modes. For a flat comoving geometry, the natural tool for achieving this is via Fourier analysis. How do we make a Fourier expansion of the density field in an infinite universe? If the field were periodic within some box of side L, then we would just have a sum over wave modes:

$$F(\mathbf{x}) = \sum F_{\mathbf{k}} e^{-i\mathbf{k}\cdot\mathbf{x}}. \tag{228}$$

The requirement of periodicity restricts the allowed wavenumbers to **harmonic boundary conditions**

$$k_x = n\,\frac{2\pi}{L}, \qquad n = 1, 2\cdots, \tag{229}$$

with similar expressions for k_y and k_z. Now, if we let the box become arbitrarily large, then the sum will go over to an integral that incorporates the density of states in k-space, exactly as in statistical mechanics. The Fourier relations in n dimensions are thus

$$\boxed{\begin{aligned} F(x) &= \left(\frac{L}{2\pi}\right)^n \int F_k(k)\exp(-i\mathbf{k}\cdot\mathbf{x})\; d^n k \\ F_k(k) &= \left(\frac{1}{L}\right)^n \int F(x)\exp(i\mathbf{k}\cdot\mathbf{x})\; d^n x. \end{aligned}} \tag{230}$$

Correlation functions and power spectra As an immediate example of the Fourier machinery in action, consider the important quantity

$$\boxed{\xi(\mathbf{r}) \equiv \langle \delta(\mathbf{x})\delta(\mathbf{x}+\mathbf{r})\rangle\,,} \tag{231}$$

which is the autocorrelation function of the density field – usually referred to simply as the **correlation function**. The angle brackets indicate an averaging over the normalization volume V. Now express δ as a sum and note that δ is real, so that we can replace one of the two δ's by its complex conjugate, obtaining

$$\xi = \left\langle \sum_{\mathbf{k}}\sum_{\mathbf{k}'} \delta_{\mathbf{k}}\delta^*_{\mathbf{k}'} e^{i(\mathbf{k}'-\mathbf{k})\cdot\mathbf{x}} e^{-i\mathbf{k}\cdot\mathbf{r}} \right\rangle. \tag{232}$$

Alternatively, this sum can be obtained without replacing $\langle\delta\delta\rangle$ by $\langle\delta\delta^*\rangle$, from the relation between modes with opposite wavevectors that holds for any real field: $\delta_{\mathbf{k}}(-\mathbf{k}) = \delta^*_{\mathbf{k}}(\mathbf{k})$. Now, by the periodic boundary conditions, all the cross terms with $\mathbf{k}' \neq \mathbf{k}$ average to zero. Expressing the remaining sum as an integral, we have

$$\boxed{\xi(\mathbf{r}) = \frac{V}{(2\pi)^3}\int |\delta_{\mathbf{k}}|^2 e^{-i\mathbf{k}\cdot\mathbf{r}} d^3k.} \tag{233}$$

In short, the correlation function is the Fourier transform of the **power spectrum**. This relation has been obtained by volume averaging, so it applies to the specific mode amplitudes and correlation function measured in any given realization of the density field. Taking ensemble

averages of each side, the relation clearly also holds for the ensemble average power and correlations – which are really the quantities that cosmological studies aim to measure. We shall hereafter often use the alternative notation

$$\boxed{P(k) \equiv \langle |\delta_k|^2 \rangle} \tag{234}$$

for the ensemble-average power. The distinction between the ensemble average and the actual power measured in a realization is clarified below in the section on Gaussian fields.

In an isotropic universe, the density perturbation spectrum cannot contain a preferred direction, and so we must have an **isotropic power spectrum**: $\langle |\delta_{\mathbf{k}}|^2(\mathbf{k}) \rangle = |\delta_k|^2(k)$. The angular part of the k-space integral can therefore be performed immediately: introduce spherical polars with the polar axis along $\mathbf{k}$, and use the reality of ξ so that $e^{-i\mathbf{k}\cdot\mathbf{x}} \to \cos(kr\cos\theta)$. In three dimensions, this yields

$$\xi(r) = \frac{V}{(2\pi)^3} \int P(k) \, \frac{\sin kr}{kr} \, 4\pi k^2 \, dk. \tag{235}$$

The 2D analogue of this formula is

$$\xi(r) = \frac{A}{(2\pi)^2} \int P(k) \, J_0(kr) \, 2\pi k \, dk. \tag{236}$$

We shall usually express the power spectrum in dimensionless form, as the variance per $\ln k$ ($\Delta^2(k) = d\langle\delta^2\rangle/d\ln k \propto k^3 P[k]$):

$$\boxed{\Delta^2(k) \equiv \frac{V}{(2\pi)^3} \, 4\pi k^3 \, P(k) = \frac{2}{\pi} k^3 \int_0^\infty \xi(r) \, \frac{\sin kr}{kr} \, r^2 \, dr.} \tag{237}$$

This gives a more easily visualizable meaning to the power spectrum than does the quantity $VP(k)$, which has dimensions of volume: $\Delta^2(k) = 1$ means that there are order-unity density fluctuations from modes in the logarithmic bin around wavenumber k. $\Delta^2(k)$ is therefore the natural choice for a Fourier-space counterpart to the dimensionless quantity $\xi(r)$.

Power-law spectra The above shows that the power spectrum is a central quantity in cosmology, but how can we predict its functional form? For decades, this was thought to be impossible, and so a minimal set of assumptions was investigated. In the absence of a physical theory, we should not assume that the spectrum contains any preferred length scale, otherwise we should then be compelled to explain this feature. Consequently, the spectrum must be a featureless power law:

$$\boxed{\left\langle |\delta_k|^2 \right\rangle \propto k^n} \tag{238}$$

The index n governs the balance between large- and small-scale power. The meaning of different values of n can be seen by imagining the results of filtering the density field by passing over it a box of some characteristic comoving size x and averaging the density over the box. This will filter out waves with $k \gtrsim 1/x$, leaving a variance $\langle\delta^2\rangle \propto \int_0^{1/x} k^n 4\pi k^2 dk \propto x^{-(n+3)}$. Hence, in terms of a mass $M \propto x^3$, we have

$$\delta_{\rm rms} \propto M^{-(n+3)/6}. \tag{239}$$

Similarly, a power-law spectrum implies a power-law correlation function. If $\xi(r) = (r/r_0)^{-\gamma}$, with $\gamma = n + 3$, the corresponding 3D power spectrum is

$$\Delta^2(k) = \frac{2}{\pi}\,(kr_0)^\gamma\,\Gamma(2-\gamma)\,\sin\frac{(2-\gamma)\pi}{2} \equiv \beta(kr_0)^\gamma \tag{240}$$

$(= 0.903(kr_0)^{1.8}$ if $\gamma = 1.8)$. This expression is only valid for $n < 0$ ($\gamma < 3$); for larger values of n, ξ must become negative at large r (because $P(0)$ must vanish, implying $\int_0^\infty \xi(r)\,r^2\,dr = 0$). A cutoff in the spectrum at large k is needed to obtain physically sensible results.

The Zeldovich spectrum Most important of all is the **scale-invariant spectrum**, which corresponds to the value $n = 1$, *i.e.* $\Delta^2 \propto k^4$. To see how the name arises, consider a perturbation $\delta\Phi$ in the gravitational potential:

$$\nabla^2\delta\Phi = 4\pi G\rho_0\delta \quad \Rightarrow \quad \delta\Phi_k = -4\pi G\rho_0\delta_k/k^2. \tag{241}$$

The two powers of k pulled down by ∇^2 mean that, if $\Delta^2 \propto k^4$ for the power spectrum of density fluctuations, then Δ^2_Φ is a constant. Since potential perturbations govern the flatness of spacetime, this says that the scale-invariant spectrum corresponds to a metric that is a **fractal**: spacetime has the same degree of 'wrinkliness' on each resolution scale. The total curvature fluctuations diverge, but only logarithmically at either extreme of wavelength.

Another way of looking at this spectrum is in terms of perturbation growth balancing the scale dependence of δ: $\delta \propto x^{-(n+3)/2}$. We know that δ viewed on a given comoving scale will increase with the size of the horizon: $\delta \propto r_{\rm H}^2$. At an arbitrary time, though, the only natural length provided by the universe (in the absence of non-gravitational effects) is the horizon itself:

$$\delta(r_{\rm H}) \propto r_{\rm H}^2 r_{\rm H}^{-(n+3)/2} = r_{\rm H}^{-(n-1)/2}. \tag{242}$$

Thus, if $n = 1$, the growth of both $r_{\rm H}$ and δ with time cancels out so that the universe always looks the same when viewed on the scale of the horizon; such a universe is self-similar in the sense of always appearing the same under the magnification of cosmological expansion. This spectrum is often known as the **Zeldovich spectrum** (sometimes hyphenated with Harrison and Peebles, who invented it independently).

Filtering and moments A common concept in the manipulation of cosmological density fields is that of **filtering**, where the density field is convolved with some **window function**: $\delta \to \delta * f$. Many observable results can be expressed in this form. Some common 3D filter functions areGaussian filtertop-hat filter

$$\begin{aligned} &\text{Gaussian}: f = \frac{V}{(2\pi)^{3/2}R_{\rm G}^3}e^{-r^2/2R_{\rm G}^2} \;\Rightarrow\; f_k = e^{-k^2R_{\rm G}^2/2} \\ &\text{top}-\text{hat}: f = \frac{3V}{4\pi R_{\rm T}^3} \quad (r < R_{\rm T}) \;\Rightarrow\; f_k = \frac{3}{y^3}(\sin y - y\cos y) \quad (y \equiv kR_{\rm T}). \end{aligned} \tag{243}$$

Note the factor of V in the definition of f; this is needed to cancel the $1/V$ in the definition of convolution. For some power spectra, the difference in these filter functions at large k is unimportant, and we can relate them by equating the expansions near $k = 0$, where $1-|f_k|^2 \propto k^2$. This equality requires

$$R_{\rm T} = \sqrt{5}\;R_{\rm G}. \tag{244}$$

We are often interested not in the convolved field itself, but in its variance, for use as a statistic (*e.g.* to measure the rms fluctuations in the number of objects in a cell). By the

convolution theorem, this means we are interested in a **moment** of the power spectrum times the squared filter transform. We shall generally use the following notation:

$$\boxed{\sigma_n^2 \equiv \frac{V}{(2\pi)^3} \int P(k)\,|f_k|^2\; k^{2n}\; d^3k;} \tag{245}$$

the filtered variance is thus σ_0^2 (often denoted by just σ^2). Clustering results are often published in the form of **cell variances** of δ as a function of scale, σ^2, using either cubical cells of side ℓ (Efstathiou *et al.* 1990) or Gaussian spheres of radius $R_{\rm G}$ (Saunders *et al.* 1991). For a power-law spectrum ($\Delta^2 \propto k^{n+3}$), we have for the Gaussian sphere

$$\sigma^2 = \Delta^2 \left(k = \left[\tfrac{1}{2} \left(\tfrac{n+1}{2} \right)! \right]^{1/(n+3)} R_{\rm G}^{-1} \right). \tag{246}$$

For $n \lesssim 0$, this formula also gives a good approximation to the case of cubical cells, with $R_{\rm G} \to \ell/\sqrt{12}$. The result is rather insensitive to assumptions about the power spectrum, and just says that the variance in a cell is mainly probing waves with $\lambda \simeq 2\ell$.

Moments may also be expressed in terms of the correlation function over the sample volume:

$$\sigma^2 = \iint \xi(|\mathbf{x} - \mathbf{x}'|)\; f(\mathbf{x})\, f(\mathbf{x}')\; d^3x\; d^3x'. \tag{247}$$

To prove this, it is easiest to start from the definition of σ^2 as an integral over the power spectrum times $|f_k|^2$, write out the Fourier representations of P and f_k and use $\int \exp[i\mathbf{k}\cdot(\mathbf{x}-\mathbf{x}'+\mathbf{r})]\; d^3k = (2\pi)^3 \delta_{\rm D}^{(3)}(\mathbf{x}-\mathbf{x}'+\mathbf{r})$. Finally, it is also sometimes convenient to express things in terms of derivatives of the correlation function at zero lag. Odd derivatives vanish, but even derivatives give

$$\xi^{(2n)}(0) = (-1)^n\; \frac{\sigma_n^2}{2n+1}. \tag{248}$$

Normalization For scale-invariant spectra, a natural amplitude measure is the variance in gravitational potential per unit $\ln k$, which is a constant, independent of scale:

$$\epsilon^2 \equiv \frac{\Delta_\Phi^2}{c^4} = \frac{9}{4} \left(\frac{ck}{H_0} \right)^{-4}\; \Delta^2(k). \tag{249}$$

Two further commonly encountered measures relate to the clustering field around 10 Mpc. One is σ_8, the rms density variation when smoothed with a **top-hat filter** (sphere of uniform weight) of radius $8h^{-1}$ Mpc; this is observed to be very close to unity. The other is an integral over the correlation function:J_3

$$J_3 \equiv \int_0^r \xi(y)\; y^2 dy = \int \Delta^2(k) W(k)\; \frac{dk}{k}, \tag{250}$$

where $W(k) = (\sin kr - kr\cos kr)/k^3$. The canonical value of this is $J_3(10\,h^{-1}\,{\rm Mpc}) = 277 h^{-3}\,{\rm Mpc}^3$ (from the CfA survey; see Davis & Peebles 1983). It is sometimes more usual to use instead the dimensionless **volume-averaged correlation function** $\bar{\xi}$:

$$\bar{\xi}(r) = \frac{3}{4\pi r^3} \int_0^r \xi(x)\; 4\pi x^2\; dx = \frac{3}{r^3} J_3(r). \tag{251}$$

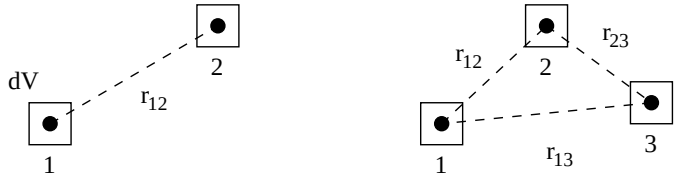

Fig. 12: The N-point correlation functions of a density field consisting of a set of particles are calculated by looking at a set of cells of volume dV (so small that they effectively only ever contain 0 or 1 particles). The Poisson probability that two cells at separation r_{12} are both occupied is $\rho_0^2 dV_1 dV_2$; with clustering, this is modified by a factor $1+\xi(r_{12})$, where ξ is the two-point correlation function. Similarly, the probability of finding a triplet of occupied cells is a factor $1+\xi(r_{12}, r_{13}, r_{23})$ times the random probability; this defines the three-point correlation function.

The canonical value then becomes $\bar{\xi}(10\,h^{-1}\,\mathrm{Mpc}) = 0.83$; this measure is clearly very close in content to $\sigma_8 = 1$.

A point to beware of is that the normalization of a theory is often quoted in terms of a value of these parameters extrapolated according to *linear* time evolution. Since the observed values are clearly nonlinear, there is no reason why theory and observation should match exactly. Even more confusingly, it is quite common in the literature to find the linear value of σ_8 called $1/b$, where b is a **bias parameter**. The implication is that $b \neq 1$ means that light does not follow mass; this may well be true in reality, but with this definition, nonlinearities will produce $b \neq 1$ even in models where mass traces light exactly. Use of this convention is not recommended.

7.2 N-point correlations

An alternative definition of the autocorrelation function is as the **two-point correlation function**, which gives the excess probability for finding a neighbour a distance r from a given galaxy (see figure 12). By regarding this as the probability of finding a pair with one object in each of the volume elements dV_1 and dV_2,

$$\boxed{dP = \rho_0^2\,[1+\xi(r)]\,dV_1\,dV_2,} \tag{252}$$

this is easily seen to be equivalent to the autocorrelation definition of ξ: $\xi = \langle \delta(x_1)\delta(x_2)\rangle$. A related quantity is the **cross-correlation function**. Here, one considers two different classes of object (a and b, say), and the cross-correlation function ξ_{ab} is defined as the (symmetric) probability of finding a pair in which dV_1 is occupied by an object from the first catalogue and dV_2 by one from the second:

$$dP = \rho_a \rho_b\,[1+\xi_{ab}(r)]\,dV_1\,dV_2. \tag{253}$$

In terms of density fields, clearly $\xi_{ab} = \langle \delta_a(x_1)\delta_b(x_2)\rangle$. Cross-correlations give information about the density profile around objects; for example, ξ_{gc} between galaxies and clusters measures the average galaxy density profile around clusters (at least out to radii where clusters overlap).

7.3 Gaussian density fields

Apart from statistical isotropy of the fluctuation field, there is another reasonable assumption we might make: that the phases of the different Fourier modes δ_k are uncorrelated and random. This corresponds to treating the initial disturbances as some form of random noise, analogous

to Johnson noise in electrical circuits; indeed, many mathematical tools that have become invaluable in cosmology were first established with applications to communication circuits in mind (*e.g.* Rice 1954). The random-phase approximation has a powerful consequence, which derives from the **central limit theorem**: loosely, the sum of a large number of independent random variables will tend to be normally distributed. This will be true not just for the field δ; all quantities that are derived from linear sums over waves (such as field derivatives) will be have a joint Normal distribution. The result is a **Gaussian random field**, whose properties are characterised entirely by its power spectrum.

Clustering of peaks Another important calculation that can be performed with density peaks is to estimate the clustering of cosmological objects. Peaks have some inbuilt clustering as a result of the statistics of the linear density field: they are 'born clustered'. For galaxies, this clustering amplitude is greatly altered by the subsequent dynamical evolution of the density field, but this is not true for clusters of galaxies, which are the largest nonlinear systems at the current epoch. We recognize clusters simply because they are the most spectacularly large galaxy systems to have undergone gravitational collapse; this has an important consequence, as first realized by Kaiser (1984). The requirement that these systems have become nonlinear by the present means that they must have been associated with particularly high peaks in the initial conditions. If we thus confine ourselves to peaks above some **density threshold** in ν, the statistical correlations can be very strong – especially for the richer clusters corresponding to high peaks.

The main effect is easy to work out, using the **peak–background split**. Here, one conceptually decomposes the density field into short-wavelength terms, which generate the peaks, plus terms of much longer wavelength, which modulate the peak number density. Consider the large-wavelength field as if it were some extra perturbation δ_+; if we select all peaks above a threshold ν in the final field, this corresponds to taking all peaks above $\delta = \nu\sigma_0 - \delta_+$ in the initial field. This varying effective threshold will now produce more peaks in the regions of high δ_+, leading to amplification of the clustering pattern. For high peaks, $P(>\nu) \propto \nu^2 e^{-\nu^2/2}$; the exponential is the most important term, leading to a perturbation $\delta P/P \simeq \nu(\delta_+/\sigma_0)$. Hence, we obtain the high-peak amplification factor for the correlation function:

$$\boxed{\xi_{\rm pk}(r) \simeq \frac{\nu^2}{\sigma_0^2}\,\xi_{\rm mass}(r).} \tag{254}$$

It is important to realize that the process as described need have nothing to do with biased galaxy formation; it works perfectly well if galaxy light traces mass exactly in the universe. Clusters occur at special places in the mass distribution, so there is no reason to expect their correlations to be the same as those of the mass field.

In more detail, the exact clustering of peaks is just an extension of the calculation of the number density of peaks. We want to find the density of peaks of height ν_2 at a distance r from a peak of height ν_1. This involves a 6×6 covariance matrix for the fields and first and second derivatives even in 1D (20×20 in 3D). Moreover, most of the elements in this matrix are non-zero, so that the analytical calculation of ξ is sadly not feasible (see Lumsden, Heavens & Peacock 1989). However, a closely related calculation is easier to solve: the correlations of **thresholded regions**. Assume that objects form with unit probability in all regions whose density exceeds some threshold value, so that we need to deal with the correlation function of a

modified density field that is constant above the threshold and zero elsewhere. This is

$$1+\xi_{>\nu}(r) = \frac{1}{[P(>\nu)]^2}\int_\nu^\infty\int_\nu^\infty \frac{dx\ dy}{2\pi[1-\psi^2(r)]^{1/2}} \times \exp\left(-\frac{x^2+y^2-2xy\psi(r)}{2[1-\psi^2(r)]}\right) \tag{255}$$

(Kaiser 1984). For high thresholds, this should be very close to the correlation function of peaks. The complete solution of this equation is given by Jensen & Szalay (1986) (see also Kashlinsky 1991 for the extension to the cross-correlation of fields above different thresholds). A good approximation, which extends Kaiser's original result, is

$$1+\xi_{>\nu} \simeq \exp\left(1+\frac{\nu^2}{\sigma_0^2}\xi_{\rm mass}\right). \tag{256}$$

There remains the question of the inclusion of dynamics into the above treatment. As the density field evolves, density peaks will move from their initial locations, and the clustering will alter. The general problem is rather nasty (see Bardeen *et al.* 1986), but things are relatively straightforward in the linear regime where the mass fluctuations are small. If the statistical enhancement of correlations produces a fractional perturbation in the numbers of thresholded objects of $\delta_{\rm statistical} = f\,\delta_{\rm mass}$, then the effect of allowing weak dynamical evolution is just

$$\delta_{\rm obs} = \delta_{\rm statistical} + \delta_{\rm mass}. \tag{257}$$

To see this, think of density perturbations arising as in the Zeldovich approximation, via objects moving closer together. Density peaks will be convected with the flow and compressed in number density in the same way as for any other particle. Thus, the effective enhancement ends up as $f \to f+1$. We can deduce the value of f for Kaiser's model by looking at the expression for the correlation function in the limit of small correlations: $f \simeq \nu/\sigma_0$. So, for large-scale correlations of high peaks, we expect

$$\xi_{\rm pk} \simeq \left(1+\frac{\nu}{\sigma_0}\right)^2\ \xi_{\rm mass}. \tag{258}$$

This idea of obtaining enhanced correlations by means of a threshold in density has been highly influential in cosmology. As well as the original application to clusters, attempts have also been made to use this mechanism to explain why galaxies might have clustering properties that differ from those of the mass.

Application to galaxy clusters This is the class of object that forms the main application of the peak clustering method. In order to model these systems as density peaks, it is necessary to specify a filter radius and a threshold; once we choose a filter radius to select cluster-sized fluctuations, the threshold is then fixed mainly by the number density (although altering the power-spectrum model also has a slight influence through γ). For Gaussian filtering, the conventional choice of R_f for clusters is $5h^{-1}$ Mpc. For $h = 1/2$ cold dark matter with $\gamma = 0.74$ on this scale, the required threshold is $\nu = 2.81$. These figures seem quite reasonable: Abell clusters are the rare high peaks of the mass distribution, and collapsed only recently. The reason for setting any threshold at all is the requirement of gravitational collapse by the present, so it is inevitable that $\nu \sim 1$.

The observations of the spatial correlations of clusters are somewhat controversial. The correlation function found by most workers is consistent with a scaled version of the galaxy

function, $\xi = (r/r_0)^{-1.8}$, but values of r_0 vary. The original value found by Bahcall & Soneira (1983) was $25\,h^{-1}$ Mpc, but later work favoured values in the range $15-20\,h^{-1}$ Mpc (Sutherland 1988; Dalton *et al.* 1992; Peacock & West 1992). The enhancement with respect to ξ for galaxies is thus a factor $\simeq 10$. Since σ_0 is close to unity for this smoothing, the simple asymptotic scaling would imply a threshold $\nu \simeq 3$, which is reasonable for moderately rare peaks.

7.4 Nonlinear clustering evolution

Observations of galaxy clustering extend into the highly nonlinear regime, $\xi \lesssim 10^4$, so it is essential to understand how this nonlinear clustering relates to the linear-theory initial conditions. A useful trick for dealing with this problem is to think of the density field under full nonlinear evolution as consisting of a set of collapsed, virialized clusters. What is the density profile of one of these objects? At least at separations smaller than the clump separation, the density profile of the clusters is directly related to the correlation function, since this just measures the number density of neighbours of a given galaxy. For a very steep cluster profile, $\rho \propto r^{-\epsilon}$, most galaxies will lie near the centres of clusters, and the correlation function will be a power law, $\xi(r) \propto r^{-\gamma}$, with $\gamma = \epsilon$. In general, because the correlation function is the convolution of the density field with itself, the two slopes differ. In the limit that clusters do not overlap, the relation is $\gamma = 2\epsilon - 3$ (for $3/2 < \epsilon < 3$; see Peebles 1974 or McClelland & Silk 1977). In any case, the critical point is that the correlation function may be be thought of as arising directly from the density profiles of clumps in the density field.

In this picture, it is easy to see how ξ will evolve with redshift, since clusters are virialized objects that do not expand. The hypothesis of **stable clustering** states that, although the separation of clusters will alter as the universe expands, their internal density structure will stay constant with time. This hypothesis clearly breaks down in the outer regions of clusters, where the density contrast is small and linear theory applies, but it should be applicable to small-scale clustering. Regarding ξ as a density profile, its small-scale shape should therefore be fixed in *proper* coordinates, and its amplitude should scale as $(1+z)^{-3}$ owing to the changing mean density of unclustered galaxies, which dilute the clustering at high redshift. Thus, with $\xi \propto r^{-\gamma}$, we obtain the comoving evolution

$$\boxed{\xi(r,z) \propto (1+z)^{\gamma-3} \quad \text{(nonlinear)}.} \tag{259}$$

Since the observed $\gamma \simeq 1.8$, this implies slower evolution than is expected in the linear regime:

$$\boxed{\xi(r,z) \propto (1+z)^{-2} g(\Omega) \quad \text{(linear)}.} \tag{260}$$

This argument does not so far give a relation between the nonlinear slope γ and the index n of the linear spectrum. However, the linear and nonlinear regimes match at the scale of quasilinearity, *i.e.* $\xi(r_0) = 1$; each regime must make the same prediction for how this break point evolves. The linear and nonlinear predictions for the evolution of r_0 are respectively $r_0 \propto (1+z)^{-2/(n+3)}$ and $r_0 \propto (1+z)^{-(3-\gamma)/\gamma}$, so that $\gamma = (3n+9)/(n+5)$. In terms of an effective index $\gamma = 3 + n_{\rm NL}$, this becomes

$$\boxed{n_{\rm NL} = -\frac{6}{5+n}.} \tag{261}$$

The power spectrum resulting from power-law initial conditions will evolve self-similarly with this index. Note the narrow range predicted: $-2 < n_{\rm NL} < -1$ for $-2 < n < +1$, with an $n = -2$ spectrum having the same shape in both linear and nonlinear regimes.

Indications from the angular clustering of faint galaxies (Efstathiou *et al.* 1991) and directly from redshift surveys (Le Fèvre *et al.* 1996) are that the observed clustering of galaxies evolves at about the linear-theory rate, rather more rapidly than the scaling solution would indicate. However, any interpretation of such data needs to assume that galaxies are unbiased tracers of the mass, whereas the observed high amplitude of clustering of quasars at $z \simeq 1$ ($r_0 \simeq 7h^{-1}\,$Mpc; see Shanks *et al.* 1987, Shanks & Boyle 1994) warns that at least some high-redshift objects have clustering that is apparently not due to gravity alone.

For many years it was thought that only these limiting cases of extreme linearity or nonlinearity could be dealt with analytically, but in a marvelous piece of alchemy, Hamilton *et al.* (1991; HKLM) suggested a general way of understanding the linear $\leftrightarrow$ nonlinear mapping. The conceptual basis of their method can be understood with reference to the spherical collapse model. For $\Omega = 1$, a spherical clump virializes at a density contrast of order 100 when the linear contrast is of order unity. The trick now is to think about the density contrast in two distinct ways. To make a connection with the statistics of the density field, the correlation function $\xi(r)$ may be taken as giving a typical clump profile. What matters for collapse is that the integrated overdensity within a given radius reaches a critical value, so one should work with the volume-averaged correlation function $\bar{\xi}(r)$:

$$\bar{\xi}(R) \equiv \frac{3}{4\pi R^3} \int_0^R \xi(r)\, 4\pi r^2\, dr. \tag{262}$$

A density contrast of $1+\delta$ can also be thought of as arising through collapse by a factor $(1+\delta)^{1/3}$ in radius, which suggests that a given nonlinear correlation $\bar{\xi}_{\rm NL}(r_{\rm NL})$ should be thought of as resulting from linear correlations on a linear scale:

$$r_{\rm L} = [1 + \bar{\xi}_{\rm NL}(r_{\rm NL})]^{1/3} r_{\rm NL}. \tag{263}$$

This is the first part of the **HKLM procedure**. Having performed this translation of scales, the second step is to conjecture that the nonlinear correlations are a universal function of the linear ones:

$$\bar{\xi}_{\rm NL}(r_{\rm NL}) = f_{\rm NL}(\bar{\xi}_{\rm L}(r_{\rm L})). \tag{264}$$

The asymptotics of the function can be deduced readily. For small arguments $x \ll 1$, $f_{\rm NL}(x) \simeq x$; the spherical collapse argument suggests $f_{\rm NL}(1) \simeq 10^2$. Following collapse, $\bar{\xi}_{\rm NL}$ depends on scale factor as a^3 (stable clustering), whereas $\bar{\xi}_{\rm L} \propto a^2$; the large-$x$ limit is therefore $f_{\rm NL}(x) \propto x^{3/2}$. HKLM deduced from numerical experiments a numerical fit that interpolated between these two regimes, in a manner that empirically showed negligible dependence on the power spectrum.

However, these equations are often difficult to use stably for numerical evaluation; it is better to work directly in terms of power spectra. The key idea here is that $\bar{\xi}(r)$ can often be thought of as measuring the power at some effective wavenumber: it is obtained as an integral of the product of $\Delta^2(k)$, which is often a rapidly rising function, and a window function that cuts off rapidly at $k \gtrsim 1/r$:

$$\bar{\xi}(r) = \Delta^2(k_{\rm eff}), \qquad k_{\rm eff} \simeq 2/r, \tag{265}$$

where n is the effective power-law index of the power spectrum. This approximation for the effective wavenumber is within 20 per cent of the exact answer over the range $-2 < n < 0$. In most circumstances, it is therefore an excellent approximation to use the HKLM formulae directly to scale wavenumbers and powers:

$$\boxed{\begin{aligned} &\Delta^2_{\rm NL}(k_{\rm NL}) = f_{\rm NL}(\Delta^2_{\rm L}(k_{\rm L})) \\ &k_{\rm L} = [1 + \Delta^2_{\rm NL}(k_{\rm NL})]^{-1/3} k_{\rm NL}. \end{aligned}} \tag{266}$$

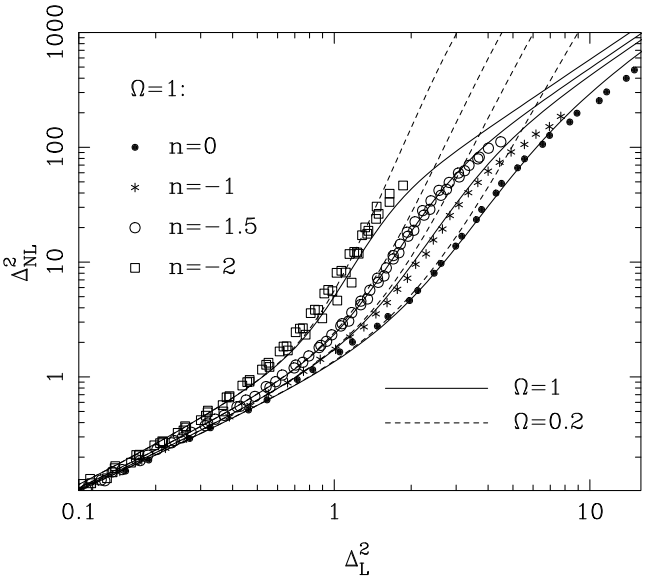

Fig. 13: The generalization of the HKLM function relating nonlinear power to linear power, for the cases $n = 0$, -1, -1.5 and -2. Data points are shown for the case $\Omega = 1$ only, with the corresponding fitting formulae shown as solid lines. This diagram clearly displays three regimes: (i) linear ($\Delta^2_{\rm NL} \lesssim 1$); (ii) quasilinear ($1 \lesssim \Delta^2_{\rm NL} \lesssim 100$); (iii) stable-clustering ($\Delta^2_{\rm NL} \gtrsim 100$). For a given linear power, the nonlinear power increases for more negative n. There is also a greater nonlinear response in the case of an open universe with $\Omega = 0.2$, indicated by the dashed lines. The fitting formula is shown for models with zero vacuum energy only, but what matters in general is the Ω-dependent linear growth suppression factor $g(\Omega)$.

What about models with $\Omega \neq 1$? The argument that leads to the $f_{\rm NL}(x) \propto x^{3/2}$ asymptote in the nonlinear transformation is just that linear and nonlinear correlations behave as a^2 and a^3 respectively following collapse. If collapse occurs at high redshift, then $\Omega = 1$ may be assumed at that time, and the nonlinear correlations still obey the a^3 scaling to low redshift. All that has changed is that the linear growth is suppressed by some Ω-dependent factor $g(\Omega)$. According to Carroll, Press & Turner (1992), the required factor may be approximated almost exactly by

$$g(\Omega) = \tfrac{5}{2}\Omega_m \left[\Omega_m^{4/7} - \Omega_v + (1 + \tfrac{1}{2}\Omega_m)(1 + \tfrac{1}{70}\Omega_v)\right]^{-1}. \tag{267}$$

where we have distinguished matter (m) and vacuum (v) contributions to the density parameter explicitly. It then follows that the large-x asymptote of the nonlinear function is

$$f_{\rm NL}(x) \propto [g(\Omega)]^{-3}\, x^{3/2}. \tag{268}$$

This says that the amplitude of highly nonlinear clustering is greater for low-density universes.

The suggestion of HKLM was that $f_{\rm NL}$ might be independent of the form of the linear spectrum, but Jain, Mo & White (1995) showed that this is not true, especially when the linear spectrum is rather flat ($n \lesssim -1.5$). Peacock & Dodds (1996) suggested that the HKLM method should be generalized by using the following fitting formula for the n-dependent nonlinear function (strictly, the one that applies to the power spectrum, rather than to $\bar{\xi}$):

$$f_{\rm NL}(x) = x\ \left[\frac{1 + B\beta x + [Ax]^{\alpha\beta}}{1 + ([Ax]^{\alpha} g^3(\Omega)/[Vx^{1/2}])^{\beta}}\right]^{1/\beta}. \tag{269}$$

B describes a second-order deviation from linear growth; A and α parameterize the power law that dominates the function in the quasilinear regime; V is the virialization parameter that gives

the amplitude of the $f_{\rm NL}(x) \propto x^{3/2}$ asymptote; β softens the transition between these regimes. An excellent fit to N-body data (illustrated in figure 13) is given by the following spectrum dependence of the expansion coefficients:

$$\begin{aligned} A &= 0.542\,(1+n/3)^{-0.685} \\ B &= 0.097\,(1+n/3)^{-0.224} \\ \alpha &= 3.235\,(1+n/3)^{-0.236} \\ \beta &= 0.659\,(1+n/3)^{-0.356} \\ V &= 11.54\,(1+n/3)^{-0.371}. \end{aligned} \tag{270}$$

The more general case of curved spectra can be dealt with very well by using the tangent spectral index at each linear wavenumber:

$$n_{\rm eff} \equiv \frac{d\ln P}{d\ln k}, \tag{271}$$

although evaluating $n_{\rm eff}$ at a wavenumber of $k/2$ gives even better results.

Evolution of clustering The discussion of nonlinear evolution has revealed that in practice the regime $1 \lesssim \Delta^2 \lesssim 100$ is dominated by the steep **quasilinear transition** where $f_{\rm NL}(x) \propto x^\alpha$, $\alpha \simeq 3.5$–4.5. This turns out to predict a rate of evolution that is very different from the extremes of linear evolution or stable clustering. For $\Delta^2 \gg 1$, the transitional spectrum scales as

$$\begin{aligned} k_{\rm NL} &\simeq [\Delta^2_{\rm NL}]^{1/3} k_{\rm L} \\ \Delta^2_{\rm NL} &\propto [D^2(a)\,\Delta^2_{\rm L}]^{1+\alpha}, \end{aligned} \tag{272}$$

where $D(a)$ is the linear growth law for density perturbations. For a power-law linear spectrum, $\Delta^2 \propto k^{3+n}$, this predicts a quasilinear power law

$$\Delta^2_{\rm NL} \propto D^{(6-2\gamma)(1+\alpha)/3}\, k^\gamma_{\rm NL}, \tag{273}$$

where the nonlinear power-law index depends as follows on the slope of the linear spectrum:

$$\gamma = \frac{3(3+n)(1+\alpha)}{3+(3+n)(1+\alpha)}. \tag{274}$$

For the observed index of $\gamma \simeq 1.8$, this would require $n \simeq -2.2$, very different from the $n = 0$ that would give $\gamma = 1.8$ in the virialized regime.

We can now summarize the rate of evolution of clustering in the three different regimes:

$$\boxed{\begin{aligned} \text{linear}: &\quad \xi(r,z) \propto [D(z)]^2 \\ \text{quasilinear}: &\quad \xi(r,z) \propto [D(z)]^{(6-2\gamma)(1+\alpha)/3} \\ \text{nonlinear}: &\quad \xi(r,z) \propto (1+z)^{-(3-\gamma)}, \end{aligned}} \tag{275}$$

where γ is the power-law slope in the relevant regime. For $\alpha = 4$, $\gamma = 1.7$, this gives $\xi \propto D^{4.3}$ for the quasilinear evolution; this is more than twice as fast as the linear evolution, and over three times the rate of stable-clustering evolution if $\Omega = 1$, so that $D(z) = 1/(1+z)$. The conclusion is that clustering in the regime where most data exist is expected to evolve very rapidly with redshift, unless Ω is low. We discuss below whether this effect has been seen.

7.5 Real-space clustering

It is possible to avoid the complications of redshift space. One can deal with pure two-dimensional projected clustering, as discussed in the next section. Alternatively, peculiar velocities may be dealt with by using the correlation function evaluated explicitly as a 2D function of transverse ($r_\perp$) and radial ($r_\parallel$) separation. Integrating along the redshift axis then gives the **projected correlation function**, which is independent of the velocities

$$\boxed{w_p(r_\perp) \equiv \int_{-\infty}^{\infty} \xi(r_\perp, r_\parallel)\, dr_\parallel = 2\int_{r_\perp}^{\infty} \xi(r)\, \frac{r\, dr}{(r^2 - r_\perp^2)^{1/2}}.} \tag{276}$$

In principle, this statistic can be used to recover the real-space correlation function by using the inverse relation for the **Abel integral equation**:

$$\xi(r) = -\frac{1}{\pi}\int_r^{\infty} w_p'(y)\, \frac{dy}{(y^2 - r^2)^{1/2}}. \tag{277}$$

An alternative notation for the projected correlation function is $\Xi(r_\perp)$ (Saunders, Rowan-Robinson & Lawrence 1992). Note that the projected correlation function is not dimensionless, but has dimensions of length. The quantity $\Xi(r_\perp)/r_\perp$ is more convenient to use in practice as the projected analogue of $\xi(r)$.

The reason that $w_p(r_\perp)$ is independent of redshift-space distortions is that peculiar velocities simply move pairs of points in $r_\parallel$, but not in $r_\perp$, and the expected pair count is just proportional to $2\pi r_\perp\, dr_\perp\, dr_\parallel$. Suppose we ignore the linear-theory velocities (which are more easily treated in Fourier space as above), and just consider the effect of a small-scale velocity dispersion. The correlation function is then convolved in the radial direction:

$$\begin{aligned} \xi(r_\perp, r_\parallel) &= \int_{-\infty}^{\infty} \xi_{\rm true}(r_\perp, r)\, f(r_\parallel - r)\, dr \\ &= \frac{r_0^\gamma}{\sqrt{2\pi}\,\sigma_v}\int_{-\infty}^{\infty} [r_\perp^2 + (r_\parallel - x)^2]^{-\gamma/2}\, e^{-x^2/2\sigma_v^2}\, dx, \end{aligned} \tag{278}$$

where the latter expression applies for power-law clustering and a Gaussian velocity dispersion. Looking at the function in the redshift direction thus allows the pairwise velocity dispersion to be estimated; this is the origin of the above estimate of σ_p. See Fisher (1995) for more discussion of this method.

Sometimes these complications are neglected, and correlations are calculated in redshift space assuming isotropy. The result is a small increase in scalelength, as power on small scales is transferred to separations of order the velocity smearing. The result is a scale length around $7h^{-1}$ Mpc for the redshift-space $\xi(s)$ as opposed to the $5h^{-1}$ Mpc that applies for $\xi(r)$.

Projection on the sky A more common situation is where we lack any distance data; we then deal with a projection on the sky of a magnitude-limited set of galaxies at different depths. The statistic that is observable is the angular correlation function, $w(\theta)$, or its angular power spectrum Δ_θ^2. If the sky were flat, the relation between these would be the usual **Hankel transform** pair:

$$\boxed{\begin{aligned} w(\theta) &= \int_0^{\infty} \Delta_\theta^2\, J_0(K\theta)\, dK/K \\ \Delta_\theta^2 &= K^2 \int_0^{\infty} w(\theta)\, J_0(K\theta)\, \theta\, d\theta. \end{aligned}} \tag{279}$$

For power-law clustering, $w(\theta) = (\theta/\theta_0)^{-\epsilon}$, this gives

$$\Delta^2_\theta(K) = (K\theta_0)^\epsilon \, 2^{1-\epsilon} \, \frac{\Gamma(1-\epsilon/2)}{\Gamma(\epsilon/2)}, \tag{280}$$

which is equal to $0.77(K\theta_0)^\epsilon$ for $\epsilon = 0.8$. At large angles, these relations are not quite correct. We should really expand the sky distribution in **spherical harmonics**:

$$\delta(\hat{\mathbf{q}}) = \sum a^m_\ell Y_{\ell m}(\hat{\mathbf{q}}), \tag{281}$$

where $\hat{\mathbf{q}}$ is a unit vector that specifies direction on the sky. The functions $Y_{\ell m}$ are the eigenfunctions of the angular part of the ∇^2 operator: $Y_{\ell m}(\theta,\phi) \propto \exp(im\phi) P^m_\ell(\cos\theta)$, where P^m_ℓ are the **associated Legendre polynomials** (see *e.g.* section 6.8 of Press *et al.* 1992). Since the spherical harmonics satisfy the orthonormality relation $\int Y_{\ell m} Y^*_{\ell' m'} \, d^2q = \delta_{\ell\ell'}\delta_{mm'}$, the inverse relation is

$$a^m_\ell = \int \delta(\hat{\mathbf{q}}) \, Y^*_{\ell m} \, d^2q. \tag{282}$$

The analogues of the Fourier relations for the correlation function and power spectrum are

$$\boxed{\begin{aligned} w(\theta) &= \frac{1}{4\pi} \sum_\ell \sum_{m=-\ell}^{m=+\ell} |a^m_\ell|^2 \, P_\ell(\cos\theta) \\ |a^m_\ell|^2 &= 2\pi \int_{-1}^{1} w(\theta) \, P_\ell(\cos\theta) \, d\cos\theta. \end{aligned}} \tag{283}$$

For small θ and large ℓ, these go over to a form that looks like a flat sky, as follows. Consider the asymptotic forms for the Legendre polynomials and the J_0 Bessel function:

$$\begin{aligned} P_\ell(\cos\theta) &\simeq \sqrt{\frac{2}{\pi\ell\sin\theta}} \cos\left[\left(\ell+\tfrac{1}{2}\right)\theta - \tfrac{1}{4}\pi\right] \\ J_0(z) &\simeq \sqrt{\frac{2}{\pi z}} \cos\left[z - \tfrac{1}{4}\pi\right], \end{aligned} \tag{284}$$

for respectively $\ell \to \infty$, $z \to \infty$; see chapters 8 & 9 of Abramowitz & Stegun 1965. This shows that, for $\ell \gg 1$, we can approximate the small-angle correlation function in the usual way in terms of an angular power spectrum Δ^2_θ and angular wavenumber K:

$$w(\theta) = \int_0^\infty \Delta^2_\theta(K) \, J_0(K\theta) \, \frac{dK}{K}, \qquad \Delta^2_\theta\left(K = \ell + \tfrac{1}{2}\right) = \frac{2\ell+1}{8\pi} \sum_m |a^m_\ell|^2. \tag{285}$$

An important relation is that between the angular and spatial power spectra. In outline, this is derived as follows. The perturbation seen on the sky is

$$\delta(\hat{\mathbf{q}}) = \int_0^\infty \delta(\mathbf{y}) \, y^2 \phi(y) \, dy, \tag{286}$$

where $\phi(y)$ is the **selection function**, normalized such that $\int y^2\phi(y)\,dy = 1$, and y is comoving distance. The function ϕ is the comoving density of objects in the survey, which is given by the integrated luminosity function down to the luminosity limit corresponding to the limiting flux of the survey seen at different redshifts; a flat universe ($\Omega = 1$) is assumed for now. Now write down the Fourier expansion of δ. The plane waves may be related to spherical harmonics via

the expansion of a plane wave in **spherical Bessel functions** $j_\ell(x) = (\pi/2x)^{1/2} J_{n+1/2}(x)$ (see chapter 10 of Abramowitz & Stegun 1965 or section 6.7 of Press *et al.* 1992):

$$e^{ikr\cos\theta} = \sum_0^\infty (2\ell+1)\; i^\ell\, P_\ell(\cos\theta)\; j_\ell(kr), \tag{287}$$

plus the spherical harmonic addition theorem

$$P_\ell(\cos\theta) = \frac{4\pi}{2\ell+1} \sum_{m=-\ell}^{m=+\ell} Y^*_{\ell m}(\hat{\mathbf{q}}) Y_{\ell m}(\hat{\mathbf{q}}'); \quad \hat{\mathbf{q}}\cdot\hat{\mathbf{q}}' = \cos\theta. \tag{288}$$

These relations allow us to take the angular correlation function $w(\theta) = \langle \delta(\hat{\mathbf{q}})\delta(\hat{\mathbf{q}}')\rangle$ and transform it to give the angular power spectrum coefficients. The actual manipulations involved are not as intimidating as they may appear, but they are left as an exercise and we simply quote the final result (Peebles 1973):

$$\boxed{\left\langle |a_\ell^m|^2 \right\rangle = 4\pi \int \Delta^2(k)\; \frac{dk}{k}\; \left[\int y^2 \phi(y)\, j_\ell(ky)\; dy\right]^2.} \tag{289}$$

What is the analogue of this formula for small angles? Rather than manipulating large-ℓ Bessel functions, it is easier to start again from the correlation function. By writing as above the overdensity observed at a particular direction on the sky as a radial integral over the spatial overdensity, with a weighting of $y^2\phi(y)$, we see that the angular correlation function is

$$\langle \delta(\hat{\mathbf{q}}_1)\delta(\hat{\mathbf{q}}_2)\rangle = \iint \langle \delta(\mathbf{y}_1)\delta(\mathbf{y}_2)\rangle\; y_1^2 y_2^2 \phi(y_1)\phi(y_2)\; dy_1\; dy_2. \tag{290}$$

We now change variables to the mean and difference of the radii, $y \equiv (y_1+y_2)/2$; $x \equiv (y_1 - y_2)$. If the depth of the survey is larger than any correlation length, we only get a signal when $y_1 \simeq y_2 \simeq y$. If the selection function is a slowly varying function, so that the thickness of the shell being observed is also of order the depth, the integration range on x may be taken as being infinite. For small angles, we then obtain **Limber's equation**:

$$\boxed{w(\theta) = \int_0^\infty y^4\phi^2\, dy \int_{-\infty}^\infty \xi\left(\sqrt{x^2 + y^2\theta^2}\right)\; dx} \tag{291}$$

(see sections 51 and 56 of Peebles 1980). Theory usually supplies a prediction about the linear density field in the form of the power spectrum, and so it is convenient to recast Limber's equation:

$$w(\theta) = \int_0^\infty y^4\phi^2\, dy \int_0^\infty \pi\, \Delta^2(k)\, J_0(ky\theta)\, dk/k^2. \tag{292}$$

The form $\phi \propto y^{-1/2}\exp[-(y/y^*)^2]$ is often taken as a reasonable approximation to the Schechter function, and this gives

$$w(\theta) = \frac{\pi}{2\Gamma^2\left(\frac{5}{4}\right)} \int_0^\infty \Delta^2(k)\, e^{-(k\theta y^*)^2/8} \left[1 - \tfrac{1}{8}(k\theta y^*)^2\right] \frac{dk}{k^2 y^*}. \tag{293}$$

The power-spectrum version of Limber's equation is already in the form required for the relation to the angular power spectrum ($w = \int \Delta_\theta^2 J_0(K\theta)\, dK/K$), and so we obtain the direct small-angle relation between spatial and angular power spectra:

$$\boxed{\Delta_\theta^2 = \frac{\pi}{K} \int \Delta^2(K/y)\; y^5 \phi^2(y)\; dy.} \tag{294}$$

This is just a convolution in log space, and is considerably simpler to evaluate and interpret than the $w - \xi$ version of Limber's equation.

Finally, note that it is not difficult to make allowance for spatial curvature in the above discussion. Write the Robertson–Walker metric in the form

$$c^2 d\tau^2 = c^2 dt^2 - R^2 \left[\frac{dr^2}{1 - kr^2} + r^2\theta^2 \right]; \tag{295}$$

for $k = 0$, the notation $y = R_0 r$ was used for comoving distance, where $R_0 = (c/H_0)|1 - \Omega|^{-1/2}$. The radial increment of comoving distance was $dx = R_0 dr$, and the comoving distance between two objects was $(dx^2 + y^2\theta^2)^{1/2}$. To maintain this version of Pythagoras's theorem, we clearly need to keep the definition of y and redefine radial distance: $dx = R_0 dr C(y)$, where $C(y) = [1 - k(y/R_0)^2]^{-1/2}$. The factor $C(y)$ appears in the non-Euclidean comoving volume element, $dV \propto y^2 C(y) dy$, so that we now require the normalization equation for ϕ to be

$$\int_0^\infty y^2 \phi(y) C(y) dy = 1. \tag{296}$$

The full version of Limber's equation therefore gains two powers of $C(y)$, but one of these is lost in converting between $R_0 dr$ and dx:

$$w(\theta) = \int_0^\infty [C(y)]^2 y^4 \phi^2\, dy \int_{-\infty}^\infty \xi\left(\sqrt{x^2 + y^2\theta^2}\right) \frac{dx}{C(y)}. \tag{297}$$

The net effect is therefore to replace $\phi^2(y)$ by $C(y)\phi^2(y)$, so that the full power-spectrum equation is

$$\Delta_\theta^2 = \frac{\pi}{K} \int \Delta^2(K/y)\; C(y) y^5 \phi^2(y)\; dy. \tag{298}$$

It is also straightforward to allow for evolution. The power version of Limber's equation is really just telling us that the angular power from a number of different radial shells adds incoherently, so we just need to use the actual evolved power at that redshift. These integral equations can be inverted numerically to obtain the real-space 3D clustering results from observations of 2D clustering; see Baugh & Efstathiou (1993; 1994).

7.6 Measuring the clustering spectrum

The history of attempts to quantify galaxy clustering goes back to Hubble's demonstration that the distribution of galaxies on the sky was non-uniform. The major post-war landmarks were the angular analysis of the Lick catalogue, described in Peebles (1980), and the analysis of the CfA redshift survey (Davis & Peebles 1983). It has taken some time to obtain data on samples that greatly exceed these in depth, but several pieces of work appeared around the start of the 1990s that clarified many of the discrepancies between different surveys and which paint a relatively consistent picture of large-scale structure. Perhaps the most significant of these surveys have been the **APM survey** (automatic plate-measuring machine survey; see Maddox

et al. 1990 and Maddox *et al.* 1996), the **CfA survey** (Center for Astrophysics survey; see Huchra *et al.* 1990) and the **LCRS** (Las Campanas redshift survey; see Shectman *et al.* 1996), together with a variety of surveys based on galaxies selected at 60μm by theinfrared astronomy satellite infrared astronomy satellite, **IRAS**IRAS satellite (Saunders *et al.* 1991; Fisher *et al.* 1993). These surveys are sensitive to rather different galaxy populations: the APM, CfA and LCRS surveys select in blue light and are sensitive to stellar populations of intermediate age; the IRAS emission originates from hot dust, associated with bursts of active star formation. A compilation of clustering results for a variety of tracers was given by Peacock & Dodds (1994); the results are shown in figure 14.

There is a wide range of power measured, ranging over perhaps a factor 20 between the real-space APM galaxies and the rich Abell clusters. Are these measurements all consistent with one Gaussian power spectrum for mass fluctuations? Corrections for redshift-space distortions and nonlinearities can be applied to these data to reconstruct the linear mass fluctuations, subject to an unknown degree of bias. The simplest assumption for this is that the bias is a linear response of the galaxy-formation process, and may be taken as independent of scale:

$$\Delta^2_{\rm tracer} = b^2 \Delta^2_{\rm mass}. \tag{299}$$

There thus exist five free parameters that can be adjusted to optimize the agreement between the various estimates of the linear power spectrum; these are Ω and the four bias parameters for Abell clusters, radio galaxies, optical galaxies and IRAS galaxies ($b_{\rm A}$, $b_{\rm R}$, $b_{\rm O}$, $b_{\rm I}$); however, only two of these really matter: Ω and some measure of the overall level of fluctuations. For now, we take the IRAS bias parameter to play this latter role. Once these two are specified, the other bias parameters are well determined, principally from the linear data at small k, and have the approximate ratios

$$b_{\rm A} : b_{\rm R} : b_{\rm O} : b_{\rm I} = 4.5 : 1.9 : 1.3 : 1 \tag{300}$$

(Peacock & Dodds 1994). The reasons why different galaxy tracers may show different strengths of clustering are discussed above. Rich clusters are inevitably biased with respect to the mass, simply through the statistics of rare high-density regions. Massive ellipticals such as radio galaxies share some of this bias through the effect of morphological segregation, which says that the E/S0 fraction rises in clusters to almost 100%, by comparison with a mean of 20%. At high overdensities, the fraction of optical galaxies that are IRAS galaxies declines by a factor $\simeq 3$–5 (Strauss *et al.* 1992), reflecting the fact that IRAS galaxies are mainly spirals. Generally, the analysis of galaxies of a given type assumes that the luminosity function is independent of environment so that bias is independent of luminosity. This is not precisely true, and the amplitude of ξ does appear to rise slightly for galaxies of luminosity above several times L^*. (Valls-Gabaud *et al.* 1989; Loveday *et al.* 1995; Benoist *et al.* 1996). However, for the bulk of the galaxies in a given population, it is a good approximation to say that **luminosity segregation** can be neglected.

The various reconstructions of the linear power spectrum for the case $\Omega = b_{\rm I} = 1$ are shown superimposed in figure 14, and display an impressive degree of agreement. This argues very strongly that what we measure from large-scale galaxy clustering has a direct relation to mass fluctuations, rather than being an optical illusion caused by non-uniform galaxy-formation efficiency (Bower *et al.* 1993). If effects other than gravity were dominant, the shape of spectrum inferred from clusters would have a very different shape at large scales, contrary to observation.

Large-scale power-spectrum data and models It is interesting to ask whether the power spectrum contains any features or whether it is consistent with a single smooth curve. A convenient description is in terms of the CDM power spectrum, which is $\Delta^2(k) \propto k^{n+3} T_k^2$. The CDM

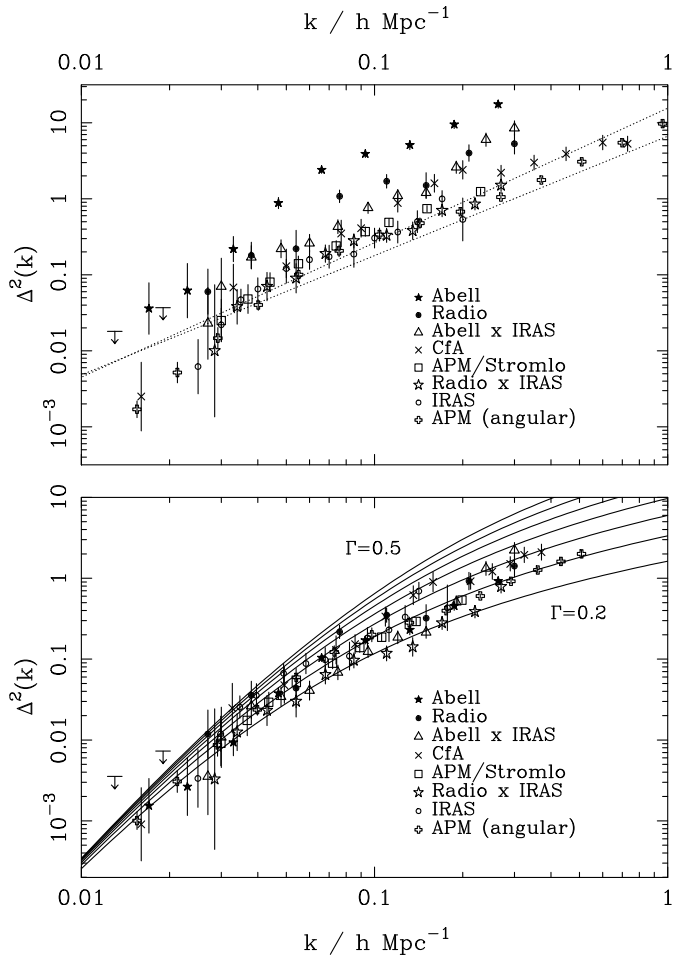

Fig. 14: A compilation of power-spectrum data, adapted from Peacock & Dodds (1994). The upper panel shows raw power-spectrum data in the form $\Delta^2 \equiv d\sigma^2/d\ln k$; all data with the exception of the APM power spectrum are in redshift space. The two dotted lines shown for reference are the transforms of the canonical real-space correlation functions for optical and IRAS galaxies ($r_0 = 5\,h^{-1}$ Mpc and $3.78\,h^{-1}$ Mpc and slopes of 1.8 and 1.57 respectively). The lower panel shows the results of correcting these datasets for different degrees of bias and for nonlinear evolution. There is an excellent degree of agreement, particularly in the detection of a break around $k = 0.03h$. The data are compared to various CDM models, assuming scale-invariant initial conditions, with the same large-wavelength normalization. Values of the fitting parameter $\Gamma = 0.5,\ 0.45,\ \ldots 0.25,\ 0.2$ are shown. The best-fit model has $\Gamma = 0.25$.

spectrum is very commonly used as a basis for comparison with cosmological observations, and it is essential to realize that this can be done in two ways. The CDM physics can be accepted, in which case the best-fitting value of Γ constrains Ω, h and $\Omega_{\rm B}$. Alternatively, the CDM spectrum can be used as a completely empirical fitting formula, which is assumed to approximate some different set of physics. For example, the MDM model includes CDM and an admixture of massive neutrinos; over a limited range of k, this will appear similar to a CDM spectrum, but with an effective value of Γ that is very much less than Ωh, because of the way in which neutrinos remove small-scale power in this case.

The normalization of the spectrum is specified by the rms variation in the fractional density contrast, averaged over $8\,h^{-1}\,$Mpc spheres; for CDM-like spectra, this measures power at an effective wavenumber well approximated by

$$\sigma_8^2 = \Delta^2(k_{\rm eff}), \qquad k_{\rm eff}/\,h\,{\rm Mpc}^{-1} = 0.172 + 0.011\,[\ln(\Gamma/0.34)]^2. \tag{301}$$

Fitting this spectrum to the large-scale linearised data of figure 14 requires the parameters

$$\Gamma \simeq 0.25 + 0.3(1/n - 1), \tag{302}$$

in agreement with many previous arguments suggesting that an apparently low-density model is needed; the linear transfer function does not bend sharply enough at the break wavenumber if a 'standard' high-density $\Gamma = 0.5$ model is adopted. For any reasonable values of h and baryon density, a high-density CDM model is not viable. Even a high degree of 'tilt' in the primordial spectrum (Cen *et al.* 1992) does not help change this conclusion unless n is set so low that major difficulties result when attempting to account for microwave-background anisotropies.

An important general lesson can also be drawn from the lack of large-amplitude features in the power spectrum. This is a strong indication that collisionless matter is deeply implicated in forming large-scale structure. Purely baryonic models contain large bumps in the power spectrum around the Jeans' length prior to recombination ($k \sim 0.03\Omega h^2$ Mpc^{-1}), whether the initial conditions are isocurvature or adiabatic. It is hard to see how such features can be reconciled with the data, beyond a 'visibility' in the region of 20%.

These ideas can be illustrated with a simple empirical model. Consider a spectrum in the form of a break between two power laws:

$$\Delta^2(k) = \frac{(k/k_0)^\alpha}{1 + (k/k_1)^{\alpha-\beta}}. \tag{303}$$

As shown in figure 16, the nonlinear power that results from this linear spectrum matches the data very nicely, if we choose the parameters

$$\begin{aligned} k_0 &= 0.3\ h\,{\rm Mpc}^{-1} \\ k_1 &= 0.05\ h\,{\rm Mpc}^{-1} \\ \alpha &= 0.8 \\ \beta &= 4.0. \end{aligned} \tag{304}$$

A value of $\beta = 4$ corresponds to a scale-invariant spectrum at large wavelengths, whereas the effective small-scale index is $n = -2.2$. We consider below the physical ways in which a spectrum of this shape might arise.

Finally, it is important to note that bias of either sign may have to be considered. A high-density universe with the cluster-normalized value of σ_8 predicts clustering well below that observed. However, the opposite is true for low Ω. If we simply take the APM power spectrum and ignore nonlinear corrections, the apparent value of σ_8 is about 0.9. Contrast this with the

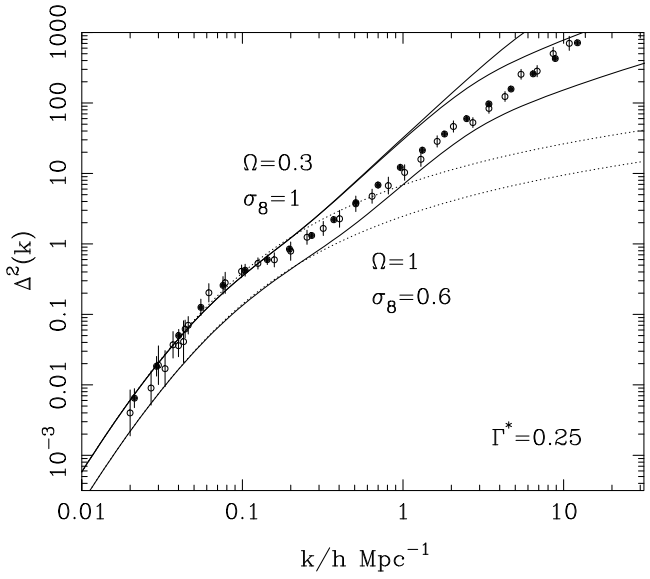

Fig. 15: The clustering data for optical galaxies, compared to three models with $\Gamma = 0.25$: $\Omega = 1$, $\sigma_8 = 0.6$; $\Omega_m = 0.3$, $\Omega_v = 0$, $\sigma_8 = 1$; $\Omega_m = 0.3$, $\Omega_v = 0.7$, $\sigma_8 = 1$. The linear spectra are shown dotted; solid lines denote evolved nonlinear spectra. All these models are chosen with a normalization that is approximately correct for the rich-cluster abundance and large-scale peculiar velocities. In all cases, the shape of the predicted spectrum fails to match observation. The high-density model would require a bias that is not a monotonic function of scale, whereas the low-density models exceed the observed small-scale clustering (figure adapted from Peacock 1997).

prediction of the cluster-normalization formula, which requires $\sigma_8 = 1.4$ for $\Omega = 0.2$, or 2.1 for $\Omega = 0.1$. Thus, low-density models inevitably require significant **antibias**, and we would have to consider the possibility that galaxy formation was suppressed in high-density regions. Bias in this sense has one advantage over positive bias, since it will tend to make the predicted small-scale spectrum less steep, which figure 15 suggests may be required in order to match the data. However, as discussed above, it is implausible that the scale dependence of the bias will be very extreme; a model that matches the data at $k \simeq 1\,h\,\mathrm{Mpc}^{-1}$ will probably significantly undershoot the 'bump' at $k \simeq 0.1\,h\,\mathrm{Mpc}^{-1}$. This large-scale feature therefore has a critical importance in the interpretation of large-scale structure. If it is correct, then the simplest CDM models fail and must be replaced by something more complicated. However, if future observations should yield lower power values at this point, then a low-density CDM model with antibias would provide a model for large-scale structure that is attractive in many ways (Jing *et al.* 1998).

7.7 Peculiar velocity fields

A research topic that has assumed increasing importance since about 1986 is the subject of deviations from the Hubble flow. Although the relation $v = Hr$ is a good approximation, it has long been known that individual galaxies have random velocities of a few hundred $\mathrm{km\,s}^{-1}$ superimposed on the general expansion. An exciting development has been the realization that these peculiar velocities display large-scale coherence in the form of **bulk flows** or **streaming flows**, which gives us the chance to probe very large-scale density fluctuations in the universe, and perhaps even to measure its mean density. Detailed reviews of these developments are given by Dekel (1994) and Strauss & Willick (1995).

In linear perturbation theory, the peculiar velocity is parallel to the peculiar gravitational

acceleration $\mathbf{g} = \boldsymbol{\nabla}\delta\Phi/a$:

$$\delta\mathbf{v} = \frac{2f(\Omega)}{3H\Omega}\mathbf{g}, \tag{305}$$

where

$$f(\Omega) \equiv \left(\frac{a}{\delta}\right)\frac{d\delta}{da} \simeq \Omega^{0.6}. \tag{306}$$

Alternatively, we can work in Fourier terms, where

$$\delta\mathbf{v}_\mathbf{k} = -\frac{iHf(\Omega)a}{k}\delta_k\hat{\mathbf{k}}. \tag{307}$$

In either case, the linear peculiar-velocity field satisfies the continuity relation

$$\boldsymbol{\nabla}\cdot\mathbf{v} = -H\,f(\Omega)\,\delta \tag{308}$$

(where the divergence is in terms of proper coordinates). Since the fractional density perturbation (denoted by δ) is unobservable directly, one makes a connection with the galaxy distribution via the linear bias parameter

$$\delta_{\rm light} = b\,\delta_{\rm mass}. \tag{309}$$

The combination

$$\beta \equiv \Omega^{0.6}/b \tag{310}$$

can therefore be measured in principle, given the observed velocity field plus a large deep redshift survey from which the density perturbation field can be estimated.

Cosmological dipoles The well-determined absolute motion of the Earth with respect to the microwave background provides one of the possible general methods of estimating the cosmological density parameter Ω. Given a galaxy redshift survey, an estimate can be made of the gravitational acceleration of the local group produced by large-scale galaxy clustering. In perturbation theory, the relation between $\mathbf{v}$ and $\mathbf{g}$ can then be used to derive the peculiar velocity at a point in terms of the surrounding density field:

$$\mathbf{v} = \frac{H_0}{4\pi}\,\Omega^{0.6}\int\frac{\delta}{r^2}\,\hat{\mathbf{r}}\,d^3r \tag{311}$$

(Peebles 1980).

Weighing the universe The power of velocity fields is that they sample scales large enough that density perturbations are fully in the linear regime. In combination with large redshift surveys to define the spatial distribution of light, this has allowed not only a test of the assumption that large-scale clustering reflects gravitational instability but also a much more powerful extension of velocity-based methods for estimating the global density. We showed above how a knowledge of the density field surrounding the local group could be used to estimate the motion of the local group and hence predict the CMB dipole. Given a deep enough redshift survey, it is possible to use the same method to predict the peculiar velocity for any point in our local region. If we know the galaxy density perturbation field δ_g, then the peculiar velocity of a point is given by

$$\boxed{\mathbf{v} = \beta\frac{H_0}{4\pi}\int\frac{\delta_g}{r^2}\,\hat{\mathbf{r}}\,d^3r,} \tag{312}$$

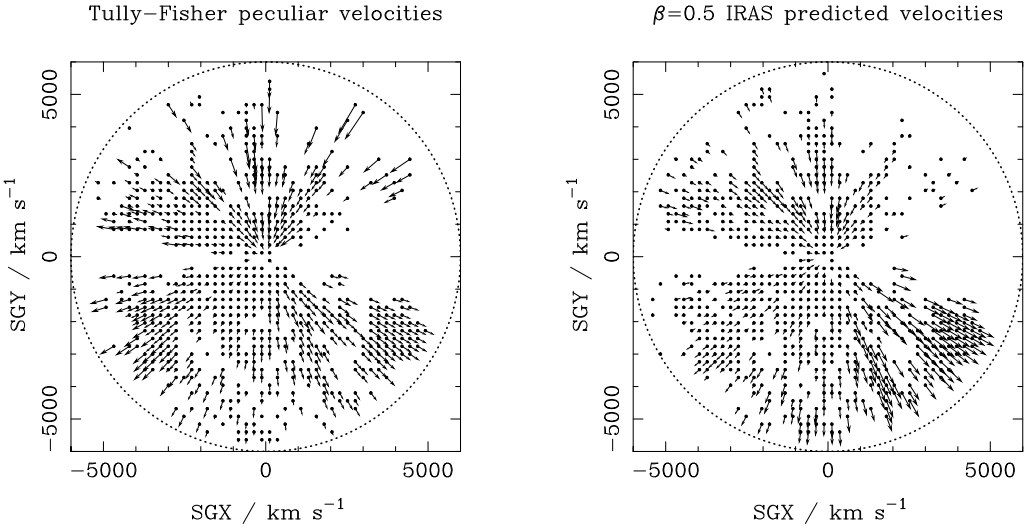

Fig. 16: A comparison of peculiar velocities inferred from the infrared Tully–Fisher method with predictions from the IRAS galaxy density field, for galaxies within 30° of the supergalactic plane and within $6000\,\mathrm{km\,s^{-1}}$, assuming $\beta\Omega^{0.6}/b = 0.5$. The data of Davis, Nusser & Willick (1996) have been averaged onto a grid for clarity in regions of high galaxy density.

where the **r** coordinate system is centred on the point of interest. The only subtlety is that δ_g is not observed directly in real space, but is deduced in redshift space. In practice, this can be corrected in an iterative way: δ_g is used to predict **v**, the galaxy redshifts are corrected for the peculiar velocities, and the exercise is repeated until a stable real-space estimate of δ_g is obtained. Both this correction and the final prediction for **v** depend on β, and so it should be possible to estimate β from a comparison between the predicted velocity field and the peculiar velocities derived using distance indicators such as the Tully-Fisher method. Figure 16 shows a recent result from one of these studies (Davis, Nusser & Willick 1997). Since most galaxies are concentrated towards the supergalactic plane defined by the local supercluster, it makes sense to plot the velocity vectors projected onto this plane. Since thousands of galaxies are involved, the velocity field is shown averaged onto a grid. The right-hand panel shows the field predicted from the gravity field due to the local density distribution.

This figure shows that, within $5000\,\mathrm{km\,s^{-1}}$, the velocity field appears to be dominated by a few distinct regions in which the flow is nearly coherent. Of these features, the one that has received the most attention is the outward flow seen near SGX $= -4000\,\mathrm{km\,s^{-1}}$, SGY $= -1000\,\mathrm{km\,s^{-1}}$. This suggests the existence of a single large mass concentration at somewhat larger radii, which has been dubbed the **great attractor** (Dressler *et al.* 1987). Popular discussions of this object have sometimes given the impression of some mysterious concentration of mass that is detected only through its gravitational attraction. However, it should be clear from figure 16 that this is implausible; the overwhelming impression is that the observed and predicted velocity fields follow each other with reasonable fidelity, strongly suggesting that it should be possible to see the great attractor. This is not a totally straightforward process, since the long-range nature of gravity leaves some ambiguity over the distance at which the mass responsible for the peculiar velocities may lie. However, it is clear that the region of sky towards the great attractor contains many particularly rich superclusters, so there is no shortage of candidates (see *e.g.* Hudson 1993). The general agreement between the observed flows and the predictions of gravitational instability is enormously encouraging; furthermore, the amplitude of the prediction scales with β, and $\beta = 0.5$ seems to give a reasonable overall match (Davis,

Nusser & Willick 1996; Willick *et al.* 1997; see figure 16). This is very satisfying, as it agrees with the determinations from clustering anisotropy discussed earlier.

Could we make the comparison the other way around, and predict the density from the velocities? At first sight, this seems impossible, since observations only reveal *radial* components of velocity. However, in linear theory, vorticity perturbations become negligible relative to the growing mode for times sufficiently long after the perturbations are created. It is therefore very tempting to make the assumption that the linear velocity field should be completely irrotational at the present epoch. Furthermore, **Kelvin's circulation theorem** (see section 8 of Landau & Lifshitz 1959) guarantees that the flow will remain irrotational even in the presence of non-linearities, provided these are not so large as to cause dissipative processes. Dissipation does of course operate on the smallest scales (galaxies rotate, after all), but this should not affect the large-scale motions. We are therefore driven to write

$$\mathbf{v} = -\boldsymbol{\nabla}\psi. \tag{313}$$

The problem is now solved in principle: the **velocity potential** ψ can be estimated by integrating the peculiar velocities along radial lines of sight. The unobservable transverse components can then be recovered by differentiation of the potential. In practice, this is a nontrivial problem, given that we are dealing with a limited number of galaxies, each of which has a rather noisy velocity estimate, of 20% precision at best. Nevertheless, by averaging over large numbers of galaxies to produce a smoothed representation of the radial velocity field on a grid, it is possible to use this method. The practical application goes by the name of **POTENT** (Bertschinger *et al.* 1990).

8 COSMIC BACKGROUND FLUCTUATIONS

8.1 Mechanisms for primary fluctuations

At the last-scattering redshift ($z \simeq 1000$), gravitational instability theory says that fractional density perturbations $\delta \gtrsim 10^{-3}$ must have existed in order for galaxies and clusters to have formed by the present. A long-standing challenge in cosmology has been to detect the corresponding fluctuations in brightness temperature of the cosmic microwave background (CMB) radiation, and it took over 25 years of ever more stringent upper limits before the first detections were obtained, in 1992. The study of CMB fluctuations has subsequently blossomed into a critical tool for pinning down cosmological models.

This can be a difficult subject; the treatment given here is intended to be the simplest possible. For technical details see *e.g.* Bond (1997), Efstathiou (1990), Hu & Sugiyama (1995), Seljak & Zaldarriaga (1996); for a more general overview, see White, Scott & Silk (1994) or Partridge (1995). The exact calculation of CMB anisotropies is complicated because of the increasing photon mean free path at recombination: a fluid treatment is no longer fully adequate. For full accuracy, the Boltzmann equation must be solved to follow the evolution of the photon distribution function. A convenient means for achieving this is provided by the public domain **CMBFAST** code (Seljak & Zaldarriaga 1996). Fortunately, these exact results can usually be understood via a more intuitive treatment, which is quantitatively correct on large and intermediate scales. This is effectively what would be called local thermodynamic equilibrium in stellar structure: imagine that the photons we see each originated in a region of space in which the radiation field was a Planck function of a given characteristic temperature. The observed brightness temperature field can then be thought of as arising from a superposition of these fluctuations in thermodynamic temperature.

We distinguish **primary anisotropies** (those that arise due to effects at the time of recombination) from **secondary anisotropies**, which are generated by scattering along the

line of sight. There are three basic primary effects, illustrated in figure 17, which are important on respectively large, intermediate and small angular scales:

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

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

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

To make quantitative progress, the next step is to see how to predict the size of these effects in terms of the spectrum of mass fluctuations.

The temperature power spectrum The statistical treatment of CMB fluctuations is very similar to that of spatial density fluctuations. We have a 2D field of random fluctuations in brightness temperature, and this can be analysed by the same tools that are used in the case of 2D galaxy clustering.

Suppose that the fractional temperature perturbations on a patch of sky of side L are Fourier expanded:

$$\begin{aligned} \frac{\delta T}{T}(\mathbf{X}) &= \frac{L^2}{(2\pi)^2} \int T_K \, \exp(-i\mathbf{K}\cdot\mathbf{X}) \, d^2K \\ T_K(\mathbf{K}) &= \frac{1}{L^2} \int \frac{\delta T}{T}(\mathbf{X}) \, \exp(i\mathbf{K}\cdot\mathbf{X}) \, d^2X, \end{aligned} \tag{314}$$

where $\mathbf{X}$ is a 2D position vector on the sky, and $\mathbf{K}$ is a 2D wavevector. This is only a valid procedure if the patch of sky under consideration is small enough to be considered flat; we give the full machinery below. We will normally take the units of length to be angle on the sky, although they could also in principle be $h^{-1}\,$Mpc at a given redshift. The relation between angle and comoving distance on the last-scattering sphere requires the comoving angular-diameter distance to the last-scattering sphere; because of its high redshift, this is effectively identical to the horizon size at the present epoch, $R_{\rm H}$:

$$\begin{aligned} R_{\rm H} &= \frac{2c}{\Omega_m H_0} \quad \text{(open)} \\ R_{\rm H} &\simeq \frac{2c}{\Omega_m^{0.4} H_0} \quad \text{(flat);} \end{aligned} \tag{315}$$

the latter approximation for models with $\Omega_m + \Omega_v = 1$ is due to Vittorio & Silk (1991).

As with the density field, it is convenient to define a dimensionless power spectrum of fractional temperature fluctuations,

$$\boxed{\mathcal{T}^2 \equiv \frac{L^2}{(2\pi)^2} \, 2\pi K^2 \, |T_K|^2,} \tag{316}$$

so that $\mathcal{T}^2$ is the fractional variance in temperature from modes in unit range of $\ln K$. The corresponding dimensionless spatial statistic is the two-point correlation function

$$\boxed{C(\theta) = \left\langle \frac{\delta T}{T}(\psi) \frac{\delta T}{T}(\psi+\theta) \right\rangle,} \tag{317}$$

which is the Fourier transform of the power spectrum, as usual:

$$C(\theta) = \int \mathcal{T}^2(K)\, J_0(K\theta)\, \frac{dK}{K}. \tag{318}$$

Here, the Bessel function comes from the angular part of the Fourier transform:

$$\int \exp(ix\cos\phi)\, d\phi = 2\pi J_0(x). \tag{319}$$

Now, in order to predict the observed anisotropy of the microwave background, the problem we must solve is to integrate the temperature perturbation field through the **last-scattering shell**. In order to do this, we assume that the sky is flat; we also neglect curvature of the 3-space, although this is only strictly valid for flat models with $k = 0$. Both these restrictions mean that the results are not valid for very large angles. Now, introducing the Fourier expansion of the 3D temperature perturbation field (with coefficients $T_k^{\rm 3D}$) we can construct the observed 2D temperature perturbation field by integrating over k space and optical depth:

$$\frac{\delta T}{T} = \frac{V}{(2\pi)^3} \iint T_k^{\rm 3D}\; e^{-i\mathbf{k}\cdot\mathbf{r}}\; d^3k\; e^{-\tau}\, d\tau. \tag{320}$$

A further simplification is possible if we approximate $e^{-\tau}d\tau$ by a Gaussian in comoving radius:

$$\exp(-\tau)\, d\tau \propto \exp[-(r - r_{\rm LS})^2/2\sigma_r^2]\; dr. \tag{321}$$

This says that we observe radiation from a last-scattering shell centred at comoving distance $r_{\rm LS}$ (which is very nearly identical to $r_{\rm H}$, since the redshift is so high), with a thickness σ_r. The section on recombination showed that the appropriate value of σ_r is approximately

$$\sigma_r = 7\,(\Omega h^2)^{-1/2}\ {\rm Mpc}. \tag{322}$$

An intuitively useful way of thinking about the integral for the observed temperature perturbation is as a two-stage process: produce a temperature field that is convolved in the radial direction, and then say that we observe a single shell that slices through this convolved field at the radius of last scattering. If the observed CMB is a slice in the (x, y) plane, the effect of the last-scattering convolution in the z direction is $T_k^{\rm 3D} \to T_k^{\rm 3D} \exp[-k_z^2\sigma_r^2/2]$ (z will briefly denote the Cartesian coordinate in the redshift direction, not redshift itself). As a result of this radial convolution and the angular dependence of the Doppler scattering term, the temperature spatial power spectrum is anisotropic. Nevertheless, we can still write down 2D and 3D Fourier-transform expressions for the correlation function in the plane $z = 0$ (taking the origin to be in the centre of the last-scattering shell):

$$\begin{aligned} C_{\rm 3D} &= \int \frac{\mathcal{T}_{\rm 3D}^2}{4\pi k^3}\, e^{-i\mathbf{k}\cdot\mathbf{x}}\, dk_x\, dk_y\, dk_z \\ C_{\rm 2D} &= \int \frac{\mathcal{T}_{\rm 2D}^2}{2\pi K^2}\, e^{-i\mathbf{K}\cdot\mathbf{x}}\, dK_x\, dK_y. \end{aligned} \tag{323}$$

Note the distinction between k and K – wavenumbers in 3D and 2D respectively. The definition of $\mathcal{T}_{3D}^2$ as the dimensionless power spectrum of spatial variations in temperature is analogous to the 3D spatial power spectrum:

$$\mathcal{T}_{\rm 3D}^2 = \frac{V}{(2\pi)^3}\, 4\pi k^3\, |T_k^{\rm 3D}|^2 \tag{324}$$

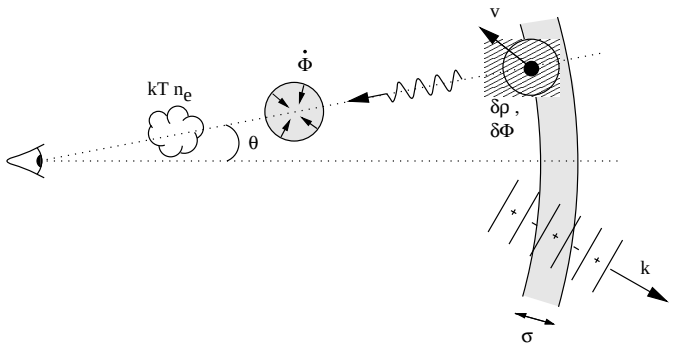

Fig. 17: Illustrating the physical mechanisms that cause CMB anisotropies. The shaded arc on the right represents the last-scattering shell; an inhomogeneity on this shell affects the CMB through its potential, adiabatic and Doppler perturbations. Further perturbations are added along the line of sight 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.

(the 2D equivalent was written above just as $\mathcal{T}^2$, but sometimes it will be convenient for clarity to add an explicit subscript 2D). Equating the two expressions for $C(\mathbf{x})$ gives the usual expression relating 2D and 3D power spectra, which we shall write in the slightly different form

$$\boxed{\mathcal{T}^2_{\rm 2D}(K) = K^2 \int_0^\infty \mathcal{T}^2_{\rm 3D}\left(\sqrt{K^2+w^2}\right)\, e^{-w^2\sigma_r^2}\, \frac{dw}{(w^2+K^2)^{3/2}},} \tag{325}$$

This simple expression gives the 2D spectrum as a projection, to which all modes with wavelength shorter than the projected wavelength of interest contribute; short-wavelength modes that run nearly towards the observer have a much longer apparent wavelength on the sky; see figure 17. The integral will generally be dominated by the contribution around $w = 0$, unless $\mathcal{T}^2_{\rm 3D}$ is a very rapidly increasing function, in which case what matters will be the small-scale cutoff governed by the width of the last-scattering shell.

The 2D power spectrum is thus a smeared version of the 3D one: any feature that appears at a particular wavenumber in 3D will cause a corresponding feature at the same wavenumber in 2D. A particularly simple converse to this rule arises when there are *no* features: the 3D power spectrum is scale-invariant ($\mathcal{T}^2_{\rm 3D}$ = constant). In this case, for scales large enough that we can neglect the radial smearing from the last-scattering shell,

$$\mathcal{T}^2_{\rm 2D} = \mathcal{T}^2_{\rm 3D} \tag{326}$$

so that the pattern on the CMB sky is scale invariant also. To apply the above machinery for a general spectrum, we now need quantitative expressions for the spatial temperature anisotropies.

Sachs–wolfe effect This is the dominant large-scale effect, and arises from potential perturbations at last scattering. These have two effects: (i) they redshift the photons we see, so that an overdensity *cools* the background as the photons climb out, $\delta T/T = \delta\Phi/c^2$; (ii) they cause 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 $\delta t/t = \delta\Phi/c^2$; since

the time dependence of the scale factor is $a \propto t^{2/3}$ and $T \propto 1/a$, this produces the counterterm $\delta T/T = -(2/3)\delta\Phi/c^2$. The net effect is thus one-third of the gravitational redshift:

$$\boxed{\frac{\delta T}{T} = \frac{\delta\Phi}{3c^2}.} \tag{327}$$

This effect was originally derived by Sachs & Wolfe (1967) and bears their nameSW effect (SW effect). It is common to see the first argument alone, with the factor 1/3 attributed to some additional complicated effect of general relativity. However, in weak fields, general relativistic effects should already be incorporated within the concept of gravitational time dilation; the above argument shows that this is indeed all that is required to explain the full result.

To relate to density perturbations, use Poisson's equation $\nabla^2\delta\Phi_k = 4\pi G\rho\delta_k$. The effect of ∇^2 is to pull down a factor of $-k^2/a^2$ (a^2 because k is a comoving wavenumber). Eliminating ρ in terms of Ω and $z_{\rm LS}$ gives

$$T_k = -\frac{\Omega(1+z_{\rm LS})}{2}\left(\frac{H_0}{c}\right)^2 \frac{\delta_k(z_{\rm LS})}{k^2}. \tag{328}$$

Doppler source term The effect here is just the Doppler effect from the scattering of photons by moving plasma:

$$\boxed{\frac{\delta T}{T} = \frac{\delta\mathbf{v}\cdot\hat{\mathbf{r}}}{c}.} \tag{329}$$

Using the standard expression for the linear peculiar velocity, the corresponding k-space result is

$$T_k = -i\sqrt{\Omega(1+z_{\rm LS})}\left(\frac{H_0}{c}\right)\frac{\delta_k(z_{\rm LS})}{k}\,\hat{\mathbf{k}}\cdot\hat{\mathbf{r}}. \tag{330}$$

Adiabatic source term This is the simplest of the three effects mentioned earlier:

$$\boxed{T_k = \frac{\delta_k(z_{\rm LS})}{3},} \tag{331}$$

because $\delta n_\gamma/n_\gamma = \delta\rho/\rho$ and $n_\gamma \propto T^3$. However, this simplicity conceals a paradox. Last scattering occurs only when the universe recombines, which occurs at roughly a fixed temperature: $kT \sim \chi$, the ionization potential of hydrogen. Surely, then, we should just be looking back to a surface of constant temperature? Hot and cold spots should normalize themselves away, so that the last-scattering sphere appears uniform. The solution is that a denser spot recombines *later*: it is therefore less redshifted and appears hotter. In algebraic terms, the observed temperature perturbation is

$$\left(\frac{\delta T}{T}\right)_{\rm obs} = -\frac{\delta z}{1+z} = \frac{\delta\rho}{\rho}, \tag{332}$$

where the last expression assumes linear growth, $\delta \propto (1+z)^{-1}$. Thus, even though a more correct picture for the temperature anisotropies seen on the sky is of a crinkled surface at

constant temperature, thinking of hot and cold spots gives the right answer. Any observable cross-talk between density perturbations and delayed recombination is confined to effects of order higher than linear.

We now draw the above results together to form the spatial power spectrum of CMB fluctuations in terms of the power spectrum of mass fluctuations at last scattering:

$$\boxed{\mathcal{T}_{3\mathrm{D}}^2 = \left[(f_{\mathrm{A}} + f_{\mathrm{SW}})^2(k) + f_{\mathrm{V}}^2(k)\mu^2\right]\Delta_k^2(z_{\mathrm{LS}}).} \tag{333}$$

There is no cross term between the adiabatic and Sachs–Wolfe terms proportional to δ and the Doppler term proportional to $i\delta$: $|a\delta + ib\delta|^2 = (a\delta + ib\delta)(a\delta^* - ib\delta^*)$. The dimensionless factors can be written most simply as

$$\begin{aligned} f_{\mathrm{SW}} &= -\frac{2}{(kD_{\mathrm{LS}})^2} \\ f_{\mathrm{V}} &= \frac{2}{kD_{\mathrm{LS}}} \\ f_{\mathrm{A}} &= 1/3, \end{aligned} \tag{334}$$

where

$$D_{\mathrm{LS}} = \frac{2c}{\Omega_m^{1/2} H_0}\,(1+z_{\mathrm{LS}})^{-1/2} = 184(\Omega h^2)^{-1/2}\ \mathrm{Mpc} \tag{335}$$

is the comoving horizon size at last scattering (a result that is independent of whether there is a cosmological constant).

We can see immediately from these expressions the relative importance of the various effects on different scales. The Sachs–Wolfe effect dominates for wavelengths $\gtrsim 1h^{-1}$ Gpc; Doppler effects then take over but are almost immediately dominated by adiabatic effects on the smallest scales.

Small-scale fluctuations The above expressions apply to perturbations for which only gravity has been important up till last scattering, *i.e.* those larger than the horizon at z_{eq}. For smaller wavelengths, a variety of additional physical processes act on the radiation perturbations, generally reducing the predicted anisotropies. An accurate treatment of these effects is not really possible without a more complicated analysis, as is easily seen by considering the thickness of the last-scattering shell, $\sigma_r = 7(\Omega h^2)^{-1/2}$ Mpc. This clearly has to be of the same order of magnitude as the photon mean free path at this time; on any smaller scales, a fluid approximation for the radiation is inadequate and a proper solution of the Boltzmann equation is needed. Nevertheless, some qualitative insight into the small-scale processes is possible. The radiation fluctuations will be damped relative to the baryon fluid by photon diffusion, characterised by the Silk-damping scale, $\lambda_{\mathrm{S}} = 2.7(\Omega\Omega_{\mathrm{B}}h^6)^{-1/4}$ Mpc. Below the horizon scale at z_{eq}, $16(\Omega h^2)^{-1}$ Mpc, there is also the possibility that dark-matter perturbations can grow while the baryon fluid is still held back by radiation pressure, which results in adiabatic radiation fluctuations that are less than would be predicted from the dark-matter spectrum alone. In principle, this suggests a suppression factor of $(1+z_{\mathrm{eq}})/(1+z_{\mathrm{LS}})$, or roughly a factor 10. In detail, the effect is an oscillating function of scale, since we have seen that baryonic perturbations oscillate as sound waves when they come inside the horizon:

$$\delta_b \propto (3c_{\mathrm{S}})^{1/4} \exp\left(\pm i \int kc_{\mathrm{S}}\, d\tau\right); \tag{336}$$

here, τ stands for conformal time. There is thus an oscillating signal in the CMB, depending on the exact phase of these waves at the time of last scattering. These oscillations in the fluid of baryons plus radiation cause a set of **acoustic peaks** in the small-scale power spectrum of the CMB fluctuations (see below).

It is clear that small-scale CMB anisotropies are a complex area, because of the near-coincidence between $z_{\rm eq}$ and $z_{\rm LS}$, and between σ_r, $r_{\rm H}(z_{\rm eq})$ and $\lambda_{\rm S}$. To some extent, these complications can be ignored, because the finite thickness of the last-scattering shell smears out small-scale perturbations in any case. However, the damping is exponential in $(k\mu)^2$ and so modes with low μ receive little damping; averaging over all directions gives a reduction in power that goes only $\propto k^{-1}$. In the absence of the other effects listed above, small-scale adiabatic fluctuations would still dominate the anisotropy pattern.

Large-scale fluctuations The flat-space formalism becomes inadequate for very large angles; the proper basis functions to use are the spherical harmonics:

$$\frac{\delta T}{T}(\hat{\mathbf{q}}) = \sum a_\ell^m Y_{\ell m}(\hat{\mathbf{q}}), \tag{337}$$

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

$$a_\ell^m = \int \frac{\delta T}{T}\, Y^*_{\ell m}\, d^2q. \tag{338}$$

The analogues of the Fourier relations for the correlation function and power spectrum are

$$\boxed{\begin{aligned} C(\theta) &= \frac{1}{4\pi}\sum_\ell \sum_{m=-\ell}^{m=+\ell} |a_\ell^m|^2\, P_\ell(\cos\theta) \\ |a_\ell^m|^2 &= 2\pi \int_{-1}^{1} C(\theta)\, P_\ell(\cos\theta)\, d\cos\theta. \end{aligned}} \tag{339}$$

These are exact relations, governing the actual correlation structure of the observed sky. However, the sky we see is only one of infinitely many possible realizations of the statistical process that yields the temperature perturbations; as with the density field, we are more interested in the **ensemble average power**. A common notation is to define C_ℓ as the expectation value of $|a_\ell^m|^2$:

$$C(\theta) = \frac{1}{4\pi}\sum_\ell (2\ell+1)\, C_\ell\, P_\ell(\cos\theta), \qquad C_\ell \equiv \left\langle |a_\ell^m|^2 \right\rangle, \tag{340}$$

where now $C(\theta)$ is the ensemble-averaged correlation. For small θ and large ℓ, the exact form reduces to a Fourier expansion:

$$C(\theta) = \int_0^\infty \mathcal{T}^2(K)\, J_0(K\theta)\, \frac{dK}{K}, \qquad \mathcal{T}^2(K=\ell+\tfrac{1}{2}) = \frac{(\ell+\frac{1}{2})(2\ell+1)}{4\pi}\, C_\ell. \tag{341}$$

The effect of filtering the microwave sky with the beam of a telescope may be expressed as a multiplication of the C_ℓ, as with convolution in Fourier space:

$$C_{\rm S}(\theta) = \frac{1}{4\pi}\sum_\ell (2\ell+1)\, W_\ell^2\, C_\ell\, P_\ell(\cos\theta). \tag{342}$$

When the telescope beam is narrow in angular terms, the Fourier limit can be used to deduce the appropriate ℓ-dependent filter function. For example, for a Gaussian beam of **FWHM** (full-width to half maximum) 2.35σ, the filter function is $W_\ell = \exp(-\ell^2\sigma^2/2)$.

For the large-scale temperature anisotropy, we have already seen that what matters is the Sachs–Wolfe effect, for which we have derived the spatial anisotropy power spectrum. The spherical harmonic coefficients for a spherical slice through such a field can be deduced using the results for large-angle galaxy clustering, in the limit of a selection function that goes to a delta function in radius:

$$\boxed{C_\ell^{\rm SW} = 16\pi \int (kD_{\rm LS})^{-4} \Delta_k^2(z_{\rm LS})\, j_\ell^2(kR_{\rm H})\, \frac{dk}{k},} \tag{343}$$

where the j_ℓ are **spherical Bessel functions** (see chapter 10 of Abramowitz & Stegun 1965). This formula, derived by Peebles (1982), strictly applies only to spatially flat models, since the Fourier expansion of the density field is invalid in an open model. Nevertheless, since the curvature radius R_0 subtends an angle of $\Omega/[2(1-\Omega)^{1/2}]$, even the lowest few multipoles are not seriously affected by this point, provided $\Omega \gtrsim 0.1$.

For simple mass spectra, the integral for the C_ℓ can be performed analytically. The case of most practical interest is a scale-invariant spectrum ($\Delta_k^2 \propto k^4$), for which the integral scales as

$$C_\ell = \frac{6}{\ell(\ell+1)}\, C_2 \tag{344}$$

(see equation 6.574.2 of Gradshteyn & Ryzhik 1980). The direct relation between the mass fluctuation spectrum and the multipole coefficients of CMB fluctuations mean that either can be used as a measure of the normalization of the spectrum. One measure that has become common is to work in terms of the amplitude of the quadrupole ($\ell = 2$), by means of the rms temperature fluctuation $Q_{\rm rms}$ produced just by the $\ell = 2$ term(s) in the spherical harmonic expansion:

$$Q_{\rm rms}^2 = \frac{1}{4\pi} \sum_{m=-2}^{m=+2} |a_2^m|^2; \tag{345}$$

Unfortunately, although the quadrupole is the largest-scale intrinsic anisotropy signal (the intrinsic dipole is unobservable, owing to the Earth's motion), it is not a good choice as a reference point, for several reasons. First, the large-scale temperature pattern is subject to corruption by emission from the Milky Way, and it is better to work at galactic latitudes $|b| \gtrsim 20°$; second, the intrinsic quadrupole is badly affected by **cosmic variance**. The C_ℓ coefficients are the average of $|a_\ell^m|^2$ over an ensemble, and so the $Q_{\rm rms}^2$ value seen by a given observer is distributed like χ^2 with five degrees of freedom. A more useful quantity is the ensemble-averaged quadrupole, since this relates directly to the power spectrum:

$$Q_{\rm rms-ps}^2 \equiv \frac{5}{4\pi}\, C_2. \tag{346}$$

However, in practice power is measured over a range of multipoles, centred at $\ell > 2$, so that the value at $\ell = 2$ is really an extrapolation that assumes a specific index for the spectrum. The most common choice is a scale-invariant spectrum, and so the clumsy quantity $Q_{\rm rms-ps|n=1}$ is used as a way of expressing the normalization of a scale-invariant spectrum. More generally, following

our discussion of small-scale anisotropies, it makes sense to define a broad-band measure of the 'power per log ℓ':

$$\boxed{\mathcal{T}^2(\ell) = \frac{\ell(\ell+1)}{2\pi} C_\ell,} \tag{347}$$

so that $\mathcal{T}^2$ is constant for a scale-invariant spectrum. This is a close relation of another measure that is sometimes encountered: $Q^2(\ell) = (5/12)\mathcal{T}^2(\ell)$ is an obvious generalization of the Q notation for the quadrupole amplitude. Finally, whichever measure is adopted, there is still the choice of 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 numbers like Q in units of μK.

8.2 Characteristics of CMB anisotropies

We are now in a position to understand the characteristic angular structure of CMB fluctuations. The change-over from scale-invariant Sachs–Wolfe fluctuations to fluctuations dominated by Doppler scattering has been shown to occur at $k \simeq D_{\rm LS}$. This is one critical angle (call it θ_1); its definition is $\theta_1 = D_{\rm LS}/R_{\rm H}$, and for a matter-only model it takes the value

$$\theta_1 = 1.8\,\Omega^{1/2} \text{ degrees}. \tag{348}$$

For flat low-density models with significant vacuum density, $R_{\rm H}$ is smaller; θ_1 and all subsequent angles would then be larger by about a factor $\Omega^{-0.6}$ (*i.e.* θ_1 is roughly independent of Ω in flat Λ-dominated models).

The second dominant scale is the scale of last-scattering smearing set by $\sigma_r = 7(\Omega h^2)^{-1/2}$ Mpc. This subtends an angle

$$\theta_2 = 4\,\Omega^{1/2} \text{ arcmin}. \tag{349}$$

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

$$\theta_3 = 9h^{-1} \text{ arcmin}, \tag{350}$$

independent of Ω. This is quite close to θ_2, so that alterations in the transfer function are an effect of secondary importance in most models.

We therefore expect that all scale-invariant models will have similar CMB power spectra: a flat Sachs–Wolfe portion down to $K \simeq 1\,\text{degree}^{-1}$, followed by a bump where Doppler and adiabatic effects come in, which turns over on arcminute scales through damping and smearing. This is illustrated well in figure 18, which shows some detailed calculations of 2D power spectra, generated with the CMBFAST package. From these plots, the key feature of the anisotropy spectrum is clearly the peak at $\ell \sim 100$. This is often referred to as the **Doppler peak**, but it is not so clear that this name is accurate. Our simplified analysis suggests that Sachs–Wolfe anisotropy should dominate for $\theta > \theta_1$, with Doppler and adiabatic terms becoming of comparable importance at θ_1, and adiabatic effects dominating at smaller scales. There are various effects that cause the simple estimate of adiabatic effects to be too large, but they clearly cannot be neglected for $\theta < \theta_1$. A better name, which is starting to gain currency, is the **acoustic peak**. In any case, it is clear that the peak is the key diagnostic feature of the CMB anisotropy spectrum: its height above the SW 'plateau' is sensitive to $\Omega_{\rm B}$ and its angular location depends on Ω and Λ. It is therefore no surprise that many experiments are

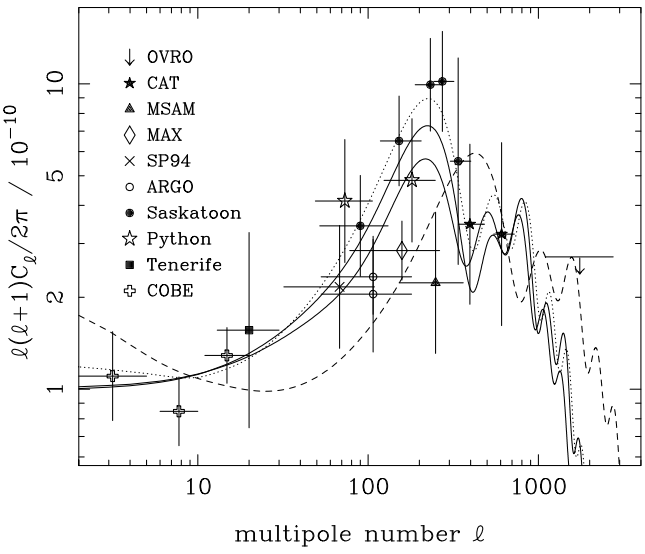

Fig. 18: Angular power spectra $\mathcal{T}^2(\ell) = \ell(\ell+1)C_\ell/2\pi$ for the CMB, plotted against angular wavenumber ℓ in radians^{-1}. The experimental data are an updated version of the compilation described in White, Scott & Silk (1994), communicated by M. White; see also Hancock *et al.* (1997). Various model predictions for adiabatic scale-invariant CDM fluctuations are shown. The two solid lines correspond to $(\Omega, \Omega_{\rm B}, h) = (1, 0.05, 0.5)$ and (1,0.1,0.5), with the higher $\Omega_{\rm B}$ increasing power by about 20% at the peak. The dotted line shows a flat Λ-dominated model with $(\Omega, \Omega_{\rm B}, h) = (0.3, 0.05, 0.65)$; the dashed line shows an open model with the same parameters. Note the very similar shapes of all the curves. The normalisation has been set to the large-scale amplitude, and so any dependence on Ω is quite modest. The main effects are that open models shift the peak to the right, and that the height of the peak increases with $\Omega_{\rm B}$ and h.

currently attempting accurate measurements of this feature. Furthermore, it is apparent that sufficiently accurate experiments will be able to detect higher 'harmonics' of the peak, in the form of smaller oscillations of amplitude perhaps 20% in power, around $\ell \simeq 500-1000$. These features arise because the matter–radiation fluid undergoes small-scale oscillations, the phase of which at last scattering depends on wavelength, since the density oscillation varies roughly as $\delta \propto \exp(ic_{\rm S}k\tau)$. Accurate measurement of these oscillations would pin down the sound speed at last scattering, and help give an independent measurement of the baryon density.

Ω dependence and normalization It is not uncommon to encounter the claim that the level of CMB fluctuations is inconsistent with a low-density universe, and in particular that a high density in collisionless dark matter is required. In fact, this statement is something of a fallacy, and it is worth examining the issue of density dependence in some detail.

Suppose we perform calculations assuming some mass power spectrum and $\Omega = 1$. If we now change Ω while keeping the shape and normalization of the power spectrum fixed, there are two effects: the power spectrum is translated both horizontally and vertically. The horizontal translation is quite simple: the main angles of importance scale as $\Omega^{1/2}$ so the CMB pattern shifts to smaller scales as Ω is reduced (unless Λ is important, in which case the shift is almost negligible; see above). To predict the vertical shift, it will suffice to consider the Sachs–Wolfe portion of the spectrum. This is

$$\mathcal{T}^2_{\rm SW} = \frac{4}{(kD_{\rm LS})^4}\,\Delta^2(z_{\rm LS}). \tag{351}$$

To relate δ at last scattering to its present value, we need the Ω-dependent growth-suppression factor for density perturbations:

$$\delta(z_{\rm LS}) \simeq \frac{\delta_0}{1+z_{\rm LS}}\,[g(\Omega)]^{-1} \tag{352}$$

i.e. there is less growth in low-density universes. Including the Ω dependence of $D_{\rm LS}$ gives

$$\mathcal{T}^2_{\rm SW} = \frac{1}{4}\left(\frac{ck}{H_0}\right)^{-4} \Delta_0^2\,\Omega^2\,[g(\Omega)]^{-2}. \tag{353}$$

The approximate power-law dependence of g is $\Omega^{0.65}$ for open models or $\Omega^{0.23}$ for flat models, so it appears that low-density universes predict lower fluctuations. This is clearly contrary to the common idea that low-density universes are ruled out owing to the freeze-out of density-perturbation growth requiring higher fluctuations at last scattering. The fractional density fluctuations are indeed higher, but the potential fluctuations that are observable depend on the *total* density fluctuation, which is lower.

Predictions from galaxy clustering However, Δ_0^2 here is the *mass* power spectrum, and the normalization we deduce from the light will generally depend on Ω. The effect this has depends on the scale at which we normalize. One common approach is to use σ_8, the rms density contrast in spheres of radius $8\,h^{-1}$ Mpc, since this is closely related to the abundance of rich clusters of galaxies. Alternatively, the amplitude of large-scale peculiar velocities measures the fractional density fluctuation on a somewhat larger scale – in both cases with a dependence of approximately $\delta_0 \propto \Omega^{-0.6}$. Note that neither of these determinations use galaxy clustering data at all: the present-day potential fluctuations are measurable directly, and yield $\delta_0\Omega^{0.6}$. What we deduce from galaxy clustering, conversely, is the biased quantity $b\delta_0$, and so galaxy clustering observations allow the parameter $\beta \equiv \Omega^{0.6}/b$ to be measured (as well as the shape of the spectrum, of course). The requirement for larger matter fluctuations in the case of lower Ω now makes the density dependence of the Sachs–Wolfe effect very weak: roughly $(\Delta^2_{\rm SW})^{1/2} \propto \Omega^{-0.25}$ (open) or $\Omega^{0.17}$ (flat).

How then can we constrain Ω from CMB observations? One possible route arises because the transfer function on small scales is rather sensitive to the total density, since modes with wavelengths below $r_{\rm H}(z_{\rm eq}) = 16(\Omega h^2)^{-1}$ Mpc have their amplitudes reduced. For a given primordial amplitude, this reduction clearly increases as Ω decreases, and is particularly severe for pure baryon universes where the small-scale power is removed by Silk damping. In the extreme case, low-density universes can have very little power on 8-Mpc wavelengths, and so normalizing on this scale gives silly answers. The number σ_8 measures the total rms density fluctuation after filtering with a sphere, and the only way this number can be large in models with a damping cutoff is to set the amplitude of 100-Mpc modes high. This is the real reason why low-density universes have often been associated with very large CMB fluctuations. In any case, now that we have clustering data on 100-Mpc scales, it makes much more sense to fix the normalization there, in which case the predicted amplitude of large-scale temperature fluctuations loses almost all dependence on Ω, as discussed above.

Following this discussion, it should be clear that it was possible to make relatively clear predictions of the likely level of CMB anisotropies, even in advance of the first detections. What was required was a measurement of the typical depth of large-scale potential wells in the universe, and many lines of argument pointed inevitably to numbers of order 10^{-5}. This was already clear from the existence of massive clusters of galaxies with velocity dispersions of up to $1000\,{\rm km\,s^{-1}}$:

$$v^2 \sim \frac{GM}{r} \quad\Rightarrow\quad \frac{\Phi}{c^2} \sim \frac{v^2}{c^2}, \tag{354}$$

so the potential well of a cluster is of order 10^{-5} deep. More exactly, the abundance of rich clusters is determined by the amplitude σ_8, which measures $[\Delta^2(k)]^{1/2}$ at an effective wavenumber of very nearly $0.17\,h\,\mathrm{Mpc}^{-1}$. If we assume that this is a large enough scale that what we are measuring is the amplitude of any scale-invariant spectrum, then the earlier expression for the temperature power spectrum gives

$$\sqrt{\mathcal{T}_{\mathrm{SW}}^2} \simeq 10^{-5.7}\,\Omega\,\sigma_8\,[g(\Omega)]^{-1}. \tag{355}$$

There were thus strong grounds to expect that large-scale fluctuations would be present at about the 10^{-5} level, and it was a significant boost to the credibility of the gravitational-instability model that such fluctuations were eventually seen.

8.3 Observations of CMB anisotropies

Prior to April 1992, no CMB fluctuations had been detected, but existing limits were close to the interesting 10^{-5} level. Continued non-detection of fluctuations at a much lower level would have forced a fundamental overhaul of our ideas about cosmological structure formation, so this period was a critical time for cosmology.

April 23 1992 saw the announcement by the COBE DMR team of the first detection of CMB fluctuations (Smoot *et al.* 1992). COBE is an acronym for NASA's **cosmic background explorer** satellite; launched in November 1989, this carried several experiments to probe the large-scale radiation field over the wavelength range 1μm to 1cm. The one concerned with the CMB at $\lambda \gtrsim 1$mm was the **differential microwave radiometer**. The DMRDMR experiment made a map of the sky with an angular resolution set by its 7° FWHM beam. This resolution means that only the low-order multipoles are accessible; DMR thus probes pure Sachs–Wolfe, and so it is easy to relate the DMR detection of sky fluctuations to a limit on the power spectrum.

In the case of the COBE measurements, the simplest and most robust datum in the initial detection reports was just the sky variance (convolved to 10° FWHM resolution, in order to suppress the noise a little), *i.e.* $C_{\mathrm{S}}(0)$ with $\sigma = 4.25^\circ$. The expected result for this for pure Sachs–Wolfe anisotropies can be predicted to almost perfect accuracy by using a small-angle approximation:

$$\begin{aligned} C_{\mathrm{S}}(0) &= \frac{\Omega^2}{g^2(\Omega)} \int 4[k(2c/H_0)]^{-4}\Delta^2(k)\,W^2(kR_{\mathrm{H}})\,\frac{dk}{k} \\ W^2(y) &= [1 - j_0^2(y) - 3j_1^2(y)]\,F(y\sigma)/(y\sigma), \end{aligned} \tag{356}$$

where a Gaussian beam of FWHM 2.35σ is assumed, and

$$F(x) \equiv \exp(-x^2)\int_0^x \exp(t^2)\,dt \tag{357}$$

is Dawson's integral. The terms involving Bessel functions correspond to the subtraction of the unobservable monopole and dipole terms. The window function is relatively sharply peaked and so the COBE variance essentially picks out the power at a given characteristic scale, which is well approximated as follows (σ in radians):

$$\begin{aligned} C_{\mathrm{S}}(0) &= 1.665\,\frac{\Omega^2}{g^2(\Omega)}\,4[k_{\mathrm{S}}(2c/H_0)]^{-4}\Delta^2(k_{\mathrm{S}}) \\ k_{\mathrm{S}}R_{\mathrm{H}} &= \frac{0.54}{\sigma} + 2.19(n-1) \end{aligned} \tag{358}$$

(Peacock & Dodds 1994). The original reported value was $C^{1/2}(0) = 1.10 \pm 0.18 \times 10^{-5}$ (Smoot *et al.* 1992). For scale-invariant spectra, this corresponds to an rms quadrupole $Q_{\mathrm{rms-ps}} =$

15.0 ± 2.5 μK. The final results from 4 years of data are consistent with the initial announcement: a scale-invariant fitted $Q_{\rm rms-ps} = 18.0 \pm 1.8$ μK. Only weak constraints on the spectrum index can be set, but the results are certainly consistent with the scale-invariant prejudice: $n = 1.2 \pm 0.3$ (Bennett *et al.* 1996).

This detection required a large number of systematic effects to be eliminated in order to be certain that the reported signal was indeed cosmological. The COBE experiment had the advantage of being a satellite in a stable environment, with no atmospheric fluctuations to contend with. By observing the same piece of sky many times from different parts of its orbit, any signals that related to low-level interference from the Earth or Sun could be eliminated. The most serious remaining problem was astronomical foregrounds – principally emission from the Milky Way. The DMR experiment observed with three different frequency channels: 31.5, 53 and 90 GHz, each of which had two independent receivers. At these wavelengths, the main contaminants are galactic synchrotron and bremsstrahlung emission (brightness temperatures varying roughly as $\nu^{-2.8}$ and ν^{-2}, respectively). A constant-temperature cosmological signal can thus be picked up by averaging the channels so as to make the galactic signal vanish (and analysing only high galactic latitudes for good measure). Even after the systematics have been dealt with, the individual DMR receivers had significant thermal noise; in the 4-year data set, the full-resolution combined map suffers a noise of around 30 μK at any given position. Since the sky map has 1844 pixels, the excess cosmological variance can be detected with huge significance, but some have criticised COBE on the grounds that it failed to detect individual structures in the CMB. Such comments are unwarranted, since all that any CMB experiment can do is to measure multipole components of the temperature perturbation field down to some limit. COBE was unable to measure individual multipole coefficients to its full resolution ($\ell \simeq 20$), but did perfectly well to $\ell \simeq 10$: the hot and cold spots on the CMB sky at 20° resolution were correctly identified even in the first-year data.

Small-scale experiments Eventually, the CMB sky will be revisited by satellite experiments with resolutions well below 1 degree. In the meantime, experiments that seek to improve on the COBE map have to work either from the ground or from balloons. In either case, they are forced to work with restricted patches of the sky, and have to contend with variable atmospheric emission. As a result, multiple-beam experiments designed to remove atmospheric emission are the norm. The simplest strategy is to switch rapidly between either two beams separated by an angle θ (chopping) or between three beam positions in a line, each separated by θ (chopping plus nodding). It is then possible to form combinations of these signals that reduce the atmospheric contribution: $T_2 - T_1$ in the two-beam case or $T_2 - (T_1 + T_3)/2$ in the three-beam case. The first case gives a signal insensitive to the mean atmosphere, whereas the second also cancels any contribution from a constant gradient in atmospheric emission. Squaring these expressions and taking expectation values shows that the rms fluctuations measured in such experiments are

$$\begin{aligned} \Delta T/T &= \sqrt{2[C(0) - C(\theta)]} \quad \text{(two-beam)} \\ &= \sqrt{2[C(0) - C(\theta)] - \tfrac{1}{2}[C(0) - C(2\theta)]} \quad \text{(three-beam)} \end{aligned} \tag{359}$$

Putting in the relation between power spectrum and $C(\theta)$, we see that three-beam experiments produce an effective 2D window function,

$$\begin{aligned} \left\langle (\delta T/T)^2 \right\rangle &= \int \mathcal{T}^2 \, W_K^2 \, dK/K \\ W_K^2 &= \left[\tfrac{3}{2} - 2J_0(K\theta) + \tfrac{1}{2} J_0(2K\theta) \right] e^{-K^2\theta_s^2}, \end{aligned} \tag{360}$$

where $2.35\theta_s$ is the beam FWHM and θ is the beam throw. The filter function peaks at some effective wavelength that is very close to 2θ. In general, all such experiments can be viewed in

this way as observing the CMB sky with some effective window function,

$$\boxed{\left\langle (\delta T/T)^2 \right\rangle = \frac{1}{4\pi} \sum_\ell (2\ell+1)\, W_\ell^2\, C_\ell} \tag{361}$$

(see Partridge 1995 for more complex observing strategies).

8.4 Conclusions and outlook

Having reviewed the physical mechanisms that cause anisotropies in the microwave background, and summarized the observational situation, it is time to ask what conclusions can be drawn. In order to narrow the field of possibilities, the discussion will concentrate on models with primordial fluctuations that are adiabatic and Gaussian. As well as being the simplest models, they will also turn out to be in reasonably good agreement with observation. Isocurvature models suffer from the high amplitude of the large-scale perturbations, and do not become any more attractive when modelled in detail (Hu, Bunn & Sugiyama 1995). Topological defects were for a long time hard to assess, since accurate predictions of their CMB properties were difficult to make. Recent progress does, however, indicate that these theories may have difficulty matching the main details of CMB anisotropies, even as they are presently known (Pen, Seljak & Turok 1997).

Inflationary predictions Matching the CMB sky with what we know of mass inhomogeneities today is important for physical cosmology, since we have seen that the anisotropies depend on the cosmological parameters Ω, Λ, $\Omega_{\rm B}$ and h. As if this were not already potentially a rich enough prize, CMB anisotropies also offer the chance to probe the very earliest phases of the big bang, and to test whether the expanding universe really did begin with an inflationary phase. Let us recall what predictions inflation makes for the fluctuation spectrum. Inflation is driven by a scalar field ϕ, with a potential $V(\phi)$. As well as the characteristic energy density of inflation, V, this can be characterised by two dimensionless parameters

$$\begin{aligned} \epsilon &\equiv \frac{m_{\rm P}^2}{16\pi} (V'/V)^2 \\ \eta &\equiv \frac{m_{\rm P}^2}{8\pi} (V''/V), \end{aligned} \tag{362}$$

where $m_{\rm P}$ is the Planck mass, $V' = dV/d\phi$, and all quantities are evaluated towards the end of inflation, when the present large-scale structure modes were comparable in size to the inflationary horizon. Prior to transfer-function effects, the primordial fluctuation spectrum is specified by a horizon-scale amplitude (extrapolated to the present) $\delta_{\rm H}$ and a slope n:

$$\Delta^2(k) = \delta_{\rm H}^2 \left(\frac{ck}{H_0}\right)^{3+n}. \tag{363}$$

The inflationary predictions for these numbers are

$$\begin{aligned} \delta_{\rm H} &\sim \frac{V^{1/2}}{m_{\rm P}^2\, \epsilon^{1/2}} \\ n &= 1 - 6\epsilon + 2\eta, \end{aligned} \tag{364}$$

which leaves us in the unsatisfactory position of having two observables and three parameters.

The critical ingredient for testing inflation by making further predictions is the possibility that, in addition to scalar modes, the CMB could also be affected by gravitational waves (following the original insight of Starobinsky 1985). We therefore distinguish explicitly between scalar and tensor contributions to the CMB fluctuations by using appropriate subscripts. The former category are those described by the Sachs–Wolfe effect, and are gravitational potential fluctuations that relate directly to mass fluctuations. The relative amplitude of tensor and scalar contributions depended on the inflationary parameter ϵ alone:

$$\frac{C_\ell^{\rm T}}{C_\ell^{\rm S}} \simeq 12.4\epsilon \simeq 6(1-n). \tag{365}$$

The second relation to the **tilt** (which is defined to be $1-n$) is less general, as it assumes a polynomial-like potential, so that η is related to ϵ. If we make this assumption, inflation can be tested by measuring the tilt and the tensor contribution. For simple models, this test should be feasible: $V = \lambda\phi^4$ implies $n \simeq 0.95$ and $C_\ell^{\rm T}/C_\ell^{\rm S} \simeq 0.3$. To be safe, we need one further observation, and this is potentially provided by the spectrum of $C_\ell^{\rm T}$. Suppose we write separate power-law index definitions for the scalar and tensor anisotropies:

$$C_\ell^{\rm S} \propto \ell^{n_{\rm S}-3}, \qquad C_\ell^{\rm T} \propto \ell^{n_{\rm T}-3}. \tag{366}$$

From the discussion of the Sachs–Wolfe effect, we know that, on large scales, the scalar index is the same as index in the matter power spectrum: $n_{\rm S} = n = 1-6\epsilon+2\eta$. By the same method, it is easily shown that $n_{\rm T} = 1-2\epsilon$ (although different definitions of $n_{\rm T}$ are in use in the literature; the convention here is that $n=1$ always corresponds to a constant $\mathcal{T}^2(\ell)$). Finally, then, we can write the **inflationary consistency equation**:

$$\boxed{\frac{C_\ell^{\rm T}}{C_\ell^{\rm S}} = 6.2(1-n_{\rm T}).} \tag{367}$$

The slope of the scalar perturbation spectrum is the only quantity that contains η, and so $n_{\rm S}$ is not involved in a consistency equation, since there is no independent measure of η with which to compare it.

From the point of view of an inflationary purist, the scalar spectrum is therefore an annoying distraction from the important business of measuring the tensor contribution to the CMB anisotropies. A certain degree of degeneracy exists here (see Bond *et al.* 1994), since the tensor contribution has no acoustic peak; $C_\ell^{\rm T}$ is roughly constant up to the horizon scale and then falls. A spectrum with a large tensor contribution therefore closely resembles a scalar-only spectrum with smaller Ω_b (and hence a relatively lower peak). One way in which this degeneracy may be lifted is through polarization of the CMB fluctuations. A nonzero polarization is inevitable because the electrons at last scattering experience an anisotropic radiation field. Thomson scattering from an anisotropic source will yield polarization, and the practical size of the fractional polarization P is of the order of the quadrupole radiation anisotropy at last scattering: $P \gtrsim 1\%$. Furthermore, the polarization signature of tensor perturbations differs from that of scalar perturbations (*e.g.* Seljak 1997; Hu & White 1997); the different contributions to the total unpolarized C_ℓ can in principle be disentangled, allowing the inflationary test to be carried out.

Implications of large-scale anisotropies Despite the above discussion, it will be convenient to compare the present-day mass distribution with the CMB data by considering only scalar perturbations at first. Possible complications due to tensor contributions can be brought in a little

later. The best and cleanest anisotropy measurements are those due to COBE, and we have seen above how the large-scale Sachs–Wolfe anisotropy can be calculated. It is possible to argue in both directions, either using the mass spectrum to predict the CMB, or *vice versa*; we shall start with the latter route. Górski *et al.* (1995), Bunn, Scott & White (1995), and White & Bunn (1995) discuss the large-scale normalization from the 2-year COBE data in the context of CDM-like models. The final 4-year COBE data favour very slightly lower results, and we scale to these in what follows. For scale-invariant spectra and $\Omega = 1$, the best normalization is

$$\text{COBE} \quad \Rightarrow \quad \Delta^2(k) = \left(\frac{k}{0.0737\, h\,\text{Mpc}^{-1}} \right)^4, \tag{368}$$

which is equivalent to $Q_{\rm rms-ps} = 18.0\,\mu$K, or $\delta_{\rm H} = 2.05 \times 10^{-5}$.

For low-density models, the earlier discussion suggests that the power spectrum should depend on Ω and the growth factor g as $P \propto g^2/\Omega^2$. Because of the time dependence of the gravitational potential (integrated Sachs–Wolfe effect) and because of spatial curvature, this expression is not exact, although it captures the main effect. From the data of White & Bunn (1995), a better approximation is

$$\Delta^2(k) \propto \frac{g^2}{\Omega^2}\, g^{0.7}. \tag{369}$$

This applies for low-Ω models both with and without vacuum energy, with a maximum error of 2% in density fluctuation provided $\Omega > 0.2$. Since the rough power-law dependence of g is $g(\Omega) \simeq \Omega^{0.65}$ and $\Omega^{0.23}$ for open and flat models respectively, we see that the implied density fluctuation amplitude scales approximately as $\Omega^{-0.12}$ and $\Omega^{-0.69}$ respectively for these two cases. The dependence is weak for open models, but vacuum energy implies much larger fluctuations for low Ω.

What if we consider a **tilted spectrum**? To see the effect of $n \neq 1$, we need to know the effective k at which COBE determines the spectrum. We saw earlier that the Sachs–Wolfe contribution to the rms sky fluctuations filtered with a Gaussian beam of FWHM 2.35σ effectively measured the power at $kR_{\rm H} = (0.54/\sigma) + 2.19(n-1)$. For 10° resolution, $\sigma = 0.0742$, and so the effective wavenumber is approximately

$$k_{\rm eff}^{\rm COBE} = 0.0012\, h\,\text{Mpc}^{-1} \times \begin{cases} \Omega^{1.0} & \text{(open)} \\ \Omega^{0.4} & \text{(flat)}, \end{cases} \tag{370}$$

ignoring the small n-dependent correction. This is a scale at least 20 times beyond the largest wavelength on which large-scale structure is reliably measured, and so the effects of tilt will be substantial. We will adopt the following measure of power on the largest reliable scales:

$$\Delta^2_{\rm opt}(k = 0.02\, h\,\text{Mpc}^{-1}) \simeq 0.005 \pm 0.0015. \tag{371}$$

Furthermore, we know from redshift-space distortions and the value of σ_8 inferred from the cluster abundance that the corresponding number for the mass must be lower if $\Omega = 1$. As before, $b \simeq 1.6$ seems the best guess for $\Omega = 1$. Since σ_8 from clusters scales as $\Omega^{-0.56}$, this suggests that $\Delta^2_{\rm mass}(k = 0.02\, h\,\text{Mpc}^{-1}) \simeq 0.002\Omega^{-1.1}$. We can compare this number with the COBE prediction, scaling the COBE-predicted amplitude with Ω as above and pivoting the power-law spectrum about $k_{\rm eff}$; clearly there will be a unique value of n that matches prediction and observation. The only problem is that, although $k = 0.02\, h\,\text{Mpc}^{-1}$ is a very large scale, the power measured there is not quite the primordial value. Two physical models that fit the shape of the large-scale clustering spectrum were discussed above: (i) $\Gamma = 0.25$ CDM; (ii) $\Gamma = 0.4$, $f_\nu = 0.3$ MDM. At $k = 0.02\, h\,\text{Mpc}^{-1}$, the transfer functions for these models are 0.69 and

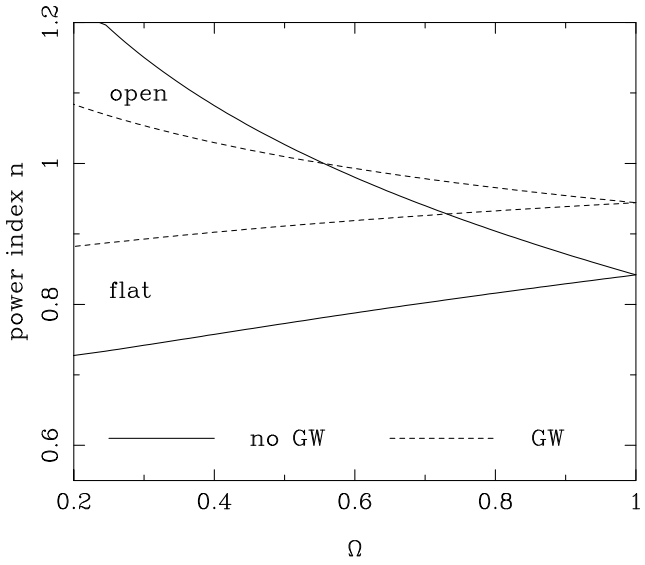

Fig. 19: A plot of the power-law index of the spectrum needed to reconcile the COBE level of CMB fluctuations with the inferred present-day mass fluctuations on the largest scales, $k = 0.02\,h\,\mathrm{Mpc}^{-1}$, allowing for the effects of the transfer function as described in the text. Both open models and flat Λ-dominated models are shown. Solid lines denote models without gravity waves; dashed lines show the effect of adding gravity waves with the usual coupling to tilt. A scale-invariant spectrum ($n = 1$) requires an open universe with $\Omega \simeq 0.6$. For lower densities, Λ-dominated flat models with gravity waves come closest to simple inflationary predictions ($n \simeq 0.9$).

0.81 respectively. We therefore adopt a mean of 0.75, and scale the inferred primordial mass fluctuations upwards by $1/(0.75)^2$.

Figure 19 shows the values of n that are required to reconcile this measure of primordial large-scale structure with COBE. Gravity waves are treated in two distinct ways: in the first case they are ignored; in the second they are added in with the above inflationary coupling to tilt, $C_\ell^{\mathrm{T}}/C_\ell^{\mathrm{S}} = 6(1-n)$. Figure 19 has several interesting features:

(1) For $\Omega = 1$, a significant tilt is needed: $n \simeq 0.84$ without tensors, rising to $n = 0.95$ if they are included.

(2) Going to low-density Λ-dominated models requires a greater degree of tilt. Even though the inferred mass fluctuations today increase for low Ω, the Ω dependence of the CMB fluctuations in Λ models is even stronger.

(3) Conversely, the weak CMB dependence on Ω for open models means that the required tilt changes rapidly with Ω. A scale-invariant spectrum with no tensor contribution is consistent with the data if $\Omega \simeq 0.6$.

Some of these results look more attractive than others. High degrees of tilt are not expected in simple models of inflation (*e.g.* $n = 0.95$ for $V = \lambda\phi^4$). Moreover, large tilt causes problems with the CMB on smaller scales. COBE normalization corresponds to $\ell \simeq 20$, whereas we have seen that there is mounting evidence for a peak at $\ell \simeq 200$. The existence of tilt thus reduces the amplitude of this peak relative to COBE by a factor 10^{1-n}, which is equal to 1.4 for $\Omega = 1$ without gravity waves. Things are just as bad if gravity waves are included: the tilt is less, but the gravity-wave component has no peak, so that the reduction of the relative height of the peak is again a factor 1.4. Looking at figure 18, we see that $n = 1$ models with a reasonable baryon

content in fact get the height of the peak about right; to allow the tilted variants, the baryon content would have to be boosted above what is allowed from nucleosynthesis.

Although the preference for $n < 1$ is thus a negative feature of flat models, it may not be fatal, since the inferred tilt depends on the accuracy of the large-scale clustering measurements. Changing the assumed large-scale power by a factor 1.5 changes n by 0.14 without gravity waves, or 0.06 with gravity waves. If flat models are to survive, the true power at $k = 0.02\,h\,\mathrm{Mpc}^{-1}$ would thus need to be larger by a factor of order 1.5, and current data do not exclude this possibility.

Implications of small-scale anisotropies Stronger diagnostics for Ω and Λ come from the intermediate-scale and small-scale CMB anisotropies. The location of the peak at $\ell \simeq 200$ is sensitive to Ω, since it measures directly the angular size of the horizon at last scattering, which scales as $\ell \propto \Omega^{-1/2}$ for open models. The cutoff at $\ell \simeq 1000$ caused by last-scattering smearing also moves to higher ℓ for low Ω; if Ω were small enough, the smearing cutoff would be carried to large ℓ, where it would be inconsistent with the upper limits to anisotropies on 10-arcminute scales. For flat models with $\Lambda \neq 0$, the Ω dependence is much weaker, which is one possible way of detecting Λ, should other arguments favour $\Omega < 1$.

This tendency for open models to violate the upper limits to arcminute-scale anisotropies is is a long-standing problem, which allowed Bond & Efstathiou (1984) to deduce the following limit on CDM universes:

$$\Omega \gtrsim 0.3h^{-4/3} \tag{372}$$

(in the absence of reionization, with a spectrum normalization that was independent of Ω, thus not allowing for the possibility of bias).

Coda The study of anisotropies in the CMB is presently one of the most exciting observational areas in cosmology, as a plethora of experiments map out the anisotropy spectrum over a wide range of scales. The fact that these detections are at the $\lesssim 10^{-5}$ level makes an amusing contrast with the early days of the subject, when fluctuations of order 10^{-3} were expected, based on the simplistic formula '$\delta T/T = (1/3)\delta\rho/\rho$ and I must make galaxies by $z = 3$'. Today, we have a much more sophisticated appreciation of the scales that are accessible to observation, plus much improved data on the inhomogeneity of the local universe.

The COBE detection and smaller-scale measurements are enormously encouraging indications of the overall correctness of the picture of structure formation via gravitational instability, but they leave open many possibilities. These will be constrained by the information which resides in the $\ell \gtrsim 100$ peaks in the anisotropy spectrum. At present, all we can say is that there are hints of a acoustic peak in the spectrum at $\ell \simeq 200$ and a sharp fall by $\ell \simeq 1000$. If confirmed, these facts would make it very difficult to sustain the idea of an open universe. As we saw earlier, the supernovae Hubble diagram strongly favours a low-density universe, if we consider only $k = 0$ models. We therefore need to consider a 'standard model' in which the majority of the energy density is in the form of vacuum energy: either a classical cosmological constant, or 'quintessence', where the scalar field continues to roll.

Definitive measurements of the CMB fluctuation spectrum will require a new generation of experiments, which are expected to yield results in the first decade of the 21st Century. As well as accurate large-scale mapping, these probes will measure the fine-scale anisotropy down to $\ell \sim 1000$. Measurements of the higher harmonics of the acoustic oscillations will be sensitive to the fine details of the physical conditions at last scattering: these will either rule out all the standard range of models – or determine the nature of dark matter and measure very accurately

the main cosmological parameters, if the present framework is correct (Bond, Efstathiou & Tegmark 1997; Zaldarriaga, Spergel & Seljak 1997).

Finally, the deepest attraction offered by CMB studies is the possibility of testing inflation. We have seen that one characteristic prediction of inflation is the existence of tensor anisotropies, and have discussed how these may in principle be detected via their contribution to the CMB anisotropy spectrum, and also through the polarization of CMB fluctuations. These will be challenging observations, but ones whose importance it would be difficult to overstate. The detection of the inflationary background of gravitational waves would give us experimental evidence on the nature of physics at almost the Planck energy. It is astonishing to realize that this might be accomplished within a mere 100 years from the first discovery of the expansion of the universe. The present is undoubtedly a golden age for cosmology, but perhaps the best is yet to come.

REFERENCES

Abramowitz M., Stegun I.A. (1965) *Handbook of Mathematical Functions*, Dover

Adler R.J. (1981) *The Geometry of Random Fields*, Wiley

Bahcall N.A., Soneira R.M. (1983) *Astrophys. J.*, **270**, 20

Baugh C.M., Efstathiou G. (1993) *Mon. Not. R. Astr. Soc.*, **265**, 145

Baugh C.M., Efstathiou G. (1994) *Mon. Not. R. Astr. Soc.*, **267**, 323

Bardeen J.M., Bond J.R., Kaiser N., Szalay A.S. (1986) *Astrophys. J.*, **304**, 15

Bennett C.L. *et al.* (1996) *Astrophys. J.*, **464**, L1

Benoist C., Maurogordato S., da Costa L.N., Cappi A., Schaeffer R. (1996) *Astrophys. J.*, **472**, 452

Bertschinger E., Dekel A., Faber S.M., Dressler A. Burstein D. (1990) *Astrophys. J.*, **364**, 370

Bond J.R., Efstathiou G. (1984) *Astrophys. J.*, **285**, L45

Bond J.R. (1997) in *Cosmology and Large-scale Structure*, proc. 60th Les Houches School, eds R. Schaeffer, J. Silk, M. Spiro & J. Zinn-Justin, Elsevier, p469

Bond J.R., Crittenden R., Davis R.L., Efstathiou G., Steinhardt P.J. (1994) *Phys. Rev. Lett.*, **72**, 13

Bond J.R., Efstathiou G., Tegmark M. (1997) *Mon. Not. R. Astr. Soc.*, **291**, L33

Bower R.G., Coles P., Frenk C.S., White S.D.M. (1993) *Astrophys. J.*, **405**, 403

Brandenberger R.H. (1990) in *Physics of the Early Universe*, proc 36th Scottish Universities Summer School in Physics, eds Peacock J.A., Heavens A.F., Davies A.T., Adam Hilger, p281

Bunn E.F., Scott D., White M. (1995) *Astrophys. J.*, **441**, 9

Burbidge E.M., Burbidge G.R., Fowler W.A., Hoyle F. (1957) *Rev. Mod. Phys.*, **29**, 547

Carroll S.M., Press W.H., Turner E.L. (1992) *Ann. Rev. Astr. Astrophys.*, **30**, 499

Cen R., Gnedin N.Y., Kofman L.A., Ostriker J.P. (1992) *Astrophys. J.*, **399**, L11

Collins P.D.B., Martin A.D., Squires E.J. (1989) *Particle Physics and Cosmology*, Wiley

Dalton G.B., Efstathiou G., Maddox S.J., Sutherland, W. (1992) *Astrophys. J.*, **390**, L1

Davis M., Peebles P.J.E. (1983) *Astrophys. J.*, **267**, 465

Davis M., Nusser A., Willick J.A. (1996) *Astrophys. J.*, **473**, 22

Dekel A. (1994) *Ann. Rev. Astr. Astrophys.*, **32**, 371

Dressler A., Lynden-Bell D., Burstein D., Davies R.L., Faber S.M., Terlevich R.J., Wegner G. (1987) *Astrophys. J.*, **313**, 42

Efstathiou G. (1990) in *Physics of the Early Universe*, proc 36th Scottish Universities Summer School in Physics, eds Peacock J.A., Heavens A.F., Davies A.T., Adam Hilger, p361

Efstathiou G. (1995) *Mon. Not. R. Astr. Soc.*, **274**, L73

Efstathiou G., Kaiser N., Saunders W., Lawrence A., Rowan-Robinson M., Ellis R.S., Frenk

C.S. (1990) *Mon. Not. R. Astr. Soc.*, **247**, 10
Efstathiou G., Bernstein G., Katz N., Tyson T., Guhathakurta P. (1991) *Astrophys. J.*, **380**, 47
Efstathiou G., Bond J.R., White S.D.M. (1992) *Mon. Not. R. Astr. Soc.*, **258**, 1P
Eisenstein D.J., Hu W. (1998) *Astrophys. J.*, **496**, 605
Eke V.R., Cole S., Frenk C.S. (1996) *Mon. Not. R. Astr. Soc.*, **282**, 263
Ellis J. (1997) in *Cosmology and Large-scale Structure*, proc. 60th Les Houches School, eds R. Schaeffer, J. Silk, M. Spiro & J. Zinn-Justin, Elsevier, p715
Felten J.E., Isaacman R. (1986) *Rev. Mod. Phys.*, **58**, 689
Fisher K.B. (1995) *Astrophys. J.*, **448**, 494
Fisher K.B., Davis M., Strauss M.A., Yahil A., Huchra J.P. (1993) *Astrophys. J.*, **402**, 42
Fixsen D.J., Cheng E.S., Gales J.M., Mather J.C., Shafer R.A., Wright E.L. (1996) *Astrophys. J.*, **473**, 576
Górski K.M., Ratra B., Sugiyama N., Banday A.J. (1995) *Astrophys. J.*, **444**, L65
Gradshteyn I.S., Ryzhik I.M. (1980) *Table of Integrals, Series and Products*, Academic Press
Grigoriev D., Shaposhnikov M., Turok N. (1992) *Phys. Lett.*, **275B**, 395
Guth A.H. (1981) *Phys. Rev. D*, **23**, 347
Hamilton A.J.S., Kumar P., Lu E., Matthews A. (1991) *Astrophys. J.*, **374**, L1
Hancock S. *et al.* (1997) *Mon. Not. R. Astr. Soc.*, **289**, 505
Heath D. (1977) *Mon. Not. R. Astr. Soc.*, **179**, 351
Hu W., Bunn E.F., Sugiyama N. (1995) *Astrophys. J.*, **447**, L59
Hu W., Sugiyama N. (1995) *Astrophys. J.*, **444**, 489
Hu W., White M. (1997) *New Astronomy*, **2**, 323
Huchra J.P., Geller M.J., de Lapparant V., Corwin H.G. (1990) *Astrophys. J. Suppl.*, **72**, 433
Hudson M.J. (1993) *Mon. Not. R. Astr. Soc.*, **265**, 43
Jain B., Mo H.J., White S.D.M. (1995) *Mon. Not. R. Astr. Soc.*, **276**, L25
Jensen L.G., Szalay A.S. (1986) *Astrophys. J.*, **305**, L5
Jing Y.P., Mo H.J., Börner G. (1998) *Astrophys. J.*, **494**, 1
Jones B.J.T., Wyse R.F.G. (1985) *Astr. Astrophys.*, **149**, 144
Kaiser N. (1984) *Astrophys. J.*, **284**, L9
Kashlinsky A. (1991) *Astrophys. J.*, **376**, L5
Kolb E.W., Turner M.S. (1990) *The Early Universe*, Addison-Wesley
Lahav O., Lilje P.B., Primack J.R., Rees M.J. (1991) *Mon. Not. R. Astr. Soc.*, **251**, 128
Landau L.D., Lifshitz E.M. (1959) *Fluid Mechanics*, Pergamon
Le Fèvre O. *et al.* (1996) *Astrophys. J.*, **461**, 534
Liddle A.R., Lyth D. (1993) *Phys. Reports*, **231**, 1
Liddle A.R., Scherrer R.J. (1998) astro-ph/9809272
Loveday J., Maddox S.J., Efstathiou G., Peterson B.A. (1995) *Astrophys. J.*, **442**, 457
Lumsden S.L., Heavens A.F., Peacock J.A. (1989) *Mon. Not. R. Astr. Soc.*, **238**, 293
McClelland J., Silk J. (1977) *Astrophys. J.*, **217**, 331
Maddox S.J., Efstathiou G., Sutherland W.J., Loveday J. (1990) *Mon. Not. R. Astr. Soc.*, **247**, 1P
Maddox S.J., Efstathiou G., Sutherland W.J. (1996) *Mon. Not. R. Astr. Soc.*, **283**, 1227
Mather J.C. *et al.* (1990) *Astrophys. J.*, **354**, L37
Mészáros P. (1974) *Astr. Astrophys.*, **37**, 225
Moore G. (1996) *Nucl. Phys.*, **B480**, 689
Mukhanov V.F., Feldman H.A., Brandenberger R.H. (1992) *Phys. Reports*, **215**, 203
Opal Consortium (1990) *Phys. Lett.*, **B240**, 497
Osterbrock D.E. (1974) *Astrophysics of Gaseous Nebulae*, Freeman
Pagel B.E.J. (1994) in *The Formation and Evolution of Galaxies*, eds C. Muñoz-Tuñón, F. Sánchez, Cambridge University Press, p151

Partridge R.B. (1995) *3K: The Cosmic Microwave Background*, Cambridge University Press
Peacock J.A. (1997) *Mon. Not. R. Astr. Soc.*, **284**, 885
Peacock J.A., Dodds S.J. (1994) *Mon. Not. R. Astr. Soc.*, **267**, 1020
Peacock J.A., Dodds S.J. (1996) *Mon. Not. R. Astr. Soc.*, **280**, L19
Peacock J.A., West M.J. (1992) *Mon. Not. R. Astr. Soc.*, **259**, 494
Peebles P.J.E. (1973) *Astrophys. J.*, **185**, 413
Peebles P.J.E. (1974) *Astrophys. J.*, **32**, 197
Peebles P.J.E. (1982) *Astrophys. J.*, **263**, L1
Peebles P.J.E. (1980) *The Large-scale Structure of the Universe*, Princeton University Press
Peebles P.J.E. (1993) *Principles of Physical Cosmology*, Princeton University Press
Pen U.-L., Seljak U., Turok N. (1997) *Phys. Rev. Lett.*, **79**, 1611
Penzias A.A., Wilson R.W. (1965) *Astrophys. J.*, **142**, 419
Perlmutter S. *et al.* (1998) astro-ph/9812133
Pogosyan D., Starobinsky A.A. (1995) *Astrophys. J.*, **447**, 465
Press W.H., Schechter P. (1974) *Astrophys. J.*, **187**, 425
Press W.H., Teukolsky S.A., Vetterling W.T., Flannery B.P. (1992) *Numerical Recipes (2nd edition)*, Cambridge University Press
Ratra B., Peebles P.J.E. (1988) *Phys. Rev. D*, **37**, 3406
Reines F., Sobel H., Pasierb E. (1980) *Phys. Rev. Lett.*, **45**, 1307
Rice S.O. (1954) in *Selected Papers on Noise and Stochastic Processes*, ed. Wax N., Dover, p133
Riess A.G. *et al.* (1998) *Astr. J.*, **116**, 1009
Rogerson J.B., York D.G. (1973) *Astrophys. J.*, **186**, L95
Sachs R.K., Wolfe A.M. (1967) *Astrophys. J.*, **147**, 73
Sakharov A.D. (1967) *JETP Lett.*, **5**, 24
Saunders W., Frenk C., Rowan-Robinson M., Efstathiou G., Lawrence A., Kaiser N., Ellis R., Crawford J., Xia X.-Y., Parry I. (1991) *Nature*, **349**, 32
Saunders W., Rowan-Robinson M., Lawrence A. (1992) *Mon. Not. R. Astr. Soc.*, **258**, 134
Seljak U., Zaldarriaga M. (1996) *Astrophys. J.*, **469**, 437
Seljak U. (1997) *Astrophys. J.*, **482**, 6
Shanks T., Fong R., Boyle B.J., Peterson B.A. (1987) *Mon. Not. R. Astr. Soc.*, **227**, 739
Shanks T., Boyle B.J. (1994) *Mon. Not. R. Astr. Soc.*, **271**, 753
Shectman S.A., Landy S.D., Oemler A., Tucker D.L., Lin H., Kirshner R.P., Schechter P.L., (1996) *Astrophys. J.*, **470**, 172
Smith M.S., Kawano L.H., Malaney R.A. (1993) *Astrophys. J. Suppl.*, **85**, 219
Smoot G.F. *et al.* (1992) *Astrophys. J.*, **396**, L1
Starobinsky A.A. (1985) *Sov. Astr. Lett.*, **11**, 133
Strauss M.A., Davis M., Yahil Y., Huchra J.P. (1992) *Astrophys. J.*, **385**, 421
Strauss M.A., Willick J.A. (1995) *Phys. Reports*, **261**, 271
Sugiyama N. (1995) *Astrophys. J. Suppl.*, **100**, 281
Sutherland W.J. (1988) *Mon. Not. R. Astr. Soc.*, **234**, 159
Valls-Gabaud D., Alimi J.-M., Blanchard A. (1989) *Nature*, **341**, 215
Vittorio N., Silk J. (1991) *Astrophys. J.*, **385**, L9
Walker T.P., Steigman G., Kang H.-S., Schramm D.M., Olive K.A. (1991) *Astrophys. J.*, **376**, 51
Weinberg S. (1989) *Rev. Mod. Phys.*, **61**, 1
White M., Scott D., Silk J. (1994) *Ann. Rev. Astr. Astrophys.*, **32**, 319
White M., Bunn E.F. (1995) *Astrophys. J.*, **450**, 477
Willick J.A., Strauss M.A., Dekel A., Kollatt T. (1997) *Astrophys. J.*, **486**, 629
Zaldarriaga M., Spergel D.N., Seljak U. (1997) *Astrophys. J.*, **488**, 1
Zlatev I., Wang L., Steinhardt P.J. (1998) astro-ph/9807002