
# Notes on cosmological structure formation
# UniverseNet Mytilene Sept 2007

**J.A. Peacock**

## Useful books

**Peacock**: Cosmological Physics (CUP)
**Dodelson**: Modern Cosmology (Wiley)
**Mukhanov**: Physical Foundations of Cosmology (CUP)
**Peebles**: Principles of Physical Cosmology (Princeton)

# 1 Overview of linear density fluctuations

## 1.1 the perturbed universe

The aim of these lectures is to cover the development of structure in the universe during more recent history, leading from the linear phase of the CMB to the emergence of complex nonlinear phenomena. The first issue we have to deal with is how to quantify departures from uniform density. Frequently, an intuitive Newtonian approach can be used, and we will adopt this wherever possible. But we should begin with a quick overview of the relativistic approach to this problem, to emphasise some of the big issues that are ignored in the Newtonian method.

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

$$S'(x^\mu) = S(x^\mu) - \epsilon^\alpha \partial S/\partial x^\alpha. \tag{1}$$

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

$$\rho' = \rho - \epsilon^0 \dot{\rho}. \tag{2}$$

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

The most direct general way of solving these difficulties is to construct perturbation variables that are explicitly independent of gauge. A comprehensive technical discussion of this method is given in chapter 7 of Mukhanov's book, and we summarize the essential elements here, largely without proof. First, we need to devise a notation that will classify the possible perturbations. Since the metric is symmetric, there are 10 independent degrees of freedom in $g^{\mu\nu}$; a convenient scheme that captures these possibilities is to write the cosmological metric as

$$d\tau^2 = a^2(\eta)\left\{(1+2\phi)d\eta^2 + 2w_i d\eta\, dx^i - [(1-2\psi)\gamma_{ij} + 2h_{ij}]\, dx^i\, dx^j\right\}. \tag{3}$$

In this equation, $\eta$ is **conformal time**,

$$d\eta = dt/a(t), \tag{4}$$

and $\gamma_{ij}$ is the comoving spatial part of the Robertson-Walker metric.

The total number of degrees of freedom here is apparently 2 (scalar fields $\phi$ and $\psi$) + 3 (3-vector field $\mathbf{w}$) + 6 (symmetric 3-tensor $h_{ij}$) = 11. To get the right number, the tensor $h_{ij}$ is required to be traceless: $\gamma^{ij}h_{ij} = 0$. The perturbations can be split into three classes: **scalar perturbations**, which are described by scalar functions of spacetime coordinate, and which correspond to growing density perturbations; **vector perturbations**, which correspond to vorticity perturbations, and **tensor perturbations**, which correspond to gravitational waves. Here, we shall concentrate mainly on scalar perturbations.

Since vectors and tensors can be generated from derivatives of a scalar function, the most general scalar perturbation actually makes contributions to all the $g_{\mu\nu}$ components in the above expansion:

$$\delta g_{\mu\nu} = a^2 \begin{pmatrix} 2\phi & -B_{,i} \\ -B_{,i} & 2[\psi\delta_{ij} - E_{,ij}] \end{pmatrix}, \tag{5}$$

where four scalar functions $\phi$, $\psi$, $E$ and $B$ are involved. It turns out that this situation can be simplified by defining variables that are unchanged by a gauge transformation:

$$\begin{aligned} \Phi &\equiv \phi + \frac{1}{a}\left[(B - E')a\right]' \\ \Psi &\equiv \psi - \frac{a'}{a}\,(B - E'), \end{aligned} \tag{6}$$

where primes denote derivatives with respect to conformal time.

These gauge-invariant 'potentials' have a fairly direct physical interpretation, since they are closely related to the Newtonian potential. The easiest way to evaluate the gauge-invariant fields is to make a specific gauge choice and work with the **longitudinal gauge** in which $E$ and $B$ vanish, so that $\Phi = \phi$ and $\Psi = \psi$. A second key result is that inserting the longitudinal metric into the Einstein equations shows that $\phi$ and $\psi$ are identical in the case of fluid-like perturbations where off-diagonal elements of the energy–momentum tensor vanish. In this case, the longitudinal gauge becomes identical to the **Newtonian gauge**, in which perturbations are described by a single scalar field, which is the gravitational potential. The conclusion is thus that the gravitational potential can for many purposes give an effectively gauge-invariant measure of cosmological perturbations, and this provides a sounder justification for the Newtonian approach that we now adopt. The Newtonian-gauge metric therefore looks like this:

$$d\tau^2 = (1 + 2\Phi)dt^2 - (1 - 2\Phi)\gamma_{ij}\,dx^i\,dx^j, \tag{7}$$

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

## 1.2 Fluctuation power spectra

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

$$\nabla^2\Phi/a^2 = 4\pi G(1+3w)\,\bar{\rho}\,\delta, \tag{8}$$

where we have defined a dimensionless fluctuation amplitude

$$\delta \equiv \frac{\rho-\bar{\rho}}{\bar{\rho}}. \tag{9}$$

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

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

$$\delta(\mathbf{x}) = \sum \delta_k e^{-i\mathbf{k}\cdot\mathbf{x}}, \tag{10}$$

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

$$k_x = n\,\frac{2\pi}{L}, \qquad n = 1, 2\cdots, \tag{11}$$

and the inverse Fourier relation would be

$$\delta_k(\mathbf{k}) = \left(\frac{1}{L}\right)^3 \int \delta(\mathbf{x})\exp\bigl(i\mathbf{k}\cdot\mathbf{x}\bigr)\,d^3x. \tag{12}$$

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

$$\Phi_k = -4\pi G(1+3w)a^2\,\bar{\rho}\,\delta_k/k^2. \tag{13}$$

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

$$\langle\delta^2\rangle = \sum |\delta_k|^2. \tag{14}$$

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

$$P(k) \equiv |\delta_k|^2 \tag{15}$$

is called the **power spectrum**. Note that, in an isotropic universe, we assume that $P$ will be independent of direction of the wavevector in the limit of a large box: the fluctuating density field is statistically **isotropic**. In applying this apparatus, we

would not want the (arbitrary) box size to appear. This happens naturally: as the box becomes big, the modes are finely spaced and a sum over modes is replaced by an integral over $k$ space times the usual density of states, $(L/2\pi)^3$:

$$\langle \delta^2 \rangle = \sum |\delta_k|^2 \rightarrow \frac{L^3}{(2\pi)^3} \int P(k)\, d^3k = \int \Delta^2(k)\, d\ln k. \tag{16}$$

In the last step, we have defined the combination

$$\Delta^2(k) \equiv \frac{L^3}{(2\pi)^3}\, 4\pi k^3\, P(k), \tag{17}$$

which absorbs the box size into the definition of a dimensionless power spectrum, which gives the contribution to the variance from each logarithmic range of wavenumber (or wavelength).

SCALE-INVARIANT SPECTRUM In a sense, it is more natural to think of potential (i.e. metric) perturbations. From Poisson's equation

$$\Delta^2_\Phi \propto k^{-4} \Delta^2_m. \tag{18}$$

A simple reference case would have a *fractal* metric with $\Delta^2_\Phi = \text{const}$. This would therefore correspond to $P(k) \propto k$.

ENSEMBLES AND ERGODICITY We interpreted the angle bracket average $\langle\rangle$ as a spatial average, but in fact we are often interested in something subtly different. The random process $\delta(\mathbf{x})$ should be homogeneous in its statistical properties, and no point in space is preferred. The actual field found in a given case is a **realization** of the statistical process, which we can imagine as having been drawn from an ensemble of possibilities.

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

$$\text{volume average} \quad \leftrightarrow \quad \text{ensemble average} \tag{19}$$

are termed **ergodic**.

As an example of this alternative approach, consider the 2-point function of Fourier amplitudes, $\langle \delta_k \delta_{k'} \rangle$ Write down the definition of $\delta_k$ twice and multiply for different wavenumbers, using the reality of $\delta$:

$$\delta_k \delta^*_{k'} = \frac{1}{V^2} \int \delta(\mathbf{r}) \delta(\mathbf{r}+\mathbf{x})\, e^{-i\mathbf{k}'\cdot\mathbf{x}}\, d^3x \int e^{i\mathbf{r}\cdot(\mathbf{k}-\mathbf{k}')}\, d^3r. \tag{20}$$

Performing the ensemble average for a stationary statistical process gives $\langle \delta(\mathbf{r})\delta(\mathbf{r}+\mathbf{x})\rangle = \xi(x)$, 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:

$$\langle \delta_k \delta_{k'} \rangle = (2\pi)^3 \, P(k) \, \delta(k + k'). \tag{21}$$

Similar relations exist for higher-order correlations, e.g.

$$\langle \delta_{k_1} \delta_{k_2} \delta_{k_3} \rangle = (2\pi)^3 \, B(k) \, \delta(k_1 + k_2 + k_3) \tag{22}$$

defines the **bispectrum** of the field.

INITIAL CONDITIONS IN THE STANDARD MODEL One of the motivations for studying late-time structure is the hope that we can learn something about the initial conditions. We will frequently approach this with the viewpoint of simple inflationary models, so it is worth summarising the terminology and predictions. In terms of the variance per $\ln k$ in potential perturbations, the predicted power spectrum is

$$\begin{aligned} \delta_{\rm H}^2 \equiv \Delta_\Phi^2(k) &= \frac{H^4}{(2\pi\dot\phi)^2} \\ H^2 &= \frac{8\pi}{3}\,\frac{V}{m_{\rm P}^2} \\ 3H\dot\phi &= -V', \end{aligned} \tag{23}$$

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

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

$$\epsilon \equiv \frac{m_{\rm P}^2}{16\pi}\,(V'/V)^2 \ll 1. \tag{24}$$

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

$$\eta \equiv \frac{m_{\rm P}^2}{8\pi}\,(V''/V) \ll 1 \tag{25}$$

(using $H \sim \sqrt{V}/m_{\rm P}$ 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.

A few other comments are in order. These perturbations 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. This holds in the case of inflation, since

modes of different wavelength behave independently and have independent quantum fluctuations. Adiabatic perturbations are ones where the matter and photon number densities are perturbed equally, which is the natural outcome of post-inflation reheating for a single field.

It is easy to see that this condition can be violated in multi-field models. The converse of an adiabatic perturbation would be an **isocurvature** perturbation. This is a fluctuation in equation of state, with $\delta = 0$ at the initial point. At early enough times, this is identical to an **isothermal** perturbation $\delta_r = 0$ (but it doesn't stay isothermal). A simple case would be the **curvaton** model, where a second field decays into radiation, thus creating a coupled mixture of adiabatic and isocurvature modes of similar amplitude. Such models are firmly ruled out by joint CMB+LSS studies.

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

$$\text{tilt} \equiv 1 - n \equiv -\frac{d \ln \delta_{\rm H}^2}{d \ln k}. \tag{26}$$

We then want to express the tilt in terms of parameters of the inflationary potential, $\epsilon$ and $\eta$. These are of order unity when inflation terminates; $\epsilon$ and $\eta$ 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,

$$a/k = H^{-1}. \tag{27}$$

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

$$\frac{d}{d \ln k} = a \frac{d}{da} = \frac{\dot\phi}{H} \frac{d}{d\phi} = -\frac{m_{\rm P}^2}{8\pi} \frac{V'}{V} \frac{d}{d\phi}. \tag{28}$$

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

$$\delta_{\rm H}^2 = \frac{H^4}{(2\pi\dot\phi)^2} = \frac{128\pi}{3} \left( \frac{V^3}{m_{\rm P}^6 V'^2} \right), \tag{29}$$

and derivatives of $V$ can be expressed in terms of the dimensionless parameters $\epsilon$ and $\eta$. The tilt of the density perturbation spectrum is thus predicted to be

$$1 - n = 6\epsilon - 2\eta \tag{30}$$

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

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

$$1 - n = 2/p \tag{31}$$

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

$$60 = \int H\, dt = \frac{8\pi}{m_{\rm P}^2} \int_{\phi_{\rm end}}^{\phi} \frac{V}{V'}\, d\phi. \tag{32}$$

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

$$60 = \frac{8\pi}{m_{\rm P}^2} \int_{\phi_{\rm end}}^{\phi} \frac{\phi}{\alpha}\, d\phi = \frac{4\pi}{m_{\rm P}^2\alpha}(\phi^2 - \phi_{\rm end}^2). \tag{33}$$

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

$$1 - n = (2+\alpha)/120. \tag{34}$$

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.

### 1.3 Newtonian equations of motion

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

$$\begin{aligned} \mathbf{x}(t) &= a(t)\mathbf{r}(t) \\ \delta\mathbf{v}(t) &= a(t)\mathbf{u}(t), \end{aligned} \tag{35}$$

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

$$\dot{\mathbf{x}} = \dot a\mathbf{r} + a\dot{\mathbf{r}} = H\mathbf{x} + a\dot{\mathbf{r}}, \tag{36}$$

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

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

$$\ddot{\mathbf{x}} = a\dot{\mathbf{u}} + 2\dot a\mathbf{u} + \frac{\ddot a}{a}\mathbf{x} = \mathbf{g}_0 + \mathbf{g}. \tag{37}$$

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

$$\dot{\mathbf{u}} + 2(\dot{a}/a)\mathbf{u} = \mathbf{g}/a. \tag{38}$$

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

The peculiar velocity is directly related to the evolution of the density field, through conservation of mass. This is expressed via the continuity equation, which takes the form

$$\frac{d}{dt}\rho_0(1+\delta) = -\rho_0(1+\delta)\,\boldsymbol{\nabla}\cdot\mathbf{u}. \tag{39}$$

As usual, spatial derivatives are with respect to comoving coordinates:

$$\boldsymbol{\nabla} \equiv a\,\boldsymbol{\nabla}_{\rm proper}, \tag{40}$$

and the time derivative is a convective one:

$$\frac{d}{dt} = \frac{\partial}{\partial t} + \mathbf{u}\cdot\boldsymbol{\nabla}, \tag{41}$$

i.e. the time derivative measured by an observer who follows a particle's trajectory. Finally, when using a comoving frame, the background density $\rho_0$ is unaffected by $d/dt$, and so the full continuity equation can be written as

$$\frac{d}{dt}\delta = -(1+\delta)\boldsymbol{\nabla}\cdot\mathbf{u}. \tag{42}$$

LINEAR APPROXIMATION The equation for $\delta$ is not linear in the perturbations $\delta$ and $\mathbf{u}$. To cure this, we restrict ourselves to the case $\delta \ll 1$ and linearize the equation, neglecting second-order terms like $\delta \times \mathbf{u}$, which removes the distinction between convective and partial time derivatives. The linearized equations for conservation of momentum and matter as experienced by fundamental observers moving with the Hubble flow are then:

$$\begin{aligned} \dot{\mathbf{u}} + 2\frac{\dot{a}}{a}\,\mathbf{u} &= \frac{\mathbf{g}}{a} \\ \dot{\delta} &= -\boldsymbol{\nabla}\cdot\mathbf{u}, \end{aligned} \tag{43}$$

where the peculiar gravitational acceleration $-\boldsymbol{\nabla}\Phi/a$ is denoted by $\mathbf{g}$.

The solutions of these equations can be decomposed into modes either parallel to $\mathbf{g}$ or independent of $\mathbf{g}$ (these are the homogeneous and inhomogeneous solutions to the equation of motion). The homogeneous case corresponds to no peculiar gravity – i.e. zero density perturbation. This is consistent with the linearized continuity equation, $\boldsymbol{\nabla}\cdot\mathbf{u} = -\dot{\delta}$, which says that it is possible to have **vorticity modes** with $\boldsymbol{\nabla}\cdot\mathbf{u} = 0$ for which $\dot{\delta}$ vanishes, so there is no growth of structure in this case. The proper velocities

of these vorticity modes decay as $v = au \propto a^{-1}$, as with the kinematic analysis for a single particle.

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

$$\ddot{\delta} + 2\frac{\dot{a}}{a}\dot{\delta} = 4\pi G \rho_0 \, \delta. \tag{44}$$

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

$$\delta(t) \propto t^{2/3} \quad \text{or} \quad t^{-1}. \tag{45}$$

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

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

$$\mathbf{g} = -\boldsymbol{\nabla}\Phi/a - \boldsymbol{\nabla}p/(a\rho). \tag{46}$$

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

$$c_s^2 \equiv \frac{\partial p}{\partial \rho}. \tag{47}$$

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

$$\ddot{\delta} + 2\frac{\dot{a}}{a}\dot{\delta} = \delta\big(4\pi G\rho_0 - c_s^2 k^2/a^2\big). \tag{48}$$

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

$$\lambda_{\rm J}^{\rm proper} = \frac{2\pi}{k_{\rm J}^{\rm proper}} = c_s\sqrt{\frac{\pi}{G\rho}}. \tag{49}$$

It is interesting to work out the value of this critical length. Consider a universe with a coupled photon-baryon fluid and ignore dark matter (which we can do at high redshifts, near matter-radiation equality). The sound speed, $c_{\rm s}^2 = \partial p/\partial\rho$, may be found by thinking about the response of matter and radiation to small adiabatic compressions:

$$\delta p = (4/9)\rho_r c^2(\delta V/V), \qquad \delta\rho = [\rho_m + (4/3)\rho_r](\delta V/V), \tag{50}$$

implying

$$c_{\rm S}^2 = c^2\left(3+\frac{9}{4}\frac{\rho_m}{\rho_r}\right)^{-1} = c^2\left[3+\frac{9}{4}\left(\frac{1+z_{\rm rad}}{1+z}\right)\right]^{-1}. \quad (51)$$

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

$$\lambda_{\rm J} = \frac{c}{H_0}\left(\frac{32\pi^2}{27\Omega_{\rm B}(1+z_{\rm rad})}\right)^{1/2} = 50\,(\Omega_{\rm B}h^2)^{-1}\ \ {\rm Mpc}. \quad (52)$$

This is of order the horizon size at matter-radiation equality. Smaller-scale fluctuations in the photon-baryon fluid will not have undergone steady gravitational growth, but will have oscillated with time as standing waves. We will see later that the imprint of these oscillations is visible in the microwave background.

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

$$\delta(a) \propto a\, f[\Omega(a)], \quad (53)$$

where the factor $f$ expresses a deviation from the simple growth law. The case of matter + cosmological constant is of the most common practical interest, and a very good approximation to the answer is given by Carroll et al. (1992):

$$f(\Omega) \simeq \tfrac{5}{2}\Omega_m\left[\Omega_m^{4/7} - \Omega_v + (1+\tfrac{1}{2}\Omega_m)(1+\tfrac{1}{70}\Omega_v)\right]^{-1}. \quad (54)$$

This is accurate, but still hard to remember. For flat models with $\Omega_m + \Omega_v = 1$, a simpler approximation is $f \simeq \Omega_m^{0.23}$, which is less marked than $f \simeq \Omega_m^{0.65}$ in the $\Lambda = 0$ case. This reflects the more rapid variation of $\Omega_v$ with redshift; if the cosmological constant is important dynamically, this only became so very recently, and the universe spent more of its history in a nearly Einstein–de Sitter state by comparison with an open universe of the same $\Omega_m$.

## 1.4 Radiation-dominated universe

At early enough times, the universe was radiation dominated ($c_s = c/\sqrt{3}$) and the analysis so far does not apply. It is common to resort to general relativity perturbation theory at this point. However, the fields are still weak, and so it is possible to generate the results we need by using special relativity fluid mechanics and Newtonian gravity with a relativistic source term:

$$\nabla^2\Phi = 4\pi G(\rho + 3p/c^2), \quad (55)$$

in Eulerian units.

The special-relativity fluid mechanics is not hard in principle, but we lack the time to go through it here. The logic is fairly straightforward, since the conservation equation we need comes from the 4-divergence of the energy-momentum tensor: $\dot{T}^{\mu\nu}/dx^\mu = 0$. In the rest frame, $T^{\mu\nu} = \mathrm{diag}(\rho c^2, p, p, p)$, so we just need to apply a shift to the lab frame: $\mathbf{T} \to \tilde{\mathbf{M}} \cdot \mathbf{T} \cdot \mathbf{M}$, where $\mathbf{M}$ is the matrix of Lorentz transformation coefficients. An easier way of achieving this is to use **manifest covariance**:

$$T^{\mu\nu} = (\rho + p/c^2)U^\mu U^\nu - p g^{\mu\nu}, \tag{56}$$

where $U^\mu$ is the 4-velocity. This is clearly a tensor, so we immediately have all the components when the velocity is non-zero, and it is straightforward to generate the equations of relativistic fluid mechanics. These equations look quite similar to the nonrelativistic equations, but with extra terms where the pressure is non-negligible. The main change from the previous treatment come from a factor of 2 from the $(\rho + 3p/c^2)$ term in Poisson's equation, and other contributions of the pressure to the relativistic equation of motion, such as $(\rho + p/c^2) = (4/3)\rho$. The resulting evolution equation for $\delta$ has a driving term on the rhs that is a factor 8/3 higher than in the matter-dominated case

$$\ddot{\delta} + 2\frac{\dot{a}}{a}\dot{\delta} = \frac{32\pi}{3} G\rho_0 \delta, \tag{57}$$

(see e.g. Section 15.2 of Peacock 1999 for the details).

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

$$\begin{aligned} \text{matter domination } (a \propto t^{2/3}): &\quad 4\pi G\rho_0 = \frac{2}{3t^2} \\ \text{radiation domination } (a \propto t^{1/2}): &\quad 32\pi G\rho_0/3 = \frac{1}{t^2}. \end{aligned} \tag{58}$$

Every term in the equation for $\delta$ is thus the product of derivatives of $\delta$ and powers of $t$, and a power-law solution is obviously possible. If we try $\delta \propto t^n$, then the result is $n = 2/3$ or $-1$ for matter domination; for radiation domination, this becomes $n = \pm 1$. For the growing mode, these can be combined rather conveniently using the **conformal time** $\eta \equiv \int dt/a$:

$$\delta \propto \eta^2. \tag{59}$$

The quantity $\eta$ is proportional to the comoving size of the cosmological particle 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 $\delta/a^2$ respectively, so that $\Phi$ is independent of $a$ in either case. In other words, the metric fluctuations resulting from potential perturbations are frozen, at least for perturbations with wavelengths greater than the horizon size.

MÉSZÁROS EFFECT What about the case of collisionless matter in a radiation background? The fluid treatment is not appropriate here, since the two species of particles can interpenetrate. A particularly interesting limit is for perturbations well inside the horizon: the radiation can then be treated as a smooth, unclustered

background that affects only the overall expansion rate. This is analogous to the effect of $\Lambda$, but an analytical solution does exist in this case. The perturbation equation is as before

$$\ddot{\delta} + 2\frac{\dot{a}}{a}\dot{\delta} = 4\pi G\rho_m\delta, \tag{60}$$

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

$$\delta'' + \frac{2+3y}{2y(1+y)}\,\delta' - \frac{3}{2y(1+y)}\,\delta = 0 \tag{61}$$

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

$$\delta \propto y + 2/3. \tag{62}$$

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 $\delta$ is frozen at a constant value. At late times, the radiation becomes negligible, and the growth increases smoothly to the Einstein–de Sitter $\delta \propto a$ behaviour (Mészáros 1974). The overall behaviour is therefore reminiscent to the effects of pressure on a coupled fluid, where growth is suppressed below the Jeans scale. However, the two phenomena are really quite 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.

This effect is critical in shaping the late-time power spectrum. For scales greater than the horizon, perturbations in matter and radiation can grow together, so fluctuations at early times grow at the same rate, independent of wavenumber. But this growth ceases once the perturbations 'enter the horizon' – i.e. when the horizon grows sufficiently to exceed the perturbation wavelength. At this point, growth ceases, so the universe preserves a 'snapshot' of the amplitude of the mode at horizon crossing. For a scale-invariant spectrum, this implies a dimensionless power $\delta^2(k) \simeq \delta_{\rm H}^2$ on small scales, breaking to the initial $\delta^2(k) \propto k^4$ on large scales. Observing this break and using it to measure the density of the universe has been one of the great success stories in recent cosmological research.

### 1.5 Transfer functions and characteristic scales

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

$$\delta(a) \propto g(a)T_k. \tag{63}$$

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

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

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

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

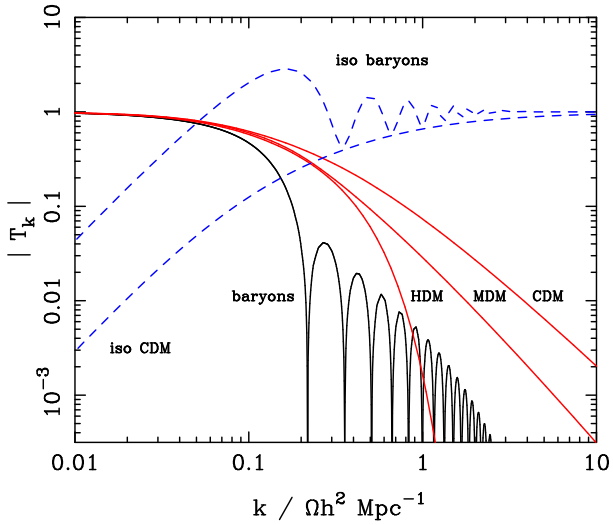


**Figure 1.** A plot of transfer functions for various adiabatic models, in which $T_k \to 1$ at small $k$. A number of possible matter contents are illustrated: pure baryons; pure CDM; pure HDM. For dark-matter models, the characteristic wavenumber scales proportional to $\Omega_m h^2$, marking the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

**(1) Horizon length at matter-radiation equality.** The main bend visible in all transfer functions is due to the Mészáros effect, which arises because the universe is radiation dominated at early times. The relative diminution in fluctuations at high $k$ is the amount of growth missed out on between horizon entry and $z_{\rm eq}$, which would be $\delta \propto D_{\rm H}^2$ in the absence of the Mészáros effect. Perturbations with larger $k$ enter the horizon when $D_{\rm H} \simeq 1/k$; they are then frozen until $z_{\rm eq}$, at which point they can grow again. The missing growth factor is just the square of the change in $D_{\rm H}$ during this period, which is $\propto k^2$. The approximate limits of the CDM transfer function are therefore

$$T_k \simeq \begin{cases} 1 & kD_{\rm H}(z_{\rm eq}) \ll 1 \\ [kD_{\rm H}(z_{\rm eq})]^{-2} & kD_{\rm H}(z_{\rm eq}) \gg 1. \end{cases} \quad (64)$$

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

**(2) Free-streaming length.** This relatively gentle filtering away of the initial fluctuations is all that applies to a universe dominated by Cold Dark Matter, in which random velocities are negligible. A CDM universe thus contains fluctuations in the dark matter on all scales, and structure formation proceeds via hierarchical process in which nonlinear structures grow via mergers. Examples of CDM would be thermal relic WIMPs with masses of order 100 GeV, but a more interesting case arises when thermal relics have lower masses. For collisionless dark matter, perturbations can be erased simply by free streaming: random particle velocities cause blobs to disperse. At early times ($kT > mc^2$), the particles will travel at $c$, and so any perturbation that has entered the horizon will be damped. This process switches off when the particles become non-relativistic, so that perturbations are erased up to proper lengthscales of $\simeq ct(kT = mc^2)$. This translates to a comoving horizon scale ($2ct/a$ during the radiation era) at $kT = mc^2$ of

$$L_{\rm free-stream} = 112\,(m/{\rm eV})^{-1}\,{\rm Mpc} \quad (65)$$

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

$$\Omega_\nu h^2 = m/94.1\,{\rm eV} \quad (66)$$

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

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

**(3) Acoustic horizon length.** The horizon at matter-radiation equality also enters in the properties of the baryon component. Since the sound speed is of order $c$, the largest scales that can undergo a single acoustic oscillation are of order the horizon. The transfer function for a pure baryon universe shows large modulations, reflecting the number of oscillations that have been completed before the universe becomes matter

dominated and the pressure support drops. The lack of such large modulations in real data is one of the most generic reasons for believing in collisionless dark matter. Acoustic oscillations persist even when baryons are subdominant, however, and can be detectable as lower-level modulations in the transfer function. We will say more about this later.

**(4) Silk damping length.** Acoustic oscillations are also damped on small scales, where the process is called Silk damping: the mean free path of photons due to scattering by the plasma is non-zero, and so radiation can diffuse out of a perturbation, convecting the plasma with it. The typical distance of a random walk in terms of the diffusion coefficient $D$ is $x \simeq \sqrt{Dt}$, which gives a damping length of

$$\lambda_{\rm S} \simeq \sqrt{\lambda D_{\rm H}}, \tag{67}$$

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

$$\lambda_{\rm S} = 16.3\,(1+z)^{-5/4}(\Omega_b^2\Omega_m h^6)^{-1/4}\ {\rm Gpc}. \tag{68}$$

This becomes close to the horizon length by the time of last scattering, $1+z \simeq 1100$. The resulting damping effect can be seen in figure 1 at $k \sim 10k_{\rm H}$.

The overall transfer function thus contains a number of features that can be probed observationally. However, the most robust signatures are the horizon-scale features, since they exist on scales where nonlinear evolution has the least impact on the shape of the spectrum:

$$\begin{aligned} \textbf{Matter}-\textbf{radiation horizon}&: 123(\Omega_m h^2/0.13)^{-1}\ {\rm Mpc} \\ \textbf{Acoustic horizon at last scattering}&: 147(\Omega_m h^2/0.13)^{-0.25}(\Omega_b h^2/0.024)^{-0.08}\ {\rm Mpc} \end{aligned} \tag{69}$$

.

SPECTRUM NORMALIZATION We now have a full recipe for specifying the matter power spectrum:

$$\Delta^2(k) \propto k^{3+n}\,T_k^2. \tag{70}$$

For completeness, we need to mention how the normalization of the spectrum is to be specified. Historically, this is done in a slightly awkward way. First suppose we wanted to consider smoothing the density field by convolution with some **window**. One simple case is to imagine averaging within a sphere of radius $R$. For the effect on the power spectrum, we need the Fourier transform of this filter:

$$\sigma^2(R) = \int \Delta^2(k)\,|W_k|^2\,d\ln k; \quad W_k = \frac{3}{(kR)^3}(\sin kR - kR\cos kR). \tag{71}$$

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

## 2 Formation of nonlinear structures

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

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

$$\frac{d}{dt}\mathbf{u} = -2\frac{\dot{a}}{a}\mathbf{u} - \frac{1}{a^2}\boldsymbol{\nabla}\Phi, \tag{72}$$

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

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

Computing lives by the 'garbage in, garbage out' rule, so how are the initial conditions in the simulation set? This can be understood by thinking of density fluctuations in **Lagrangian** terms (also known as the **Zeldovich approximation**). The proper coordinate of a given particle can be written as

$$\mathbf{x}(t) = a(t)\left(\mathbf{q} + \mathbf{f}(\mathbf{q}, t)\right), \tag{73}$$

where $\mathbf{q}$ is the usual comoving position, and the **displacement field** $\mathbf{f}(\mathbf{q}, t)$ tends to zero at $t = 0$. The comoving peculiar velocity is just the time derivative of this displacement:

$$\mathbf{u} = \frac{\partial \mathbf{f}}{\partial t} \tag{74}$$

(partial time derivative because each particle is labelled by an unchanging value of $q$ – this is what is meant by a Lagrangian coordinate).

By conservation of particles, the density at a given time is just the Jacobian determinant between $q$ and $x$:

$$\rho\,/\,\bar{\rho} = \left|\frac{\partial \mathbf{q}}{\partial \mathbf{x}/a}\right|. \tag{75}$$

When the displacement is small, this is just

$$\rho \,/\, \bar{\rho} = 1 - \boldsymbol{\nabla} \cdot \mathbf{f}(\mathbf{q}, t), \tag{76}$$

so the linear density perturbation $\delta$ is just (minus) the divergence of the displacement field. All this can be handled quite simply if we define a **displacement potential**:

$$\mathbf{f} = -\boldsymbol{\nabla}\psi(\mathbf{q}), \tag{77}$$

from which we have $\delta = \nabla^2\psi$ in the linear regime. The displacement potential $\psi$ is therefore proportional to the gravitational potential, $\Phi$. These equations are easily manipulated in Fourier space: given the amplitudes of the Fourier modes, $\delta_k$, we can obtain the potential

$$\psi_k = -\delta_k/k^2, \tag{78}$$

and hence the displacement and velocity

$$\begin{aligned} \mathbf{f}_k &= i\mathbf{k}\,\psi_k \\ \mathbf{u}_k &= i\mathbf{k}\,\dot{\psi}_k. \end{aligned} \tag{79}$$

Thus, given the density power spectrum to specify $|\delta_k|$ and the assumption of random phases, we can set up a field of consistent small displacements and consistent velocities. These are applied to a uniform particle 'load', and then integrated forward into the nonlinear regime.

The efficient way of performing the required Fourier transforms is by averaging the density field 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**, in which particles at large distances from the point at which the force is required are grouped into groups of successively coarser resolution as distance increases. The most popular public-domain $N$-body code is of this type (Volker Springel's GADGET).

THE SPHERICAL MODEL $N$-body models can yield evolved density fields that are nearly exact solutions to the equations of motion, but working out what the results mean is then more a question of data analysis than of deep insight. Where possible, it is important to have analytic models that guide the interpretation of the numerical results. The most important model of this sort is the spherical density perturbation, which can be analysed immediately using the tools developed for the Friedmann models, since Birkhoff's theorem tells us that such a perturbation behaves in exactly the same

way as part of a closed universe. The equations of motion are the same as for the scale factor, and we can therefore write down the **cycloid solution** immediately. For a matter-dominated universe, the relation between the proper radius of the sphere and time is

$$\begin{aligned} r &= A(1-\cos\theta) \\ t &= B(\theta-\sin\theta). \end{aligned} \tag{80}$$

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

$$r \simeq \frac{A}{2}\left(\frac{6t}{B}\right)^{2/3}\left[1-\frac{1}{20}\left(\frac{6t}{B}\right)^{2/3}\right], \tag{81}$$

and we can identify the density perturbation within the sphere:

$$\delta \simeq \frac{3}{20}\left(\frac{6t}{B}\right)^{2/3}. \tag{82}$$

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

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

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

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

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

These calculations are the basis for a common 'rule of thumb', whereby one assumes that linear theory applies until $\delta_{\rm lin}$ is equal to some $\delta_c$ a little greater than unity, at which point virialization is deemed to have occurred. Although the above only applies for $\Omega = 1$, analogous results can be worked out from the full $\delta_{\rm lin}(z,\Omega)$ and $t(z,\Omega)$ relations. These indicate that $\delta_{\rm lin} \simeq 1$ is a good criterion for collapse for any value of $\Omega$ likely to be of practical relevance. The density contrast at virialization tends to be higher in low-density universes, where the faster expansion means that, by the time a perturbation has turned round and collapsed to its final radius, a larger density

contrast has been produced. For real non-spherical systems, it is not clear that this effect is meaningful, and in practice a fixed density contrast of around 200 is used to define the **virial radius** that marks the boundary of an object.

PRESS–SCHECHTER AND THE HALO MASS FUNCTION What relevance does the spherical model have to the real world? Despite the lack of spherical symmetry, we can still use the model to argue that nonlinear collapse will occur whenever we have a region within which the mean linear-theory density contrast is of order unity. This has an interesting consequence in the context of the CDM model, where there is power on all scales: the sequence of structure formation must be **hierarchical**. This means that we expect the universe to fragment into low-mass clumps at high redshift, following which a number of clumps **merge** into larger units at later times. This process is controlled by the density variance as a function of smoothing scale, $\sigma^2(R)$. In a hierarchical model, this increases without limit as $R \to 0$, so there is always a critical scale at which $\sigma \simeq 1$. As the density fluctuations grow, this critical scale grows also. These collapsed systems are known as **dark-matter haloes**; a name that dates back to the 1970s, when the existence of extended dark matter around galaxies was first firmly established. The largest such haloes, forming today, are the rich clusters of galaxies. Galaxy-scale haloes formed earlier, and this process effectively dictates the era of galaxy formation.

We can improve on this outline, and calculate the distribution of halo masses that exist at any one time, using a method pioneered by Press & Schechter (1974). If the density field is Gaussian, the probability that a given point lies in a region with $\delta > \delta_c$ (the **critical overdensity** for collapse) is

$$p(\delta > \delta_c \mid R) = \frac{1}{\sqrt{2\pi}} \int_{\delta_c}^{\infty} \exp\left(-\delta^2/2\sigma^2(R)\right)\ d\delta, \tag{83}$$

where $\sigma(R)$ is the linear rms in the filtered version of $\delta$. The PS argument now takes this probability to be proportional to the probability that a given point has ever been part of a collapsed object of scale $> R$. This is really assuming that the only objects that exist at a given epoch are those that have only just reached the $\delta = \delta_c$ collapse threshold; if a point has $\delta > \delta_c$ for a given $R$, 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: PS therefore simply multiplying the probability by a factor 2. This fudge can be given some justification, but we just accept it for now. The fraction of the universe condensed into objects with mass $> M$ can then be written in the universal form

$$F(> M) = \sqrt{\frac{2}{\pi}} \int_{\nu}^{\infty} \exp(-\nu^2/2)\ d\nu, \tag{84}$$

where $\nu = \delta_c/\sigma(M)$ is the threshold in units of the rms density fluctuation.

Here, we have converted from spherical radius $R$ to mass $M$, using just

$$M = \frac{4\pi}{3}\,\bar{\rho}\,R^3. \tag{85}$$

In other words, $M$ is the mass contained in a sphere of comoving radius $R$ in a homogeneous universe. This is the linear-theory view, before the object has collapsed; its final vivialized radius will be $R/200^{1/3}$. The integral collapse 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

$$Mf(M)/\rho_0 = |dF/dM|, \tag{86}$$

where $\rho_0$ is the total comoving density. Thus,

$$\frac{M^2 f(M)}{\rho_0} = \frac{dF}{d\ln M} = \left|\frac{d\,\ln\sigma}{d\,\ln M}\right| \sqrt{\frac{2}{\pi}}\,\nu\,\exp\left(-\frac{\nu^2}{2}\right). \tag{87}$$

We have expressed the result in terms of the **multiplicity function**, $M^2 f(M)/\rho_0$, which is the fraction of the mass carried by objects in a unit range of $\ln M$.

Remarkably, given the dubious assumptions, this expression matches very well to what is found in direct N-body calculations, when these are analysed in order to pick out candidate haloes: connected groups of particles with density about 200 times the mean. The PS form is imperfect in detail, but the idea of a mass function that is universal in terms of $\nu$ seems to hold, and a good approximation is

$$F(>\nu) = (1 + a\,\nu^b)^{-1}\exp(-c\,\nu^2), \tag{88}$$

where $(a, b, c) = (1.529, 0.704, 0.412)$. Empirically, one can use $\delta_c = 1.686$ independent of the density parameter. A plot of the mass function according to this prescription is given in figure 2, assuming what we believe to be the best values for the cosmological parameters. This shows that the Press-Schechter formula captures the main features of the evolution, even though it is inaccurate in detail. We see that the richest clusters of galaxies, with $M \simeq 10^{15}\,h^{-1}M_\odot$, are just coming into existence now, whereas at $z = 5$ even a halo with the mass of the Milky Way, $M \simeq 10^{12}\,h^{-1}M_\odot$ was similarly rare. It can be seen that the abundance of low-mass haloes declines with redshift, reflecting their destruction in the merging processes that build up the large haloes.

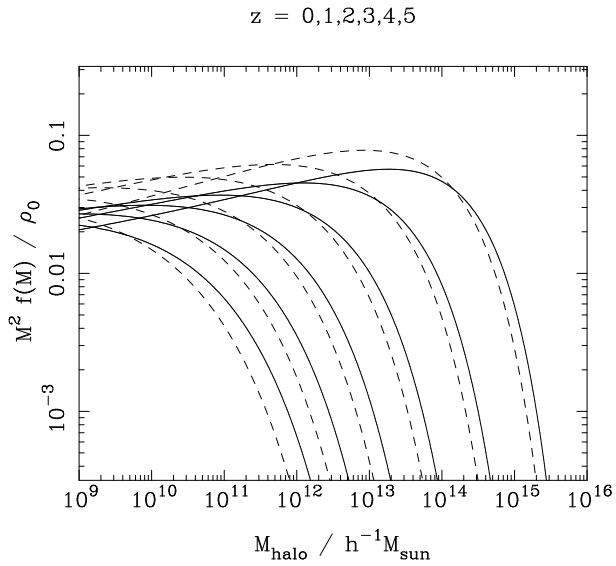


**Figure 2.** The mass function in the form of the **multiplicity function**: fraction of mass in the universe found in virialized haloes per unit range in $\ln M$. The solid lines show a fitting formula to $N$-body data and the dashed lines contrast the original Press-Schechter formula.

## 2.1 Recipes for galaxy formation

Since the appreciation in the 1970s that galaxies seemed to be embedded in haloes of dark matter, it has been clear that one should be able to construct an approximate theory for the assembly of galaxies based on the assumption that everything is dominated by the dark matter. Therefore, once we understand the history of the haloes, we should be able to make plausible guesses about how the baryonic material will behave. Over the years, this route has been followed to the point where there now exists an elaborate apparatus known as **semianalytic galaxy formation**. This is not yet a fully satisfactory theory, in that it is not able to make robust predictions of the properties that the galaxy population should have. However, it has succeeded in illuminating the main issues that need to be understood in a complete theory. In essence, semianalytic models include the following elements:

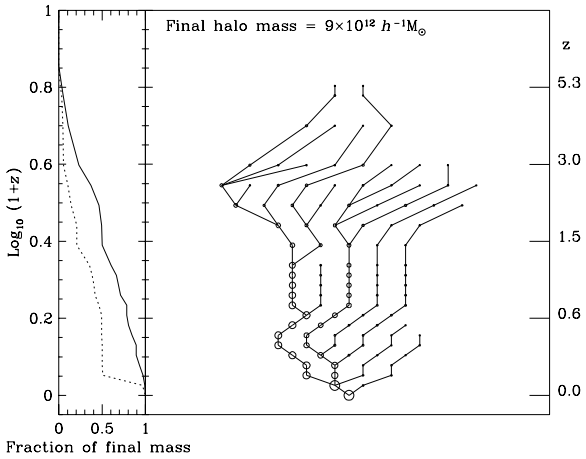


**Figure 3.** An example of a merger tree for a halo of $M \simeq 10^{13}\,M_\odot$ at $z = 0$, from Helly et al. (2002). The size of circle is proportional to halo mass, and the leftmost panel shows the fraction of the total mass in resolved progenitors (solid) and the mass of the largest progenitor (dashed).

(1) **Merger trees**. A halo that exists at a given time will have been constructed by the merging of smaller fragments over time. We need to be able to predict this history. This is most directly estimated from $N$-body results, although one can use approximate **extended Press-Schechter theory** (Bond et al. 1991).

(2) **Fate of subhaloes**. When haloes merge, they do not instantly lose their identity. Their cores survive as distinct subhaloes for some time. In group/cluster scale haloes, these will mark the locations of the galaxies. In general, subhaloes will eventually merge within the parent halo, and sink to the centre. Thus there is always a tendency to have a dominant central galaxy (e.g. the Milky Way is surrounded by the much smaller Magellanic Cloud dwarfs).

(3) **Accounting of gas and stars**. The first generation of haloes is assumed to start life with gas distributed along with the dark matter in the universal ratio $\Omega_b/\Omega_{dm}$. From the density of the gas, the cooling rate can be calculated. Whatever gas reaches a temperature below $10^4$ K is deemed to be a reservoir of cold gas suitable for

star formation. Some empirical relation based on the amount and density of this gas is then used to predict a star-formation rate. When haloes merge, their contents of stars, cold gas, and hot gas are added.

(4) **Feedback**. As we will show below, the above recipe fails to match observation, as it predicts that stars should form most efficiently in the smallest galaxies – so that a system of the size of the Milky Way should be just a collection of globular clusters, rather than predominately a giant gaseous disk. Therefore, the critical (and so far unsolved) problem in galaxy formation is to make gas cool less efficiently. The idea here is that energy is put back into the hot gas as a result of the nonlinear events that happen inside galaxies. These are principally of two kinds: supernova explosions and nuclear activity around a central black hole.

VIRIAL TEMPERATURE There is no time here to dig very deeply into the component parts of this recipe, but a few points are worth making. First consider the characteristic density of a virialized halo. We have argued that this is some multiple $f_c \simeq 200$ of the background density at virialization (or 'collapse'):

$$\rho_c = f_c\,\rho_0\,(1+z_c)^3. \tag{89}$$

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 $\rho_0 = 2.78\times10^{11}\Omega_m h^2\,M_\odot\mathrm{Mpc}^{-3}$, we obtain the **virial temperature**:

$$T_{\rm virial}/\mathrm{K} = 10^{5.1}(M/10^{12}M_\odot)^{2/3}\,(f_c\Omega_m h^2)^{1/3}\,(1+z_c). \tag{90}$$

This is an illuminating expression. It tells us that the most massive systems forming today, with $M \simeq 10^{15}M_\odot$, will have temperatures of $10^7$ – $10^8$ K. The intergalactic medium in clusters is thus very hot, and emits in X-rays. It also cools very inefficiently, since such hot plasmas emit only bremsstrahlung. Conversely, pregalactic systems with $M \lesssim 10^9 M_\odot$ at $z \simeq 10$ have a virial temperature that is barely at the level of $10^4$ K required for ionization. Their gas is thus predominately neutral, and should form stars with maximum efficiency. This is the cooling paradox referred to above.

But the same formula allows us to see how to escape from the paradox. The virial temperature is equivalent to a velocity dispersion, which is essentially the velocity at which particles orbit in the dark-matter potential well. This velocity therefore also gives the order of magnitude of the escape velocity for the system. Haloes with a virial temperature of only $\sim 10^4$ K thus constitute very shallow potential wells and will lose any of their gas that becomes heated to $\gtrsim 10^5$ K. This is liable to happen as soon as any supernovae from the first generation of star formation explode. For type II supernovae associated with massive stars, this can be virtually instantaneous ($\lesssim 10^7$ years). Star formation in these early dwarf galaxies might well be expected to be self-quenching. Indications that this process did happen can be found when measuring HI rotation curves of dwarfs: the typical baryon fraction is only about 1% (as opposed to something close to the global 20% in clusters).

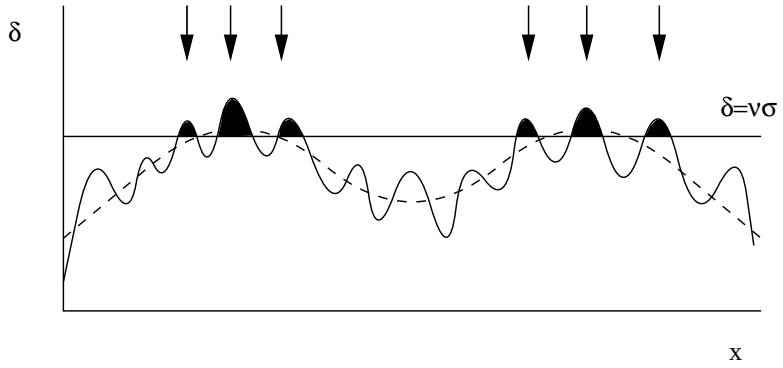


**Figure 4.** The high-peak bias model. If we decompose a density field into a fluctuating component on galaxy scales, together with a long-wavelength 'swell' (shown dashed), then those regions of density that lie above a threshold in density of $\nu$ times the rms will be strongly clustered. If proto-objects are presumed to form at the sites of these high peaks (shaded, and indicated by arrows), then this is a population with Lagrangian bias – i.e. a non-uniform spatial distribution even prior to dynamical evolution of the density field. The key question is the physical origin of the threshold; for massive objects such as clusters, the requirement of collapse by the present imposes a threshold of $\nu \gtrsim 2$. For galaxies, there will be no bias without additional mechanisms to cause star formation to favour those objects that collapse first.

## 2.2 Biased clustering and halo mass

In order to make full use of the cosmological information encoded in large-scale structure, it is essential to understand the relation between the number density of galaxies and the mass density field. It was first appreciated during the 1980s that these two fields need not be strictly proportional, starting with attempts to reconcile the $\Omega_m = 1$ Einstein–de Sitter model with observations. Although $M/L$ ratios in rich clusters argued for dark matter, as first shown by Zwicky (1933), typical blue values of $M/L \simeq 300h$ implied only $\Omega_m \simeq 0.2$ if they were taken to be universal. Those who argued that the value $\Omega_m = 1$ 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**.

An argument for bias at the opposite extreme of density arose through the discovery of large **voids** in the galaxy distribution (Kirshner et al. 1981). 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.

What seemed to be required was a galaxy correlation function that 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. (1985).

As shown below, 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**:

$$\left(\frac{\delta\rho}{\rho}\right)_{\rm galaxies} = b\left(\frac{\delta\rho}{\rho}\right)_{\rm mass}. \tag{91}$$

This seems a reasonable assumption when $\delta\rho/\rho \ll 1$, although it leaves open the question of how the effective value of $b$ would be expected to change on nonlinear scales. Galaxy clustering on large scales therefore allows us to determine mass fluctuations only if we know the value of $b$. When we observe large-scale galaxy clustering, we are only measuring $b^2\xi_{\rm mass}(r)$ or $b^2\Delta^2_{\rm mass}(k)$.

THE PEAK-BACKGROUND SPLIT We now consider the central mechanism of biased clustering, in which a rare high density fluctuation, corresponding to a massive object, collapses sooner if it lies in a region of large-scale overdensity. This 'helping hand' from the long-wavelength modes means that overdense regions contain an enhanced abundance of massive objects with respect to the mean, so that these systems display enhanced clustering. The basic mechanism can be immediately understood via the diagram in figure 4; it was first clearly analysed by Kaiser (1984) in the context of rich clusters of galaxies. What Kaiser did not do was consider the degree of bias that applies to more typical objects; the generalization to consider objects of any mass was made by Cole & Kaiser (1989; see also Mo & White 1996 and Sheth et al. 2001).

The key ingredient of this analysis is the mass function of dark-matter haloes. The universe fragments into virialized systems such that $f(M)\,dM$ is the number density of haloes in the mass range $dM$; conservation of mass requires that $\int M\,f(M)\,dM = \rho_0$. A convenient related dimensionless quantity is therefore the **multiplicity function**, $M^2 f(M)/\rho_0$, which gives the fraction of the mass of the universe contained in haloes of a unit range in $\ln M$. The simplest analyses of the mass function rest on the concept of a density threshold: collapse to a virialized object is deemed to have occurred where linear-theory $\delta$ averaged over a box containing mass $M$ reaches some critical value $\delta_c$. Generally, we shall assume the value $\delta_c = 1.686$ appropriate for spherical collapse in an Einstein–de Sitter universe. Now imagine that this situation is perturbed, by adding some constant shift $\epsilon$ to the density perturbations over some large region. The effect of this is to perturb the threshold: fluctuations now only need to reach $\delta = \delta_c - \epsilon$ in order to achieve collapse.

This change of threshold will increase the total collapse fraction on a given mass scale:

$$F \to F - (dF/d\delta_c)\,\epsilon. \tag{92}$$

Since $\nu = \delta_c/\sigma(M)$, $d/d\delta_c = (1/\sigma)(d/d\nu) = (\nu/\delta_c)(d/d\nu)$, this is

$$F \to F - (dF/d\ln\nu)\,\epsilon/\delta_c. \tag{93}$$

To find the impact on the clustering of objects at a given mass scale, differentiate again with respect to $\ln\nu$ and use the relation to the multiplicity function

$$M^2 f(M)/\rho_0 = (d\ln\nu/d\ln M)\,|dF/d\ln\nu|. \tag{94}$$

If we denote $-dF/d\ln\nu$ by the positive quantity $G$, then

$$G \to G + (dG/d\ln\nu)\,\epsilon/\delta_c. \tag{95}$$

This gives a bias in the number density of haloes in Lagrangian space: $\delta G/G = b_{\rm L}\epsilon$, where the Lagrangian bias is

$$b_{\rm L} = \frac{d\ln G}{d\ln \nu}/\delta_c. \tag{96}$$

In addition to this modulation of the halo properties, the large-scale disturbance will move haloes closer together where $\epsilon$ is large, giving a density contrast of $1+\epsilon$. If $\epsilon \ll 1$, the overall fractional density contrast of haloes is therefore the sum of the dynamical and statistical effects: $\delta_{\rm halo} = \epsilon + b_{\rm L}\epsilon$. The overall bias in Eulerian space ($b = \delta_{\rm halo}/\epsilon$) is therefore

$$b = 1 + \frac{d\ln G}{d\ln \nu}/\delta_c. \tag{97}$$

Of course, the field $\epsilon$ can hardly be imposed by hand; instead, we make the **peak-background split**, in which $\delta$ is mentally decomposed into a small-scale and a large-scale component – which we identify with $\epsilon$. The scale above which the large-scale component is defined does not matter so long as it lies between the sizes of collapsed systems and the scales at which we wish to measure correlations.

To apply this, we need an explicit expression for the mass function. The simplest alternative is the original expression of Press & Schechter (1974), which can be written in terms of the parameter $\nu = \delta_c/\sigma(M)$:

$$G(\nu) = \sqrt{\frac{2}{\pi}}\,\nu\,\exp\left(-\frac{\nu^2}{2}\right). \tag{98}$$

We now use $d/d\delta_c = \sigma(M)^{-1}(d/d\nu) = (\nu/\delta_c)(d/d\nu)$, since $M$ is not affected by the threshold change, which yields

$$b(\nu) = 1 + \frac{\nu^2 - 1}{\delta_c}. \tag{99}$$

This says that $M^*$ haloes are unbiased, low-mass haloes are antibiased and high-mass haloes are positively biased, eventually reaching the $b = \nu/\sigma$ value expected for high peaks. The corresponding expression can readily be deduced for more accurate fitting formulae for the mass function, such as that of Sheth & Tormen (1999):

$$G(\nu) = 0.21617[1 + (\sqrt{2}/\nu^2)^{0.3}]\,\nu\,\exp[-\nu^2/(2\sqrt{2})]. \tag{100}$$

As stated earlier, an even better approximation is

$$F(>\nu) = (1 + a\,\nu^b)^{-1}\exp(-c\,\nu^2), \tag{101}$$

where $(a, b, c) = (1.529, 0.704, 0.412)$

We can now understand the observation that Abell clusters are much more strongly clustered than galaxies in general: regions of large-scale overdensity contain systematically more high-mass haloes than expected if the haloes traced the mass. This phenomenon was dubbed **natural bias** by White et al. (1987). However, applying the idea to galaxies is not straightforward: we have shown that enhanced clustering is only expected for massive fluctuations with $\sigma \lesssim 1$, but galaxies at $z = 0$ fail this

criterion. The high-peak idea applies will at high redshift, where massive galaxies are still assembling, but today there has been time for galaxy-scale haloes to collapse in all environments. The large bias that should exist at high redshifts is erased as the mass fluctuations grow: if the Lagrangian component to the biased density field is kept unaltered, then the present-day bias will tend to unity as

$$b(\nu) = 1 + \frac{\nu^2 - 1}{(1 + z_f)\delta_c}. \tag{102}$$

(Fry 1986; Tegmark & Peebles 1998). Strong galaxy bias at $z = 0$ therefore requires some form of selection that locates present-day galaxies preferentially in the rarer haloes with $M > M^*$ (Kauffmann, Nusser & Steinmetz 1997).

This dilemma forced the introduction of the idea of **high-peak bias**: bright galaxies form only at the sites of high peaks in the initial density field (Bardeen et al. 1986; Davis et al. 1985). This idea is commonly, but incorrectly, attributed to Kaiser (1984), but it needs an extra ingredient, namely a non-gravitational threshold. Attempts were therefore made to argue that the first generation of objects could propagate disruptive signals, causing neighbours in low-density regions to be 'still-born'. It is then 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. 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). These difficulties were only removed when the standard model became a low-density universe, in which the dynamical argument for high galaxy bias no longer applied.

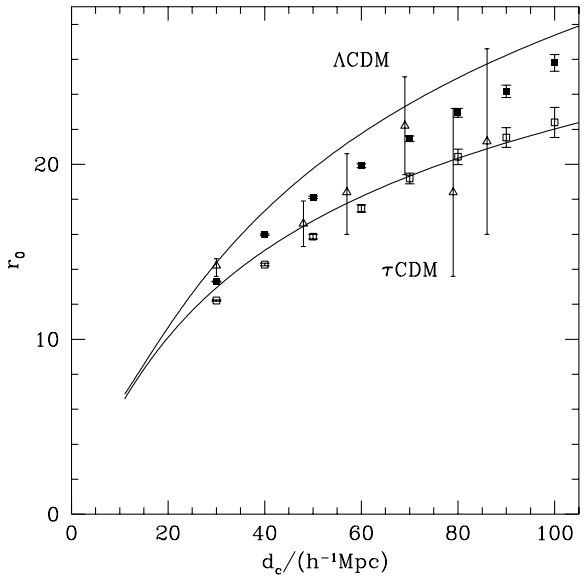


**Figure 5.** The correlation length for clusters of galaxies, $r_0$, as a function of mean intercluster separation, $d_c$, taken from Colberg et al. (2000). Results are shown for $\tau$CDM (open squares) and $\Lambda$CDM (filled squares) simulations. The predictions of Sheth et al. (2001) are shown as solid lines. Also shown are data from the APM cluster catalogue (open triangles), taken from Croft et al. (1997).

## 2.3 Observations of biased clustering

As indicated above, the first strong indications of biased clustering came from measurements of the correlation function of Abell clusters, which showed a far greater amplitude than for galaxies in general (Klypin & Kopylov 1983; Bahcall & Soneira 1983). Following Kaiser (1984), Cole & Kaiser (1989) etc., our explanation for this is that massive haloes show clustering that is an increasing function of mass. This is illustrated in figure 5, which shows that the rarest and most rich clusters (as measured by the intercluster separation) have the highest clustering, and that the trend is in agreement with the theoretical predictions.

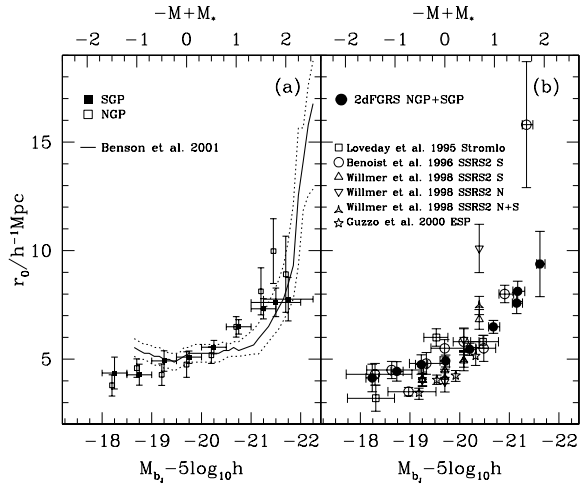


**Figure 6.** (a) The correlation length of galaxies in real space as a function of absolute magnitude. The solid line shows the predictions of the semi-analytic model of Benson et al. (2001), computed in a series of overlapping bins, each 0.5 magnitudes wide. The dotted curves show an estimate of the errors on this prediction, including the relevant sample variance for the survey volume. (b) The real space correlation length estimated combining the NGP and SGP (filled circles). The open symbols show a selection of recent data from other studies.

Because galaxy halo masses are less extreme, it is not so clear a priori that any trend of this sort should be expected for galaxies. However, our empirical knowledge of luminosity functions and morphological segregation did argue for an effect. It has been clear for many years that elliptical galaxies display a higher correlation amplitude than spirals (Davis & Geller 1976), and this makes sense in terms of the preference of ellipticals for cluster environments. Since ellipticals are also more luminous on average than spirals, some trend with luminosity is to be expected, but the challenge is to detect it. For a number of years, the existence of any effect was controversial (e.g. Loveday et al. 1995; Benoist et al. 1996), but Norberg et al. (2001) were able to use the 2dFGRS to demonstrate very clearly that the effect existed, as shown in Figure 6. The results can be described by a linear dependence of effective bias parameter on luminosity:

$$b/b^* = 0.85 + 0.15\,(L/L^*), \tag{103}$$

and the scale-length of the real-space correlation function for $L^*$ galaxies is approximately $r_0 = 4.8\,h^{-1}\,\mathrm{Mpc}$. Finally, with spectral classifications, it is possible to measure

the dependence of clustering both on luminosity and on spectral type, to see to what extent morphological segregation is responsible for this result. Norberg et al. (2002) show that, in fact, the principal effect seems to be with luminosity: $\xi(r)$ increases with $L$ for all spectral types.

Finally, we can look at high-redshift clustering. At high enough redshift, $M^*$ is of order a galaxy mass and galaxies could be strongly biased relative to the mass at that time. Indeed, there is good evidence that this is the case. Steidel *et al.* (1997) have used the Lyman-limit technique to select galaxies around redshifts $2.5 \lesssim z \lesssim 3.5$ and found their distribution to be highly inhomogeneous. The apparent value of $\sigma_8$ for these objects is of order unity (Adelberger et al. 1998), whereas the present value of $\sigma_8 \simeq 0.8$ should have evolved to about 0.26 at these redshifts (for $\Omega_m = 0.3$, $k = 0$). This suggests a bias parameter of $b \simeq 4$, or $\nu \simeq 2.5$, which requires a halo mass of about $10^{12.1} h^{-1} M_\odot$ for concordance ΛCDM. The masses of these high-redshift galaxies can be estimated directly through their stellar masses, which are typically $10^{10} h^{-2} M_\odot$ (Papovich, Dickinson & Ferguson 2001), and thus only 1% of what is required in order to explain the clustering. This is an unreasonably small baryon fraction, so the correct explanation is more plausibly that each $10^{12} h^{-1} M_\odot$ halo at $z = 3$ contains a number of Lyman-break galaxies. This theme is pursued below.

### 2.4 The halo model – I: mass

We can now assemble some of the above ingredients into heuristic model that captures the main processes at work in the full semianalytic models. The following section describes an approach of this sort (Peacock & Smith 2000; Seljak 2000; Cooray & Sheth 2002). The formation of galaxies must be a non-local process to some extent, and the modern paradigm was introduced by White & Rees (1978): galaxies form through the cooling of baryonic material in virialized haloes of dark matter. The virial radii of these systems are in excess of 0.1 Mpc, so there is the potential for large differences in the correlation properties of galaxies and dark matter on these scales. The 'halo model' addresses this by creating a density field in which dark-matter haloes are superimposed. The key feature that allows bias to be included is to encode all the complications of galaxy formation via the halo occupation number: the number of galaxies found above some luminosity threshold in a virialized halo of a given mass.

To some extent, this is a very old idea: one of the earliest suggested models for the galaxy correlation function was to consider a density field composed of randomly-placed independent clumps with some universal density profile (Neyman, Scott & Shane 1953; Peebles 1974). Since the clumps are placed at random (with number density $n$), the only excess neighbours to a given mass point arise from points in the same clump, and the correlation function is straightforward to compute in principle. For the case where the clumps have a power-law density profile,

$$\rho = nBr^{-\epsilon}, \tag{104}$$

truncated at $r = R$, the small-$r$ behaviour of the correlation function is $\xi \propto r^{3-2\epsilon}$, provided $3/2 < \epsilon < 3$. For smaller values of $\epsilon$, $\xi(r)$ tends to a constant as $r \to 0$. In the isothermal $\epsilon = 2$ case, the correlation function for $r \ll R$ is

$$\xi(r) = \frac{\pi^2 B}{4rR} = \frac{\pi N}{16 r R^2 n}, \tag{105}$$

where $N$ is the total number of particles per clump (Peebles 1974).

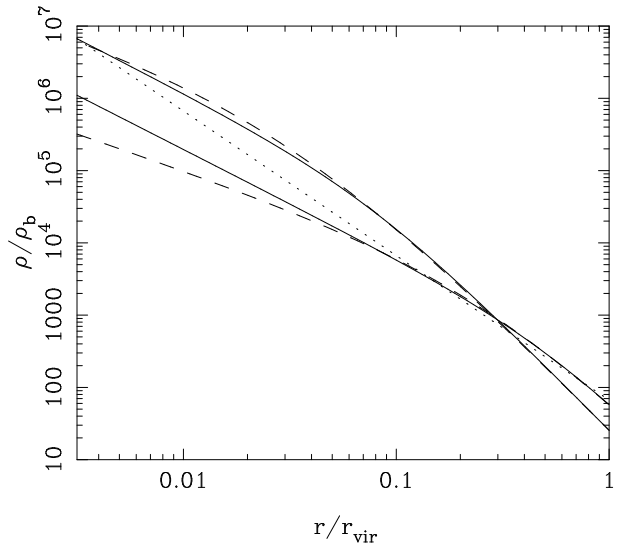


**Figure 7.** A comparison of various possible density profiles for virialized haloes. The dotted line is a singular isothermal sphere. The solid lines show haloes with formation redshifts of 0 and 5 according to NFW ($\Omega = 1$) and M99.

The general result is that the correlation function is less steep at small $r$ than the clump density profile, which is inevitable because an autocorrelation function involves convolving the density field with itself. A long-standing problem for this model is therefore that the predicted correlation function is much flatter than is observed for galaxies: $\xi \propto r^{-1.8}$ is the canonical slope, apparently requiring clumps with very steep density profiles, $\epsilon = 2.4$. This is not in agreement with the profiles of dark-matter haloes as 'observed' in numerical simulations.

Traditionally, virialized systems have been found by a criterion based on percolation ('friends-of-friends'), such that the mean density is about 200 times the mean. Sometimes, the criterion is taken as a density of 200 times the critical value. We shall use the former definition:

$$r_v = \left( \frac{3M}{800\pi\rho_b} \right)^{1/3}. \tag{106}$$

Thus $r_v$ is related to the Lagrangian radius containing the mass via $r_v = R/200^{1/3}$. Of course, the density contrast used to define the boundary of an object is somewhat arbitrary. Fortunately, much of the mass resides at smaller radii, near a 'core radius'. These core radii are relatively insensitive to the exact definition of virial radius.

The simplest model for the density structure of the virialized system is the singular isothermal sphere: $\rho = \sigma_v^2/(2\pi G r^2)$, or

$$\rho/\rho_b = \frac{200}{3y^2}; \quad (y < 1); \quad y \equiv r/r_v. \tag{107}$$

A more realistic alternative is the profile proposed by Navarro, Frenk & White (1996; NFW):

$$\rho/\rho_b = \frac{\Delta_c}{y(1+y)^2}; \quad (r < r_v); \quad y \equiv r/r_c. \tag{108}$$

The parameter $\Delta_c$ is related to the core radius and the virial radius via

$$\Delta_c = \frac{200c^3/3}{\ln(1+c) - c/(1+c)}; \quad c \equiv r_v/r_c \tag{109}$$

(we change symbol from NFW's $\delta_c$ to avoid confusion with the linear-theory density threshold for collapse, and also because our definition of density is relative to the mean, rather than the critical density). NFW showed that $\Delta_c$ is related to collapse redshift via

$$\Delta_c \simeq 3000(1+z_c)^3, \tag{110}$$

An advantage of the definition of virial radius used here is that $\Delta_c$ is independent of $\Omega$ (for given $z_c$), whereas NFW's $\delta_c$ is $\propto \Omega$.

The above equations determine the concentration, $c = r_v/r_c$ implicitly, hence in principle giving $r_c$ in terms of $r_v$ once $\Delta_c$ is known. NFW give a procedure for determining $z_c$. A simplified argument would suggest a typical formation era determined by $D(z_c) = 1/\nu$, where $D$ is the linear-theory growth factor between $z = z_c$ and the present, and $\nu$ is the dimensionless fluctuation amplitude corresponding to the system in units of the rms: $\nu \equiv \delta_c/\sigma(M)$, where $\delta_c \simeq 1.686$. For very massive systems with $\nu \gg 1$, only rare fluctuations have collapsed by the present, so $z_c$ is close to zero. This suggests the interpolation formula

$$D(z_c) = 1 + 1/\nu; \tag{111}$$

The NFW formula is actually of this form, except that the $1/\nu$ term is multiplied by a spectrum-dependent coefficient of order unity. It has been claimed by Moore et al. (1999; M99) that the NFW density profile is in error at small $r$. M99 proposed the alternative form

$$\rho/\rho_b = \frac{\Delta_c}{y^{3/2}(1+y^{3/2})}; \quad (r < r_v); \quad y \equiv r/r_c. \tag{112}$$

It is straightforward to use this in place of the NFW profile. In fact, a third expression $\rho \propto \exp(-(r/r_c)^\beta)$ with $\beta \simeq 0.25$ seems to work better than either of these, so that the asymptotic slope of the density profile is possibly not divergent.

We now compute the power spectrum for the halo model. Start by distributing point seeds throughout the universe with number density $n$, in which case the power spectrum of the resulting density field is just shot noise:

$$\Delta^2(k) = \frac{4\pi}{n}\left(\frac{k}{2\pi}\right)^3. \tag{113}$$

Here, we use a dimensionless notation for the power spectrum: $\Delta^2$ is the contribution to the fractional density variance per unit interval of $\ln k$. In the convention of Peebles (1980), this is

$$\Delta^2(k) \equiv \frac{\mathrm{d}\sigma^2}{\mathrm{d}\ln k} = \frac{V}{(2\pi)^3}\, 4\pi\, k^3\, |\delta_k|^2 \tag{114}$$

($V$ being a normalization volume), and the relation to the correlation function is

$$\xi(r) = \int \Delta^2(k)\, \frac{dk}{k}\, \frac{\sin kr}{kr}. \tag{115}$$

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

$$\Delta^2(k) = \frac{4\pi}{n}\, \left(\frac{k}{2\pi}\right)^3\, |W_k|^2, \tag{116}$$

where

$$W_k = \frac{\int \rho(r)\, \frac{\sin kr}{kr}\, 4\pi\, r^2\, dr}{\int \rho(r)\, 4\pi\, r^2\, dr}. \tag{117}$$

For a practical calculation, we should also use the fact that hierarchical models are expected to contain a distribution of masses of clumps. If we use the notation $f(M)\, dM$ to denote the number density of haloes in the mass range $dM$, the effective number density in the shot noise formula becomes

$$\frac{1}{n_{\mathrm{eff}}} = \frac{\int M^2\, f(M)\, dM}{\left[\int M\, f(M)\, dM\right]^2}. \tag{118}$$

The window function also depends on mass, so the overall power spectrum is

$$\Delta^2_{\mathrm{halo}}(k) = 4\pi\, \left(\frac{k}{2\pi}\right)^3\, \frac{\int M^2\, |W_k(M)|^2\, f(M)\, dM}{\left[\int M\, f(M)\, dM\right]^2}. \tag{119}$$

The normalization term $\int M\, f(M)\, dM$ just gives the total background density, $\rho_b$, so there is only a single numerical integral to perform. Using this model, it is then possible to calculate the correlations of the nonlinear density field, neglecting only the large-scale correlations in halo positions. The power spectrum determined in this way is shown in figure 8, and turns out to agree very well with the exact nonlinear result on small and intermediate scales. The lesson here is that a good deal of the nonlinear correlations of the dark matter field can be understood as a distribution of random clumps, provided these are given the correct distribution of masses and mass-dependent density profiles.

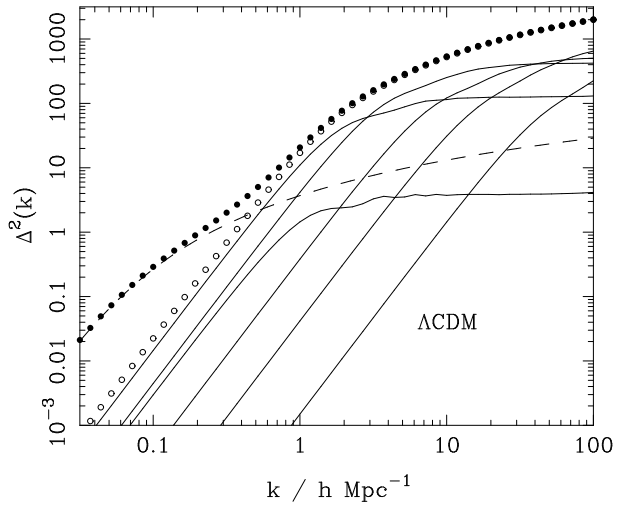


**Figure 8.** The decomposition of the mass power spectrum according to the halo model, for the flat $\Omega_m = 0.3$, $\Gamma = 0.2$, $\sigma_8 = 0.8$ case. The dashed line shows linear theory, and the open circles show the predicted 1-halo contribution. Adding in linear theory to produce the correct large-scale clustering yields the solid points. The full lines show the contribution of different mass ranges to the 1-halo term: bins of width a factor 10 in width, starting at $10^{10} - 10^{11}h^{-1}M_\odot$ and ending at $10^{15} - 10^{16}h^{-1}M_\odot$. The more massive haloes have larger virial radii and hence filter the power spectrum on progressively larger scales. The majority of the quasilinear power is contributed by the haloes near the peak in the mass function at $10^{14} - 10^{15}h^{-1}M_\odot$.

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

$$\Delta^2_{\rm tot} = \Delta^2_{\rm halo} + \Delta^2_{\rm linear}. \tag{120}$$

The justification for this is that the extra small-scale power introduced by nonlinear evolution is associated with the internal structure of the haloes. In practice, this model works extremely well, giving an almost perfect description of the power spectrum on all scales. This is a novel way of looking at the features in the nonlinear spectrum, particularly the steep rise between $k \simeq 0.5\,h\,{\rm Mpc}^{-1}$ and $k \simeq 5\,h\,{\rm Mpc}^{-1}$, and the flattening on smaller scales. According to the ideas presented here, the flat small-scale spectrum arises because haloes have central density profiles rising as $r^{-1.5}$, but not much faster. The sharp fall in power at smaller $k$ reflects the cutoff at the virial radii of the haloes that dominate the correlation signal.

It might be objected that this model is still not completely realistic, since we have treated haloes as smooth objects and ignored any substructure. At one time, it was generally believed that collisionless evolution would lead to the destruction of galaxy-scale haloes when they are absorbed into the creation of a larger-scale nonlinear system such as a group or cluster. However, it turns out that this 'overmerging problem' was only an artefact of inadequate resolution (see e.g. van Kampen 2000). When a simulation is carried out with $\sim 10^6$ particles in a rich cluster, the cores of galaxy-scale haloes can still be identified after many crossing times (Ghigna et al. 1998). This substructure must have some effect on the correlations of the density field, and indeed

Valageas (1999) has argued that the high-order correlations of the density field seen in $N$-body simulations are inconsistent with a model where the density file is composed of smooth virialized haloes. Nevertheless, substructure seems to be unimportant at the level of two-point correlations.

### 2.5 The Halo model – II: biased galaxy populations

In relating the distribution of galaxies to that of the mass, there are two distinct ways in which a degree of bias is inevitable:

(1) Halo occupation numbers. For low-mass haloes, the probability of obtaining an $L^*$ galaxy must fall to zero. For haloes with mass above this lower limit, the number of galaxies will in general not scale linearly with halo mass.

(2) Nonlocality. Galaxies can orbit within their host haloes, so the probability of forming a galaxy depends on the overall halo properties, not just the density at a point. Also, the galaxies can occupy special places within the haloes: for a halo containing only one galaxy, the galaxy will clearly mark the halo centre. In general, we will *assume* one central galaxy and a number of satellites.

The first mechanism leads to large-scale bias, because large-scale halo correlations depend on mass, and are some biased multiple of the mass power spectrum: $\Delta_h^2 = b^2(M)\Delta^2$. As discussed earlier, the linear bias parameter for a given class of haloes, $b(M)$, depends on the rareness of the fluctuation and the rms of the underlying field, as discussed above.

If we do not wish to assume that the number of galaxies in a halo of mass $M$ is strictly proportional to $M$, we are in effect giving haloes a mass-dependent weight, as was first considered by Jing, Mo & Börner (1998). This weight is just

$$w(M) \propto N(M)/M, \tag{121}$$

where $N(M)$ is the **halo occupation number** – i.e. the number of galaxies (above some limiting observational luminosity threshold). A simple but instructive model for the occupation number is

$$N(M) = \begin{cases} 0 & (M < M_c) \\ (M/M_c)^\alpha & (M > M_c) \end{cases} \tag{122}$$

A model in which mass traces light would have $M_c \to 0$ and $\alpha = 1$. We will show below that, empirically, we should choose $\alpha < 1$.

The bias formula applies to haloes of a given $\nu$, i.e. of a given mass, so the effect of mass-dependent weights is

$$b_{\rm tot} = 1 + \frac{\int_\nu^\infty b(\nu)\, w(\nu)\, \frac{dF}{d\nu}\, d\nu}{\int_\nu^\infty w(\nu)\, \frac{dF}{d\nu}\, d\nu,} \tag{123}$$

Where $F(>\nu)$ is the fraction of the mass in haloes exceeding a given $\nu$; $dF/d\nu \propto \exp(-\nu^2/2)$ according to Press-Schechter theory. The total model for the galaxy power spectrum is then

$$\Delta_g^2 = \left\langle \Delta_{\rm halo}^2 \right\rangle + b_{\rm tot}^2 \Delta_{\rm lin}^2 \tag{124}$$

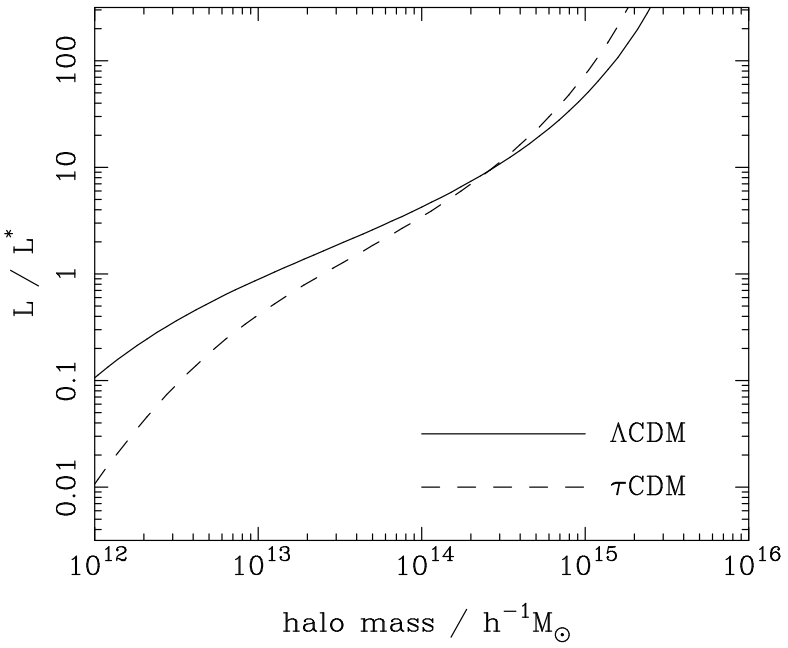


**Figure 9.** The empirical luminosity–mass relation required to reconcile the observed AGS luminosity function with two variants of CDM. $L^*$ is the characteristic luminosity in the AGS luminosity function ($L^* = 7.6 \times 10^{10} h^{-2} L_\odot$). Note the rather flat slope around $M = 10^{13}$ to $10^{14} h^{-1} M_\odot$, especially for ΛCDM.

where

$$\left\langle \Delta^2_{\rm halo}(k) \right\rangle = 4\pi \left(\frac{k}{2\pi}\right)^3 \frac{\int M^2\, w^2(M)\, |W_k(M)|^2\, f(M)\, dM}{\left[\int M\, w(M)\, f(M)\, dM\right]^2}. \tag{125}$$

The key ingredient needed to make this machinery work is the occupation number, which in principle needs to be calculated via a detailed numerical model of galaxy formation. However, for a given assumed background cosmology, the answer may be determined empirically. Galaxy redshift surveys have been analyzed via grouping algorithms similar to the 'friends-of-friends' method widely employed to find virialized clumps in $N$-body simulations. With an appropriate correction for the survey limiting magnitude, the observed number of galaxies in a group can be converted to an estimate of the total stellar luminosity in a group. This allows a determination of the All Galaxy System (AGS) luminosity function: the distribution of virialized clumps of galaxies as a function of their total luminosity, from small systems like the Local Group to rich Abell clusters.

The AGS function for the CfA survey was investigated by Moore, Frenk & White (1993), who found that the result in blue light was well described by

$$d\phi = \phi^* \left[(L/L^*)^\beta + (L/L^*)^\gamma\right]^{-1} dL/L^*, \tag{126}$$

where $\phi^* = 0.00126 h^3 {\rm Mpc}^{-3}$, $\beta = 1.34$, $\gamma = 2.89$; the characteristic luminosity is $L^* = 7.6 \times 10^{10} h^{-2} L_\odot$. One notable feature of this function is that it is rather flat at

low luminosities, in contrast to the mass function of dark-matter haloes (see Sheth & Tormen 1999). It is therefore clear that any fictitious galaxy catalogue generated by randomly sampling the mass is unlikely to be a good match to observation. The simplest cure for this deficiency is to assume that the stellar luminosity per virialized halo is a monotonic, but nonlinear, function of halo mass. The required luminosity–mass relation is then easily deduced by finding the luminosity at which the integrated AGS density $\Phi(>L)$ matches the integrated number density of haloes with mass $>M$. The result is shown in figure 9.

We can now calculate the halo-based galaxy power spectrum and use semi-realistic occupation numbers, $N$, as a function of mass. This is needs a little care at small numbers, however, since the number of haloes with occupation number unity affects the correlation properties. These haloes contribute no correlated pairs, so they simply dilute the signal from the haloes with $N \geq 2$. This means that we need in principle to use different weights for the large-scale bias and the halo term:

$$w_i = \frac{\langle N_i \rangle}{M} \qquad w_i = \frac{\langle N_i(N_i-1)\rangle^{1/2}}{M} \tag{127}$$

respectively (Seljak 2000). In practice, this correction has a rather small effect, provided the relation between $N$ and $M$ has no scatter. If, in contrast, the distribution of $N$ for given $M$ is assumed to obey a Poisson distribution, the small-scale clustering properties are strongly affected, and do not match the data well (Benson et al. 2000a).

The overall result of this exercise is that the shape of the galaxy spectrum is expected to differ from that of the mass; it can be very close to a power law, which has been a long-standing puzzle to explain. Nevertheless, the power-law should not be perfect; small deviations have long been suspected, and were confirmed by Hawkins et al. (2002) and Zehavi et al. (2003). The inflection is at a scale of $\sim 0.5\,h\,\mathrm{Mpc}^{-1}$, as expected from the halo model.

# 3 Bibilography

With apologies, these notes do not include a consistent bibliography. The papers referred to above can be located efficiently by searching on ADS: `adswww.harvard.edu/ads_abstracts.html`