Abstract: These lectures cover the basics of inflationary models for the early universe, concentrating particularly on the generation of density fluctuations from scalar-field dynamics. The subsequent gravitational dynamics of these fluctuations in dark matter in a Friedmann model are described, leading to a review of the current situation in confronting inflationary models with the latest data on the clustering of galaxies and other measures of large-scale structure.

1 General arguments for inflation

1.1 The problems of classical cosmology

The standard isotropic cosmology is a very successful framework for interpreting observations, but there are certain questions which, prior to the early 1980s, had to be avoided. These are encapsulated in a set of classical 'problems', as follows.

The horizon problem

Standard cosmology contains a particle horizon of comoving radius

$$r_{\rm H} = \int_0^t \frac{c\,dt}{R(t)}, \tag{1}$$

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

$$D_{\rm H} = R_0\, r_{\rm H} \simeq \frac{6000}{\sqrt{\Omega z}}\, h^{-1}\,{\rm Mpc}. \tag{2}$$

The horizon at last scattering 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:

$$[1 - 1/\Omega(z)] = f(z)\,[1 - 1/\Omega)], \tag{3}$$

where $f(z) = (1+z)^{-1}$ in the matter-dominated era, $\propto (1+z)^{-2}$ for radiation domination, so that $f(z) \simeq (1+z_{\rm eq})/(1+z)^2$ at early times. To get $\Omega \simeq 1$ now 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 supposed to have been imposed. Ignoring annihilation effects, $1+z = T_{\rm init}/2.7$ K; $1+z_{\rm eq} \simeq 10^4$, and so the required fine-tuning is

$$|\Omega(t_{\rm init}) - 1| \lesssim 10^{-22}\,[E_{\rm init}/{\rm GeV}]^2. \tag{4}$$

At the Planck epoch, which is the natural initial time, this requires a deviation of only 1 part in 10^{60}. This is satisfied if $\Omega = 1$ exactly, but a mechanism is still required to set up such an initial state. This equation is especially puzzling if $\Omega \neq 1$ today: how could the universe 'know' to start with a deviation from $\Omega = 1$ just so tuned that the curvature starts to become important only now after so many e-foldings of the expansion?

The antimatter problem

At $kT \gtrsim m_p c^2$, there exist in equilibrium roughly equal numbers of photons, protons and antiprotons. Today, $N_p/N_\gamma \sim 10^{-9}$, but $N_{\bar{p}} \simeq 0$. Conservation of baryon number would imply that $N_p/N_{\bar{p}} = 1 + O(10^{-9})$ at early times Where did this initial asymmetry come from?

The structure problem

The Universe is not precisely homogeneous. We generally presume that galaxies and clusters grew via gravitational instability from some initial perturbations. What is the origin of these?

This list can be extended to problems which come closer to astrophysics than cosmology *per se.* There is the question of the dark matter and its composition, for example. However, the above list encompasses problems which go back to the initial conditions of the Big Bang. It seems clear that conventional cosmological models need to be set up in an extremely special configuration, and this is certainly a deficiency of the theory. Critics can point out with some force that the big bang model explains nothing about the origin of the universe as we now perceive it, because all the most important features are 'predestined' by virtue of being built into the assumed initial conditions near to $t = 0$.

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. The gravitational dynamics of $R(t)$ 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 it is hardly a satisfying explanation for the motion of a cannonball which was 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 of the catalogue of classical cosmological problems. Certainly, early generations of cosmologists were convinced that some specific mechanism was required in order to explain how the universe was set in motion.

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, and this is what lies behind modern developments in **quantum cosmology**. This field is not really suitable for treatment at the present level, and it deals with the subtlest possible questions concerning the meaning of the wave function for the entire universe. The eventual aim is to understand if there could be a way in which the universe could be spontaneously created as a quantum-mechanical fluctuation, and if so whether it would have the initial properties which observations seem to require. Since this programme has to face up to the challenge of quantizing gravity, it is fair to say that there are as yet no definitive answers, despite much thought-provoking work. See chapter 11 of Kolb & Turner (1990) for an introduction.

However, 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 of 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).

1.2 An overview of inflation

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 modified 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 $t = 0$: $R \propto t^{\alpha}$; $\alpha > 1$. If such a phase 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 of 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 $p = (\Gamma - 1)\epsilon$ 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

$$\boxed{\Gamma < 2/3 \Rightarrow \rho c^2 + 3p < 0.} \tag{5}$$

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:

$$\dot{R}^2 = \frac{8\pi G \rho R^2}{3} - kc^2. \tag{6}$$

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 which 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:

$$\begin{aligned} R &\propto \sinh Ht \quad k = -1 \\ &\propto \cosh Ht \quad k = +1 \\ &\propto \exp Ht \quad k = 0, \end{aligned} \tag{7}$$

where $H = \sqrt{\Lambda c^2/3} = \sqrt{8\pi G\rho_{\rm vac}/3}$. Thus, 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

$$\Omega = 1 + O(f^{-2}). \tag{8}$$

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

$$\boxed{\ln f \gtrsim 60} \tag{9}$$

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 being effectively exactly unity, 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

$$\boxed{\text{inflation} \Rightarrow k = 0.} \tag{10}$$

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.

Reheating from inflation

The discussion so far indicates a possible route to solving the problems with initial conditions in conventional cosmology, but has a critical missing ingredient. The idea of inflation is to set the universe expanding towards an effective $k = 0$ state by using the repulsive gravitational force of vacuum energy or some other unknown state of matter which satisfies $p < -\rho c^2/3$. There remains the difficulty of returning to a normal equation of state: the universe is required to undergo a **cosmological phase transition**. Such a suggestion would have seemed highly *ad hoc.* in the 1960s when the horizon and flatness problems were first clearly articulated by Dicke. The invention of inflation by Guth (1981) had to await developments in quantum field theory which provided a plausible basis for this phase transition. This mechanism

will be described below; it has been deliberately put off so far to emphasise the general character of many of the arguments for inflation.

What will emerge is that it is possible for inflation to erase its tracks in a very neat way. If we are dealing with quantum fields at a temperature T, then an energy density $\sim T^4$ in natural units is expected in the form of vacuum energy. The vacuum-driven expansion produces a universe which is essentially devoid of normal matter and radiation; these are all redshifted away by the expansion, and so the temperature of the universe becomes $\ll T$. A phase transition to a state of zero vacuum energy, if instantaneous, would transfer the energy T^4 to normal matter and radiation as a latent heat. The universe would therefore be **reheated**: it returns to the temperature T at which inflation was initiated, but with the correct special initial conditions for the expansion. The transition in practical models is not instantaneous, however, and so the reheating temperature is lower than the temperature prior to inflation.

Quantum fluctuations

Note that de Sitter space contains an **event horizon**, in that the comoving distance that particles can travel between a time t_0 and $t = \infty$ is finite

$$r_{\rm EH} = \int_{t_0}^{\infty} \frac{c\,dt}{R(t)} \tag{11}$$

(do not confuse this with the particle horizon, where the upper limit for the integral would be t_0). With $R \propto \exp Ht$, the proper radius of the horizon is $R_0 r_{\rm EH} = c/H$. Figure 1 illustrates the situation. The exponential expansion literally makes distant regions of space move faster than light, so that points separated by $> c/H$ can never communicate with each other.

As with black holes, it therefore follows that thermal **Hawking radiation** will be created. These quantum fluctuations in de Sitter spacetime provide the seeds for what will eventually become galaxies and clusters. The second main prediction of inflation is that such fluctuations should exist in all fields, in particular that there should exist a background of gravitational waves left as a relic of inflation.

This idea of obtaining all structure in the universe (including ourselves) from quantum fluctuations is a magical idea of tremendous appeal. If it could be shown to be correct, it would rank as one of the greatest possible intellectual advances. We now have to look at some of the practical details to see how this concept might be made to function in practice, and how it may be tested.

2 Inflation field dynamics

The general concept of inflation rests on being able to achieve a negative-pressure equation of state. This can be realised in a natural way by quantum fields in the early universe. In order to understand what is going on, it is necessary to summarize some of the most important concepts and jargon from this field.

Fig. 1. The event horizon in de Sitter space. Particles outside the sphere at $r = c/H$ can never receive light signals from the origin, not can an observer at the origin receive information from outside the sphere. The exponential expansion quickly accelerates any freely-falling observers to the point where their recession from the origin is effectively superluminal.

2.1 Quantum fields and potentials

Lagrangians and fields

Consider the formulation of classical mechanics in terms of an **action principle**. We write the variational equation

$$\boxed{\delta \int L \, dt = 0,} \tag{12}$$

where the **Lagrangian** L is the difference of the kinetic and potential energies, $L = T - V$, and the integral $\int L \, dt$ is the **action**. What this says is that, for particles described by coordinates $q_i(t)$, the path travelled by each particle between given starting and finishing positions is such that the action is extremal (not necessarily a minimum, even though one often speaks of the principle of least action) with respect to small variations in the paths. If we take the 'positions' $q_i(t)$ and 'velocities' $\dot{q}_i(t)$ to be independent variables, then expansion of L in terms of small variations in the paths and integration by parts leads to **Euler's equation** for each particle

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right) = \frac{\partial L}{\partial q_i}. \tag{13}$$

A field may be regarded as a dynamical system, but with an infinite number of coordinates, q_i, which are the field values at each point in space. How do we handle this? A hint is provided by electromagnetism, where we are familiar with writing

the total energy in terms of a density which, as we are dealing with generalized mechanics, we may formally call the Hamiltonian density. This suggests that we write the Lagrangian in terms of a **Lagrangian density** $\mathcal{L}$: $L = \int \mathcal{L}\, dV$. This quantity is of such central importance in quantum field theory, that it is usually referred to (incorrectly) simply as 'the Lagrangian'. In these terms, our action principle now takes the pleasingly relativistic form

$$\delta \int \mathcal{L}\, d^4x^\mu = 0, \tag{14}$$

although note that to be correct in general relativity, the Lagrangian needs to take the form of a invariant scalar times the Jacobian $\sqrt{-g}$.

We can now apply the variational principle as before, considering $\mathcal{L}$ to be a function of a general coordinate, ϕ which is the field, and the '4-velocity' $\partial_\mu \phi$. This yields the **Euler-Lagrange equation**

$$\boxed{\frac{\partial}{\partial x^\mu}\left(\frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)}\right) = \frac{\partial \mathcal{L}}{\partial \phi}.} \tag{15}$$

A Lagrangian then immediately gives a wave equation, as the following examples illustrate. The forms of most Lagrangians are often quite similar: if we want a wave equation linear in second derivatives of the field, then the Lagrangian must contain terms quadratic in first derivatives of the field.

(1) Waves on a string. A good classical example of the Euler-Lagrange formalism is provided by waves on a string. If the density per unit length is σ, the tension is T, and we call the transverse displacement of the string y, then the (one-dimensional) Lagrangian density is

$$\mathcal{L} = \tfrac{1}{2}\sigma \dot{y}^2 - \tfrac{1}{2} T y'^2 \tag{16}$$

(at least for small displacements). The potential term comes from the work done in stretching the string. Inserting this in the Euler-Lagrange equation yields a familiar result:

$$\sigma \ddot{y} - T y'' = 0. \tag{17}$$

This is just the wave equation, and it tells us that the speed of sound on a plucked string is $\sqrt{T/\sigma}$.

(2) Complex scalar field. Here we want a Lagrangian which will yield the Klein-Gordon equation $(\Box + \mu^2)\phi = 0$. The required form is

$$\mathcal{L} = (\partial^\mu \phi)(\partial_\mu \phi)^* - \mu^2 \phi \phi^*, \tag{18}$$

where the only subtlety is that *phi* and ϕ^* are treated as independent variables (rather than the real and imaginary parts of ϕ). For a real field, the Lagrangian becomes

$$\mathcal{L} = \tfrac{1}{2}(\partial^\mu \phi \partial_\mu \phi - \mu^2 \phi^2). \tag{19}$$

Noether's theorem

The existence of global symmetries of the Lagrangian is closely connected with conservation laws in physics. In classical mechanics, conservation of energy and momentum arise by considering Euler's equation

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}_i}\right) - \frac{\partial L}{\partial x_i} = 0. \tag{20}$$

If L is independent of position, then we obtain conservation of momentum (or angular momentum, if x is an angular coordinate): $p_i \equiv \partial L/\partial \dot{x}_i =$ constant. If L has no explicit dependence on t, then

$$\frac{dL}{dt} = \sum\left(\frac{\partial L}{\partial q_i}\dot{q}_i + \frac{\partial L}{\partial \dot{q}_i}\ddot{q}\right) = \sum(\dot{p}_i\dot{q}_i + p_i\ddot{q}_i), \tag{21}$$

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

$$H \equiv \sum p_i\dot{q}_i - L = \text{constant}. \tag{22}$$

Something rather similar happens in the case of quantum (or classical) field theory: the existence of a global symmetry leads directly to a conservation law. The difference between discrete dynamics and field dynamics where the Lagrangian is a *density* is that the result is expressed as a **conserved current** rather than a simple constant of the motion. In what follows, we symbolize the field by ϕ, but this is not to imply that there is any restriction to scalar fields. If several fields are involved (*e.g.* ϕ and ϕ^* for a complex scalar field), they should be summed over.

Suppose the Lagrangian is independent of explicit dependence on spacetime (*i.e.* depends on x^μ only implicitly through the fields and their 4-derivatives). The algebra is similar to that above, and we obtain

$$\boxed{\frac{d}{dx_\nu}\left[\frac{\partial \mathcal{L}}{\partial(\partial^\nu\phi)}\,\frac{\partial\phi}{\partial x^\mu} - \mathcal{L}\,g_{\mu\nu}\right] = 0.} \tag{23}$$

This has produced a conserved tensor: the term in square brackets is to be identified with the energy-momentum tensor of the field, $T_{\mu\nu}$.

Natural units

To simplify the appearance of equations, it is universal practice in quantum field theory to adopt **natural units** where we take

$$\boxed{\hbar = c = \mu_0 = \epsilon_0 = 1.} \tag{24}$$

This convention makes the meaning of equations clearer by reducing the algebraic clutter, and is also useful in the construction of intuitive arguments for the order of magnitude of quantities in field theory.

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

$$\begin{aligned} [E] &= [m] \\ [L] &= [m]^{-1} \end{aligned} \tag{25}$$

Hence, the dimensions of energy density are

$$[\mathcal{L}] = [m]^4, \tag{26}$$

with units usually quoted in GeV^4. Thus, when we deal with quadratic derivative terms $\partial^\mu\phi\partial_\mu\phi$ and quadratic mass terms ($m^2\phi^2$, $m^2A^\mu A_\mu/2$ *etc.*), the dimensions of the fields must clearly be

$$[\phi] = [A^\mu] = [m]. \tag{27}$$

2.2 Equations of motion

Quantum fields at high temperatures

The critical fact we shall need from quantum field theory is that quantum fields can produce an energy density which 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:

$$\mathcal{L} = \tfrac{1}{2}\partial_\mu\phi\,\partial^\mu\phi - V(\phi). \tag{28}$$

In familiar examples of quantum fields, the potential would be

$$V(\phi) = m^2\,\phi^2/2, \tag{29}$$

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

$$T^{\mu\nu} = \partial^\mu\phi\partial^\nu\phi - g^{\mu\nu}\mathcal{L}. \tag{30}$$

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

$$\begin{aligned} \rho &= \tfrac{1}{2}\dot{\phi}^2 + V(\phi) + \tfrac{1}{2}(\nabla\phi)^2 \\ p &= \tfrac{1}{2}\dot{\phi}^2 - V(\phi) - \tfrac{1}{6}(\nabla\phi)^2. \end{aligned} \tag{31}$$

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:

$$V(\phi) = -\mu^2|\phi|^2 + \lambda|\phi|^4. \tag{32}$$

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 which 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 the 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 which 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 vanished or diverged); 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 get the constant of proportionality.

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

$$V_{\rm eff}(\phi, T) = V(\phi, 0) + aT^2|\phi|^2. \tag{33}$$

The effect of this on the symmetry-breaking potential is dramatic, as illustrated in Figure 2. At very high temperatures, the potential will be parabolic, with a minimum at $|\phi| = 0$, whereas at $T = 0$, the ground state is at $|\phi| = \sqrt{\mu^2/(2\lambda)}$ and the symmetry is broken. In between, there must be three critical temperatures: 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**. Finally, at $T = T_3$, the curvature around the origin changes sign and there is only one minimum in the potential. The universe can no longer be 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:

$$\Delta V = \frac{\mu^4}{2\lambda} \tag{34}$$

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 displayed an effective cosmological constant. On dimensional grounds, this must be an energy density $\sim m^4$ in natural units,

Fig. 2. The temperature-dependent effective potential illustrated at several temperatures. For $T > T_1$, only the false vacuum is available; for $T \simeq T_2$ the true vacuum is energetically favoured and the potential approaches the zero-temperature form. For $T < T_3$ the true vacuum is the only minimum.

where m is the energy at which the phase transition occurs. For GUTs, $m \simeq 10^{15}$ GeV; in laboratory units, this implies

$$\rho_{\rm vac} = \frac{10^{60} {\rm GeV}^4}{\hbar^3 c^5} \simeq 10^{80}\ {\rm kg\,m^{-3}}. \tag{35}$$

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 which violate baryon-number conservation can generate the observed matter/antimatter asymmetry.

However, while a workable inflationary cosmology will very likely deploy these 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

$$\boxed{\ddot{\phi} + 3H\dot{\phi} - \nabla^2\phi + dV/d\phi = 0.} \tag{36}$$

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 which arises from a stationary action ($\sqrt{-g} = R^3(t)$ for an FRW model, and this is the origin of the Hubble drag $3H\dot{\phi}$ term).

The solution of the equation of motion becomes tractable if we both ignore spatial inhomogeneities in ϕ and make the **slow-rolling approximation** that $|\ddot{\phi}| \ll |3H\dot{\phi}|$, $|dV/d\phi|$. 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:

$$V \gg \phi^2/T^2, \quad V \gg \phi^2/X^2, \tag{37}$$

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

$$\boxed{3H\dot{\phi} = -dV/d\phi.} \tag{38}$$

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

$$\boxed{\epsilon \equiv \frac{m_{\rm P}^2}{16\pi}(V'/V)^2 \ll 1.} \tag{39}$$

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

$$\boxed{\eta \equiv \frac{m_{\rm P}^2}{8\pi}(V''/V) \ll 1} \tag{40}$$

(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 which was historically important in motivating the invention of inflation – the **monopole problem**. Monopoles are point-like topological defects which 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 (see *e.g.* section 7.6 of Kolb & Turner 1990). 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, but with the oscillations becoming damped by the $3H\dot{\phi}$ friction term. Eventually, we will be left with a stationary field which either continues to inflate without end (if $V(\phi = 0) > 0$) or which simply has zero density. This would be rather a 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 as much 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 realise 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 which 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 to today is

$$a_{\rm GUT}^{-1} \simeq E_{\rm GUT}/E_\gamma. \tag{41}$$

The comoving horizon size at the end of inflation was therefore

$$d_{\rm H}(t_{\rm GUT}) \simeq a_{\rm GUT}^{-1}\,[c/H_{\rm GUT}] \simeq [E_{\rm P}/E_\gamma]\,E_{\rm GUT}^{-1}, \tag{42}$$

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. We need enough e-foldings to have stretched this to the present horizon size

$$\boxed{N_{\rm obs} = \ln\left[\frac{3000h^{-1}\,{\rm Mpc}}{[E_{\rm P}/E_\gamma]E_{\rm GUT}^{-1}}\right] \simeq 60.} \tag{43}$$

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

$$N \gtrsim \tfrac{1}{2} \ln \left[\frac{a_{\rm eq}}{a_{\rm GUT}^2} \right]. \tag{44}$$

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 the one for N in the case of the horizon problem: the same number of e-folds will always solve both.

Realizing that the observational regime corresponds only to the terminal phases of inflation is both depressing and stimulating. Depressing, because ϕ may well not move very much during the last phases: our observations relate only to a small piece of the potential, and we cannot hope to recover its form without substantial *a priori* knowledge. Stimulating because observations even on very large scales must relate to a period where the simple concepts of exponential inflation and scale-invariant density fluctuations were coming close to breaking down. This opens the possibility of testing inflation theories in a way that would not be possible with data relating to only the simpler early phases. These tests take the form of tilt and gravitational waves in the final perturbation spectrum, to be discussed further below.

2.3 Inflationary models

Early inflation models

These general principles contrast sharply with Guth's initial idea, where the potential was trapped at $\phi = 0$, and eventually underwent a first-order phase transition. This model suffers from the problem that it predicts far too large residual inhomogeneities after inflation is over. This is easily seen: because the transition is first-order, it proceeds by **bubble nucleation** where the vacuum tunnels between false and true vacuum. However, the extent of these bubbles will spread as a causal process, whereas outside the bubbles the exponential expansion of inflation is continuing. This means that it is very difficult for the bubbles to percolate and eliminate the false vacuum everywhere, as is needed for an end to inflation. Instead, inflation continues indefinitely, with the bubbles of true vacuum having only a small filling factor at any time. This **graceful exit problem** motivated variants in which the potential is flatter near the origin, so that the phase transition is second order and can proceed smoothly everywhere.

However, there is also a more general problem with Guth's model and its variants. If the initial conditions are at a temperature $T_{\rm GUT}$, we expect thermal fluctuations in ϕ; the potential should generally differ from its minimum by an amount $V \sim T_{\rm GUT}^4$, which is of the order of the difference between true and false vacua. How then is the special case needed to trap the potential near $\phi = 0$ to arise? We have returned to the sort of fine-tuned initial conditions from which inflation was designed to save us.

Combined with the difficulties in achieving small inhomogeneities after inflation is over, Guth's original inflation model thus turned out to have insuperable difficulties. However, for many cosmologists the main concepts of inflation have been

too attractive to give up. The price one pays for this is to decouple inflation from standard particle physics (taking the liberty of including GUTs in this category): inflation can in principle be driven by the vacuum energy of any scalar field. The ideological inflationist will then take the position that such a field (the **inflaton**) must have existed, and that it is our task to work out its properties from empirical cosmological evidence, rather than *a priori* particle-physics considerations. Barring a credible alternative way of understanding the peculiarities of the initial conditions of the big bang, there is much to be said for this point of view.

Chaotic inflation models

Most attention is currently paid to the more general models where the field finds itself some way from its potential minimum. This idea is termed **chaotic inflation** (see *e.g.* Linde 1989). The name originates because this class of models is also quite philosophically different from other inflation models; it does not require that there is a single Friedmann model containing an inflation-driving scalar field. Rather, there could be some primordial chaos, within which conditions might vary. Some parts may attain the conditions needed for inflation, in which case they will expand hugely, leaving a universe inside a single bubble – which we inhabit. In principle this bubble has an edge, but if inflation persists for sufficiently long, the distance to this nastiness is so much greater than the current particle horizon that its existence has no testable consequences.

A wide range of inflation models of this kind is possible, illustrating the freedom which arises once the parameters of the theory are constrained only by the requirement that inflation be produced. Things become even less constrained once it is realized that inflation need not correspond to de Sitter space, even though this was taken for granted in early discussions. As discussed earlier, it is only necessary that the universe enter a phase of 'superluminal' expansion in which the equation of state satisfies $p < -\rho c^2/3$.

For a pure static field, we will have the usual $p = -\rho c^2$ vacuum equation of state, and so a significant deviation from de Sitter space requires a large contribution from $\dot{\phi}$ terms (although the slow-roll conditions will often still be satisfied). Intuitively, this corresponds to a potential which must be steep in some sense that is determined by the desired time dependence of the scale factor. Three special cases are of particular interest:

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

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

$$V(\phi) \propto \exp\left[\sqrt{\frac{16\pi}{p\, m_{\rm P}^2}}\,\phi\right]. \qquad (45)$$

(iii) **Intermediate inflation.** Another simple time-dependence which suffices for inflation is $a(t) \propto \exp[(t/t_0)^f]$. In the slow-roll approximation, the required potential here is $V(\phi) \propto \phi^{-\beta}$, where $\beta = 4(f^{-1} - 1)$.

There are in addition a plethora of more specific models with various degrees of particle-physics motivation. Since at the time of writing none of these seem likely to become permanent fixtures, these will mostly not be described in detail. The above examples are more than enough to illustrate the wide range of choice available.

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

$$N = \int H\, dt = -\frac{8\pi}{m_{\rm P}^2} \int_{\phi_1}^{\phi_2} \frac{V}{V'}\, d\phi \tag{46}$$

For any potential which 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 the start. The criterion for successful inflation is thus that the initial value of the field exceeds the Planck scale:

$$\boxed{\phi_{\rm start} \gg m_{\rm P}.} \tag{47}$$

By the same argument, it is easily seen that this is also the criterion needed to make the slow-roll parameters ϵ and η be $\ll 1$. To summarise, 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 our condition becomes $m \ll m_{\rm P}$; 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 inflated by a large factor in the meantime.

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

3 Relic fluctuations from inflation

3.1 Motivation

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 of 2 higher:

$$kT_{\rm d-s} = \frac{\hbar H}{2\pi}. \tag{48}$$

The details of how these fluctuations translate into density perturbations after inflation are somewhat technical. However, we can immediately note that a natural prediction will be a spectrum of perturbations which are *scale invariant*. This means that the metric fluctuations of spacetime receive equal levels of distortion from each decade of wavelength of perturbation, and may be quantified in terms of the fluctuations in Netwonian gravitational potential, Φ ($c = 1$):

$$\boxed{\delta_{\rm H}^2 \equiv \Delta_\Phi^2 \equiv \frac{d\,\sigma^2(\Phi)}{d\,\ln k}}. \tag{49}$$

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 scale-invariant prediction 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 which exist on this scale are the same at all times. After inflation ceases, the resulting fluctuations (constant amplitude on the scale of the horizon) give us the **Zeldovich** or **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 which 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), but there is no space to do this justice here. Rather, we shall rely on simply motivating the 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. More formally, we shall show below that small potential perturbations preserve their value, provided they are on scales where pressure effects can be neglected, and that this critical scale corresponds

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

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

$$\boxed{\delta_{\rm H} = \frac{H^2}{2\pi\,\dot\phi}} \tag{50}$$

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

$$\delta t = \frac{\delta\phi}{\dot\phi}. \tag{51}$$

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

$$\delta_{\rm H} \simeq H\ \delta t = \frac{H^2}{2\pi\,\dot\phi}, \tag{52}$$

where the last step uses the crucial input of quantum field theory, which is to say that the rms $\delta\phi = H/2\pi$. This result will be derived below, but it is immediately reasonable on dimensional grounds (in natural units, the field has the dimension of a temperature).

3.2 The fluctuation spectrum

We now need to go over this vital result in rather more detail. First, consider the equation of motion obeyed by perturbations in the inflaton field. The basic equation of motion is

$$\ddot\phi + 3H\dot\phi - \nabla^2\phi + V'(\phi) = 0, \tag{53}$$

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 amplitude A: $\delta\phi = A\exp[i\mathbf{k}\cdot\mathbf{x} - ikt/a]$; the perturbed field $\delta\phi$ obeys the same equation of motion as the main field. If the slow-roll conditions are also assumed, so that V' may be treated as a constant, we get

$$[\ddot{\delta\phi}] + 3H[\dot{\delta\phi}] + (k/a)^2\delta\phi = 0, \tag{54}$$

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 which 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

$$\phi_k = \omega_k a_k + \omega_k^* a_k^\dagger, \tag{55}$$

and the field variance would be

$$\langle 0|\ |\phi_k|^2\ |0\rangle = |\omega_k|^2. \tag{56}$$

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 $k/a \ll H$).

Before finishing the calculation, note the critical assumption that the initial state is the vacuum: in the modes that will eventually be relevant for observational cosmology, we start with not even one quantum of the inflaton field present. Is this smuggling fine-tuning of the initial conditions back in through the back door? Given our ignorance of the exact conditions in the primordial chaos from which the inflationary phase is supposed to emerge, it is something of a matter of taste whether this is seen as being a problem. Certainly, if the initial state is close to equilibrium at temperature T, this is easily understood, since all initial scales are given in terms of T. In natural units, the energy density in $V(\phi)$ and radiation will be $\sim T^4$, and the proper size of the horizon will be $\sim T^{-2}$. In the initial state, the inflaton occupation number will be 1/2 for very long wavelengths, and will fall for proper wavelengths $\lesssim T^{-1}$. Now, remember that T is in units of the Planck temperature, so that $T \sim 10^{-4}$ for GUT-scale inflation. That means that perturbations of scale smaller than T times the horizon would start with $n \simeq 0$ for a thermal state. However, this is only the initial state, and we expect that occupation number will be an adiabatic invariant which is constant for a given comoving wavelength. Thus, after $\ln(1/T)$ e-foldings of inflation (*i.e.* a few), every mode that remains inside the horizon will have the required zero occupation number. Of course, the motivation for a thermal initial state is weak, but the main point can be made in terms of energy density. If inflation is to happen at all, $V(\phi)$ must dominate, and it cannot do this if the inflaton fluctuations exist down to zero wavelength because the effective radiation density would then diverge. It thus seems reasonable to treat any initial state that inflates as rapidly entering a vacuum state.

It is also worth noting in passing that these fluctuations in the scalar field can in principle affect the progress of inflation itself. They can be thought of as adding a random-walk element to the classical rolling of the scalar field down the trough defined by $V(\phi)$. In cases where ϕ is too close to the origin for inflation to persist for sufficiently long, it is possible for the quantum fluctuations to push ϕ further out – creating further inflation in a self-sustaining process. This is the concept of **stochastic inflation** (Linde 1986, 1989).

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 (expanding) flat-space result

$$\omega_k = a^{-3/2}\,[2k/a]^{-1/2}\,e^{-ikt/a}. \tag{57}$$

The powers of 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:

$$\boxed{\omega_k = a^{-3/2}\,[2k/a]^{-1/2}\,e^{-ik/aH}\,(1 + iaH/k)} \tag{58}$$

(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

$$\langle 0|\;|\phi_k|^2\;|0\rangle = \frac{H^2}{2k^3}. \tag{59}$$

The initial quantum zero-point fluctuations in the field have been transcribed to a constant classical fluctuation which 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:

$$\frac{d\,(\delta\phi)^2}{d\,\ln k} = \frac{4\pi k^3}{(2\pi)^3}\,\langle 0|\;|\phi_k|^2\;|0\rangle = \left(\frac{H}{2\pi}\right)^2. \tag{60}$$

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

$$\boxed{\begin{aligned} \delta_{\rm H}^2 \equiv \Delta_\Phi^2(k) &= \frac{H^4}{[2\pi\dot{\phi}]^2} \\ H^2 &= \frac{8\pi}{3}\,\frac{V}{m_{\rm P}^2} \\ 3H\dot{\phi} &= -V', \end{aligned}} \tag{61}$$

where we have 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 made 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 should 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. Finally, the principal result of the inflationary fluctuations in their late-time classical guise is as a perturbation to curvature, and it is not easy to see how to produce the separation in behaviour between photon and matter perturbations which is needed for isocurvature modes. Towards the end of inflation, the universe contains nothing but scalar field and whatever mechanisms that generate the matter/antimatter asymmetry have yet to operate. When they do, the result will be a universal photon/baryon ratio but with a total density modulated by the residual inflationary fluctuations – adiabatic initial conditions, in short.

Inflation thus makes a relatively firm prediction about the statistical character of the initial density perturbations, plus a somewhat less firm prediction for their power spectrum. With sufficient ingenuity, the space of predictions can be widened; isocurvature perturbations can be produced at the price of introducing additional inflation fields and carefully adjusting the coupling between them (Kofman & Linde 1987), and breaking the Gaussian character of the fluctuations is also possible in such multi-field models (Yi & Vishniac 1993) – essentially because all modes in field 2 can respond coherently to a fluctuation in field 1, in much the same way as non-Gaussian perturbations are generated by cosmic strings. However, a nearly scale-free Gaussian adiabatic spectrum is an inevitable result in the simplest models with a single inflaton; if the theory is to have any predictive power and not to appear contrived, this is the clear prediction of inflation. As we shall see in the observational sections, the true state of affairs seems to be close to this state.

Inflaton coupling

The calculation of density inhomogeneities sets an important limit on the inflation potential. From the slow-rolling equation, we know that the number of e-foldings of inflation is

$$N = \int H\, dt = \int H\, d\phi/\dot{\phi} = \int 3H^2\, d\phi/V'. \tag{62}$$

Suppose $V(\phi)$ takes the form $V = \lambda\phi^4$, so that $N = H^2/(2\lambda\phi^2)$. The density perturbations can then be expressed as

$$\delta_{\rm H} \sim \frac{H^2}{\dot{\phi}} = \frac{3H^3}{V'} \sim \lambda^{1/2} N^{3/2}. \tag{63}$$

Since $N \gtrsim 60$, the observed $\delta_{\rm H} \sim 10^{-5}$ requires

$$\boxed{\lambda \lesssim 10^{-15}.} \tag{64}$$

Alternatively, in the case of $V = m^2\phi^2$, $\delta_{\rm H} = 3H^3/(2m^2\phi)$. Since $H \sim \sqrt{V}/m_{\rm P}$, this gives $\delta_{\rm H} \sim m\phi^2/m_{\rm P}^3 \sim 10^{-5}$. Since we have already seen that $\phi \gtrsim m_{\rm P}$ is needed for inflation, this gives

$$\boxed{m \lesssim 10^{-5} m_{\rm P}.} \tag{65}$$

These constraints appear to suggest a defect in inflation, in that we should be able to us the theory to *explain* why $\delta_{\rm H} \sim 10^{-5}$, rather than using this observed fact to constrain the theory. The amplitude of $\delta_{\rm H}$ is one of the most important numbers in cosmology, and it is vital to know if there is a simple explanation for its magnitude. Such an explanation does exist for theories based on topological defects, where we would have

$$\delta_{\rm H} \sim [E_{\rm GUT}/E_{\rm P}]^2. \tag{66}$$

In fact, the situation in inflation is similar, since another way of expressing the horizon-scale amplitude is

$$\delta_{\rm H} \sim \frac{V^{1/2}}{m_{\rm P}^2\, \epsilon^{1/2}}. \tag{67}$$

We have argued that inflation will end with ϵ of order unity; if the potential were to have the characteristic value $V \sim E_{\rm GUT}^4$ then this would give the same prediction for $\delta_{\rm H}$ as in defect theories. The appearance of a tunable 'knob' in inflation theories really arises because we need to satisfy $\phi \sim m_{\rm P}$ (for enough inflation), while dealing with the characteristic value $V \sim E_{\rm GUT}^4$ (to be fair, this is likely to apply only at the start of inflation, but the potential does not change by a large factor 60 e-folds from the end of inflation unless the total number of e-folds is $\gg 60$). It is therefore reasonable to say that a much smaller horizon-scale amplitude would need $V \ll E_{\rm GUT}^4$, *i.e.* a smaller $E_{\rm GUT}$ than the conventional value.

This section has demonstrated the cul-de-sac in which inflationary models now find themselves: the field which drives inflation must be very weakly coupled – and effectively undetectable in the laboratory. Instead of Guth's original heroic vision of a theory motivated by particle physics, we have had to introduce a new entity into particle physics which exists only for cosmological purposes. In a sense, then, inflation is a failure. However, the hope of a consistent scheme eventually emerging (plus the lack of any alternative), means that inflationary models continue to be explored with great vigour.

3.3 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 $\phi = (m_{\rm P}/4\sqrt{\pi})\, h_{\mu\nu}$, so the expected rms gravity-wave amplitude is

$$h_{\rm rms} \sim H/m_{\rm P}. \tag{68}$$

The fluctuations in ϕ are transmitted 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

$$\left.\frac{\delta T}{T}\right|_{\rm GW} \sim h_{\rm rms} \lesssim \delta_{\rm H} \sim 10^{-5} \tag{69}$$

(where the second step follows because the tensor modes can make 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

$$\rho_{\rm GW} \sim m_{\rm P}^2\, h_{\rm rms}^2\, H^2(a_{\rm entry}). \tag{70}$$

After horizon entry, the waves redshift away like radiation, as a^{-4}, and generate a present-day energy spectrum per $\ln k$ which is constant for modes which entered the horizon while the universe was radiation dominated (because $a \propto t^{1/2} \Rightarrow H^2 a^4 = {\rm const}$). 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

$$\boxed{\Omega_{\rm GW} \sim \Omega_r \, [H/m_{\rm P}]^2 \sim 10^{-4} V/m_{\rm P}^4.} \tag{71}$$

Therefore, just as with density perturbations in dark matter, the gravity-wave spectrum displays a break between constant metric fluctuations on super-horizon scales to constant density fluctuations on small scales. If gravity waves make an important contribution to CMB anisotropies, $h_{\rm rms} \sim 10^{-5}$ and $\Omega_{\rm GW} \sim 10^{-14}$ is expected.

A gravity-wave background of a similar flat spectrum is also predicted from cosmic strings (see section 10.4 of Vilenkin & Shellard 1994). Here, the prediction is

$$\Omega_{\rm GW} \sim 100 \, [G\mu/c^2] \, \Omega_r, \tag{72}$$

where μ is the mass per unit length. A viable string cosmology requires $G\mu/c^2 \sim 10^{-5}$, so $\Omega_{\rm GW} \sim 10^{-7}$ is expected – much higher than the inflationary prediction.

The part of the spectrum with periods of order years would perturb the emission from pulsars through fluctuating gravitational redshifts, and the absence of this modulation sets a bound of $\Omega \lesssim 10^{-7}$, so that $V \lesssim 10^{-2} m_{\rm P}^4$. However, this is still a very long way from the interesting inflationary level of $\Omega_{\rm GW} \lesssim 10^{-14}$. Although the strains implied by this level of relic gravity-wave background are tiny, it is not completely inconceivable that space-based versions of the same interferometer technology being used to search for kHz-period gravity waves on Earth might eventually reach the required sensitivity. A direct detection of the gravity-wave background at the expected level would do much the same for the credibility of inflation as was achieved for the Big Bang itself by Penzias & Wilson in 1965.

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). Express the fractional temperature variance as the contribution of a given spherical harmonic, C_ℓ; for a scale-invariant spectrum, $\ell^2 C_\ell$ is a constant. The tensor and scalar contributions are respectively

$$\ell^2 \, C_\ell^{\rm T} \sim h_{\rm rms}^2 \sim [H^2/m_{\rm P}^2] \sim V/m_{\rm P}^4. \tag{73}$$

$$\ell^2 \, C_\ell^{\rm S} \sim \delta_{\rm H}^2 \sim H^2/\dot{\phi} \sim H^6/(V')^2 \sim \frac{V^3}{m_{\rm P}^6 \, V'^2}. \tag{74}$$

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

$$\frac{C_\ell^{\rm T}}{C_\ell^{\rm S}} \simeq 12.4 \, \epsilon, \tag{75}$$

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 one

of the dimensionless inflation parameters, ϵ. 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 which have spectra that deviate from scale invariance, this suggest a potential test of inflation. Define the **tilt** of the fluctuation spectrum as:

$$\text{tilt} = (1-n) = -\frac{d\ln\delta_{\rm H}^2}{d\,\ln k}. \tag{76}$$

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

$$a/k = H^{-1}. \tag{77}$$

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

$$\frac{d}{d\,\ln k} = a\,\frac{d}{da} = \frac{\dot\phi}{H}\,\frac{d}{d\phi} = \frac{m_{\rm P}^2}{8\pi}\,\frac{V'}{V}\,\frac{d}{d\phi}. \tag{78}$$

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

$$\delta_{\rm H}^2 = \frac{H^4}{[2\pi\dot\phi]^2} = \frac{128\pi}{3}\,\frac{V^3}{m_{\rm P}^6\,V'^2}, \tag{79}$$

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

$$\boxed{(1-n) = 6\epsilon - 2\eta} \tag{80}$$

For most models in which the potential is a smooth polynomial-like function, $|\eta| \simeq |\epsilon|$. Since ϵ has the larger coefficient and is positive by definition, a general but not unavoidable prediction of inflation is that the spectrum of scalar perturbations should be slightly tilted in the sense that n is slightly less than unity. With a similar level of confidence, one can state that there is a coupling between this tilt and the level of the gravity-wave contribution to CMB anisotropies:

$$\frac{C_\ell^{\rm T}}{C_\ell^{\rm S}} \simeq 6(1-n) \tag{81}$$

In principle, this is a distinctive prediction of inflation, but it is a test which loses power the more closely the fluctuations approach scale invariance.

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

$$(1-n) = 2/p \tag{82}$$

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

$$60 = \int H\, dt = \frac{8\pi}{m_{\rm P}^2} \int_{\phi_{\rm end}}^{\phi} \frac{V}{V'}\, d\phi. \tag{83}$$

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

$$60 = \frac{8\pi}{m_{\rm P}^2} \int_{\phi_{\rm end}}^{\phi} \frac{\phi}{\alpha}\, d\phi = \frac{4\pi}{m_{\rm P}^2\alpha}[\phi^2 - \phi_{\rm end}^2]. \tag{84}$$

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

$$(1-n) = (2+\alpha)/120. \tag{85}$$

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

4 Gravitational dynamics of fluctuations

4.1 Gravitational perturbation theory

We now summarize briefly some of the basics relating to the growth of density perturbations in the post-inflationary phase. The study of perturbations in general relativity can be a rather complicated and messy subject. Fortunately, most of the essential physics can be extracted from a Newtonian approach (not so surprising when one remembers that small perturbations mean weak gravitational fields). We start by writing down the fundamental equations governing fluid motion (non-relativistic for now):

$$\begin{aligned} \text{Euler}: &\quad \frac{D\mathbf{v}}{Dt} = -\frac{\boldsymbol{\nabla} p}{\rho} - \boldsymbol{\nabla}\Phi \\ \text{Energy}: &\quad \frac{D\rho}{Dt} = -\rho\boldsymbol{\nabla}\cdot\mathbf{v} \\ \text{Gauss}: &\quad \nabla^2\Phi = 4\pi G\rho, \end{aligned} \tag{86}$$

where $D/Dt = \partial/\partial t + \mathbf{v}\cdot\boldsymbol{\nabla}$ is the usual comoving derivative. We now **linearize** these by collecting terms of first order in perturbations about a homogeneous background: $\rho = \rho_0 + \delta\rho$ *etc.* The result looks simpler if we define the fractional density perturbation

$$\delta \equiv \frac{\delta\rho}{\rho_0}. \tag{87}$$

Also, when dealing with time derivatives of perturbed quantities, the full comoving time derivative D/Dt can 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

$$\begin{aligned}
&\frac{d}{dt}\delta\mathbf{v} = -\frac{\boldsymbol{\nabla}\,\delta p}{\rho_0} - \boldsymbol{\nabla}\,\delta\Phi - (\delta\mathbf{v}\cdot\boldsymbol{\nabla})\mathbf{v_0} \\
&\frac{d}{dt}\delta = -\boldsymbol{\nabla}\cdot\delta\mathbf{v} \\
&\nabla^2\delta\Phi = 4\pi G\rho_0\,\delta.
\end{aligned} \tag{88}$$

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

$$\mathbf{x}(t) = a(t)\mathbf{r}(t) \quad \Rightarrow \quad \boldsymbol{\nabla} \rightarrow \frac{1}{a}\boldsymbol{\nabla}, \tag{89}$$

and to make a similar transformation for peculiar velocity:

$$\delta\mathbf{v} = a\mathbf{u} \tag{90}$$

(although note that $\mathbf{u}$ is still a function of time, unlike $\mathbf{r}$).

When the equations are recast in these variables, there is 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:

$$[(\delta\mathbf{v}\cdot\boldsymbol{\nabla})\mathbf{v_0}]_j = [\delta v]_i \nabla_i [v_0]_j = H\,[\delta v]_j, \tag{91}$$

where the last step follows because $\mathbf{v_0} = H\,\mathbf{x_0} \Rightarrow \nabla_i[v_0]_j = H\,\delta_{ij}$. The equations for conservation of momentum and matter then take the following simple forms in comoving units:

$$\boxed{\begin{aligned}
&\dot{\mathbf{u}} + 2\frac{\dot{a}}{a}\,\mathbf{u} = \frac{\mathbf{g}}{a} - \frac{\boldsymbol{\nabla}\,\delta p}{\rho_0} \\
&\dot{\delta} = -\boldsymbol{\nabla}\cdot\mathbf{u}.
\end{aligned}} \tag{92}$$

The peculiar gravitational acceleration is denoted by $\mathbf{g}$. Note that, in the absence of peculiar accelerations and pressure forces, velocities redshift away through the 'Hubble drag' term $2H\mathbf{u}$. This behaviour is reasonable: if we shoot a bullet away from us with a proper peculiar velocity v, then after time t it is vt away, and its near neighbours have a recessional velocity $H\,vt$. The proper velocity thus decays as $\dot{v} + Hv = 0$ or $\dot{u} + 2Hu = 0$, because the bullet is always having to overtake distant galaxies with progressively higher speeds.

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

$$c_s^2 \equiv \frac{\partial p}{\partial\rho}. \tag{93}$$

If we now think of a plane-wave disturbance $\delta \propto e^{i\mathbf{k}\cdot\mathbf{r}}$ (so that $\mathbf{k}$ is a comoving wave-vector), then an equation for δ can be obtained by eliminating $\mathbf{u}$ (take the divergence of the perturbed Euler equation and the time derivative of the continuity equation and eliminate $\boldsymbol{\nabla}\cdot\dot{\mathbf{u}}$):

$$\ddot{\delta} + 2\frac{\dot{a}}{a}\dot{\delta} = \delta\left[4\pi G\rho_0 - c_s^2 k^2/a^2\right]. \tag{94}$$

This equation is the one which governs gravitational amplification of density perturbations. If we had linearized about a stationary background and taken $\mathbf{k}$ to be a *proper* wave-vector, then the equation would have been much easier to derive, and the result would be just $\ddot{\delta} = \delta(4\pi G\rho_0 - c_s^2 k^2)$, which has the solutions

$$\delta(t) = e^{\pm t/\tau}; \quad \tau = 1/\sqrt{4\pi G\rho_0 - c_s^2 k^2}. \tag{95}$$

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

$$\lambda_{\rm J} = c_s\sqrt{\frac{\pi}{G\rho}} \tag{96}$$

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. These effects help to govern the form of the perturbation spectrum which is propagated to the present Universe from early times, and will be considered in detail shortly.

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 conventional 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 linearized Einstein gravity. For simplicity, assume that pressure gradients are negligible (*i.e.* restrict ourselves to $\lambda \gg \lambda_{\rm J}$ from the start). The basic equations are then

$$\begin{aligned}
&\text{Euler}: \quad \frac{D\mathbf{v}}{Dt} = -\boldsymbol{\nabla}\Phi \\
&\text{Energy}: \quad \frac{D}{Dt}(\rho + p/c^2) = \frac{\partial}{\partial t}p/c^2 - (\rho + p/c^2)\boldsymbol{\nabla}\cdot\mathbf{v} \\
&\text{Gauss}: \quad \nabla^2\Phi = 4\pi G(\rho + 3p/c^2).
\end{aligned} \tag{97}$$

For total radiation domination, $p = \rho c^2/3$ and we can linearize as before, to obtain

$$\boxed{\ddot{\delta} + 2\frac{\dot{a}}{a}\dot{\delta} = \frac{32\pi}{3}G\rho_0\delta,} \tag{98}$$

so the net result of all the relativistic corrections is a driving term which is a factor 8/3 higher.

Solutions for $\delta(t)$

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

$$\begin{aligned} &\text{Matter domination } (a \propto t^{2/3}) : 4\pi G\rho_0 = \frac{2}{3t^2} \\ &\text{Radiation domination } (a \propto t^{1/2}) : 32\pi G\rho_0/3 = \frac{1}{t^2} \end{aligned} \tag{99}$$

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

$$\boxed{\delta \propto \eta^2.} \tag{100}$$

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

One further way of stating this result is that gravitational potential perturbations are independent of time (at least while $\Omega = 1$). Poisson's equation tells us that $-k^2\Phi/a^2 \propto \rho\,\delta$; since $\rho \propto a^{-3}$ for matter domination or a^{-4} for radiation, that gives $\Phi \propto \delta/a$ or δ/a^2 respectively – independent of a in either case. In other words, the metric fluctuations resulting from potential perturbations are frozen, at least for perturbations which are outside the horizon. This conclusion demonstrates the self-consistency of the basic set of fluid equations which were linearized. The equations for momentum and energy conservation are always valid, but the correct relativistic description of linear gravity should be a wave equation, including time derivatives of Φ as well as spatial derivatives. Since we have used only Newtonian gravity, this implies that any solutions of the perturbation equations in which Φ varies will not be valid on super-horizon scales. This criticism does not apply to the growing mode, where Φ is constant, but it does apply to decaying modes (Press & Vishniac 1980). This difficulty concerning perturbations on scales greater than the horizon is related to gauge freedom in general relativity and the fact that the value of density perturbations δ can be altered by a coordinate transformation. Exchange of light signals can be used to establish a 'sensible' coordinate system, but only on sub-horizon scales. Otherwise, the results obtained are gauge dependent. We are implicitly using here what might be called the Newtonian gauge, where the the metric is expressed as the FRW form with perturbation factors $(1 \pm 2\Phi/c^2)$ in the time and spatial parts respectively.

In models with $\Omega < 1$, the growth is slowed. It is possible to write explicit expressions for $\delta(a)$, but it is more convenient in practice to use the following accurate approximation, due to Carroll *et al.* (1992), which also allows for the effects of vacuum energy:

$$\frac{\delta(z=0,\Omega)}{\delta(z=0,\Omega=1)} \simeq \tfrac{5}{2}\Omega_m \left[\Omega_m^{4/7} - \Omega_v + (1+\tfrac{1}{2}\Omega_m)(1+\tfrac{1}{70}\Omega_v)\right]. \tag{101}$$

For models without vacuum energy, the growth is reduced by a factor of approximately $\Omega^{0.65}$; for flat models with $\Omega_m + \Omega_v = 1$, the growth suppression is less marked – approximately $\Omega^{0.23}$.

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 which affects only the overall expansion rate. The perturbation equation is as before

$$\ddot{\delta} + 2\frac{\dot{a}}{a}\dot{\delta} = 4\pi G\rho_m\delta, \tag{102}$$

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

$$\delta'' + \frac{2+3y}{2y(1+y)}\delta' - \frac{3}{2y(1+y)}\,\delta \tag{103}$$

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

$$\boxed{\delta \propto y + 2/3.} \tag{104}$$

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. At late times, the radiation becomes negligible, and the growth picks up smoothly to the Einstein-de Sitter $\delta \propto a$ behaviour. 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; this growth ceases once the perturbations enter the horizon. However, the explanations 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 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.

Solutions for u

To discuss velocities, go back to the basic equation for $\mathbf{u}$:

$$\dot{\mathbf{u}} + \frac{2\dot{a}}{a}\mathbf{u} = \frac{\mathbf{g}}{a}; \tag{105}$$

$\mathbf{g} = -\boldsymbol{\nabla}\delta\Phi/a$ is the peculiar gravitational acceleration, and pressure terms are neglected, so $\lambda \gg \lambda_{\rm J}$. The peculiar velocity can be decomposed into modes either parallel or perpendicular to $\mathbf{g}$. The latter are **vorticity modes** which decay. For the former, we know from the continuity equation ($\boldsymbol{\nabla}\cdot\mathbf{u} = -\dot{\delta}$) that $\dot{\mathbf{u}} = (\ddot{\delta}/\dot{\delta})\mathbf{u}$. Hence, the solution of the above equation for $\mathbf{u}$ has $\mathbf{u} \propto \mathbf{g}$, and may be expressed as

$$\delta\mathbf{v} = \frac{2f(\Omega)}{3H\Omega}\,\mathbf{g}, \tag{106}$$

where the function $f(\Omega) \equiv (a/\delta)d\delta/da$. A very good approximation to this (Peebles 1980) is $f \simeq \Omega^{0.6}$. Alternatively, we can work in Fourier terms. This is easy, as $\mathbf{g}$ and $\mathbf{k}$ are parallel, so that $\boldsymbol{\nabla}\cdot\mathbf{u} = ik\mathbf{u}$. Thus, directly from the continuity equation,

$$\boxed{\delta\mathbf{v}_{\mathbf{k}} = -\frac{iHf(\Omega)a}{k}\,\delta_k\,\hat{\mathbf{k}}.} \tag{107}$$

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

4.2 Transfer functions

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

(i) Jeans' mass effects. Prior to matter-radiation equality, we have already seen that perturbations inside the horizon are prevented from growing by radiation pressure. This leads to an effective 'break' of $\Delta n = 4$ in the power spectrum at this point:

$$\begin{aligned} \delta_k &\propto \lambda^{-(n+3)/2} \quad \lambda > \lambda_{\rm J} \\ &\propto \lambda^{-(n-1)/2} \quad \lambda < \lambda_{\rm J} \end{aligned} \tag{108}$$

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, $c_s^2 = \partial p/\partial\rho$, may be found by thinking about the response of matter and radiation to

small adiabatic compressions: $\delta p = (4/9)\rho_r c^2(\delta V/V)$, $\delta\rho = [\rho_m + (4/3)\rho_r](\delta V/V)$, implying

$$c_s^2 = c^2\left(3 + \frac{9}{4}\frac{\rho_m}{\rho_r}\right)^{-1} = c^2\left(3 + \frac{9}{4}\frac{1+z_{\rm rad}}{1+z}\right)^{-1}. \quad (109)$$

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 $c_s \propto \sqrt{1+z}$. Since $\rho = (1+z)^3 3\Omega_{\rm B}H_0^2/(8\pi G)$, the *comoving* Jeans' length is constant at

$$\Lambda_{\rm J} = \frac{c}{H_0}\left(\frac{32\pi^2}{27\Omega_{\rm B}(1+z_{\rm rad})}\right)^{1/2} = 50\,(\Omega_{\rm B}h^2)^{-1}\ \ {\rm Mpc}. \quad (110)$$

Thus, in either case, one of the critical length scales for the power spectrum will be the horizon distance at $z_{\rm eq}$ ($= 25000\Omega h^2$ for $T = 2.7$ K, counting neutrinos as radiation). In the matter-dominated approximation, we get

$$d_{\rm H} = \frac{2c}{H_0}(\Omega z)^{-1/2} = 29\,(\Omega h^2)^{-1}{\rm Mpc}. \quad (111)$$

The exact answer including radiation is a factor $\sqrt{2}-1$ times this: $15.7\,(\Omega h^2)^{-1}$ Mpc.

(ii) Damping. In addition to having their growth retarded, very small perturbation 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 which has entered the horizon will be damped. This process switches off when the particles become non-relativistic; for massive particles, this happens long before $z_{\rm eq}$ (**Cold Dark Matter**). 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**. In a pure baryon 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

$$\lambda_{\rm S} \simeq \sqrt{\lambda d_{\rm H}} \quad (112)$$

– 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

$$\lambda_{\rm S} = 16.3\,(1+z)^{-5/4}(\Omega_{\rm B}^2\Omega h^6)^{-1/4}\ {\rm Gpc}. \quad (113)$$

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

Real power spectra thus result from modifications of any primordial power by a variety of processes: growth under self-gravitation, effects of pressure and 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. The detailed result can be hard to calculate, mainly because we

have a mixture of matter (both collisionless dark particles and baryonic plasma) and relativistic particles (collisionless neutrinos and collisional photons) which does not behave as a simple fluid. Particular problems are caused by the change in the photon component from being a fluid tightly coupled to the baryons by Thomson scattering, to being collisionless after recombination. Accurate results require a solution of the Boltzmann equation to follow the evolution in detail. The transfer function is thus a by-product of elaborate numerical calculations of microwave background fluctuations. Nevertheless, once we possess the transfer function, it is a most valuable tool. The evolution of linear perturbations back to last scattering obeys the simple relations summarised above, and it is easy to see how structure in the Universe will have changed during the matter-dominated epoch.

It is thus invaluable in practice to have some accurate analytic formulae which fit the numerical results for transfer functions. We give below results for some common models in the form of the transfer function needed to produce a scale-invariant power spectrum at large wavelength, $\Delta^2 \propto k^4 T_k^2$. For adiabatic models, T_k is the true transfer function; for isocurvature models, this is not the case and T_k is proportional to $1/k^2$ times the true transfer function. We assume $\Omega_{\rm B} \ll \Omega$, so that all lengths scale with the horizon size at matter-radiation equality, leading to the definition $q \equiv k/(\Omega h^2 {\rm Mpc}^{-1})$. We consider the cases of (A) Adiabatic CDM; (B) Adiabatic massive neutrinos (1 massive, 2 massless); (C) Isocurvature CDM; these expressions come from Bardeen *et al.* (1986; BBKS).

$$\begin{aligned}
&\text{(A)} \quad T_k = \frac{\ln(1+2.34q)}{2.34q}[1+3.89q+(16.1q)^2+(5.46q)^3+(6.71q)^4]^{-1/4} \\
&\text{(B)} \quad T_k = \exp(-3.9q-2.1q^2) \\
&\text{(C)} \quad T_k = (1+[15.0q+(0.9q)^{3/2}+(5.6q)^2]^{1.24})^{-1/1.24}
\end{aligned} \tag{114}$$

These models contain some hidden variables. 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 which is 0.68 that of the photons.

There is also the question of the baryon contribution: 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 for the transfer function, Γ^*, where

$$\boxed{q \equiv (k/h\ {\rm Mpc}^{-1})/\Gamma^*,} \tag{115}$$

and $\Gamma^* = \Omega h$ in a model with zero baryon content. Peacock & Dodds (1994) showed that the effect of increasing $\Omega_{\rm B}$ 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):

$$\Gamma^* = \Omega h \exp[-\Omega_{\rm B}(1+1/\Omega)]. \tag{116}$$

4.3 N-body models

This is a good place to discuss how to use the above equations of motion to carry out a nonlinear evolution of the density field. This is usually done by means of the **N-body simulation**, in which the density field is represented by the sum of a set of fictitious discrete particles. The equations of motion for each particle depend on solving for the gravitational field due to all the other particles, finding the change in particle positions and velocities over some small time step, moving and accelerating the particles, and finally re-calculating the gravitational field to start a new iteration. Using comoving units for length and velocity ($\mathbf{v} = a\mathbf{u}$), we have from above the equation of motion

$$\frac{d}{dt}\mathbf{u} = -2\frac{\dot{a}}{a}\mathbf{u} - \frac{1}{a^2}\boldsymbol{\nabla}\Phi, \tag{117}$$

where Φ is the Newtonian potential. The time derivative is already in the form of a total (or convective) derivative, as required for particle motions, rather than the partial $\partial/\partial t$. If we change time variable from t to a, this becomes

$$\frac{d}{d\,\ln a}[a^2\mathbf{u}] = \frac{a}{H}\,\mathbf{g} = \frac{G}{aH}\,\sum_i m_i\,\frac{\mathbf{x_i} - \mathbf{x}}{|\mathbf{x_i} - \mathbf{x}|^3}. \tag{118}$$

Here, the gravitational acceleration has been written exactly by summing over all particles, but this becomes prohibitive for very large numbers of particles. Since the problem is to solve Poisson's equation, a faster approach is to use Fourier methods, since this allows the use of the FFT algorithm. If the density perturbation field (not assumed small) is expressed as $\delta = \sum \delta_k \exp[-i\mathbf{k}\cdot\mathbf{x}]$, then Poisson's equation becomes $-k^2\Phi_k = 4\pi G a^2 \bar{\rho}\,\delta_k$, and the required k-space components of $\boldsymbol{\nabla}\Phi$ are just

$$[\boldsymbol{\nabla}\Phi]_k = -i\Phi_k\mathbf{k}. \tag{119}$$

If we finally eliminate density in terms of Ω, the equation of motion for a given particle is

$$\begin{aligned} \frac{d}{d\,\ln a}[a^2\mathbf{u}] &= \sum \mathbf{F}_k \exp[-i\mathbf{k}\cdot\mathbf{x}]; \\ \mathbf{F}_k &= -i\mathbf{k}\,\frac{3\Omega H a^2}{2k^2}\,\delta_k. \end{aligned} \tag{120}$$

Boxes and grids

The efficient way of performing the required Fourier transforms is by averaging the data onto a grid and using the FFT algorithm, both to perform the transformation of density, and to perform the (three) inverse transforms to obtain the real-space force components from their k-space counterparts. This leads to the simplest N-body algorithm: the **particle-mesh (PM) code**. The only complicated part of the algorithm is the procedure for assigning mass to gridpoints, and interpolating the force as evaluated on the grid back onto the particles (for consistency, the same procedure must be used for both these steps). The most naive method is simply to bin the data: *i.e.* associate a given particle with whatever gridpoint happens to be

nearest. There are a variety of more subtle approaches (see Hockney & Eastwood 1988; Efstathiou *et al.* 1985), but whichever strategy is used, the resolution of a PM code is clearly limited to about the size of the mesh. To do better, one can use a **particle-particle-particle-mesh ($\mathbf{P^3M}$) code**, also discussed by the above authors. Here, the direct forces are evaluated between particles in the neighbouring cells, with the grid estimate being used only for particles in more distant cells. A similar effect, although without the use of the FFT, is achieved by **tree codes** (*e.g.* Hernquist Bouchet & Suto 1991).

In practice, however, the increase in resolution from these methods is limited to a factor of a few. This is because each particle in a cosmological N-body simulation in fact stands for a large number of less massive particles. Close encounters of these spuriously large particles can lead, through three-body processes, to the formation of unphysical close massive binaries. To prevent this, the forces must be **softened**: set to a constant below some critical separation, rather than rising as $1/r^2$. If there are already few particles per PM cell, the softening must be some significant fraction of the cell size, so there is a limit to the gain over pure PM. For example, consider a box of side $50h^{-1}$ Mpc, which is the smallest that can be used to simulate the observed universe without serious loss of power from the omitted long-wavelength modes. A typical size of calculation might use 128^3 particles on a Fourier mesh of the same size, so that the mean density is one particle per cell, and the cell size is of order that of the core of a rich cluster. To use such a simulation to study cluster cores means we are interested in overdensities of $10^3 - 10^4$, or typical interparticle separations of $0.05 - 0.1$ of a cell. To avoid collisional effects, the pairwise interaction must be softened on this scale. A much larger improvement in resolution is only justified in regions of huge overdensity ($\sim 10^6$ for a 100-fold increase in resolution over PM). The overall message is that N-body simulations in cosmology are severely limited by mass resolution, and that this limits the spatial resolution that can be achieved while still modelling the evolution of the true collisionless fluid.

Units

From a practical point of view, it is convenient to change to a new set of units which incorporate the size of the computational box, and allow the simulation to be rescaled to different physical situations. Let the side of the box be L; it is clearly convenient measure length in terms of L and velocities in terms of the expansion velocity across the box:

$$\begin{aligned} \mathbf{X} &= \mathbf{x}/L \\ \mathbf{U} &= \delta\mathbf{v}/(HLa) = \mathbf{u}/HL. \end{aligned} \tag{121}$$

Since, for N particles the density is $\rho = Nm/(aL)^3$, the mass of the particles and the gravitational constant can be eliminated and the equation of motion can be cast in an attractively dimensionless form:

$$\boxed{\frac{d}{d\,\ln a}[f(a)\mathbf{U}] = \frac{3}{8\pi}\,\Omega(a) f(a)\,\frac{1}{N}\sum_i \frac{\mathbf{X_i} - \mathbf{X}}{|\mathbf{X_i} - \mathbf{X}|^3}.} \tag{122}$$

The function $f(a)$ is proportional to $a^2 H(a)$, and has an arbitrary normalization – *e.g.* unity at the initial epoch. If the forces are instead evaluated on a grid, dimensionless wavenumbers $\mathbf{K} = \mathbf{k}\,L$ are used, and the corresponding equation becomes

$$\frac{d}{d\,\ln a}[f(a)\mathbf{U}] = \sum \mathbf{F}_K \exp[-i\mathbf{K}\cdot\mathbf{X}];$$
$$\mathbf{F}_K = -i\mathbf{K}\,\frac{3}{2}\,\frac{\delta_K}{K^2}\,\Omega(a)f(a). \tag{123}$$

Particles are now moved according to $d\mathbf{x} = \mathbf{u}\,dt$, which becomes

$$d\mathbf{X} = \mathbf{U}\,d\ln a \tag{124}$$

in our new units. It only remains to set up the initial conditions; this is easy to do if the initial epoch is at high enough redshift that $\Omega = 1$, since then $\mathbf{U} \propto a$ and the initial displacements and velocities are related by

$$\Delta\mathbf{X} = \mathbf{U}. \tag{125}$$

In the case where the initial conditions are specified late enough that Ω is significantly different from unity, this can be modified by using the linear relation between density and velocity perturbations: for a given Δ, the corresponding velocity scales as $\Omega^{0.6}$ (see above). It is not in fact critical that the density fluctuations be very small at this time: this is related to a remarkable approximation for nonlinear dynamics due to Zeldovich.

4.4 Hierarchical density fields

We have seen that the perturbations which survive recombination are of two distinct classes: in some (such as cold dark matter), primordial fluctuations survive on very small scales (**small-scale damping**); in other cases, such as hot dark matter or adiabatic baryons, the perturbation field is dominated by fluctuations on scales $\sim$ the horizon at $z_{\rm eq}$ (**large-scale damping**). The consequences for galaxy formation are radically different. In the former case, non-linear collapse of sub-galactic mass units can be the first event which occurs after recombination. These will then cluster together in a **hierarchy**, forming successively more massive systems as time progresses. Hierarchies are also known as **bottom-up** pictures for galaxy formation. Conversely, in large-scale damping, pancakes of cluster or supercluster size are the first structures to form. Galaxies must be presumed to form through dissipative processes occurring in the shocked gas which forms at pancake collapse (**top down**).

As described, it is clear that top-down pictures are unappealing in that the physics of galaxy formation is likely to be very complex and messy. Reality *is* often like this: the formation of stars is a good example of a fundamental process where is is hard to understand what is going on. In contrast, the computational simplicity of hierarchies has led to much more detailed work being performed on them. Theoretical prejudice aside, however, the Universe *looks* like a hierarchy, displaying many small groups of galaxies (*e.g.* the Milky Way's own Local Group), which exist within superclusters which are only mildly nonlinear.

It may seem that such a situation cannot be analysed within the bounds of linear theory, but a way forward was identified by Press & Schechter (1974; PS). The

critical assumption in the PS analysis is that, even if the field is non-linear, the amplitude of large-wavelength modes in the final field will be close to that predicted from linear theory. For this to be true requires the 'true' large-scale power to exceed that generated via non-linear coupling of small-scale modes, which turns out to require a spectral index $n < 1$ (Williams *et al.* 1991a). We now proceed by recognising that, for a massive clump to undergo gravitational collapse, the average overdensity in a volume containing that mass should (as usual) exceed some threshold, δ_c, of order unity. The location and properties of these bound objects can thus be estimated by an artificial smoothing (or filtering) of the initial linear density field. If the filter function has some characteristic length R_f, then the typical size of filtered fluctuations will be $\sim R_f$ and they can be assigned a mass $M \sim \rho_0 R_f^3$. The exact analytic form of the filter function is arbitrary and is often taken to be a Gaussian for analytic convenience.

The argument now proceeds in integral terms. For a given R_f, the probability that a given point lies in a region with $\delta > \delta_c$ (the critical overdensity for collapse) is

$$p(\delta > \delta_c \,|R_f) = \tfrac{1}{2}\left[1 - \mathrm{erf}\left(\frac{\delta_c}{\sqrt{2}\,\sigma(R_f)}\right)\right], \tag{126}$$

where $\sigma(R_f)$ is the linear rms in the filtered version of δ. The PS argument now takes this to be proportional to the probability that a given point has ever been processed through a collapsed object of scale $> R_f$. This is really assuming that the only objects which exist at a given epoch are those which have just collapsed: if a point has $\delta > \delta_c$ for a given R_f, then it will have $\delta = \delta_c$ when filtered on some larger scale and will be counted as an object of the larger scale. The problem with this argument is that half the mass remains unaccounted for: this was amended by PS simply by multiplying the probability by a factor 2. Note that this procedure need not be confined to Gaussian fields; all we need is the functional form of $p(\delta > \delta_c \,|R_f)$. The analogue of the factor of 2 problem remains: what to do with the points having $\delta < 0$.

This integral probability is related to the mass function $f(M)$ (defined such that $f(M)dM$ is the comoving number density of objects in the range dM) via

$$Mf(M)/\rho_0 = |dp/dM|, \tag{127}$$

where ρ_0 is the total comoving density. Thus,

$$\boxed{\frac{M^2 f(M)}{\rho_0} = \frac{2\delta_c}{\sqrt{2\pi}\,\sigma}\left|\frac{d\,\ln\sigma}{d\,\ln M}\right| \exp(-\tfrac{1}{2}\delta_c^2/\sigma^2).} \tag{128}$$

We have expressed the result in terms of the **multiplicity function**: $M^2 f(M)/\rho_0$ is the fraction of the mass which is carried by objects in a unit range of $\ln M$. For power-law spectra, this function takes a very simple form:

$$\frac{M^2 f(M)}{\rho_0} = \frac{n+3}{6}\sqrt{\frac{2}{\pi}}\;\nu\, e^{-\nu^2/2}, \tag{129}$$

Where ν is the threshold in units of the rms density fluctuation. The multiplicity function thus always has the same shape (a skew-negative hump around $\nu \simeq 1$); changing the spectral index only alters the mass scale via $\nu = (M/M_c)^{(n+3)/6}$.

Random walks and conditional mass functions

The factor of 2 'fudge' has long been recognized as the crucial weakness of the PS analysis. What one has in mind is that the mass from lower-density regions accretes onto collapsed objects, but it does not seem correct for this to cause a doubling of the total number of objects. Recent work has shed some light on the origin of this problem (Peacock & Heavens 1990; Bond *et al.* 1991). To see where the error crept in, consider the **random trajectory** taken by the filtered field at some fixed point as a function of filtering radius. This starts at $\delta = 0$ at $R = \infty$, and develops fluctuations of increasing amplitude as we move to smaller R. Thus, if $\delta < \delta_c$ at a given point, it is quite possible that it will exceed the threshold at some other point – indeed, if the field variance diverges as $R \to 0$, it is inevitable that the threshold will be exceeded. So, instead of ignoring all points below threshold at a given R, we should find the **first upcrossing** of the random trajectory: the largest value of R for which $\delta = \delta_c$.

This analysis is most easily performed for one particular choice of filter: sharp truncation in k-space. Decreasing R then corresponds to adding in new k-space shells, all of which are independent for a Gaussian field. The trajectory is then just a random walk, and the solution is very easy. Consider a point on the walk which has reached the threshold; its subsequent motion will be symmetric, so that it is equally likely to be found above the threshold as below at some smaller R. The probability of never having crossed the threshold (the **survival probability**) is then obtained by reflection of the Gaussian above threshold:

$$\frac{dP_{\rm S}}{d\delta} = \frac{1}{\sqrt{2\pi}\,\sigma}\left[\exp\left(-\frac{\delta^2}{2\sigma^2}\right) - \exp\left(-\frac{(\delta - 2\delta_c)^2}{2\sigma^2}\right)\right]. \tag{130}$$

Integrating this up to get the probability of having crossed the threshold at least once gives

$$1 - P_{\rm S} = 1 - \mathrm{erf}\left(\frac{\delta_c}{\sqrt{2}\,\sigma}\right), \tag{131}$$

which is just twice the unconstrained probability of lying above the threshold – thus supplying the missing factor of 2. Unfortunately, the above analysis is only valid for this special choice of filter: using k-space filters which are differentiable leads to just the original PS form (without the factor 2) at high mass, with the surplus probability being shifted to low masses, so that the shape of the function changes. Which filter is the best choice, we cannot say in advance; we are left with the empirical fact that the PS formula does fit N-body results quite well.

One useful extension of the random-walk model is that it allows a calculation of the **conditional multiplicity function**: given a particle in a system of mass M_0 at some epoch a_0, what was the distribution of masses where that particle resided at some earlier epoch a_1? This is now just the random walk with two absorbing barriers, at δ_c/a_0 and δ_c/a_1; we want the probability of not having crossed the

second, subject to having crossed the first at M_0. The solution for the integral mass distribution is

$$P(>\mu) = 1 - \mathrm{erf}\left[\frac{\nu(t^{-1}-1)}{\sqrt{2(\mu^{-1}-1)}}\right], \tag{132}$$

where $\nu \equiv \delta_c/a_0\sigma_0$; $t \equiv a_1/a_0$; $\mu \equiv \sigma^2(M_0)/\sigma^2(M)$. This function tells us that the early histories of particles which end up in different mass systems are markedly different over a large range of expansion factor. This may provide a way of understanding many of the systematics of galaxy systems in terms of their merger histories: see Bower (1991); Lacey & Cole (1993).

The way to deduce a merger *rate* from the above is to consider the limit of the conditional mass function at two nearly equal epochs: the time derivative of the conditional function must be related to the merger rate. First, we need to invert the above reasoning, which gives $P(M_1, z \mid M_0, z = 0)$. Conditional probability definitions then imply

$$P(M_0 \mid M_1) = \frac{P(M_1 \mid M_0)\, P(M_0)}{P(M_1)}. \tag{133}$$

We now have the probability that an object of mass M_1 at redshift z becomes incorporated into one of mass M_0 at redshift $z = 0$. The merger rate $\mathcal{M}$, at which an object of mass M_0 accretes objects of mass ΔM, is then

$$\mathcal{M}(M_0, \Delta M) = \left.\frac{d^2P(M_0 \mid M_1)}{dM_0\, dt}\right|_{t_1 \to t_0}. \tag{134}$$

These analytical descriptions of how mass gathers into clumps in hierarchical clustering have been shown to describe the results of numerical simulations extremely well (Bond *et al.* 1991) and provide a powerful tool for understanding the growth of cosmological structure. For example, it is now possible to answer the apparent paradox of the simple Press-Schechter theory, where all collapsed objects considered are those which formed only at the instant considered (*i.e.* sit exactly at the threshold $\delta = \delta_c$), even though common sense says that some objects must have formed at high redshift and survived unchanged to the present day. The conditional mass function indeed allows us to define something close to the typical formation epoch for a clump, which we might take as the time when the probability is 0.5 that the precursor mass is at least half the final mass. The conditional mass function says that this time depends on the final mass of the clump; if it is $\gg M^*$, the formation epoch will be in the recent past, whereas low-mass clumps survive for longer. As a specific example, this calculation implies that massive clusters form extremely late in an Einstein-de Sitter model. About 30% of Abell clusters will have doubled their mass since as recently as $z = 0.2$ (as against only 5% if $\Omega = 0.2$). The observed frequency of substructure in clusters is used on this basis by Lacey & Cole to argue that $\Omega \gtrsim 0.5$ (although lower Ω is allowed in models with vacuum energy, since what matters for this analysis is the linear growth suppression factor).

Application to galaxy clusters

The most important practical application of the PS formalism is to rich clusters. As already discussed, these are the most massive non-linear systems in the current Universe, so a study of their properties should set constraints on the shape and normalization of the power spectrum on large scales. These issues are discussed by *e.g.* Henry & Arnaud (1991), who sidestep the issue of what mass to assign to a cluster by using the observed distribution of temperatures (see below for the relation between mass and virial temperature). Fitting their data with the PS form for top-hat filtering and $\delta_c = 1.69$ they deduce a *linear-theory* rms in $8h^{-1}$ Mpc spheres of $\sigma_8 = 0.59 \pm 0.02$ for $\Omega = 1$. Similarly, White, Efstathiou & Frenk (1993) deduce $\sigma_8 = 0.57 \pm 0.06$, although they rely on the central masses of Abell clusters being known when calibrating this number. These calculations have been checked against numerical simulations, so the result therefore seems rather robust and well-determined. For models with low density, the required normalization rises: for a lower mean density, more collapsed systems are needed to yield the observed number of objects of a given mass. The approximate scaling with Ω is approximately the same as is required in the cosmic virial theorem to keep velocity dispersions constant (reasonably enough, given that these determine the cluster masses):

$$\boxed{\sigma_8 \simeq (0.5 - 0.6)\ \Omega^{-0.5}.} \tag{135}$$

This knowledge of σ_8 as a measure of the true mass inhomogeneity is a fundamental cosmological datum.

4.5 Galaxy formation

Cooling and galaxy formation

This discussion of mass functions really applies only to 'haloes' of collisionless dark matter. Things are more complex for baryonic matter, where we must ask if the matter has been able to **dissipate** and turn into stars. This question was analyzed in a classic paper by Rees & Ostriker (1977), and has been reconsidered in the context of CDM by Blumenthal *et al.* (1984).

Mergers will heat gas up to the virial temperature via shocks; in order for the gas to form stars, it must be able to undergo **radiative cooling** – to dispose of this thermal energy. Clearly, the redshift of collapse clearly needs to be sufficiently large that there is time for an object to cool between its formation at redshift $z_{\rm cool}$ (when $\delta\rho/\rho \simeq \delta_c$) and the present epoch. We shall show below that $z_{\rm cool}$ is a function of mass; it is therefore possible to put cooling into the Press-Schechter machinery simply by using the *mass-dependent* threshold $\nu(M) = \delta_c[1 + z_{\rm cool}(M)]/\sigma_0(M)$ in the mass function.

The cooling function for a plasma in thermal equilibrium has been calculated by Raymond, Cox & Smith (1976). For an H + He plasma with $Y = 0.25$ and some admixture of metals, their results for the cooling time ($t_{\rm cool} \equiv 3kT/2\Lambda(T)n$) may be approximated as roughly

$$t_{\rm cool}/{\rm years} = 1.8 \times 10^{24} \left(\frac{\rho_B}{M_\odot {\rm Mpc}^{-3}}\right)^{-1} \left(T_8^{-1/2} + 0.5 f_m T_8^{-3/2}\right)^{-1}, \tag{136}$$

where $T_8 \equiv T/10^8 K$. The $T^{-1/2}$ term represents bremsstrahlung cooling and the $T^{-3/2}$ term approximates the effects of recombination radiation. The parameter f_m governs the metal content: $f_m = 1$ for solar abundances; $f_m \simeq 0.03$ for no metals. In this model where so far dissipation has not been considered, the baryon density is proportional to the total density, the collapse of both resulting from purely gravitational processes. $\rho_{\rm B}$ is then a fraction $\Omega_{\rm B}/\Omega$ of the virialized total density. This is itself some multiple f_c of the background density at virialization (which we refer to as 'collapse'):

$$\rho_c = f_c\, \rho_0\, (1+z_c)^3. \tag{137}$$

The virialized potential energy for constant density is $3GM^2/(5r)$, where the radius satisfies $4\pi\rho_c r^3/3 = M$. This energy must equal $3MkT/(\mu m_p)$, where $\mu = 0.59$ for a plasma with 75% Hydrogen by mass. Hence, using $\rho_0 = 2.78\times10^{11}\Omega h^2\, M_\odot {\rm Mpc}^{-3}$, we obtain

$$\boxed{T_{\rm virial}/K = 10^{5.1}(M/10^{12}M_\odot)^{2/3}\,(f_c\Omega h^2)^{1/3}\,(1+z_c).} \tag{138}$$

So, for $\Omega = 1$, we must solve $f_t\, t_{\rm cool} = \frac{2}{3}H_0^{-1}[1-(1+z_c)^{-3/2}]$. If only recombination cooling was important, the solution to this would be

$$\begin{aligned} (1+z_c) &= (1+M/M_{\rm cool})^{2/3} \\ M_{\rm cool}/M_\odot &= 10^{13.1} f_t^{-1} f_m f_c^{1/2} \Omega_{\rm B} \Omega^{-1/2} \end{aligned} \tag{139}$$

For high metallicity, where bremsstrahlung only dominates at $T \gtrsim 10^8 K$, this equation for z_c will be a reasonable approximation up to $z_c \simeq 10$, at which point Compton cooling will start to operate. Given that we expect at least some enrichment rather early in the progress of the hierarchy, we shall keep things simple by using just the above expression for z_c.

We see that cooling is rapid for low masses, where the luminous and dark mass functions are expected to coincide. Given that cooling of massive objects is ineffective, probability in the mass function must therefore accumulate at intermediate masses: the numbers of faint galaxies relative to bright are decreased. If $M_{\rm cool} \ll M_c$, then there is a power-law region between these two masses which differs from the PS slope: $M^2 f(M) \propto M^{\left(\frac{n+3}{6}\right)+\frac{2}{3}}$; *i.e.* there is an effective change in n to $n+4$. This may be relevant for the galaxy luminosity function, where the faint-end slope is close to constant numbers per magnitude. For constant mass-to light ratio, this implies $M^2 f(M) \propto M$, and apparently requires $n = 3$. Alternatively, a spectral index more in accord with large-scale structure observations of $n \lesssim -1$ gives a Press-Schechter slope much steeper than the observed galaxy luminosity function. There have been many full and complicated studies of galaxy formation that show in detail why this is not really as much of a paradox as it initially seems, but the above simplified discussion of cooling illustrates the main way in which the naive analysis goes wrong.

5 Testing inflation against galaxy clustering

5.1 Clustering statistics

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. For other models, plane waves are not a complete set and one should use instead the eigenfunctions of the wave equation in a curved space. Normally this complication is neglected: even in an open Universe, the difference only matters on scales of order the present-day horizon.

How do we make a Fourier expansion of an infinite density field? If the field was periodic within some box of side L, then we would just have a sum over wave modes:

$$F(\mathbf{x}) = \sum F_{\mathbf{k}} e^{-i\mathbf{k}\cdot\mathbf{x}}. \tag{140}$$

Now, if we let the box become arbitrarily large, then the sum will go over to an integral which incorporates the density of states in k-space – exactly as in statistical mechanics: The Fourier relations in n dimensions are thus

$$\boxed{\begin{aligned} F(x) &= \left(\frac{L}{2\pi}\right)^n \int F_k(k) \exp -i\mathbf{k}\cdot\mathbf{x}\, d^n k \\ F_k(k) &= \left(\frac{1}{L}\right)^n \int F(x) \exp i\mathbf{k}\cdot\mathbf{x}\, d^n x. \end{aligned}} \tag{141}$$

One advantage of this particular Fourier convention is that the definition of convolution is just a simple volume average, with no gratuitous factors of $(2\pi)^{-1/2}$:

$$f * g \equiv \frac{1}{L^n} \int f(\mathbf{x}-\mathbf{y}) g(\mathbf{y}) d^n y. \tag{142}$$

Although one can make all manipulations on density fields which follow using either the integral or sum formulations, it is usually easier to use the sum. This saves having to introduce δ-functions in k-space. For example, if we have $f = \sum f_k \exp(-ikx)$, the obvious way to extract f_k is via $f_k = (1/L) \int f \exp(ikx)\ dx$: because of the harmonic boundary conditions, all oscillatory terms in the sum integrate to zero, leaving only f_k to be integrated from 0 to L. There is less chance of committing errors of factors of 2π in this way than considering $f = (L/2\pi) \int f_k \exp(-ikx)\ dk$ and then using $\int \exp[i(k-K)x]\ dx = 2\pi\delta_{\rm D}(k-K)$.

Correlation functions and power spectra

As an immediate example of the Fourier machinery in action, consider the important quantity

$$\xi(\mathbf{r}) \equiv \langle \delta(\mathbf{x}) \delta(\mathbf{x}+\mathbf{r}) \rangle\,, \tag{143}$$

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. If we express δ as a sum, and note that reality means we can replace one of the two δ's by its complex conjugate, then we obtain

$$\xi = \left\langle \sum_{\mathbf{k}} \sum_{\mathbf{k}'} \delta_{\mathbf{k}} \delta^*_{\mathbf{k}'} e^{i(\mathbf{k}'-\mathbf{k})\cdot\mathbf{x}} e^{-i\mathbf{k}\cdot\mathbf{r}} \right\rangle . \tag{144}$$

An alternative way of obtaining this is to use the relation between modes with opposite wavevectors which holds for any real field: $\delta_{\mathbf{k}}(-\mathbf{k}) = \delta^*_{\mathbf{k}}(\mathbf{k})$. By the periodic boundary conditions, however, all the cross terms with $\mathbf{k}' \neq \mathbf{k}$ average to zero. Expressing the remaining sum as an integral, we have

$$\boxed{\xi(\mathbf{r}) = \frac{V}{(2\pi)^3} \int |\delta_{\mathbf{k}}|^2 e^{-i\mathbf{k}\cdot\mathbf{r}} d^3k.} \tag{145}$$

In short, the correlation function is the Fourier transform of the **power spectrum**. We shall hereafter often use the alternative notation $P(k) \equiv |\delta_k|^2$.

Now, in an isotropic universe, the density perturbation spectrum should contain no preferred direction, and so we must have an **isotropic power spectrum**: $|\delta_{\mathbf{k}}|^2(\mathbf{k}) = |\delta_k|^2(k)$. We can therefore perform the angular integral: introduce spherical polars with the polar axis along $\mathbf{k}$, and use the reality of ξ so that $e^{-i\mathbf{k}\cdot\mathbf{x}} \rightarrow \cos(kr\cos\theta)$. In three dimensions, this yields

$$\xi(r) = \frac{V}{(2\pi)^3} \int |\delta_k|^2 \frac{\sin kr}{kr}\, 4\pi k^2\, dk. \tag{146}$$

The 2D analogue of this formula is

$$\xi(r) = \frac{A}{(2\pi)^2} \int |\delta_k|^2\, J_0(kr)\, 2\pi k\, dk. \tag{147}$$

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

$$\boxed{\Delta^2(k) \equiv \frac{V}{(2\pi)^3}\, 4\pi k^3\, P(k) = \frac{2}{\pi} k^3 \int_0^\infty \xi(r)\, \frac{\sin kr}{kr}\, r^2\, dr.} \tag{148}$$

This gives a more easily visualisable 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.

Power-law spectra

The above shows that the power spectrum is a vital 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 it. This means that the spectrum must be a featureless power law:

$$\boxed{|\delta_k|^2 \propto k^n} \tag{149}$$

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 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

$$\delta_{\rm rms} \propto M^{-(n+3)/6}. \tag{150}$$

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

$$\Delta^2(k) = \frac{2}{\pi} (kr_0)^\gamma \, \Gamma(2-\gamma) \, \sin \frac{(2-\gamma)\pi}{2} \equiv \beta (kr_0)^\gamma \tag{151}$$

$(= 0.903(kr_0)^{1.8}$ 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.

What general constraints can we set on the value of n? Asymptotic homogeneity clearly requires $n > -3$. An upper limit of $n < 4$ comes from an argument due to Zeldovich. Suppose we begin with a totally uniform matter distribution and then group it into discrete chunks as uniformly as possible. It can be shown that conservation of momentum in this process means that we cannot create a power spectrum which goes to zero at small wavelengths more rapidly than $\delta_k \propto k^2$. Thus, discreteness of matter produces the **minimal spectrum**: $n = 4$.

More plausible alternatives lie between these extremes. The value $n = 0$ corresponds to **white noise**: the same power at all wavelengths. This is also known as the **Poissonian** power spectrum, because it corresponds to fluctuations between different cells which scale as $1/\sqrt{M_{\rm cell}}$ (see below). A density field created by throwing down a large number of point masses at random would therefore consist of white noise. Particles placed at random within cells, one per cell, create an $n = 2$ spectrum on large scales. Practical spectra in cosmology, conversely, often have negative effective values of n over a large range of wavenumber. In this sense, large-scale structure is much more 'real' than the simple white-noise fluctuations familiar in other contexts.

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 where the name arises, consider perturbations in gravitational potential:

$$\nabla^2 \delta\Phi = 4\pi G \rho_0 \delta \Rightarrow \delta\Phi_k = -4\pi G \rho_0 \delta_k / k^2. \tag{152}$$

The two powers of k pulled down by ∇^2 mean that, if $\Delta^2 \propto k^4$ for matter, then δ_Φ^2 is a constant. Since potential perturbations govern the flatness of spacetime, this says that the scale-invariant spectrum corresponds to a metric which is fractal-like: it 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:

$$\delta(r_{\rm H}) \propto r_{\rm H}^{-(n-1)/2}. \tag{153}$$

Thus, if $n = 1$, the growth of $r_{\rm H}$ and δ with time cancel out so that the universe always looks the same when viewed on the scale of the horizon; such a universe is a fractal 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 also invented it independently).

Filtering and moments

A common concept in the manipulation of cosmological density fields is that of **filtering**: convolution of the density field with some **window function**: $\delta \rightarrow \delta * f$. Many observable results can be expressed in this form. Some common 3D filter functions are

$$\begin{aligned} &\text{Gaussian}: f = \frac{V}{(2\pi)^{3/2} R_{\rm G}^3} e^{-r^2/2R_{\rm G}^2} \Rightarrow f_k = e^{-k^2 R_{\rm G}^2/2} \\ &\text{Top}-\text{hat}: f = \frac{3V}{4\pi R_{\rm T}^3} (r < R_{\rm T}) \Rightarrow f_k = \frac{3}{y^3}[\sin y - y\cos y] \quad (y \equiv kR_{\rm T}) \end{aligned} \tag{154}$$

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

$$R_{\rm T} = \sqrt{5}\, R_{\rm G}. \tag{155}$$

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:

$$\boxed{\sigma_n^2 \equiv \frac{V}{(2\pi)^3} \int P(k)\, |f_k|^2\, k^{2n}\, d^3k;} \tag{156}$$

the filtered variance is thus σ_0^2 (which we shall often denote by just σ^2). Moments may also be expressed in terms of the correlation function over the sample volume:

$$\sigma^2 = \int\int \xi(|\mathbf{x}-\mathbf{x}'|)\, f(\mathbf{x})\, f(\mathbf{x}')\, d^3x\, d^3x'. \tag{157}$$

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

$$\xi^{(2n)}(0) = (-1)^n\, \frac{\sigma_n^2}{2n+1}. \tag{158}$$

Normalization

For scale-invariant spectra, a natural amplitude measure is the (constant) gravitational potential variance per unit $\ln k$:

$$\epsilon^2 \equiv \frac{V}{(2\pi)^3}\, 4\pi k^3 |\Phi_k|^2/c^4 = \frac{9}{4}\left(\frac{ck}{H_0}\right)^{-4} \Delta^2(k). \tag{159}$$

Two more commonly encountered measures relate to the clustering field around 10 Mpc. One is σ_8: the rms density variation when smoothed in spheres of radius $8h^{-1}$ Mpc; this is observed to be very close to unity. The other is an integral over the correlation function:

$$J_3 \equiv \int_0^r \xi(y)\, y^2 dy = \int \Delta^2(k) W(k)\, \frac{dk}{k}, \tag{160}$$

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

$$\bar{\xi}(r) = \frac{3}{4\pi r^3}\int_0^r \xi(x)\, 4\pi x^2\, dx = \frac{3}{r^3} J_3(r). \tag{161}$$

The canonical value then becomes $\bar{\xi}(10h^{-1}{\rm Mpc}) = 0.83$; 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 non-linear, there is no reason why these two definitions 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.

5.2 Non-linear evolution

How is the power spectrum altered by nonlinear evolution? As with most nonlinear questions, this cannot be answered analytically in general, but some aspects of the scaling of the problem are well understood.

A useful trick 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 will be of the same form as the correlation function, since this just measures the number density of neighbours to a given galaxy. Thus, a power-law correlation $\xi(r) \propto r^{-\gamma}$ may be thought of as arising from the $\rho \propto r^{-\gamma}$ haloes 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 which do not change as the universe expands. We have **stable clustering**: ξ is fixed in *proper* terms apart from a $(1+z)^{-3}$ scaling 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

$$\xi(r,z) \propto (1+z)^{\gamma-3} \quad (\mathrm{non-linear}). \tag{162}$$

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

$$\xi(r,z) \propto (1+z)^{-2} \quad (\mathrm{linear}). \tag{163}$$

This argument does not so far give a relation between the non-linear slope γ and the index n of the linear spectrum. However, these two rates of evolution must match to give the same prediction for the evolution of r_0 – the length-scale of nonlinearity – since this is where ξ will break from the linear $\xi \propto r^{-(n+3)}$ to the nonlinear $\xi \propto r^{-\gamma}$. The linear and non-linear 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 $\gamma = (3n+9)/(n+5)$, or in terms of an effective index $\gamma = 3 + n_{\rm eff}$:

$$\boxed{n_{\rm eff} = -\frac{6}{5+n}.} \tag{164}$$

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

Whether this evolution has been seen or not is presently controversial. Efstathiou *et al.* (1991) have observed a very low amplitude for the angular clustering of galaxies at $B \simeq 26$, inferring that (if $\Omega = 1$) the clustering must evolve very rapidly – at about the linear-theory rate. However, their fields are very small and it is possible their small result is not representative. An indication that clustering may not decline this rapidly is given by the observed clustering of quasars at $z \simeq 1$; Shanks & Boyle (1994) find the relatively high value $r_0 = 7h^{-1}$ Mpc.

For many years it was thought that only the limiting cases of extreme linearity or nonlinearity could be dealt with analytically, but in a marvelous piece of alchemy,

Hamilton *et al.* (1991; HKLM) gave a universal analytical formula for accomplishing the linear $\leftrightarrow$ nonlinear mapping. The conceptual basis of their method can be understood with reference to the spherical collapse model. For $\Omega = 1$ (the only case they considered), 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 reaches a critical value, so one should work with the volume-averaged correlation function $\bar{\xi}(r)$. 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 non-linear correlation $\bar{\xi}_{\rm NL}(r_{\rm NL})$ should be thought of as resulting from linear correlations on a linear scale

$$r_{\rm L} = [1 + \bar{\xi}_{\rm NL}(r_{\rm NL})]^{1/3} r_{\rm NL}. \tag{165}$$

This is one part of the HKLM procedure. The second part, having translated scales as above, is to conjecture that the nonlinear correlations are a universal function of the linear ones:

$$\bar{\xi}_{\rm NL}(r_{\rm NL}) = f_{\rm NL}[\bar{\xi}_{\rm L}(r_{\rm L})]. \tag{166}$$

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 power spectrum.

To use this method with power spectra, we can use the relations between $\bar{\xi}(r)$ and $\xi(r)$

$$\begin{aligned} \bar{\xi}(r) &= \frac{3}{r^3} \int_0^r \xi(x)\, x^2\, dx \\ \xi(r) &= \frac{d\,[r^3\, \bar{\xi}(r)]}{d[r^3]}, \end{aligned} \tag{167}$$

followed by the Fourier relations between $\xi(r)$ and $\Delta^2(k)$, to obtain

$$\begin{aligned} \bar{\xi}(r) &= \int_0^\infty \Delta^2(k)\, \frac{dk}{k}\, \frac{3}{(kr)^3} [\sin kr - kr \cos kr] \\ \Delta^2(k) &= \frac{2k^3}{3\pi} \int_0^\infty \bar{\xi}(r)\, r^2\, dr\, \frac{1}{(kr)} [\sin kr - kr \cos kr], \end{aligned} \tag{168}$$

where the last relation holds provided that $\bar{\xi}(r) \to 0$ faster than r^{-2} at large r (i.e. a spectrum which asymptotically has $n > -1$, a valid assumption for spectra of practical interest).

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 which cuts off rapidly at $k \gtrsim 1/r$:

$$\begin{aligned} \bar{\xi}(r) &= \Delta^2(k_{\rm eff}) \\ k_{\rm eff} &\simeq 2/r, \end{aligned} \tag{169}$$

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:

$$\boxed{\begin{aligned} \Delta^2_{\rm NL}(k_{\rm NL}) &= f_{\rm NL}[\Delta^2_{\rm L}(k_{\rm L})] \\ k_{\rm L} &= [1+\Delta^2_{\rm NL}(k_{\rm NL})]^{-1/3} k_{\rm NL}. \end{aligned}} \tag{170}$$

Even better, it is not necessary that the number relating $1/x$ and $k_{\rm eff}$ be a constant over the whole spectrum. All that matters is that the number can be treated as constant over the limited range $r_{\rm NL}$ to $r_{\rm L}$. This means that the deviations of the above formulae from the exact transformation of the HKLM procedure are only noticeable in cases where the power spectrum deviates markedly from a smooth monotonic function, or where either the linear or nonlinear spectra are very flat ($n \lesssim -2$).

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 $g(\Omega)$. It then follows that the large-x asymptote of the nonlinear function is

$$f_{\rm NL}(x) \propto [g(\Omega)]^{-3}\, x^{3/2}. \tag{171}$$

According to Carroll, Press & Turner (1992), the required growth-suppression factor may be approximated almost exactly by

$$g(\Omega) = \tfrac{5}{2}\Omega_m \left[\Omega_m^{4/7} - \Omega_v + (1+\Omega_m/2)(1+\Omega_v/70)\right]^{-1}, \tag{172}$$

where we have distinguished matter (m) and vacuum (v) contributions to the density parameter explicitly.

Peacock & Dodds (1996) suggested the following generalization of the HKLM method, using the following fitting formula for the nonlinear function (strictly, the one which applies to the power spectrum, rather than to $\bar{\xi}$):

$$f_{\rm NL}(x) = x\, \left[\frac{1 + B\beta x + [Ax]^{\alpha\beta}}{1 + ([Ax]^{\alpha} g^3(\Omega)/[Vx^{1/2}])^{\beta}}\right]^{1/\beta}. \tag{173}$$

B describes a second-order deviation from linear growth; A and α parameterise the power-law which dominates the function in the quasilinear regime; V is the virialization parameter which gives the amplitude of the $f_{\rm NL}(x) \propto x^{3/2}$ asymptote; β softens the transition between these regimes.

Fig. 3. The generalization of the HKLM function relating nonlinear power to linear power, for the cases $n = 0$, -1 and -2, and with $\Omega = 1$ and 0.2 (dotted lines). The nonlinear power increases for lower Ω and for more negative n, but in a nearly universal way for $n \geq -1$. The fitting formula is shown for models with zero vacuum energy only, but what matters in general is just the Ω-dependent linear growth suppression factor.

HKLM's suggestion 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) find that an excellent fit (illustrated in Figure 3) is given by the following spectrum dependence of the expansion coefficients:

$$A = 0.542\,(1 + n/3)^{-0.685} \tag{174}$$

$$B = 0.097\,(1 + n/3)^{-0.224} \tag{175}$$

$$\alpha = 3.235\,(1 + n/3)^{-0.236} \tag{176}$$

$$\beta = 0.659\,(1 + n/3)^{-0.356} \tag{177}$$

$$V = 11.54\,(1 + n/3)^{-0.371}. \tag{178}$$

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

$$n_{\rm eff} \equiv \frac{d\ln P}{d\ln k} \tag{179}$$

Note that the cosmological model does not enter anywhere in these parameters. It is present in the fitting formula only through the growth factor g, which governs the amplitude of the virialized portion of the spectrum. This says that all the quasilinear features of the power spectrum are independent of the cosmological model, and only know about the overall level of power. This is not surprising to the extent that quasilinear evolution is well described by the Zeldovich approximation, in which the final positions of particles are obtained by extrapolating their initial displacements by some universal time-dependent factor.

5.3 Redshift-space effects

Although a huge amount of astronomical effort has been invested in galaxy redshift surveys, with the aim of mapping the three-dimensional galaxy distribution, the results are not a true 3D picture. The radial axis of redshift is modified by the Doppler effects of peculiar velocity: $1+z \rightarrow (1+z)(1+v/c)$. Since the peculiar velocities arise from the clustering itself, the apparent clustering pattern in **redshift space**, where redshift is assumed to arise from ideal Hubble expansion only, differs systematically from that in **real space**.

The distortions caused by working in redshift space are relatively simple to analyse if we assume we are dealing with a distant region of space which subtends a small angle, so that radial distortions can be considered as happening along one Cartesian axis. In this case, the apparent amplitude of any linear density disturbance is readily obtained from the usual linear relation between displacement and velocity (Kaiser 1987)

$$\boxed{\delta_{\rm obs} = \delta \left(1 + \frac{\Omega^{0.6}\mu^2}{b}\right),} \tag{180}$$

where μ is the cosine of the angle between the wavevector and the line of sight ($\mu = \hat{\mathbf{r}} \cdot \hat{\mathbf{k}}$). The parameter b allows for bias: the set of objects under study may be more clustered than the mass ($\delta = b\delta_{\rm mass}$). The function $f(\Omega) \simeq \Omega^{0.6}$ is the well-known velocity-suppression factor due to Peebles, which is in practice a function of Ω_m only, with negligible dependence on the vacuum density (Lahav *et al.* 1991). Redshift-space effects depend on the combination

$$\beta \equiv \Omega^{0.6}/b. \tag{181}$$

The anisotropy arises because mass flows from low-density regions onto high density sheets, and the apparent density contrast of the pattern is thus enhanced in redshift space if the sheets lie near the plane of the sky. If we average this anisotropic effect by integrating over a uniform distribution of μ, the net boost to the power spectrum is

$$|\delta_k|^2 \rightarrow b^2 |\delta_k|^2 \left(1 + \frac{2}{3}\beta + \frac{1}{5}\beta^2\right). \tag{182}$$

On small scales, this is not valid. The main effect here is to reduce power through the radial smearing due to virialized motions and the associated 'finger-of-God' effect. This is hard to treat exactly because of the small-scale velocity correlations. A simplified model was introduced by Peacock (1992) in which the small-scale velocity field is taken to be an incoherent Gaussian scatter with 1D rms dispersion σ. This turns out to be quite a reasonable approximation, because the observed pairwise velocity dispersion is a very slow function of separation, and is all the better if the redshift data are afflicted by significant measurement errors (which should be included in σ). This model is just a radial convolution, and so the k-space effect is

$$\delta_k \rightarrow \delta_k \exp[-k^2\mu^2\sigma^2/2]. \tag{183}$$

This effect in isolation gives an average isotropic factor of

$$|\delta_k|^2 \rightarrow |\delta_k|^2 \, \frac{\sqrt{\pi}}{2} \, \frac{\mathrm{erf}(k\sigma)}{k\sigma} \tag{184}$$

and produces only mild damping (one power of k at large k). This last feature is true whatever damping function is assumed, since at large k most surviving signal comes from $\mu \simeq 0$. An alternative model is to use the observation that the pairwise distribution of velocity differences is well-fit by an exponential (probably a superposition of Gaussians of different widths from different clumps). For this, the k-space damping function is a Lorentzian:

$$\delta_k \rightarrow \delta_k \, \left[1 + k^2\mu^2\sigma^2/2\right]^{-1} . \tag{185}$$

In either case, note that the damping depends on μ as does the Kaiser factor: both are anisotropic in k space and they interfere when averaging to get the mean power. What happens in this case is that the linear Kaiser boost to the power is lost at large k, where the result is the same as for $\beta = 0$ (because the main contribution at large k comes from small μ).

In practice, the relevant value of σ to choose is approximately $1/\sqrt{2}$ times the pairwise dispersion σ_p seen in galaxy redshift surveys (to this should be added in quadrature any errors in measured velocities). At $1h^{-1}$ Mpc separation, the pairwise dispersion is approximately

$$\sigma_p \simeq 300 - 400 \ \mathrm{kms}^{-1} \tag{186}$$

(Davis & Peebles 1983; Mo, Jing & Börner 1993; Fisher *et al.* 1994). We therefore expect wavenumbers $k \gtrsim 0.3\,h$ Mpc^{-1} to be seriously affected by redshift-space smearing.

Measuring the cosmological constant

There is an alternative potential source of clustering anisotropy in redshift space, which is geometrical in nature. We have to turn positions on the sky and redshifts into Cartesian coordinates using the following quantities

$$A(z) \equiv R_0 \frac{dr}{dz} = \frac{c}{H_0} \frac{1}{\sqrt{\Omega_v + \Omega_m(1+z)^3}} \tag{187}$$

and $B(z) \equiv R_0 S_k(r)$. We normally assume the the Einstein-de Sitter model

$$R_0 \frac{dr}{dz}(z) = \frac{c}{H_0} \frac{1}{(1+z)^{3/2}} \equiv A_0(z) \tag{188}$$

$$R_0 r(z) = 2\frac{c}{H_0}\left(1 - \frac{1}{\sqrt{1+z}}\right) \equiv B_0(z), \tag{189}$$

but if this is not correct, then contours of $\xi(r)$ will be squashed by a factor F:

$$F(z) = \frac{A/A_0}{B/B_0}. \tag{190}$$

If $\Lambda = 0$, this distortion is small, but not for significant vacuum energy: for $\Omega_m = 0.2$, $F \simeq 1.3$ for $z \gtrsim 1$. Detection of this distortion in *e.g.* quasar clustering would be an attractive means of detecting Λ. However, this effect interferes with the dynamical distortions: for a power-law spectrum, the distorted spectrum is

$$P^S(k,\mu) \propto k^n \left[1+\mu^2\left(\frac{1}{F^2}-1\right)\right]^{\frac{n-4}{2}} \times \left[1+\mu^2\left(\frac{\beta+1}{F^2}-1\right)\right]^2 D[k\mu\sigma_{\rm p}], \tag{191}$$

where $D[k\mu\sigma_{\rm p}]$ is the redshift smearing function (Ballinger, Peacock & Heavens 1996). To first order in μ^2, $\beta_{\rm eff} \simeq -0.5n(F-1)$. However, the μ^4 dependencies are different, so this method may still be attractive with large databases.

Real-space clustering

There are a number of methods available which avoid altogether the need to deal with the complications of redshift space. These deal with either pure two-dimensional clustering, as in angular correlations, or the use of projected correlations in redshift surveys.

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

$$\delta(\hat{\mathbf{q}}) = \int_0^\infty \delta(y)\, y^2 \phi(y)\, dy, \tag{192}$$

where $\phi(y)$ is the **selection function**, normalized such that $\int y^2\phi\, dy = 1$, and y is comoving distance. The form $\phi \propto y^{-1/2} \exp -(y/y^*)^2$ is often taken as a reasonable approximation to the Schechter function. A flat Universe ($\Omega = 1$) is assumed. 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_ℓ

$$e^{ikr\cos\theta} = \sum_0^\infty (2\ell+1)\, i^\ell\, P\ell(\cos\theta)\, j_\ell(kr), \tag{193}$$

plus the spherical harmonic addition theorem

$$P_\ell(\cos\theta) = \frac{4\pi}{2\ell+1} \sum_{m=-\ell}^{m=+\ell} Y^*_{\ell m}(\hat{\mathbf{q}}) Y_{\ell m}(\hat{\mathbf{q}}'), \tag{194}$$

where $\hat{\mathbf{q}} \cdot \hat{\mathbf{q}}' = \cos\theta$. These relations yield the desired result:

$$\boxed{\left\langle |a_\ell^m|^2 \right\rangle = 4\pi \int \Delta^2(k)\, \frac{dk}{k} \left[\int y^2 \phi(y)\, j_\ell(ky)\, dy\right]^2 .} \tag{195}$$

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

$$\langle\delta(\hat{\mathbf{q}}_1)\delta(\hat{\mathbf{q}}_2)\rangle = \iint \langle\delta(\mathbf{y_1})\delta(\mathbf{y_2})\rangle\ y_1^2 y_2^2 \phi(y_1)\phi(y_2)\ dy_1\ dy_2. \tag{196}$$

We now change variables to the mean and difference of the radii, $y \equiv (y_1 + y_2)/2$; $x \equiv (y1 - y2)/2$. If the depth of the survey is larger than any correlation length, we only get 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**:

$$\boxed{w(\theta) = \int_0^\infty y^4\phi^2\ dy\ \int_{-\infty}^\infty \xi(\sqrt{x^2 + y^2\theta^2})\ dx.} \tag{197}$$

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:

$$w(\theta) = \int_0^\infty y^4\phi^2\ dy\ \int_0^\infty \pi\ \Delta^2(k)\ J_0(ky\theta)\ dk/k^2. \tag{198}$$

The power-spectrum version of Limber's equation is already in the form required for 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:

$$\boxed{\Delta_\theta^2 = \frac{\pi}{4}\int \Delta^2(K/y)\ y^5\phi^2(y)\ dy.} \tag{199}$$

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 easy to make allowance for spatial curvature in the above discussion. All that is needed is to replace the $\Omega = 1$ volume element $y^2\,dy$ by its generalised counterpart, $y^2 dy/(1 - ky^2)^{1/2}$.

When working with redshift surveys and treating redshift as a radial coordinate, the presence of peculiar velocities or redshift errors causes the correlation function to be convolved in the radial direction:

$$\begin{aligned}\xi(r_p,\pi) &= \int_{-\infty}^\infty \xi_{\rm true}(r_p, r)\ f(\pi - r)\ dr \\ &= \frac{r_0^\gamma}{\sqrt{2\pi}\,\sigma_v}\int_{-\infty}^\infty [r_p^2 + (\pi - x)^2]^{-\gamma/2}\ e^{-x^2/2\sigma_v^2}\ dx,\end{aligned} \tag{200}$$

where the latter expression applies for power-law clustering and a Gaussian dispersion. Treating the convolving function $f(\Delta r)$ as a zero-mean scatter ignores large-scale streaming, which becomes more important on larger scales. However, a Fourier analysis is clearer in this regime. Looking at the elongation of $\xi(r_p, \pi)$ in the redshift

direction allows the pairwise velocity dispersion to be estimated; it comes out at a *relative* dispersion of between 300 & 400 kms^{-1} for pairs of ~ 1 Mpc separation (Davis & Peebles 1983; Fisher *et al.* 1994).

The effects of peculiar velocities may be evaded by using the correlation function evaluated explicitly as a 2D function of transverse (r_p) and radial (π) separation. The projection along the redshift axis is then independent of the velocities

$$w(r_p) = \int_{-\infty}^{\infty} \xi(r_p, \pi)\, d\pi = 2 \int_{r_p}^{\infty} \xi(r)\, \frac{r\, dr}{(r^2 - r_p^2)^{1/2}}, \tag{201}$$

and has the Abel integral inverse

$$\xi(r) = -\frac{1}{\pi} \int_r^{\infty} w'(y)\, \frac{dy}{(y^2 - r^2)^{1/2}}. \tag{202}$$

Improved signal-to-noise in projected correlations can be obtained in the case of a **sparse-sampled redshift survey** (Kaiser 1986a), where there is a large catalogue of angular positions from which redshifts are measured for some fraction (either the brighter members, or a random subset). Saunders *et al.* (1992) used the angular cross-correlation between the 1-in-6 QDOT IRAS galaxy redshift survey and its parent catalogue to obtain the statistic

$$\Xi(r) = 2 \int_0^{\infty} \xi[(r^2 + x^2)^{1/2}]\, dx = 2 \int_r^{\infty} \xi(y)\, \frac{y\, dy}{\sqrt{y^2 - r^2}}. \tag{203}$$

At first sight, it is not very attractive to use this to infer the power spectrum, because the window function involved is extremely broad:

$$\frac{1}{r}\Xi(r) = \int \Delta^2(k)\, \frac{dk}{k}\, \left[\frac{\pi}{kr}\, J_0(kr)\right]. \tag{204}$$

However, useful results were obtained by Saunders *et al.* (1992) from a multi-stage process where $\Xi(r)$ is first deprojected to obtain $\xi(r)$, which can then be integrated to yield $\bar{\xi}(r)$. This has a compact window function, and can be directly related to the power spectrum.

5.4 Bias

One of the major advances of cosmology in the 1980s was the realization that the distribution of galaxies need not trace the underlying density field. The main motivation for such a view may be traced to 1933 and Zwicky's measurement of the dark matter in the Coma cluster. A series of ever more detailed studies of cluster masses have confirmed his original numbers: if the Coma mass-to-light ratio is Universal, then the density parameter of the Universe is $\Omega = 0.1 - 0.2$. Those who argued that the value $\Omega = 1$ was more natural (a greatly increased camp after the advent of inflation) were therefore forced to postulate that the efficiency of galaxy formation was enhanced in dense environments: **biased galaxy formation**. This probably remains the strongest argument for the reality of bias.

A weaker argument surfaced at around the same time as inflation through the discovery of large voids in the galaxy distribution. There was a reluctance to believe

that such vast regions could be truly devoid of matter – although this was at a time before the discovery of large-scale velocity fields. This tendency was given further stimulus through the work of Davis, Efstathiou, Frenk & White (1985), who were the first to calculate N-body models with the CDM spectrum. Since the CDM spectrum curves slowly between effective indices of $n = -3$ and $n = 1$, the correlation function clearly steepens with time. There is therefore a unique epoch when ξ will have the observed slope of -1.8. Davis *et al.* identified this epoch as the present and then noted that, for $\Omega = 1$, it implied a rather low *amplitude* of fluctuations: $r_0 = 1.3h^{-2}$ Mpc. An independent argument for this low amplitude came from the size of the peculiar velocities in CDM models: if given an amplitude corresponding to the $\sigma_8 \simeq 1$ seen in the galaxy distribution, the pairwise dispersion was $\sigma_p \simeq 1000 - 1500\,\mathrm{km\,s^{-1}}$, some $3 - 4$ times the observed value. What seemed to be required was a galaxy correlation function which was an amplified version of that for mass. This was exactly the phenomenon analysed for Abell clusters by Kaiser (1984), and thus was born the idea of **high-peak bias**: bright galaxies form only at the sites of high peaks in the initial density field. This was developed in some analytical detail by Bardeen *et al.* (1986), and was implemented in the simulations of Davis *et al.*, leading to the conclusion that the $\Omega = 1$ $h = 1/2$ CDM model now gave a good match to observation.

Since the mid-1980s, fashion has moved in the direction of low-Ω universes, which removes many of the original arguments for bias. However, the lesson of the attempts to save the $\Omega = 1$ universe is that cosmologists have learned to be wary of assuming that light traces mass. The assumption is that the galaxy density field is guilty of bias, until it is shown to be innocent.

As was shown by Kaiser (1984), the high-peak model produces a linear amplification of large-wavelength modes. This is likely to be a general feature of other models for bias, so it is useful to introduce the **linear bias parameter**:

$$\left.\frac{\delta\rho}{\rho}\right|_{\mathrm{galaxies}} = b \left.\frac{\delta\rho}{\rho}\right|_{\mathrm{mass}} . \tag{205}$$

This seems a reasonable assumption when $\delta\rho/\rho \ll 1$. Galaxy clustering on large scales therefore allows us to determine mass fluctuations only if we know the value of b. For example, the normalization in scale-invariant models may be specified by ϵ (the rms potential fluctuation per $\ln k$), but we can only measure the combination ϵb. However, the coinage of the bias parameter has been debased by its use in the mildly non-linear regime, where misleading definitions such as $b = 1/\sigma_8$ are to be found.

Even the linear relation cannot be taken for granted, however; if galaxy formation is not an understood process, then in principle studies of galaxy clustering may tell us nothing useful about the statistics of the underlying potential fluctuations against which we would like to test inflationary theories. It is possible to construct models (*e.g.* Bower *et al.* 1993) in which the large-scale modulation of the galaxy density is entirely non-gravitational in nature.

Mechanisms for bias

Why should the galaxy distribution be biased at all? In the context of the high-peak model, attempts were made to argue that the first generation of objects could propagate disruptive signals, causing neighbours in low-density regions to be 'still-born'. However, it turned out to be hard to make such mechanisms operate: the energetics and required scale of the phenomenon are very large (Rees 1985; Dekel & Rees 1987). A more promising idea, known as **natural bias**, was introduced by White *et al.* (1987). This relied on the idea that, in an overdense region, an object of a given mass will collapse sooner and thus have a higher density and circular velocity. Application of a circular-velocity threshold then yields a bias towards high-density regions. White *et al.* argued that such an effect was to be expected owing to the **Tully-Fisher effect**: a tight correlation between luminosity and circular velocity for spiral galaxies.

However, the problem with this model is that it still apparently predicts no bias if a strict selection by mass is performed. What is needed is some way in which star formation is biased (perhaps by epoch dependent efficiency) in order to produce more stars in the galaxies which collapse earlier. A general discussion of this problem was given by Cole & Kaiser (1989): suppose an object collapsing at redshift z generates a stellar luminosity

$$L \propto M^{\alpha}(1+z)^{\beta}. \tag{206}$$

Cole & Kaiser show that a perfect Tully-Fisher relation then requires $\beta = 3\alpha/2$. This is easily proved from the usual expression for the virial velocity dispersion resulting from gravitational collapse: $V \propto M^{1/3}(1+z_c)^{1/2}$. The above condition removes any redshift dependence and leaves $V \propto L^{1/3\alpha}$. The conventional Tully-Fisher slope of 1/4 then implies $\alpha = 4/3$, $\beta = 2$. The natural bias mechanism implicitly depends on a strong epoch dependence of star-forming efficiency. In fact, Cole & Kaiser argue that a somewhat stronger epoch dependence ($\beta \gtrsim 3$) is required to achieve sufficient bias to understand cluster mass-to-light ratios. However, Peacock (1990) showed that such a high value would predict a large scatter in the Faber-Jackson relation between luminosity and velocity dispersion for ellipticals. The data are much more closely consistent with $\beta \simeq 1$.

In short, there appears to be little evidence for traditional bias schemes where one tinkers with the efficiency of star formation. There remains the alternative that galaxies were born unbiased but subsequently migrated into clusters more quickly than dark matter. These dynamical schemes are currently attracting the most attention (West & Richstone 1988; Carlberg *et al.* 1990; Carlberg 1991; Couchman & Carlberg 1992). The idea of bias being a phenomenon largely confined to clusters fits well with the emerging picture of large-scale structure: away from clusters, it seems that all types of galaxies follow the same overall 'skeleton' of large-scale structure, independent of Hubble type (Thuan *et al.* 1987; Babul & Postman 1990; Mo, McGaugh & Bothun 1994). Even luminosity segregation is a very weak and controversial effect (Valls-Gabaud *et al.* 1989; Loveday *et al.* 1995). This is further evidence against earlier pictures in which the voids were filled with mass, but failed to produce bright galaxies. Increasingly, it seems that the voids really have been largely emptied by gravity, as implied by the large-amplitude peculiar velocities on these scales. It would not be surprising if the formation of all classes of galaxy was

suppressed in these regions of very low density, but it is implausible that the voids contain more than a small fraction ($\sim 10\%$) of the total mass in the universe, so the implied degree of bias would not be large.

Outstanding issues

To sum up the present position, the pictures of the galaxy distribution obtained with different sets of tracer objects have certain features in common (mainly the behaviour in low-density regions: the existence of filaments and voids), but diverge greatly in regions of high density. This **morphological segregation** is a long-established phenomenon (*e.g.* Dressler 1980); it is known that E/S0 galaxies increase from perhaps 20% of the galaxy population in the field to almost 100% in the cores of rich clusters where the overdensities are $\sim 10^3$. This is closely related to the relative proportions of optically-selected and IRAS galaxies as a function of density. At high overdensities, the fraction of optical galaxies which are IRAS galaxies declines by a factor $\simeq 3 - 5$ from the mean (Strauss *et al.* 1992), reflecting the fact that IRAS galaxies are mainly spirals.

We need to be able to decide on physical grounds which tracer is more likely to follow the mass in regions of high density, and in some cases this is quite easy. For example, it is well understood how and why the spatial correlations of rich clusters will be an amplified version of the underlying density correlations (at least for Gaussian statistics). Given this, it is a safe bet that the clustering properties of elliptical galaxies also over-estimate the density correlations; the known phenomenon of morphological segregation means that ellipticals are closely associated with clusters. We know in both the cases of clusters and ellipticals that we are excluding the low-density universe simply through our observational selection, so an unrepresentative answer would be expected.

Much harder is the critical decision between optically-selected galaxies or IRAS galaxies. The former give dynamical determinations consistently in the region of $\Omega = 0.2$ if optical light traces mass, whereas the latter favour $\Omega = 1$, because IRAS galaxies peak up less strongly than optically-selected galaxies in high-density regions, as discussed above. So, which (if either) of these fields follows the mass? The problem for the committed believer in $\Omega = 1$ is that it is much more plausible that it is the IRAS result which is corrupted. Morphological segregation is thought to arise because spiral discs have trouble surviving in high-density environments, and it is plausible that the X-ray emitting gas in clusters is all that remains of material which might have made spiral discs in a less troubled habitat. We conclude that either $\Omega \simeq 0.2$, or that some other mechanism has operated to boost the optical light in clusters, which is then fortuitously cancelled by the suppression of IRAS emission from these regions. This interpretation is made still more contrived by an accounting of the total baryonic material in clusters. Within the central $\simeq 1$ Mpc, the masses in stars, X-ray emitting gas and total dark matter can be determined with reasonable accuracy (perhaps 20% rms), and allow a minimum baryon fraction to be determined:

$$\frac{M_{\rm Baryons}}{M_{\rm Total}} \gtrsim 0.009 + 0.050\, h^{-3/2} \qquad (207)$$

(White *et al.* 1993). This equation is often referred to as the **baryon catastrophe**, for the following reasons. Assume for now that the baryon fraction in clusters is representative of the whole universe, and adopt the primordial nucleosynthesis prediction of $\Omega_{\rm B}h^2 = 0.0125$. This gives an equation for Ω:

$$\Omega \lesssim 0.25\,[h^{1/2} + 0.18h^2]^{-1}, \tag{208}$$

which is a limit varying between 0.21 and 0.33 for h between 0.5 and 1. This is a catastrophe for the Einstein-de Sitter universe, in that clusters have to be biased not only in the light they emit, but also in the sense of containing a larger baryon fraction than the average. However, producing a large-scale separation of dark matter and baryons on scales which are little past the turn-round phase is very difficult. If the density parameter is really unity, it appears that the nucleosynthesis density must be too low by at least a factor 3.

Since there continue to be compelling reasons to expect $\Omega = 1$, a high priority in current cosmological research will continue to be to produce a convincing mechanism for bias, and to detect its traces in galaxy properties. However, this should not blind us to the fact that the simplest interpretation of the existing evidence is that the Universe has $\Omega < 1$..

5.5 Power-spectrum data

The history of attempts to quantify galaxy clustering goes back to Hubble's proof 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 which greatly exceed these in depth, but several pieces of work appeared around the start of the 1990s which clarified many of the discrepancies between different surveys, and which paint a relatively consistent picture of large-scale structure (see Peacock & Dodds 1994).

Clustering results are often published in the form of the variance (σ^2) of δ as a function of scale – using either cubical cells of side ℓ (Efstathiou *et al.* 1990b) and 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

$$\sigma^2 = \Delta^2 \left(k = \left[\tfrac{1}{2} \left(\tfrac{n+1}{2} \right)! \right]^{1/(n+3)} / R_{\rm G} \right). \tag{209}$$

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 $\lambda \simeq 2\ell$. Since we know the shape of Δ^2 reasonably well, we can get very accurate effective wavenumbers and plot the σ^2 values on the $\Delta^2 - k$ plane directly using these $k_{\rm eff}$ values. Such a compilation of results is shown in Figure 4.

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? The 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 reconstruction analysis has available eight datasets containing 91

Fig. 4. (a) The 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 lines shown for reference are the transforms of the canonical real-space correlation functions for optical and IRAS galaxies ($r_0 = 5$ and 3.78 h^{-1} Mpc and slopes of 1.8 and 1.57 respectively). (b) The power-spectrum data, individually linearized assuming $\Omega = b_{\rm I} = 1$. There is an excellent degree of agreement, particularly in the detection of a break around $k = 0.03h$.

distinct $k - \Delta^2$ pairs. The modelling has available five free parameters in the form of 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 a 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 ratios

$$b_{\rm A} : b_{\rm R} : b_{\rm O} : b_{\rm I} = 4.5 : 1.9 : 1.3 : 1, \tag{210}$$

to within 6 per cent rms.

The various reconstructions of the linear power spectrum for the case $\Omega = b_{\rm I} = 1$ are shown superimposed in Figure 4, and display an impressive degree of agreement. This argues very strongly that what we measure with galaxy clustering has a direct relation to mass fluctuations, rather than the large-scale clustering pattern being an optical illusion caused by non-uniform galaxy-formation efficiency (Bower *et al.* 1993). If this were the case, the shape of spectrum inferred from clusters should have a very different shape at large scales, contrary to observation.

Large-scale power-spectrum data and models

It is interesting to ask if 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 $\Delta^2(k) \propto k^{n+3}T_k^2$. We shall use the BBKS approximation for the transfer function:

Fig. 5. The linearized power-spectrum data of Figure 4 averaged over bins of width 0.1 in $\log_{10} k$, compared to various CDM models. These assume scale-invariant initial conditions, with the same large-wavelength normalization. Different values of the fitting parameter $\Omega h = 0.5,\ 0.45,\ \ldots 0.25,\ 0.2$ are shown. The best-fit model has $\Omega h = 0.25$ and a normalization of $\sigma_8(\mathrm{IRAS}) = 0.75$.

$$T_k = \frac{\ln(1+2.34q)}{2.34q} \times [1 + 3.89q + (16.1q)^2 + (5.46q)^3 + (6.71q)^4]^{-1/4}, \tag{211}$$

where $q \equiv k/[\Omega h^2\ \mathrm{Mpc}^{-1}]$. Since observable wavenumbers are in units of $h\,\mathrm{Mpc}^{-1}$, the shape parameter is the apparent value of Ωh. This scaling applies for models with zero baryon content, but there is an empirical scaling (Sugiyama 1995) that can account for this:

$$T_k(k) = T_{\mathrm{BBKS}}(k/[\Omega h^2\ \exp -(\Omega_{\mathrm{B}} + \Omega_{\mathrm{B}}/\Omega)]). \tag{212}$$

The symbol Γ^* is used to refer to Ωh in the BBKS spectrum as an empirical fitting parameter, on the understanding that it would mean the above combination if CDM models were taken literally. Fitting this spectrum to the large-scale linearised data of Figure 5 requires the parameters

$$\Gamma^* \simeq 0.25 + 0.3(1/n - 1), \tag{213}$$

$$\sigma_8(\mathrm{IRAS}) \simeq 0.75, \tag{214}$$

in agreement with many previous arguments suggesting that an apparently low-density model is needed. 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.

The fit of this model is illustrated in Figure 5, which makes it clear that the problem with CDM is the *shape* of the power spectrum, rather than the absolute amount of power at large scales. The linear transfer function does not bend sharply enough at the break wavenumber if a high-density $\Gamma^* = 0.5$ model is adopted. 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 (*e.g.* Section 25 of Peebles 1993). It is hard to see how such features can be reconciled with the data, beyond a 'visibility' of perhaps 20 – 30%.

Small-scale clustering data

It should clearly be possible to reach stronger conclusions by using the data at larger k. However, here the assumption of linear bias is clearly very weak, and what is needed is a model for the scale dependence of the bias. Mann, Peacock & Heavens (1996) argue that an empirical approach can be taken here, even lacking the physics of bias. We know that, if $\Omega = 1$, the density of light in high-density regions must receive an additional enhancement. There is also the possibility that galaxy formation may be suppressed in voids. General arguments (Coles 1993) indicate that such **local bias** should produce a systematic steepening of the correlation function, so that the effective bias is larger on small scales. Experiments with such local density-field modifications on numerical datasets suggests that the effect on the power spectrum is of a steepening which can be roughly approximated by

$$1 + \Delta^2(k) \rightarrow [1 + b_1\Delta^2]^{b_2}, \tag{215}$$

where $[b_1 b_2]^{1/2}$ would be the bias parameter in the linear regime. Such an expression can fit bias schemes from high-peak bias as in Davis *et al.* (1985) or the 'physical bias' seen in full hydrodynamical simulations (*e.g.* Cen & Ostriker 1992). Generally the steepening is not so extreme, and it is hard to find models where b_2 exceeds b_1.

Irrespective of *a priori* considerations, such an expression accounts well for the difference in real-space clustering of APM and IRAS galaxies, and can map one onto the other with almost uncanny precision (see Figure 6), with a relative bias of $b_1 = 1.2$, $b_2 = 1.1$ or $b_{\rm APM}/b_{\rm IRAS} \simeq 1.15$. Of particular interest is the inflection in the spectrum around $k \simeq 0.1\,h\,$Mpc^{-1}, which seems likely to be real, since it is seen in two rather different datasets, which probe different regions of space.

Attempting to fit the small-scale clustering data now complicates the picture from large scales, since the nonlinear extrapolation of the $\Gamma^* = 0.25$ model is not consistent with the small-scale clustering. Figure 7 compares three different models with the data: Einstein-de Sitter and open and flat $\Omega = 0.3$ models. There is a tension between the data and all of these models, in that it seems impossible to fit both large scales and small scales simultaneously. Without bias, the correct amplitude of small-scale clustering requires $\sigma_8 \simeq 0.7$; this greatly under-predicts the clustering at $k \simeq 0.1\,h\,$Mpc^{-1} unless $\Gamma^* \lesssim 0.1$ is adopted, in which case the power at $k \simeq 0.02\,h\,$Mpc^{-1} is greatly exceeded. The $\Omega = 1$ model needs a lower normalization, and so does not exceed the small-scale data, but it suffers from related

problems of shape. Any hypothetical bias which would scale the mass spectrum to that of light would need to be non-monotonic, with a smaller effect at $k \simeq 2\,h\,\mathrm{Mpc}^{-1}$ than on larger scales.

What then is required of a linear power spectrum that would fit the data? None of the CDM-like alternatives considered above explain the inflection at $k \simeq 0.1\,h\,\mathrm{Mpc}^{-1}$, and it is unlikely to be produced by any bias scheme, since these always tend to give a smooth scale dependence. The general conclusion is therefore that there must be a relatively sharp break in the linear spectrum around this point.

A further general point which emerges from this plot is that it is something of a puzzle that the clustering data continue as an unbroken $n \simeq -1$ power law up to $\Delta^2 \sim 10^3$. As the regime of virialized clustering is reached, a break to a flatter slope would be expected; only the open models fail to show this feature, which is a robust prediction for any smooth linear spectrum. What is needed is a small linear growth-suppression factor $g(\Omega)$, with $\Omega_m \lesssim 0.5$ for open models or $\lesssim 0.1$ for flat models This is far from being a conclusive argument for open models, but it does require a coincidence from the bias mechanism: that the galaxy correlations should be steepened just where the mass correlations are saturating.

We can implement these ideas with a simple empirical model, which works extremely well. Consider a spectrum in the form of a break between two power laws:

$$\Delta^2(k) = \frac{(k/k_0)^\alpha}{1 + (k/k_1)^{\alpha-\beta}}. \tag{216}$$

As shown in Figure 8, this matches the data very nicely, if we choose the parameters

$$\begin{aligned} k_0 &= 0.3\,h\,\mathrm{Mpc}^{-1} \\ k_1 &= 0.05\,h\,\mathrm{Mpc}^{-1} \\ \alpha &= 0.8 \\ \beta &= 4.0. \end{aligned} \tag{217}$$

A value of $\beta = 4$ corresponds to a scale-invariant spectrum at large wavelengths, whereas the effective small-scale index is $n = -2.2$. The linear spectrum is not required to be non-zero for $k \gtrsim 1\,h\,\mathrm{Mpc}^{-1}$, and so a variety of other possibilities would be made to work, including those with short-wavelength cutoffs.

CMB anisotropies

A consistent model must match the normalization of the mass fluctuations on large scales inferred from fluctuations in the Cosmic Microwave Background. In making this comparison, it is important to be clear that the CMB fluctuations depend only on the very large-scale $P \propto k^n$ portion of the spectrum. Predictions of smaller-scale fluctuations such as the amplitude σ_8 then require additional information in the form of the shape parameter Γ^*. Rather than quoting the σ_8 implied by the CMB, it is therefore clearer to give the large-scale normalization separately, with σ_8 then depending on the choice of Γ^*.

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 slightly lower results, and we scale to these in what follows. For scale-invariant spectra and $\Omega = 1$, the best normalization is

Fig. 6. (a) The real-space power spectra of APM and IRAS galaxies, as deduced by Baugh & Efstathiou (1993; 1994) and Saunders *et al.* 1992. The APM data have been boosted by a factor 1.2 because clustering evolution was not allowed for in the Baugh & Efstathiou data. (b) The same with a two-parameter scale-dependent boost to the IRAS data. The agreement is outstanding, except for the three largest-scale IRAS points, and these can be seen to be too high from a comparison with the redshift-space results.

Fig. 7. The clustering data for optical galaxies, compared to three models with $\Gamma^* = 0.25$: (a) $\Omega = 1$, $\sigma_8 = 0.5$; (b) $\Omega_m = 0.3$, $\Omega_v = 0$ $\sigma_8 = 1$. (c) $\Omega_m = 0.3$, $\Omega_v = 0.7$ $\sigma_8 = 1$. Linear spectra are shown dotted; evolved nonlinear spectra are solid lines. All of these models are chosen with a normalization which is approximately correct for the rich-cluster abundance and large-scale peculiar velocities. In all cases, the shape of the spectrum is wrong. The high-density model would require a bias which is not a monotonic function of scale, whereas the low-density models exceed the observed small-scale clustering.

Fig. 8. An empirical double power-law model for the power spectrum provides an extemely good fit to the optical-galaxy power spectrum but requires an open universe and a sharp break in the spectrum to a rather flat ($n < -2$) high-k behaviour.

$$\Delta^2(k) = (k/0.0737\,h\,\mathrm{Mpc}^{-1})^4, \tag{218}$$

equivalent to $Q_{\rm rms} = 18.0\,\mu$K, or $\epsilon = 3.97 \times 10^{-5}$ in the notation of Peacock (1991), with an rms error in density fluctuation of 8%.

For low-density models, a naive analysis as in PD suggests that the power spectrum should depend on Ω and the growth factor g as $P \propto g^2/\Omega^2$. Because of time dependence of gravitational potential (integrated Sachs-Wolfe effect) and spatial curvature, this expression is not exact, although it captures the main effect. From the data of White & Bunn (1995), a better approximation is

$$\Delta^2(k) \propto \frac{g^2}{\Omega^2}\, g^{0.7}. \tag{219}$$

This applies for low-Ω models both with and without vacuum energy, with a maximum error of 2% in density fluctuation provided $\Omega > 0.2$ (and gives the same σ_8 values as Górski *et al.* (1995), when the appropriate Γ^* corrections are made, to within 3%). Since the rough power-law dependence of g is $g \simeq \Omega^{0.65}$ 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 $\Omega^{-0.69}$ for these two cases. The dependence is very weak for open models, but vacuum energy implies very much larger fluctuations. These results are illustrated for CDM spectra in Figure 9, which shows σ_8 as a function of Γ^* for three models. For $\Gamma^* = 0.25$, open low-density models are close to the required $\sigma_8 = 1$, whereas flat models have an amplitude perhaps a factor 2 too high. Einstein-de Sitter models have $\sigma_8 = 0.65$, which is only slightly high.

What if we tilt the spectrum? For tilt, one evaluates σ_8 as in the above no-tilt case, and then scales as

Fig. 9. The clustering normalization σ_8 as a function of Γ^*, predicted from COBE assuming scale-invariant primordial fluctuations. Note that flat low-density models require a much larger normalization than do open models.

$$\sigma_8 \propto \exp[2.3(n-1)]; \tag{220}$$

since σ_8 measures the power spectrum at $k \simeq 0.2\,h\,\mathrm{Mpc}^{-1}$, this corresponds to saying that COBE determines the spatial power spectrum at an effective wavenumber of about $0.002\,h\,\mathrm{Mpc}^{-1}$. If gravity waves are included with the usual inflationary coupling between wave amplitude and tilt, the effect increases to

$$\sigma_8 \propto \exp[4.3(n-1)]. \tag{221}$$

The flat models can then have σ_8 reduced by the required factor of 2, but a substantial degree of tilt is needed ($n \simeq 0.70$ or 0.84, the latter figure including gravity waves). These are significantly larger degrees of tilt than would be expected from at least the simplest inflationary models.

In any case, the CDM model is, as we have seen, in some difficulty as a general description of the spectrum. A more robust datum is probably the power on the largest reliable scales:

$$\Delta^2_{\mathrm{opt}}(k = 0.02\,h\,\mathrm{Mpc}^{-1}) \simeq 0.005, \tag{222}$$

which is to be compared to a COBE scale-invariant prediction of 0.0054. Scaling as $\Omega^{-0.24}$ or $\Omega^{-1.38}$ boosts this by a factor of 1.5 (open $\Omega = 0.2$) or 9.2 (flat $\Omega = 0.2$). The former factor is within the plausible effect of a transfer function, but the latter is not. The required tilt to remove the additional factor 6 in power is gross: $n = 0.2$, pivoting the spectrum about $k = 0.002\,h\,\mathrm{Mpc}^{-1}$. Allowing for gravity waves improves this to $n = 0.7$, but the conclusion remains that low-density flat models require an extremely large degree of tilt in order to be viable.

What does it all mean?

What then is the interpretation of the spectrum? A CDM spectrum with $\Gamma^* \simeq 0.25$ is not consistent with $\Omega = 1$ and any plausible estimate for h. However, even the low-Γ^* spectra are probably of the wrong shape, so it is not clear if one can argue from the best-fitting Γ^* for $\Omega < 1$.

Interesting alternatives with high density are either mixed dark matter (Holtzman 1989; Klypin *et al.* 1993), or non-Gaussian pictures such as cosmic strings + HDM, where the lack of a detailed prediction for the power spectrum helps ensure that the model is not yet excluded (Albrecht & Stebbins 1992). Isocurvature CDM is also attractive in that it gives a rather sharper bend at the break scale, as the data seem to require. However, this model conflicts strongly with limits on CMB anisotropies, and cannot be correct. (Efstathiou & Bond 1986). Mixed dark matter seems rather ad hoc, but may be less so if it is possible to produce both hot and cold components from a single particle, with a Bose condensate playing the role of the cold component (Madsen 1992; Kaiser, Malaney & Starkman 1993). The main problems with an MDM model are ones generic to any model with a very flat high-k spectrum in a high-Ω universe: difficulty in forming high-redshift objects and difficulty in achieving a steep correlation function on small scales.

Alternatively, if the good fit of a low-density CDM transfer function is taken literally, then perhaps this is a hint that the epoch of matter-radiation equality needs to be delayed. An approximate doubling of the number of relativistic degrees of freedom would suffice – but this would do undesirable violence to primordial nucleosynthesis: any such boost would have to be provided by a particle which decays after nucleosynthesis. The apparent value of Ωh depends on the mass and lifetime of the particle roughly as

$$\Omega h|_{\rm apparent} = \Omega h \, [1 + (m^2_{\rm keV} \tau_{\rm years})^{2/3}]^{-1/2} \qquad (223)$$

(Bardeen, Bond & Efstathiou 1987; Bond & Efstathiou 1991), so a range of masses is possible. Apart from making the observed large-scale structure, such a model yields a small-scale enhancement of power which could lead to early galaxy formation (McNally & Peacock 1995).

Of course, the simplest alternative is to admit that the above attempts to save the Einstein-de Sitter model are too contrived, and that Ω_m is < 1. This would make a low-Γ^* model easier to understand, but it introduces few new possible ways to alter the shape of the linear power spectrum. There are of course the oscillatory features expected in the power spectra of baryon-dominated universes (*e.g.* Section 25 of Peebles 1993), but these occur at too small k. A possibility would be warm dark matter with a cutoff at $k \simeq 0.2\, h\,{\rm Mpc}^{-1}$, but it may well be that something entirely new is needed.

Apart from lessening the large-scale structure problems, low densities make life easier in two other ways: the universe is made older for a given h and strong bias need not be accounted for. The main difficulty which will need to be overcome in order for the reality of a low-density universe to be accepted is to understand the values for Ω which have emerged from attempts to us large-scale peculiar velocities to 'weigh' the universe (*e.g.* Dekel 1994. The amplitude of such motions should not be a problem, as they should scale with $\delta v \propto \Omega^{0.6}\sigma_8$, and the normalization inferred from nonlinear systems is $\sigma_8 \propto \Omega^{-0.5}$. However, there is a problem in the degree

of bias inferred from comparing such velocities with the density field. In principle, this allows one to determine $\beta \equiv \Omega^{0.6}/b$, and values of $\beta \simeq 1$ have been inferred for IRAS galaxies. If $\Omega \simeq 0.2$, this would require $b \simeq 0.5$. However, Loveday *et al.* (1996) have found the much lower value $\beta \simeq 0.5$ from redshift-space distortions in the Stromlo-APM redshift survey. If these lower values are confirmed, the strength of the case for low Ω and little bias would become overwhelming.

This leaves unresolved the distinction between an open model and one in which a significant vacuum energy keeps the inflationary $k = 0$ prediction (Efstathiou *et al.* 1990b). The general case $\Omega_v \neq 0$ and $k \neq 0$ is also a logical possibility, but not an appealing one. As constraints on h, galactic ages, and details of the CMB measurements improve, there is the happy possibility that a clear discrimination between these alternatives may soon be reached.

5.6 Gaussianity

Skewness

How are we to distinguish observationally whether the density field of the Universe is Gaussian? We need to look at more subtle statistics than just the power spectrum. In principle, one might use the higher-order n-point correlation functions, since these are directly related to the power spectrum for a Gaussian field. On small scales, nonlinear evolution must produce non-Gaussian behaviour. For any σ, a Gaussian density distribution will produce a tail to unphysical $\delta < -1$ values. The lognormal model $\delta \rightarrow \exp[\delta - \sigma^2/2]$ is the simplest analytical modification which cures this problem (Coles & Jones 1991). This distribution is skew, with a tail towards large values of δ. The question is whether the density field contains a greater degree of non-Gaussianity than that induced by gravitational evolution.

Non-Gaussian behaviour may be measured through the skewness parameter (not the skewness itself):

$$S \equiv \frac{\langle \delta^3 \rangle}{[\langle \delta^2 \rangle]^2}, \tag{224}$$

which can be calculated through second-order gravitational perturbation theory, and should be a constant of order unity. Gaztañaga (1992) showed that the skewness parameter for the APM galaxy survey was approximately constant with scale, at the value expected for nonlinear gravitational evolution of a Gaussian field. Does this mean (a) that conditions are Gaussian; (b) that $b = 1$? This is possible, but the effects of bias need to be understood first. As a simple example, consider a power-law modification of a lognormal field: $\rho' \propto \rho^b$. Since this is generated by a Gaussian field, it is easy to find the skewness parameter, which is

$$S = 2 + e^{b^2\sigma^2}. \tag{225}$$

In the linear regime, the skewness is independent of b and so this sort of model would not violate the observation that the moments $\langle \delta^3 \rangle$ and $\langle \delta^2 \rangle$ are in the correct ratio for straightforward gravitational evolution without bias.

Topology

An interesting alternative probe of Gaussianity was suggested by Gott, Melott & Dickinson (1986): the topology of the density field. To visualise the main principles, it will help to think initially about a 2D field. Two extreme non-Gaussian fields would consist either of discrete 'hotspots' surrounded by uniform density, or the opposite: discrete 'coldspots'; the picture in either case is a set of polka dots. Both of these cases are clearly non-Gaussian just by symmetry: the contours of average density will be simply-connected circles containing regions which are all either above or below the mean density, but in a Gaussian field (or any symmetric case), the numbers of hotspots and coldspots must balance.

One might think that things would be much the same in 3D: the obvious alternatives are 'meatball' or 'Swiss cheese' models. However, there is a third topological possibility: that of the **sponge**. In our two previous examples, high- and low-density regions were distinguished by their **connectivity** (whether it is possible to move continuously between all points in a given set). In contrast, a sponge has both classes of region being connected: it is possible to swim to any point through the holes, or to burrow to any point within the body of the sponge; filling a sponge with cement and etching away the sponge produces a cement sponge. Again, just by the symmetry between overdensity and underdensity, a Gaussian field in 3D must have a sponge-like topology.

The above discussion has focused on the properties of contour surfaces. These properties can be studied quantitatively via the **genus**: the number of 'holes' in a surface (zero for a sphere, one for a doughnut *etc.*). This is related to the Gaussian curvature of the surface, $K = 1/(r_1 r_2)$ (where r_1 and r_2 are the two principal radii of curvature), via the Gauss-Bonnet theorem (see *e.g.* Dodson & Poston 1977)

$$C \equiv \int K \, dA = 4\pi(1 - G), \tag{226}$$

where G is the genus. Topological results are sometimes instead quoted in terms of the **Euler-Poincaré characteristic**, which is -2 times the genus.

For Gaussian fields, the expectation value of the genus per unit volume (denoted by g) is (see Hamilton, Gott & Weinberg 1986)

$$\boxed{g = \frac{1}{4\pi^2}\left[\frac{-\xi''(0)}{\xi(0)}\right]^{3/2}(1-\nu^2)e^{-\nu^2/2}.} \tag{227}$$

For the median density contour ($\nu = 0$), the curvature is negative, implying that the surface has genus greater than unity. For $|\nu| > 1$, however, the curvature is positive – as expected if there are no holes. The contours become simply connected balls around either isolated peaks or voids.

It is interesting to note that the genus carries some information about the shape of the power spectrum, not in the behaviour with ν, but in the overall scaling. For Gaussian filtering,

$$g = \frac{1}{4\pi^2 R_{\rm G}^3}\left(\frac{3+n}{3}\right)^{3/2}(1-\nu^2)e^{-\nu^2/2}, \tag{228}$$

and so the effective spectral index can be determined in this way.

A similar procedure can be carried out in 2D (see Melott *et al.* 1989; Coles & Plionis 1991). The Gauss-Bonnet theorem is now

$$C \equiv \int K \, dA = 2\pi(1 - G), \tag{229}$$

where the meaning of $1 - G$ is the number of isolated contours minus the number of contour loops within other loops (sometimes the 2D genus is defined with the opposite sign; our convention follows that in 3D and the signs below are consistent). The result for the 2D genus per unit area is

$$\boxed{g = -\frac{1}{(2\pi)^{3/2}} \left[\frac{-\xi''(0)}{\xi(0)}\right] \nu \, e^{-\nu^2/2}.} \tag{230}$$

The 3D case is analogous but, as always, more messy; see BBKS.

Fig. 10. Results from the Genus analysis applied to 3D redshift data, taken from Gott *et al.* (1989). The small 'meatball shift' seen here is argued by the authors to be consistent with non-linear evolution from Gaussian initial conditions. It is interesting that the behaviour becomes more nearly Gaussian as we move to deeper samples which allow larger filtering lengths. Such plots constitute the strongest evidence we have that cosmic structure did indeed form via gravitational instability from Gaussian primordial fluctuations.

Applications of this method to real data (Figure 10) naturally reveal departures from Gaussian behaviour – one wishes to test whether the initial conditions were Gaussian, realising that nonlinear evolution will cause the field to become non-Gaussian. This means that either N-body simulations have to be used to predict the degree of non-Gaussian behaviour (usually in the 'meatball' direction), or one is confined to smoothing the data heavily to probe only large linear/angular scales. These should still be Gaussian, but of course by smoothing over many small regions

there is the danger that the central limit theorem will produce a Gaussian-like result in all cases.

Fourier tests

A third class of test was suggested by Feldman, Kaiser & Peacock (1994), and rests on measuring phase correlations between different mode amplitudes in the Fourier analysis of redshift surveys. First note that the two-point function in k space for a homogeneous random statistical process is always a delta-function:

$$\langle \delta_k(\mathbf{k}) \delta_k^*(\mathbf{k}') \rangle = \frac{[2\pi]^3}{V} P(k)\ \delta_{\rm D}(\mathbf{k}-\mathbf{k}'), \tag{231}$$

and that this applies for both Gaussian and non-Gaussian processes. To prove this, write down the definition of δ_k twice and multiply for different wavenumbers, using the reality of δ:

$$\delta_k \delta_{k'}^* = \frac{1}{V^2} \int \delta(\mathbf{r})\delta(\mathbf{r}+\mathbf{x})\ e^{i\mathbf{k}'\cdot\mathbf{x}}\ d^3x \int e^{i\mathbf{r}\cdot(\mathbf{k}-\mathbf{k}')}\ d^3r. \tag{232}$$

Performing the ensemble average for a stationary statistical process gives

$$\langle \delta(\mathbf{r})\delta(\mathbf{r}+\mathbf{x}) \rangle = \xi(x), \tag{233}$$

independent of r. The integral over r can now be performed, showing that $\langle \delta_k(\mathbf{k}) \delta_k^*(\mathbf{k}') \rangle$ vanishes unless $\mathbf{k} = \mathbf{k}'$ in the discrete case, or that in the continuum limit there is a delta-function in k space.

This result applies in the limit of an infinite survey. When there is a limited survey volume, delimited by the mean density $\bar{n}(\mathbf{r})$, we know that the Fourier coefficients are convolved by the transform of $\bar{n}(\mathbf{r})$. There will therefore be a coherence length in k space of order the reciprocal of the survey depth, over which length Fourier modes will have a significant two-point correlation. In the generalization where the survey galaxies may be weighted, Feldman, Kaiser & Peacock show that the exact expression for the two-point function is

$$\langle \delta_k(\mathbf{k}) \delta_k^*(\mathbf{k}+\delta\mathbf{k}) \rangle = P(k) Q(\delta\mathbf{k}) + S(\delta\mathbf{k}), \tag{234}$$

where

$$\begin{aligned} Q(\mathbf{k}) &\equiv \frac{\int w^2 \bar{n}^2\ \exp[i\mathbf{k}\cdot\mathbf{r}]\ d^3r}{\int w^2\bar{n}^2\ d^3r} \\ S(\mathbf{k}) &\equiv \frac{\int w^2 \bar{n}\ \exp[i\mathbf{k}\cdot\mathbf{r}]\ d^3r}{\int w^2\bar{n}^2\ d^3r} \end{aligned} \tag{235}$$

Furthermore, in the case of Gaussian fields only, this two-point function of amplitudes is simply related to the two-point function for the power:

$$\langle \delta P(\mathbf{k}) \delta P(\mathbf{k}+\delta\mathbf{k}) \rangle = |\langle \delta_k(\mathbf{k}) \delta_k^*(\mathbf{k}+\delta\mathbf{k}) \rangle|^2. \tag{236}$$

The significance of these results is that they allow direct constraints to be placed on a large class of non-Gaussian models in which the character of the linear density

fluctuations is Gaussian, but with a spatial modulation or **intermittency**, so that there are 'quiet' and 'noisy' parts of space:

$$\delta(\mathbf{r}) \rightarrow \delta(\mathbf{r})\,[1 + M(\mathbf{r})]. \tag{237}$$

Such a model was proposed on phenomenological grounds by Peebles (1983), but might also be realised in inflationary models with multiple scalar fields. The modulating field $M(\mathbf{r})$ acts in the same way as a mask imposed by observational selection, multiplying the effective $\bar{n}$. It therefore convolves the transform of $\bar{n}$ and broadens it. The signature of this form of non-Gaussianity is thus an extended tail of correlated power in the transform of the survey, and the agreement of the observed and expected power correlations can be used to set limits on the non-Gaussianity. Figure 11 shows the application of this analysis to the combined QDOT and 1.2-Jy surveys, and the excellent agreement provides a strong piece of evidence for the Gaussian nature of primordial fluctuations.

Fig. 11. The normalized 2-point correlation of the power measured in the combined QDOT and 1.2-Jy IRAS redshift surveys, plotted against k separation $\delta k/\ h\,\mathrm{Mpc}^{-1}$ (Stirling & Peacock 1996). Wavenumbers $k < 0.1\ h\,\mathrm{Mpc}^{-1}$ are considered. The observed correlations follow the expected form very closely, and limit any modulation of the statistical properties of the density field on 100 Mpc scales.

6 Conclusions

The testability of inflation

This brief summary of inflationary models has presented the 'party line' of many workers in the field for the simplest ways in which inflation could happen. It is a tremendous achievement to have a picture of this level of detail – but is it at all close to the truth?

What are the predictions of inflation? The simplified package is (1) $k = 0$; (2) scale-invariant, Gaussian fluctuations, and these are essentially the only tests discussed in the 1990 NATO symposium on observational test of inflation (Shanks *et al.* 1991). As the observations have hardened, however, there has been a tendency for the predictions to weaken. The flatness prediction was long taken to favour an $\Omega = 1$ Einstein-de Sitter model, but timescale problems have moved attention to models with $\Omega_{\rm matter} + \Omega_{\rm vacuum} = 1$. More recently, inflationary models have even been proposed which might yield open universes. This is achieved not by fine-tuning the amount of inflation, which would certainly be contrived, but by appealing to the details of the mechanism whereby inflation ends. It has been proposed that quantum tunnelling might create a 'bubble' of open universe in a plausible way (Bucher, Goldhaber & Turok 1995; Górski *et al.* 1995). It seems that inflation can never be disproved by the values of the global cosmological parameters.

A more characteristic inflationary prediction is the gravity-wave background. Although it is not unavoidable, the coupling between tilt and the gravity-wave contribution to CMB anisotropies is a signature expected in many of the simplest realizations of inflation. An observation of such a coupling would certainly constitute very powerful evidence that something like inflation did occur. Sadly, this test loses its power as the degree of tilt becomes smaller, which does seem to be the case observationally (Peacock & Dodds 1994). The most convincing verification of inflation would of course be the direct detection of the predicted flat spectrum of gravity waves in the local universe. However, barring some clever new technique, this will remain technologically unfeasible for the immediate future.

It therefore seems likely that the debate over the truth of inflation will continue without a clean resolution. There are certainly points of internal consistency in the theory which will attract further work: the form of the potential, and whether it can be maintained in the face of quantum corrections. Also, there is one more fundamental difficulty, which has been evaded until now.

The Λ problem

All our discussion of inflation has implicitly assumes that the zero of energy is set at $\Lambda = 0$ now, but there is no known principle of physics which requires this to be so. By experiment, $\Omega_{\rm vac} \lesssim 1$, which corresponds to a density in the region of 10^{100} times smaller than the GUT value. This fine tuning represents one of the major unsolved problems in physics (Weinberg 1989).

Other ways of looking at the origin of the vacuum energy include thinking of it as arising from a Bose-Einstein condensate, or via contributions from virtual particles. In the latter case, the energy density would be the rest mass times the number density of particles. A guess at this is to set the separation of particles at the Compton wavelength, $\hbar/mc$, yielding a density

$$\rho_{\rm vac} \simeq \frac{m^4 c^3}{\hbar^3}. \tag{238}$$

This exceeds the cosmological limit unless $m \lesssim 10^{-8} m_e$. One proposal for evading such a nonsensical result was made by Zeldovich: perhaps we cannot observe the rest mass of virtual particles directly, and we should only count the contribution of their gravitational interaction. This gives instead

$$\rho_{\rm vac} \simeq \left(\frac{Gm^2}{c\hbar}\right) \frac{m^4 c^3}{\hbar^3}, \tag{239}$$

which is acceptable if $m \lesssim 100 m_e$ – still far short of any plausible GUT-scale or Planck-scale cutoff.

We are left with the strong impression that some vital physical principle is missing. It is perhaps just as well that the average taxpayer who funds research in physics does not realise how much difficulty we have in understanding even nothing at all!

Fortunately, the existence of a non-zero vacuum density is not entirely a philosophical conundrum, but is subject to empirical verification in cosmology. If Λ exists at above the level of a few tenths of the critical density, it can be detected by a combination of geometrical tests, its effect on cosmological ages, and detailed signatures in CMB anisotropies. The challenge of confirming or ruling out a cosmologically significant vacuum energy is therefore developing into one of the dominant themes of cosmology in the 1990s, and is an area where we can reasonably hope to reach a decision within the next few years.

References

Albrecht A., Stebbins A., 1992, Phys. Rev. Lett., 69, 2615
Babul A., Postman M., 1990, ApJ, 359, 280
Ballinger W.E., Peacock J.A., Heavens A.F., 1996, MNRAS, submitted.
Bardeen J.M., Bond J.R., Efstathiou G., 1987, ApJ, 321, 28
Bardeen J.M., Bond J.R., Kaiser N., Szalay A.S., 1986, ApJ, 304, 15 (BBKS)
Baugh C.M., Efstathiou G., 1993, MNRAS, 265, 145
Baugh C.M., Efstathiou G., 1994, MNRAS, 267, 323
Blumenthal, G.R., Faber, S.M., Primack, J.R. & Rees, M.J., 1984. *Nature*, **311**, 517.
Bond J.R., Efstathiou G., 1991, Phys. Lett. B, 265, 245
Bond, J.R., Cole, S., Efstathiou, G. & Kaiser, N., 1991. *Astrophys. J.*, 379, 440
Bower R.G., Coles P., Frenk C.S., White S.D.M., 1993, ApJ, 405, 403
Bower, R.G., 1991. *Mon. Not. R. astr. Soc.*, **248**, 332.
Brandenberger, R.H., 1990. in *Physics of the Early Universe*, proc 36^{th} Scottish Universities Summer School in Physics, eds Peacock, J.A., Heavens, A.F. & Davies, A.T. (Adam Hilger), p281.
Bucher M., Goldhaber A.S., Turok N., 1995, *Phys. Rev. D*, **52**, 3314
Bunn, E.F., Scott D., White M., 1995, ApJ, 441, 9
Carlberg, R.G., 1991. *Astrophys. J.*, **367**, 385.
Carlberg, R.G., Couchman, H.M.P. & Thomas, P.A., 1990. *Astrophys. J.*, **352**, L29.
Carroll S.M., Press W.H., Turner E.L., 1992, ARAA, 30, 499
Cen R., Gnedin N.Y., Kofman L.A., Ostriker J.P., 1992, ApJ, 399, L11

Cen R., Ostriker J.P., 1992, ApJ, 399, L113
Cole, S. & Kaiser, N., 1989. *Mon. Not. R. astr. Soc.*, **237**, 1127.
Coles P., 1993, MNRAS, 262, 1065
Coles, P. & Jones, B.J.T., 1991. *Mon. Not. R. astr. Soc.*, **248**, 1.
Coles, P. & Plionis, M, 1991. *Mon. Not. R. astr. Soc.*, **250**, 75.
Couchman, H.M.P. & Carlberg, R.G., 1992. *Astrophys. J.*, 389, 453
Davis M., Efstathiou G., Frenk C.S., White S.D.M., 1985, ApJ, 292, 371
Davis, M. & Peebles, P.J.E., 1983. *Astrophys. J.*, **267**, 465.
Dekel, A. & Rees, M.J., 1987. *Nature*, **326**, 455.
Dekel, A., 1994, ARAA, 32, 371
Dodson, C.T.J. & Poston, T., 1977. *Tensor Geometry* (London; Pitman).
Dressler, A., 1980, ApJ, 236, 351.
Efstathiou G., Davis M., White S.D.M., Frenk C.S., 1995, ApJ Suppl., 57, 241
Efstathiou G., Sutherland W.J., Maddox S.J., 1990a, Nature, 348, 705
Efstathiou, G. & Bond, J.R., 1986. *Mon. Not. R. astr. Soc.*, **218**, 103.
Efstathiou, G., Bernstein, G., Katz, N., Tyson, T. & Guhathakurta, P., 1991. *Astrophys. J.*, 380, 47
Efstathiou, G., Kaiser, N., Saunders, W., Lawrence, A., Rowan-Robinson, M., Ellis, R.S. & Frenk, C.S., 1990b. *Mon. Not. R. astr. Soc.*, **247**, 10P.
Feldman H.A., Kaiser N., Peacock J.A., 1994, ApJ, 426, 23
Fisher K.B., Davis M., Strauss M.A., Yahil A., Huchra J.P., 1994, MNRAS, 267, 927
Górski K.M., Ratra B., Sugiyama N., Banday A.J., 1995, *Astrophys. J.*, 444, L65
Gaztañaga E., 1992, ApJ, 398, L17
Gott, J.R. III, Melott, A.L. & Dickinson, M., 1986. *Astrophys. J.*, **306**, 341.
Gott, J.R. III, *et al.* 1989. *Astrophys. J.*, **340**, 625.
Guth, A.H., 1981. *Phys. Rev. D*, **23**, 347.
Hamilton A.J.S., Kumar P., Lu E., Matthews A., 1991, ApJ, 374, L1 (HKLM)
Hamilton, A.J.S., Gott, J.R. III & Weinberg, D.H., 1986. *Astrophys. J.*, **309**, 1.
Henry, J.P. & Arnaud, K.A., 1991. *Astrophys. J.*, **372**, 410.
Hernquist L., Bouchet F.R., Suto Y., 1991, ApJ Suppl., 75, 231.
Hockney R. W., Eastwood J. W., 1988, "Computer Simulations Using Particles", IOP Publishing, Bristol.
Holtzman J.A., 1989, ApJs, 71, 1
Jain B., Mo H.J., White S.D.M., 1995, MNRAS, 276, L25
Kaiser N., 1984, ApJ, 284, L9
Kaiser N., 1987, MNRAS, 227, 1
Kaiser, N., 1986a. *Mon. Not. R. astr. Soc.*, **219**, 785.
Kaiser N., Malaney R.A., Starkman G.D., 1993, Phys. Rev. Lett., 71, 1128
Klypin A., Holtzman J., Primack J., Regős E., 1993, ApJ, 416, 1
Kofman, L. & Linde, A., 1987. *Nucl. Phys.*, **B282**, 555.
Kolb E.W., Turner M.S., 1990, *The Early Universe* (Addison-Wesley)
Lacey C., Cole S., 1993, MNRAS, 262, 632
Lahav O., Lilje P.B., Primack J.R., Rees M.J., 1991, MNRAS, 251, 128
Liddle A.R., Lyth D., 1993, Phys. Rep., 231, 1
Linde, A., 1986. *Phys. Lett.*, **175B**, 295.
Linde, A., 1989. *Inflation and quantum cosmology*, Academic Press, Boston.
Loveday J., Maddox S.J., Efstathiou G., Peterson B.A., 1995, ApJ, 442, 457
Loveday J., Efstathiou G., Maddox S.J., Peterson B.A., 1996, ApJ, in press
Madsen J., 1992, phys. Rev. Lett., 69, 571
Mann R.G., Peacock J.A., Heavens A.F., 1996, MNRAS, in preparation.
McNally S.J., Peacock J.A., 1995, MNRAS, 277, 143

Melott, A.L., Cohen, A.P., Hamilton, A.J.S., Gott, J.R. III & Weinberg, D.H., 1989. *Astrophys. J.*, **345**, 618.
Mo H.J., Jing Y.P., Börner G., 1993, MNRAS, 264, 825
Mo H.J., McGaugh S.S., Bothun G.D., 1994, MNRAS, 267, 129
Mukhanov V.F., Feldman H.A., Brandenberger R.H., 1992, Phys. Rep., 215, 203
Peacock J.A., 1991, MNRAS, 253, 1P
Peacock J.A., 1992, in Martinez V., Portilla M., Sáez D., eds, New insights into the Universe, Proc. Valencia summer school (Springer, Berlin), p1
Peacock J.A., Dodds S.J., 1994. MNRAS, 267, 1020
Peacock J.A., Dodds S.J., 1996. MNRAS, submitted
Peacock, J.A. & Heavens, A.F., 1990, *Mon. Not. R. astr. Soc.*, **243**, 133
Peacock, J.A., 1990. *Mon. Not. R. astr. Soc.*, **243**, 517.
Peebles P.J.E., 1980, The Large-Scale Structure of the Universe. Princeton Univ. Press, Princeton, NJ
Peebles P.J.E., 1993, Principles of physical cosmology. Princeton Univ. Press, Princeton, NJ
Peebles P.J.E., 1983, ApJ, 274, 1
Press W.H., Schechter P., 1974, ApJ, 187, 425
Press W.H., Vishniac E.T., 1980, ApJ, 239, 1
Raymond, J.C., Cox, D.P. & Smith, B.W., 1976. *Astrophys. J.*, **204**, 290.
Rees, M.J. & Ostriker, J.P., 1977. *Mon. Not. R. astr. Soc.*, **179**, 541.
Rees, M.J., 1985. *Mon. Not. R. astr. Soc.*, **213**, 75P.
Saunders W., Rowan-Robinson M., Lawrence A., 1992, MNRAS, 258, 134
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.
Shanks T., Boyle B.J., 1994, MNRAS, 271, 753
Shanks T. *et al.* (eds), 1991, *Observational tests of cosmological inflation*, NATO ASI C264, Kluwer
Starobinsky A.A., 1985, Sov. Astr. Lett., 11, 133
Stirling A.J., Peacock J.A., 1996, MNRAS, in preparation.
Strauss M.A., Davis M., Yahil A., Huchra J.P., 1992, ApJ, 385, 421
Sugiyama N., 1995, ApJ Suppl., 100, 281
Thuan, T.X., Gott, J.R. III & Schneider, S.E., 1987. *Astrophys. J.*, **315**, L93.
Valls-Gabaud, D., Alimi, J.-M., & Blanchard, A., 1989. *Nature*, **341**, 215.
Vilenkin, A. & Shellard, E.P.S., 1994. *Cosmic strings and other topological defects*, CUP.
Weinberg, S. 1989. *Rev. Mod. Phys.*, **61**, 1.
West, M.J. & Richstone, D.O., 1988. *Astrophys. J.*, **335**, 532.
White M., Bunn E.F., 1995, ApJ, 450, 477
White S.D.M., Efstathiou G., Frenk C.S., 1993, MNRAS, 262, 1023
White, S.D.M., Davis, M., Efstathiou, G. & Frenk, C.S., 1987. *Nature*, **330**, 451.
White, S.D.M., Navarro, J.F., Evrard, A.E. & Frenk, C.S., 1993. *Nature*, **366**, 429.
Williams, B.G., Heavens, A.F., Peacock, J.A. & Shandarin, S.F., 1991a. *Mon. Not. R. astr. Soc.*, **250**, 458.
Yi, I. & Vishniac, E.T., 1993. *Astrophys. J. Suppl.*, **86**, 333.

This book was processed by the author using the TEX macro package from Springer-Verlag.