The Abdus Salam
International Centre for Theoretical Physics

United Nations
Educational, Scientific and
Cultural Organization

IAEA
International Atomic Energy Agency

2156-5

Summer School in Cosmology

19 - 30 July 2010

Introduction to Cosmology

John Andrew PEACOCK

University of Edinburgh
Institute for Astronomy Royal Observatory
Blackford Hill, Edinburgh EH9 3HJ
U.K.

Introduction to Cosmology

ICTP Cosmology School; Trieste, July 2010

J.A. Peacock

Institute for Astronomy, University of Edinburgh

jap@roe.ac.uk http://www.roe.ac.uk/japwww

Outline

(1) **Spacetime in an expanding universe** FRW spacetime; Dynamics; Horizons; Observables

(2) **Structure formation** Inhomogeneities and Spherical model; Lagrangian approach; N-body simulations; Dark-matter haloes & mass function

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

(4) **Frontiers** Initial condition problems; Vacuum energy; anthropics and the multiverse

Useful textbooks:

Dodelson: Modern Cosmology (Wiley)
Kolb & Turner: The Early Universe (addison-Wesley)
Lyth & Liddle: The Primordial Density Perturbation (CUP) **Mukhanov**: Physical Foundations of Cosmology (CUP)
Peacock: Cosmological Physics (CUP)
Peebles: Principles of Physical Cosmology (Princeton)
Weinberg: Gravitation & Cosmology (Wiley)
Weinberg: Cosmology (Oxford)

A very impressive web tutorial by Ned Wright may be helpful:
`http://www.astro.ucla.edu/~wright/cosmolog.htm`

1 Spacetime in an expanding universe

These lectures concern the modern view of the overall properties of the universe. The heart of this view is that the universe is a dynamical entity that has existed for only a finite period, and which reached its present state by evolution from initial conditions that are violent almost beyond belief. Speculation about the nature of creation is older than history, of course, but the present view was arrived at only rather recently. A skeptic might therefore say that our current ideas may only be passing fashions. However, we are bold enough to say that something is now really understood of the true nature of space and time on the largest scales. This is not to claim that we are any brighter than those who went before; merely that we are fortunate enough to live when technology has finally revealed sufficient details of the universe for us to make a constrained theory. The resulting theory is strange, but it has been forced on us by observational facts that will not change.

The first key observation of the modern era was the discovery of the expanding universe. This is popularly credited to Edwin Hubble in 1929, but the honour arguably lies with Vesto Slipher, more than 10 years earlier. Slipher was measuring spectra of **nebulae**, and at that time there was a big debate about what they were. Some thought that these extended blobs of light were clouds of gas, some thought they were systems of stars at great distance. We now know that there are some of each, but stellar systems are in the majority away from the plane of the Milky Way. This was finally settled only in 1924, when Hubble discovered Cepheid variable stars in M31, establishing its distance of roughly 1 Mpc. More than a decade earlier, in 1913, Slipher had measured the spectrum of M31, and found that it was approaching the Earth at over $200\,\mathrm{km\,s^{-1}}$. Strangely, Slipher had the field to himself for another decade, by which time he had measured Doppler shifts for dozens of galaxies: with only a few exceptions, these were redshifted. Furthermore, there was a tendency for the redshift to be larger for the fainter galaxies. By the time Hubble came on the scene, the basics of relativistic cosmology were worked out and predictions existed that redshift should increase with distance. It is hard to know how much these influenced Hubble, but by 1929 he had obtained Cepheid distances for 24 galaxies with redshifts and claimed that these displayed a linear relationship:

$$v = Hd, \tag{1}$$

citing theoretical predictions as a possible explanation. At the time, Hubble estimated $H \simeq 500\,\mathrm{km\,s^{-1}Mpc^{-1}}$, in part because his calibration of Cepheid luminosities was in error. The best modern value is close to $70\,\mathrm{km\,s^{-1}Mpc^{-1}}$.

1.1 The scale factor

A very simple model that yields Hubble's law is what might be called the **grenade universe**: at time $t = 0$, set off a grenade in a big empty space. Different bits of debris fly off at different speeds, and at time t will have reached a distance $d = vt$. This is Hubble's law, with $H = 1/t$. We may therefore suspect that there was a violent event at a time about $1/H$ ago. This event is basically what we mean by the **big bang**: an origin of the expansion at a finite time in the past. The characteristic time of the expansion is called the **Hubble time**, and takes the value

$$t_{\rm H} \equiv 9.78\,{\rm Gyr}\ \times\ (H/100\,{\rm km\,s^{-1}Mpc^{-1}})^{-1}. \tag{2}$$

As we shall see, this is not the actual age of the universe, since gravity stops the expansion proceeding at uniform speed.

The grenade universe is a neat idea, but it can leave you with a seriously flawed view of the universe. First, the model has a centre, where we are presumed to live; second, the model has an edge – and the expansion proceeds to fill empty space. The real situation seems to be that we do not live in a special place, nor is there an edge to the galaxy distribution.

It is easy enough to think of an alternative model, in which the Earth need not be at the centre of the universe. Consider a distribution of galaxies that is made to expand uniformly, in the same way as if a picture of the pattern was undergoing continuous magnification. Mathematically, this means that all position vectors at time t are just scaled versions of their values at a reference time t_0:

$$\mathbf{x}(t) = R(t)\mathbf{x}(t_0). \tag{3}$$

Differentiating this with respect to t gives

$$\dot{\mathbf{x}}(t) = \dot{R}(t)\mathbf{x}(t_0) = [\dot{R}(t)/R(t)]\,\mathbf{x}(t), \tag{4}$$

or a velocity proportional to distance, as required. Writing this relation for two points 1 & 2 and subtracting shows that this expansion appears the same for any choice of origin: everyone is the centre of the universe:

$$[\dot{\mathbf{x}}_2(t) - \dot{\mathbf{x}}_1(t)] = H(t)\,[\mathbf{x}_2(t) - \mathbf{x}_1(t)]; \qquad H(t) = \dot{R}(t)/R(t). \tag{5}$$

This shows that Hubble's constant can be identified with $\dot{R}(t)/R(t)$, and that in general it is not a constant, but something that can change with time.

1.2 Cosmological spacetime

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

$$-ds^2 = c^2 d\tau^2 = c^2 dt'^2 - dx'^2 - dy'^2 - dz'^2 = g_{\mu\nu} dx^\mu dx^\nu. \tag{6}$$

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

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

$$x^2 + y^2 + z^2 + w^2 - v^2 = A^2 \tag{7}$$

where the metric is

$$ds^2 = dx^2 + dy^2 + dz^2 + dw^2 - dv^2. \tag{8}$$

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

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

$$\begin{aligned} v &= A \sinh \alpha \\ w &= A \cosh \alpha \cos \beta \\ z &= A \cosh \alpha \sin \beta \cos \gamma \\ y &= A \cosh \alpha \sin \beta \sin \gamma \cos \delta \\ x &= A \cosh \alpha \sin \beta \sin \gamma \sin \delta. \end{aligned} \tag{9}$$

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

$$ds^2 = -A^2 d\alpha^2 + A^2 \cosh^2 \alpha \left(d\beta^2 + \sin^2(\beta)[d\gamma^2 + \sin^2 \gamma d\delta^2] \right), \tag{10}$$

which by an obvious change of notation becomes

$$c^2 d\tau^2 = c^2 dt^2 - A^2 \cosh^2(t/A) \left(dr^2 + \sin^2(r)[d\theta^2 + \sin^2 \theta d\phi^2] \right). \tag{11}$$

Now we have a completely different interpretation of the metric:

$$(\text{interval})^2 = (\text{time interval})^2 - (\text{scale factor})^2 (\text{comoving interval})^2. \tag{12}$$

There is a universal **cosmological time**, which is the ticking of clocks at constant **comoving radius** r and constant angle on the sky. The spatial part of the metric expands with time, so that particles at constant r recede from the origin, and must thus suffer a Doppler redshift. This of course presumes that constant r corresponds to the actual trajectory of a free particle, which we have not proved – although it is true.

Historically, de Sitter space was extremely important in cosmology, although it was not immediately clear that the model is non-static. The metric was first derived by de Sitter in the following form:

$$c^2 d\tau^2 = (1 - r'^2/\mathcal{R}^2)\, c^2 dt'^2 - (1 - r'^2/\mathcal{R}^2)^{-1} dr'^2 - r'^2\, d\psi^2, \tag{13}$$

where now r is a proper radius, and $\mathcal{R}$ is a curvature radius. To show that this is the same metric, a fair bit of tedious algebra is required. The starting point is to identify the transverse parts involving $d\psi^2$, so that $r' = A\cosh(t/A)\sin r$, which can be differentiated to yield $dr' = \sinh(t/A)\sin r\ dt + A\cosh(t/A)\cos r\ dr$. Squaring this yields terms that include an undesired cross term involving $dr\ dt$. This can be eliminated by writing $dt' = a\ dt + b\ dr$, where a & b are two spacetime functions. With the right choice of a & b, we get the static de Sitter metric, with $\mathcal{R} = cA$.

It is not at all obvious that there is anything expanding about the second form, and historically this remained obscure for some time. Although it was eventually concluded (in 1923, by Weyl) that one would expect a redshift that increased linearly with distance in de Sitter's model, this was interpreted as measuring the constant radius of curvature of spacetime, $\mathcal{R}$. By this time, Slipher had already established that most galaxies were redshifted. Hubble's 1929 'discovery' of the expanding universe was explicitly motivated by the possibility of finding the 'de Sitter effect' (although we now know that his sample was too shallow to be able to detect it reliably).

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

1.3 The Robertson-Walker metric

The de Sitter model is only one example of an isotropically expanding spacetime, and we need to make the idea general, which involves weakening the symmetry – but only slightly.

FUNDAMENTAL OBSERVERS As with de Sitter space, we assume a **cosmological time** t, which is the time measured by the clocks of these observers – *i.e.* t is the proper time measured by an observer at rest with respect to the local matter distribution (these characters are

usually termed **fundamental observers**). We can envisage them as each sitting on a different galaxy, and so receding from each other with the general expansion. Actually this is not quite right, since each galaxy has a **peculiar velocity** with respect to its neighbours of a few hundred $\mathrm{km\,s^{-1}}$. We really need to deal with an idealized universe where the matter density is uniform.

The fundamental observers give us a way of defining a universal time coordinate, even though relativity tells us that such a thing is impossible in general. It makes sense that such a universal time exists if we accept that we are looking for models that are **homogeneous**, so that there are no preferred locations. This is obvious in de Sitter space: because it derives from a 4-sphere, all spacetime points are manifestly equivalent: the spacetime curvature and hence the matter density must be a constant. The next step is to to weaken this so that conditions can change with time, but are uniform at a given time. A cosmological time coordinate can then be defined and synchronized by setting clocks to a reference value at some standard density.

ISOTROPY AND HOMOGENEITY So far, cosmological time is not very useful, since it is not so easy to arrange to sychronize all the clocks of the different observers. The way this problem is solved is because we will consider mass distributions with special symmetries. The Hubble expansion that we see is **isotropic** – the same in all directions. Also, all large-scale properties of the universe such as the distribution of faint galaxies on the sky seem to be accurately isotropic. If this is true for us, we can make a plausible guess, called the **cosmological principle**: that conditions will be seen as isotropic around each observer. If this holds (and it can be checked observationally, so it's not just an article of religious faith), then we can prove that the mass distribution must be **homogeneous** – *i.e.* the same density everywhere at a given time. The proof is very easy: just draw a pair of intersecting spheres about two observers. The density on each sphere is a constant by isotropy, and it must be the same constant since they intersect.

Homogeneity is what allows cosmological time to be useful globally rather than locally: because the clocks can be synchronized if observers set their clocks to a standard time when the universal homogeneous density reaches some given value.

METRICS ON SPHERES We now need a way of describing the global structure of space and time in such a homogeneous space. Locally, we have said that things look like special relativity to a fundamental observer on the spot: for them, the proper time interval between two events is $c^2 d\tau^2 = c^2 dt^2 - dx^2 - dy^2 - dz^2$. Since we use the same time coordinate as they do, our only difficulty is in the spatial part of the metric: relating their dx *etc.* to spatial coordinates centred on us.

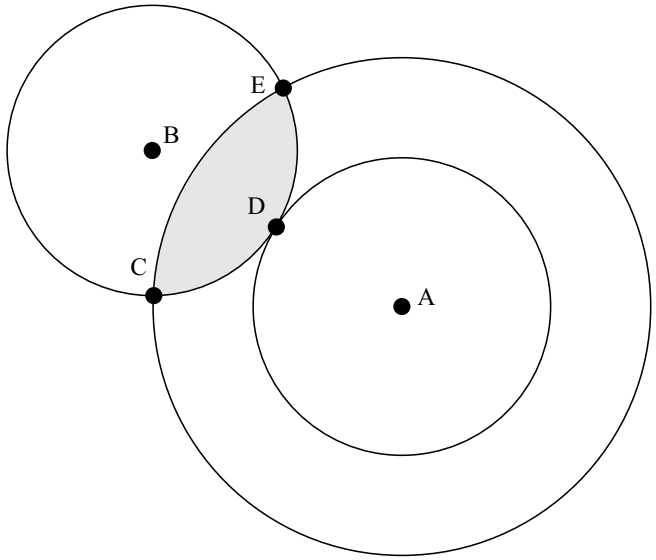

Figure 1. Isotropy about two points A and B shows that the universe is homogeneous. From isotropy about B, the density is the same at each of C,D,E. By constructing spheres of different radii about A, the shaded zone is swept out and shown to be homogeneous. By using large enough shells, this argument extends to the entire universe.

Using isotropy, we already have enough information to conclude that the metric must take the following form:

$$c^2 d\tau^2 = c^2 dt^2 - R^2(t) \left[f^2(r)\, dr^2 + g^2(r)\, d\psi^2 \right]. \tag{14}$$

Because of spherical symmetry, the spatial part of the metric can be decomposed into a radial and a transverse part (in spherical polars, $d\psi^2 = d\theta^2 + \sin^2\theta\, d\phi^2$). Distances have been decomposed into a product of a time-dependent **scale factor** $R(t)$ and a time-independent **comoving coordinate** r. The functions f and g are arbitrary; however, we can choose our radial coordinate such that either $f = 1$

or $g = r^2$, to make things look as much like Euclidean space as possible. The problem is solved if we can only determine the form of the remaining function.

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

$$d\sigma^2 = R^2 \left(dr^2 + \sin^2 r \, d\phi^2 \right). \tag{15}$$

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

$$d\sigma^2 = R^2 \left(dr^2 + \sinh^2 r \, d\phi^2 \right). \tag{16}$$

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

$$d\sigma^2 = R^2 \left(\frac{dr^2}{1 - kr^2} + r^2 \, d\phi^2 \right), \tag{17}$$

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

An isotropic universe has the same form for the comoving spatial part of its metric as the surface of a sphere. This is no accident, since it it possible to define the equivalent of a sphere in higher numbers of dimensions, and the form of the metric is always the same. Let's start with the case of the surface of a sphere, supposing that we were ants, with no concept of the third dimension away from the surface of the sphere. A higher-dimensional generalization of the circle, $x^2 + y^2 = R^2$, would be Pythagoras with one extra coordinate:

$$x^2 + y^2 + z^2 = R^2. \tag{18}$$

We can always satisfy this by defining some angles:

$$\begin{aligned} z &= R\cos\theta \\ y &= R\sin\theta\sin\phi \\ x &= R\sin\theta\cos\phi. \end{aligned} \tag{19}$$

3D beings recognize these as the usual polar angles, but we don't need this insight – other angles could have been defined that would work just as well. An element of length in this Euclidean space will be

$$d\sigma^2 = dx^2 + dy^2 + dz^2, \tag{20}$$

and we can express this in terms of the angles using

$$\begin{aligned} dx &= R(\cos\theta\cos\phi\, d\theta - \sin\theta\sin\phi\, d\phi) \\ dy &= R(\cos\theta\sin\phi\, d\theta + \sin\theta\cos\phi\, d\phi) \\ dz &= R(-\sin\theta\, d\theta). \end{aligned} \tag{21}$$

Multiplying everything out, we get $d\sigma^2 = R^2(d\theta^2 + \sin^2\theta\, d\phi^2)$. This is the expected result, but no geometrical insight was required beyond the element of length in Euclidean space.

Moving up one dimension, a **3-sphere** embedded in four-dimensional Euclidean space would be defined as the coordinate relation $x^2 + y^2 + z^2 + w^2 = R^2$. Now define the equivalent of spherical polars and write $w = R\cos\alpha$, $z = R\sin\alpha\cos\beta$, $y = R\sin\alpha\sin\beta\cos\gamma$, $x = R\sin\alpha\sin\beta\sin\gamma$, where α, β and γ are three arbitrary angles. Differentiating with respect to the angles gives a four-dimensional vector (dx, dy, dz, dw), and we need the modulus of this vector. We could work this out by brute force as before, by evaluating $dx^2+dy^2+dz^2+dw^2$; a slightly easier route is to consider the vectors generated by increments in $d\alpha$ *etc.*:

$$\begin{aligned} \mathbf{e}_\alpha &= d\alpha\, R\,(\,\cos\alpha\sin\beta\sin\gamma,\ \cos\alpha\sin\beta\cos\gamma,\ \cos\alpha\cos\beta,\ -\sin\alpha\,) \\ \mathbf{e}_\beta &= d\beta\, R\,(\,\sin\alpha\cos\beta\sin\gamma,\ \sin\alpha\cos\beta\cos\gamma,\ -\sin\alpha\sin\beta,\ 0\,) \\ \mathbf{e}_\gamma &= d\gamma\, R\,(\,\sin\alpha\sin\beta\cos\gamma,\ -\sin\alpha\sin\beta\sin\gamma,\ 0,\ 0\,). \end{aligned} \tag{22}$$

These are easily checked to be orthogonal, so the squared length of the vector is just $|\mathbf{e}_\alpha|^2 + |\mathbf{e}_\beta|^2 + |\mathbf{e}_\gamma|^2$, which gives

$$|(dx, dy, dz, dw)|^2 = R^2 \left[d\alpha^2 + \sin^2\alpha \, (d\beta^2 + \sin^2\beta \, d\gamma^2) \right] . \tag{23}$$

This is the metric for the case of positive spatial curvature, if we relabel $\alpha \to r$ and $(\beta, \gamma) \to (\theta, \phi)$ – the usual polar angles. We could write the angular part just in terms of the angle $d\psi$ that separates two points on the sky, $d\psi^2 = d\theta^2 + \sin^2\theta \, d\phi^2$, in which case the metric is the same form as for the surface of a sphere. This was inevitable: the hypersurface $x^2 + y^2 + z^2 + w^2 = R^2$ always allows two points to be chosen to have $w = 0$ (the first by choice of origin; the second via rotation), so that their separation is that of two points on the surface of a sphere.

It is possible to deal with the 3-sphere without introducing angles via the following trick: write $x^2+y^2+z^2+w^2 = R^2$ as $r^2+w^2 = R^2$, so that $w\, dw = -r\, dr$, implying $dw^2 = r^2 dr^2/(R^2 - r^2)$. The spatial part of the metric is therefore just

$$d\sigma^2 = dx^2 + dy^2 + dz^2 + r^2 dr^2/(R^2 - r^2). \tag{24}$$

Introducing 3D polar coordinates, we have

$$dx^2 + dy^2 + dz^2 = dr^2 + r^2 \left(d\theta^2 + \sin^2\theta \, d\phi^2 \right) , \tag{25}$$

so that we get the spatial part of the RW metric in its second form:

$$d\sigma^2 = \frac{dr^2}{1 - r^2/R^2} + r^2 \left(d\theta^2 + \sin^2\theta \, d\phi^2 \right) . \tag{26}$$

To convert to the form we had previously, we should replace r here by $R \sin r$. In making this argument, remember the subtlety that (x, y, z) are coordinates in the embedding space, which differ from the coordinates that a 3D observer would erect in their vicinity; to clarify things, imagine how the same arguments work for a sphere, where the two points of interest can always be chosen to lie along a great circle.

This $k = +1$ metric describes a **closed universe**, in which a traveller who sets off along a trajectory of fixed β and γ will eventually return to their starting point (when $\alpha = 2\pi$). In this respect, the positively curved 3D universe is identical to the case of the surface of a sphere: it is finite, but unbounded. By contrast, if we define a space of negative curvature via $R \to iR$ and $\alpha \to i\alpha$, then $\sin\alpha \to i \sinh\alpha$

and $\cos\alpha \to \cosh\alpha$ (so that x, y, z stay real, as they must). The new angle α can increase without limit, and (x, y, z) never return to their starting values. The $k = -1$ metric thus describes an **open universe** of infinite extent.

SUMMARY OF THE RW METRIC We can now sum up the overall metric, where the time part just comes from cosmological time: $c^2 d\tau^2 = c^2 dt^2 - d\sigma^2$. The result is the Robertson–Walker metric (**RW metric**), which may be written in a number of different ways. The most compact forms are those where the comoving coordinates are *dimensionless*. Define the very useful function

$$S_k(r) = \begin{cases} \sin r & (k = 1) \\ \sinh r & (k = -1) \\ r & (k = 0). \end{cases} \tag{27}$$

The metric can now be written in the preferred form that we shall use throughout:

$$c^2 d\tau^2 = c^2 dt^2 - R^2(t) \left[dr^2 + S_k^2(r)\, d\psi^2 \right]. \tag{28}$$

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

$$c^2 d\tau^2 = c^2 dt^2 - R^2(t) \left(\frac{dr^2}{1 - kr^2} + r^2 d\psi^2 \right). \tag{29}$$

There should of course be two different symbols for the different comoving radii, but each is often called r in the literature. We will normally stick with the first form. Alternatively, one can make the scale factor dimensionless, defining

$$a(t) \equiv \frac{R(t)}{R_0}, \tag{30}$$

so that $a = 1$ at the present.

Lastly, note that, although comoving distance is dimensionless in the above conventions, it is normal in the cosmological literature to discuss comoving distances with units of length (*e.g.* Mpc). This is because one normally considers the combination $R_0 r$ or $R_0 S_k(r)$ – *i.e.* these are the proper lengths that correspond to the given comoving separation at the current time.

1.4 Light propagation and redshift

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

$$r = \int c\,dt/R(t). \tag{31}$$

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

$$\frac{\nu_{\rm emit}}{\nu_{\rm obs}} \equiv 1 + z = \frac{R(t_{\rm obs})}{R(t_{\rm emit})}. \tag{32}$$

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

Photon wavelengths therefore stretch with the universe, as may seem intuitively reasonable. We can prove this more directly, as follows. Suppose we send a photon, which travels for a time δt until it meets another observer, at distance $d = c\,\delta t$. The recessional velocity of this galaxy is $\delta v = Hd$, so there is a fractional redshift:

$$\delta\nu\,/\,\nu = \delta v/c = -(Hd)/c = -H\delta t. \tag{33}$$

Now, since $H = \dot{R}/R$, this becomes

$$\delta\nu\,/\,\nu = -\delta R\,/\,R, \tag{34}$$

which integrates to give the main result: $\nu \propto 1/R$. The redshift is thus the accumulation of a series of infinitesimal Doppler shifts as the photon passes from observer to observer.

However, this is not the same as saying that the redshift tells us how fast the observed galaxy is receding. A common but incorrect approach is to use the special-relativistic Doppler formula and write

$$1 + z = \sqrt{\frac{1 + v/c}{1 - v/c}}. \tag{35}$$

Indeed, it is all too common to read of the latest high-redshift quasar as "receding at 95% of the speed of light". The reason the redshift cannot be interpreted in this way is because a non-zero mass density must cause gravitational redshifts. If we want to think of the redshift globally, it is better to stick with the ratio of scale factors. However, one can get somewhere by considering the distance-redshift relation to quadratic order. At this level, spacetime curvature is unimportant: r and $S_k(r)$ differ only at order r^3, so we are spared the difficult question as to which is the better definition of distance. Now consider a photon that we receive from distance d: drawing a sphere around us, it is clear that this photon falls towards us through a gravitational potential GM/d, which is proportional to d^2. Thus there is a gravitational *blueshift* contribution to z that is quadratic in distance.

Finally, note that the law that frequency of photons scales as $1/R$ actually applies to the momentum of all particles – relativistic or not: we just need to consider applying a small boost to the 4-momentum P^μ and remember that P/E is the particle velocity. Thinking of quantum mechanics, the de Broglie wavelength is $\lambda = 2\pi\hbar/p$, so this scales with the side of the universe, as if the waves were standing waves trapped in a box (see figure 2).

1.5 Cosmological dynamics

Thus almost everything in cosmology reduces to knowing $R(t)$. The equation of motion for the scale factor is something that can nearly be understood by a Newtonian approach, and this is the approach we shall try to adopt, but it is worth outlining the steps needed in a more rigorous approach.

What we need to do is insert the RW metric, $g^{\mu\nu}$ into Einstein's field equations:

$$G^{\mu\nu} \equiv R^{\mu\nu} - Rg^{\mu\nu}/2 = -(8\pi G/c^4)T^{\mu\nu}, \tag{36}$$

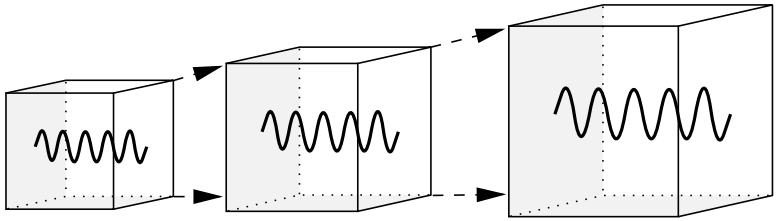

Figure 2. Suppose we trap some radiation inside a box with silvered sides, that expands with the universe. At least for an isotropic radiation field, the photons trapped in the box are statistically equivalent to any that would pass into this space from outside. Since the walls expand at $v \ll c$ for a small box, it is easily shown that the Doppler shift maintains an adiabatic invariant, which is the ratio of wavelength to box side, and so radiation wavelengths increase as the universe expands. This argument also applies to quantum-mechanical standing waves: momentum declines as $a(t)^{-1}$.

where the Ricci tensor and scalar derive from the Riemann tensor

$$R^{\mu}{}_{\alpha\beta\gamma} = \frac{\partial \Gamma^{\mu}_{\alpha\gamma}}{\partial x^{\beta}} - \frac{\partial \Gamma^{\mu}_{\alpha\beta}}{\partial x^{\gamma}} + \Gamma^{\mu}_{\sigma\beta}\Gamma^{\sigma}_{\gamma\alpha} - \Gamma^{\mu}_{\sigma\gamma}\Gamma^{\sigma}_{\beta\alpha}. \tag{37}$$

via

$$\begin{aligned} R_{\alpha\beta} &= R^{\mu}{}_{\alpha\beta\mu} \\ R &= R_{\mu}{}^{\mu} = g^{\mu\nu}R_{\mu\nu} \end{aligned} \tag{38}$$

(other conventions are possible). The components of the connection are

$$\Gamma^{\alpha}_{\lambda\mu} = \tfrac{1}{2} g^{\alpha\nu}\left(\frac{\partial g_{\mu\nu}}{\partial x^{\lambda}} + \frac{\partial g_{\lambda\nu}}{\partial x^{\mu}} - \frac{\partial g_{\mu\lambda}}{\partial x^{\nu}}\right). \tag{39}$$

So in principle we just have to perform the necessary differentiations, remembering clearly that the coordinates being used are not Cartesian:

$$x^\mu = (ct, r, \theta, \phi). \tag{40}$$

But what is the energy-momentum tensor, $T^{\mu\nu}$? For an isotropic fluid, it must be $T^{\mu\nu} = (\rho + p)U^\mu U^\nu - pg^{\mu\nu}$. Since three of our coordinates are unchanging comoving coordinates, $U^\mu \propto (1, 0, 0, 0)$. Things can be made simpler if we lower one index, so that $T^\mu_\nu = \mathrm{diag}(\rho c^2, -p, -p, -p)$, as in special relativity, because $g^\mu_\nu = \mathrm{diag}(1, 1, 1, 1)$, independent of whether the metric is curved. The two independent Einstein equations are

$$\begin{aligned} G^0_0 &= 3(\dot{R}^2 + k)/R^2 = 8\pi G\rho \\ G^1_1 &= (2R\ddot{R} + \dot{R}^2 + k)/R^2 = 8\pi G p, \end{aligned} \tag{41}$$

which are two forms of the Friedmann equation.

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

$$\dot{R}^2 - \frac{8\pi G}{3}\rho R^2 = -kc^2. \tag{42}$$

This is almost obviously true, since the Newtonian result that the gravitational field inside a uniform shell is zero does still hold in general relativity, and is known as **Birkhoff's theorem**. But there are some surprises hidden here. This energy-like equation can be turned into a force-like equation by differentiating with respect to time:

$$\ddot{R} = -4\pi G R(\rho + 3p/c^2)/3. \tag{43}$$

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

$$d[\rho c^2 R^3] = -p d[R^3]. \tag{44}$$

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

$$\nabla^2 \Phi = 4\pi G(\rho + 3p/c^2). \tag{45}$$

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

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

$$H^2 - \frac{8\pi G}{3}\rho = \frac{\text{const}}{R^2}. \tag{46}$$

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

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

$$\rho_c = \frac{3H^2}{8\pi G}. \tag{47}$$

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

$$\Omega \equiv \frac{\rho}{\rho_c} = \frac{8\pi G\rho}{3H^2}. \tag{48}$$

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

$$R_0 = \frac{c}{H_0}\left[(\Omega_0 - 1)/k\right]^{-1/2}, \tag{49}$$

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

$$h \equiv \frac{H_0}{100\,\mathrm{km\,s^{-1}Mpc^{-1}}}, \tag{50}$$

in terms of which the current density of the universe is

$$\begin{aligned} \rho_0 &= 1.878 \times 10^{-26}\,\Omega h^2\ \mathrm{kg\,m^{-3}} \\ &= 2.775 \times 10^{11}\,\Omega h^2\ M_\odot\,\mathrm{Mpc}^{-3}. \end{aligned} \tag{51}$$

1.6 Vacuum energy in cosmology

The discussion so far is general, and independent of the detailed constituents of the universe. But before proceeding, it is worth an aside on what is in many ways the key factor in determining the expansion: the energy density of empty space. This is an idea that was introduced by Einstein, soon after he arrived at the theory of general relativity. However, the main argument has a much older motivation, as follows.

EINSTEIN'S STATIC UNIVERSE The expanding universe is the solution to a problem that goes back to Newton. After expounding the idea of universal gravitation, he was asked what would happen to mass in an infinite space. If all particles with mass attracted each other, how could the heavens be stable (as they apparently were, give or take the motions of planets)? In 1917, Einstein was still facing the same problem, but he thought of an ingenious solution. Gravitation can be reduced to the potential that solved Poisson's equation: $\nabla^2\Phi = 4\pi G\rho$.

Einstein argued by symmetry that, in a universe where the density ρ is constant, the potential Φ must be also (so that the acceleration $\mathbf{a} = -\boldsymbol{\nabla}\Phi$ vanishes everywhere). Since this doesn't solve Poisson's equation, he proposed that it should be replaced:

$$\nabla^2\Phi + \lambda\Phi = 4\pi G\rho, \tag{52}$$

where λ is a new constant of nature, called the **cosmological constant** in its relativistic incarnation. This clearly lets us have a static model, with $\Phi = 4\pi G\rho/\lambda$.

The modern way of writing this is to take the new term onto the other side, defining $\rho_{\rm rep} = \lambda\Phi/4\pi G$:

$$\nabla^2\Phi = 4\pi G(\rho - \rho_{\rm rep}), \tag{53}$$

i.e. to interpret it as a constant *repulsive* density, with **antigravity** properties. By this definition, $\rho = \rho_{\rm rep}$, so the rhs vanishes, and the repulsive density cancels the effect of normal matter. This repulsion would have to be an intrinsic property of the vacuum, since it has to be present when all matter is absent. This may sound like a really stupid idea, but in fact it is the basis of much of modern cosmology.

Interestingly, Einstein's neat argument is not really consistent with the relativistic generalization, and he might equally well have written $\nabla^2\Phi + \lambda = 4\pi G\rho$; this still allows a constant-potential solution, although the absolute value of potential cannot be found. This does not matter, since only gradients will cause forces.

When looked at as a vacuum repulsion, we can see that Einstein's idea couldn't work. Suppose we increase the matter density in some part of space a little: the mutual attraction of normal matter goes up, but the vacuum repulsion stays constant and doesn't compensate. In short, Einstein's static universe is unstable, and must either expand or contract. We can stretch this only a little to say that the expanding universe could have been predicted by Newton.

ENERGY DENSITY OF THE VACUUM Nevertheless, vacuum energy is central to modern cosmology. How can a vacuum have a non-zero energy density? Surely this is zero by definition in a vacuum? It turns out that this need not be true. What we can say is that, if the vacuum has a non-zero energy density, it must also have a non-zero pressure, with a negative-pressure equation of state:

$$p_{\rm vac} = -\rho_{\rm vac}\, c^2. \tag{54}$$

In this case, $\rho c^2 + 3p$ is indeed negative: a positive vacuum density will act to cause a large-scale repulsion.

The proof of this statement comes from energy conservation: as the universe expands, the work done by the pressure is just sufficient to maintain the energy density constant (see figure 3). In effect, the vacuum acts as a reservoir of unlimited energy, which can supply as much as is required to inflate a given region to any required size at constant energy density. This supply of energy is what is used in 'inflationary' theories of cosmology to create the whole universe out of almost nothing.

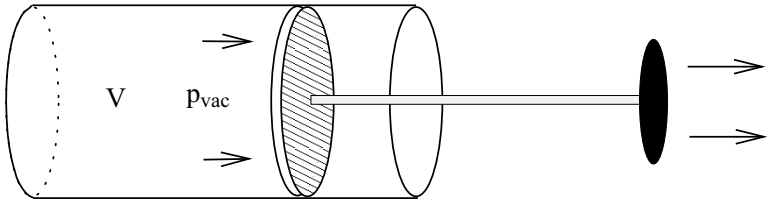

Figure 3. A thought experiment to illustrate the application of conservation of energy to the vacuum. If the vacuum density is $\rho_{\rm vac}$ then the energy created by withdrawing the piston by a volume dV is $\rho_{\rm vac}c^2\,dV$. This must be supplied by work done by the vacuum pressure $p_{\rm vac}dV$, and so $p_{\rm vac} = -\rho_{\rm vac}c^2$, as required.

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

$$\frac{8\pi G\rho}{3} = H_0^2(\Omega_v a^{-3(w+1)} + \Omega_m a^{-3} + \Omega_r a^{-4}) \tag{55}$$

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

$$H^2(a) = H_0^2\left[\Omega_v + \Omega_m a^{-3} + \Omega_r a^{-4} - (\Omega - 1)a^{-2}\right]. \tag{56}$$

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

1.7 Solving the Friedmann equation

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

$$w \equiv p/\rho c^2, \tag{57}$$

then conservation of energy says

$$d(\rho c^2 V) = -p\, dV \Rightarrow d(\rho c^2 V) = -w\rho c^2\, dV \Rightarrow d\ln\rho/d\ln a = -3(w+1), \tag{58}$$

so

$$\rho \propto a^{-3(w+1)} \tag{59}$$

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

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

$$\tau = \frac{2\sqrt{2}}{3}\left(2 + (a-2)\sqrt{1+a}\right). \tag{60}$$

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

$$\begin{aligned} \tau \ll 1: &\quad a = (\sqrt{2}\tau)^{1/2}. \\ \tau \gg 1: &\quad a = (3\tau/2\sqrt{2})^{2/3}, \end{aligned} \tag{61}$$

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

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

$$a = \Big(\sinh(\sqrt{2}\tau)\Big)^{1/2}. \tag{62}$$

Here, we move from $a \propto t^{1/2}$ at early times to an exponential behaviour characteristic of vacuum-dominated **de Sitter space**. This would be an appropriate model for the onset of a phase of inflation following a big-bang singularity. What about the case of negative vacuum density? It is easy to repeat the above exercise defining the critical era as one where ρ_r equals $|\rho_v|$, in which case the solution is the same, except with sinh $\to$ sin. A negative vacuum density always leads to eventual recollapse into a big crunch.

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

$$a = \Big(\sinh(3\tau/2\sqrt{2})\Big)^{2/3}. \tag{63}$$

This transition from the flat matter-dominated $a \propto t^{2/3}$ to de Sitter space seems to be the one that describes our actual universe (apart from the radiation era at $z \gtrsim 10^4$). It is therefore worth being explicit about how to translate units to the usual normalization at $a = 1$ today. We have to multiply the above expression by a_*, which is the usual scale factor at matter-vacuum equality, and we have to relate the Hubble parameter H_* to the usual H_0. This is easily done using the Friedmann equation:

$$\begin{aligned} a_* &= (\Omega_m/\Omega_v)^{1/3} \\ H_* &= H_0\sqrt{2\Omega_v}. \end{aligned} \tag{64}$$

CURVED MODELS We will not be very strongly concerned with highly curved models here, but it is worth knowing some basic facts, as shown in figure 4 (neglecting radiation). On a plot of the $\Omega_m - \Omega_v$ plane, the diagonal line $\Omega_m + \Omega_v = 1$ always separates open and closed models. If $\Omega_v < 0$, recollapse always occurs – whereas a positive vacuum density does not always guarantee expansion to infinity, especially when the matter density is high. For closed models with sufficiently high vacuum density, there was no big bang in the past, and the universe must have emerged from a 'bounce' at some finite minimum radius. All these statements can be deduced quite simply from the Friedmann equation.

1.8 The meaning of an expanding universe

Before going on, it is worth looking in a little more detail at the basic idea of an expanding universe. The RW metric written in comoving coordinates emphasizes that one can think of any given fundamental observer as fixed at the centre of their local coordinate system. A common interpretation of this algebra is to say that the galaxies separate "because the space between them expands", or some such phrase.

But even if 'expanding space' is a correct *global* description of spacetime, does the concept have a meaningful *local* counterpart? Is the space in my bedroom expanding, and what would this mean? Do we expect the Earth to recede from the Sun as the space between them expands? The very idea suggests some completely new physical effect that is not covered by Newtonian concepts. However, on scales much smaller than the current horizon, we should be able to ignore curvature and treat galaxy dynamics as occurring in Minkowski spacetime; this approach works in deriving the Friedmann equation. How do we relate this to 'expanding space'? It should be clear that

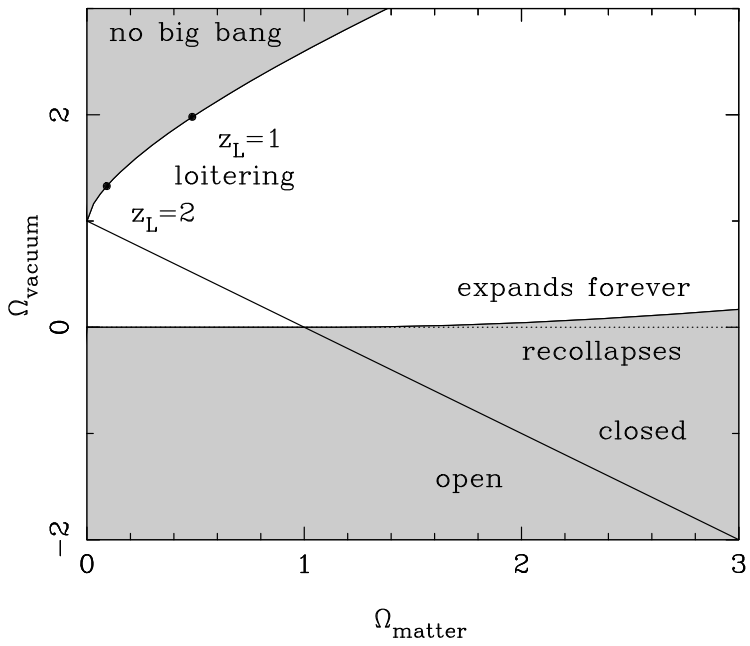

Figure 4. This plot shows the different possibilities for the cosmological expansion as a function of matter density and vacuum energy. Models with total $\Omega > 1$ are always spatially closed (open for $\Omega < 1$), although closed models can still expand to infinity if $\Omega_v \neq 0$. If the cosmological constant is negative, recollapse always occurs; recollapse is also possible with a positive Ω_v if $\Omega_m \gg \Omega_v$. If $\Omega_v > 1$ and Ω_m is small, there is the possibility of a 'loitering' solution with some maximum redshift and infinite age (top left); for even larger values of vacuum energy, there is no big bang singularity.

Minkowski spacetime does not expand – indeed, the very idea that the motion of distant galaxies could affect local dynamics is profoundly anti-relativistic: the equivalence principle says that we can always find a tangent frame in which physics is locally special relativity.

It is worth working this out in more detail by considering the case of the empty universe. The metric of uncurved Minkowski spacetime is

$$c^2 d\tau^2 = c^2 dt^2 - \left(dr^2 + r^2 d\psi^2\right), \tag{65}$$

but we can describe it as in the grenade universe, from the point of view of a set of test particles ejected from the origin at $t = 0$. The velocity of particles seen at radius r at time t is therefore a function of radius: $v = r/t$ ($t = H_0^{-1}$, as required); particles do not exist beyond the radius $r = ct$, at which point they are receding from the origin at the speed of light. If all clocks are synchronized at $t = 0$, then the cosmological time t' is just related to the background time via time dilation:

$$t' = t/\gamma = t\sqrt{1 - r^2/c^2t^2}. \tag{66}$$

If we also define $d\ell$ to be the radial separation between events measured by fundamental observers at fixed t', the metric can be rewritten as

$$c^2 d\tau^2 = c^2 dt'^2 - d\ell^2 - r^2 d\psi^2. \tag{67}$$

To complete the transition from Minkowski to fundamental-observer coordinates, we need to relate the two radial coordinates ℓ and r. These are related by length contraction:

$$d\ell = dr/\gamma. \tag{68}$$

The contraction is this way round because $d\ell$ is at constant t': fundamental observers see laboratory measuring rods of length dr moving at high speed, so they appear contracted. More explicitly, the Lorentz transformation says $dr = \gamma(d\ell - v\,dt')$, but $dt' = 0$.

We now need to re-express $\gamma = (1 - r^2/c^2t^2)^{-1/2}$ to eliminate t in terms of t'. We do this using the time-dilation relation $t' = t/\gamma$, which gives

$$\gamma = \left[1 + (r/ct')^2\right]^{1/2}. \tag{69}$$

The metric therefore becomes

$$c^2 d\tau^2 = c^2 dt'^2 - \frac{dr^2}{1 + (r/ct')^2} - r^2 d\psi^2. \tag{70}$$

Defining the comoving radius as (r/ct'), this is the $k = -1$ Robertson–Walker form, with $R = ct'$. Alternatively, we can introduce the velocity variable ω:

$$v/c = \tanh\omega \quad \Rightarrow \quad \gamma = \cosh\omega. \tag{71}$$

Now, the time-dilation equation gives r in terms of t and t' as

$$r = c\sqrt{t^2 - t'^2} = ct' \sinh\omega, \tag{72}$$

so that $d\ell = dr/\gamma$ becomes $d\ell = ct' d\omega$, yielding

$$d\tau^2 = dt'^2 - t'^2 \left(d\omega^2 + \sinh^2\omega \, d\psi^2 \right). \tag{73}$$

This is the $k = -1$ Robertson–Walker metric in its more standard form, with $R = ct'$. This is the result we needed earlier to verify the Friedmann equation for the $k = -1$ case.

TEST-PARTICLE DYNAMICS A further instructive example is to consider the effect of the expansion on the peculiar motion of a test particle; there is a neat paradox here. Suppose we take a nearby low-redshift galaxy and give it a velocity boost such that its redshift becomes zero. At a later time, will the expansion of the universe have caused the galaxy to recede from us, so that it once again acquires a positive redshift? To idealize the problem, imagine that the galaxy is a massless test particle in a homogeneous universe.

The 'expanding space' idea would suggest that the test particle should indeed start to recede from us, and it appears that one can prove this formally, as follows. Consider the peculiar velocity with respect to the Hubble flow, $\delta\mathbf{v}$. A completely general result is that this declines in magnitude as the universe expands:

$$\delta v \propto \frac{1}{a(t)}. \tag{74}$$

This is the same law that applies to photon energies, and the common link is that it is particle momentum in general that declines as $1/a$, just through the accumulated Lorentz transforms required to overtake successively more distant particles that are moving with the Hubble flow. So, at $t \to \infty$, the peculiar velocity tends to zero, leaving the particle moving with the Hubble flow, however it started out: 'expanding space' has apparently done its job.

Now look at the same situation in a completely different way. If the particle is nearby compared with the cosmological horizon, a Newtonian analysis should be valid: in an isotropic universe, Birkhoff's theorem assures us that we can neglect the effect of all matter at distances greater than that of the test particle, and all that counts is the mass between the particle and us. Call the proper separation of the particle from the origin r. Our initial conditions are that $\dot{r} = 0$ at $t = t_0$, when $r = r_0$. The equation of motion is just

$$\ddot{r} = \frac{-GM(< r \mid t)}{r^2}, \tag{75}$$

and the mass internal to r is just

$$M(< r \mid t) = \frac{4\pi}{3}\,\rho r^3 = \frac{4\pi}{3}\,\rho_0\, a^{-3} r^3, \tag{76}$$

where we assume $a_0 = 1$ and a matter-dominated universe. The equation of motion can now be re-expressed as

$$\ddot{r} = -\frac{\Omega_0 H_0^2}{2a^3}\, r. \tag{77}$$

Adding vacuum energy is easy enough:

$$\ddot{r} = -\frac{H_0^2}{2}\, r\,(\Omega_m a^{-3} - 2\Omega_v). \tag{78}$$

The -2 in front of the vacuum contribution comes from the effective mass density $\rho + 3p/c^2$.

We now show that this Newtonian equation is identical to what is obtained from $\delta v \propto 1/a$. In our present notation, this becomes

$$\delta v = \dot{r} - H(t) r = -H_0 r_0 / a; \tag{79}$$

the initial peculiar velocity is just $-Hr$, cancelling the Hubble flow. We can differentiate this equation to obtain $\ddot{r}$, which involves $\dot{H}$. This can be obtained from the standard relation

$$H^2(t) = H_0^2[\Omega_v + \Omega_m a^{-3} + (1 - \Omega_m - \Omega_v)a^{-2}]. \tag{80}$$

It is then a straightforward exercise to show that the equation for $\ddot{r}$ is the same as obtained previously (remembering $H = \dot{a}/a$).

Now for the paradox. It will suffice at first to solve the equation for the case of the Einstein-de Sitter model, choosing time units such that $t_0 = 1$, with $H_0 t_0 = 2/3$:

$$\ddot{r} = -2r/9t^2. \tag{81}$$

The acceleration is negative, so the particle moves *inwards*, in complete apparent contradiction to our 'expanding space' conclusion that the particle would tend with time to pick up the Hubble expansion. The resolution of this contradiction comes from the full solution of the equation. The differential equation clearly has power-law solutions $r \propto t^{1/3}$ or $t^{2/3}$, and the combination with the correct boundary conditions is

$$r(t) = r_0(2t^{1/3} - t^{2/3}). \tag{82}$$

At large t, this becomes $r = -r_0 t^{2/3}$. The use of a negative radius may seem suspect, but we can regard r as a Cartesian coordinate along a line that passes through the origin, and the equation of motion $\ddot{r} \propto r$ is correct for either sign of r. The solution for $r(t)$ at large t thus describes a particle moving with the Hubble flow, but it arises because the particle has fallen right through the origin and emerged on the other side.

In no sense, therefore, can 'expanding space' be said to have operated: in an Einstein-de Sitter model, a particle initially at rest with respect to the origin falls towards the origin, passes through it, and asymptotically regains its initial comoving radius on the opposite side of the sky. The behaviour can be understood quantitatively using only Newtonian dynamics.

This analysis demonstrates that there is no local effect on particle dynamics from the global expansion of the universe: the tendency to separate is a kinematic initial condition, and once this is removed, all memory of the expansion is lost. Perhaps the cleanest illustration of the point is provided by the Swiss Cheese universe, an exact model in which the mass within (non-overlapping) spherical cavities is

compressed to a black hole. Within the cavity, the metric is exactly Schwarzschild, and the behaviour of the rest of the universe is irrelevant. This avoids the small complication that arises when considering test particles in a homogeneous universe, where we still have to consider the gravitational effects of the matter between the particles. It should now be clear how to deal with the question, "does the expansion of the universe cause the Earth and Moon to separate?", and that the answer is not the commonly-encountered "it would do, if they weren't held together by gravity".

Two further cases are worth considering. In an empty universe, the equation of motion is $\ddot{r} = 0$, so the particle remains at $r = r_0$, while the universe expands linearly with $a \propto t$. In this case, $H = 1/t$, so that $\delta v = -Hr_0$, which declines as $1/a$, as required. Finally, models with vacuum energy are of more interest. Provided $\Omega_v > \Omega_m/2$, $\ddot{r}$ is initially positive, and the particle does move away from the origin. This is the criterion for $q_0 < 0$ and an accelerating expansion. In this case, there is a tendency for the particle to expand away from the origin, and this is caused by the repulsive effects of vacuum energy. In the limiting case of pure de Sitter space ($\Omega_m = 0$, $\Omega_v = 1$), the particle's trajectory is

$$r = r_0 \cosh H_0(t - t_0), \tag{83}$$

which asymptotically approaches half the $r = r_0 \exp H_0(t - t_0)$ that would have applied if we had never perturbed the particle in the first place. In the case of vacuum-dominated models, then, the repulsive effects of vacuum energy cause all pairs of particles to separate at large times, whatever their initial kinematics; this behaviour could perhaps legitimately be called 'expanding space'. Nevertheless, the effect stems from the clear physical cause of vacuum repulsion, and there is no new physical influence that arises purely from the fact that the universe expands. The earlier examples have proved that 'expanding space' is in general a dangerously flawed way of thinking about an expanding universe.

1.9 Observational cosmology

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

$$\frac{dz}{dt} = -\frac{R_0}{R^2}\frac{dR}{dt} = -(1 + z)H(z), \tag{84}$$

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

$$H^2(a) = H_0^2 \left[\Omega_v + \Omega_m a^{-3} + \Omega_r a^{-4} - (\Omega - 1)a^{-2}\right]. \tag{85}$$

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

$$H_0 t_0 \simeq \frac{2}{3}\,(0.7\Omega_m - 0.3\Omega_v + 0.3)^{-0.3}. \tag{86}$$

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

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

$$R_0 r = \int \frac{c}{H(z)}\, dz. \tag{87}$$

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

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

$$R_0 r_{\rm H} \simeq \frac{2c}{H_0}\Omega_m^{-0.4}. \tag{88}$$

At high redshift, where H increases, this tends to zero. The onset of radiation domination does not change this: even though the presently visible universe was once very small, it expanded so quickly that causal contact was not easy. The observed large-scale near-homogeneity is therefore something of a puzzle.

ANGULAR DIAMETERS Recall the RW metric:

$$c^2 d\tau^2 = c^2 dt^2 - R^2(t) \left[dr^2 + S_k^2(r)\, d\psi^2 \right]. \tag{89}$$

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

$$d\ell_\perp = d\psi\ R(z) S_k(r) = d\psi\ R_0 S_k(r)/(1+z). \tag{90}$$

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

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

$$S_\nu(\nu_0) = \frac{L_\nu([1+z]\nu_0)}{4\pi R_0^2 S_k^2(r)(1+z)} = \frac{L_\nu(\nu_0)}{4\pi R_0^2 S_k^2(r)(1+z)^{1+\alpha}}, \tag{91}$$

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

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

$$I_\nu(\nu_0) = \frac{B_\nu([1+z]\nu_0)}{(1+z)^3}, \tag{92}$$

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

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

$$\begin{aligned} \textbf{angular}-\textbf{diameter distance}: &\quad D_{\rm A} = (1+z)^{-1} R_0 S_k(r) \\ \textbf{luminosity distance}: &\quad D_{\rm L} = (1+z)\, R_0 S_k(r). \end{aligned} \tag{93}$$

Some example distance-redshift relations are shown in figure 5. Notice how a high matter density tends to make high-redshift objects brighter: stronger deceleration means they are closer for a given redshift.

Recall that the comoving distance approaches that of the particle horizon as $z \to \infty$. Here, there is a substantial dependence on the matter density, but in a way that differs depending on whether or not there is vacuum energy:

$$R_0 S_k(r) = (2c/H_0)\Omega_m^{-1} \text{ (open)}; \quad \text{(or)} \quad (2c/H_0)\Omega_m^{-0.4} \text{ (flat)}. \tag{94}$$

The first relation is exact; the second is an accurate approximation.

1.10 Absolute magnitude and K-correction

Absolute magnitude is defined as the apparent magnitude that would be observed if the source lay at a distance of 10 pc; it is just a measure of luminosity. Absolute magnitudes in cosmology are affected by a shift of the spectrum in frequency; the K-correction accounts for this effect, giving the difference between the observed dimming with redshift for a fixed observing waveband, and that expected on bolometric grounds:

$$m = M + 5\log_{10}\left(\frac{D_{\rm L}}{10\text{ pc}}\right) + K(z), \tag{95}$$

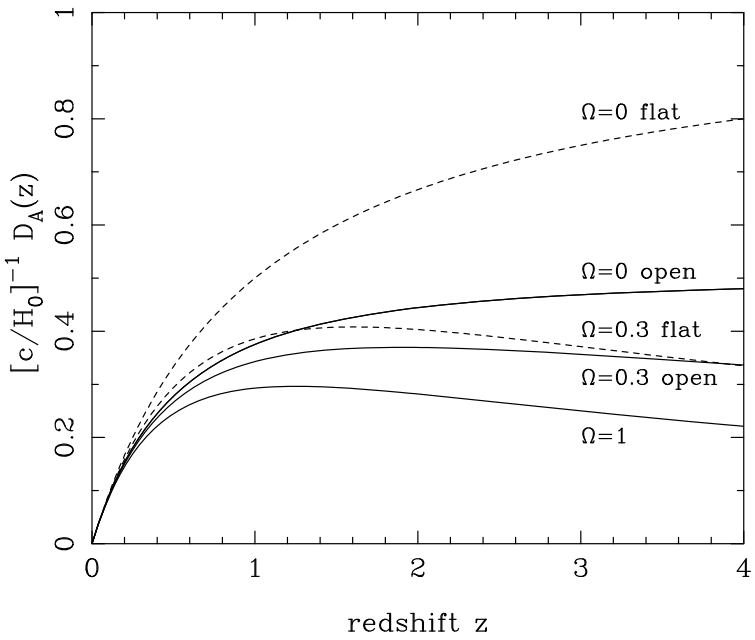

Figure 5. A plot of dimensionless angular-diameter distance versus redshift for various cosmologies. Solid lines show models with zero vacuum energy; dashed lines show flat models with $\Omega_m + \Omega_v = 1$. In both cases, results for $\Omega_m = 1, 0.3, 0$ are shown; higher density results in lower distance at high z, due to gravitational focusing of light rays.

where $D_{\rm L}$ is luminosity distance. There is a significant danger of ending up with the wrong sign here: remember that $K(z)$ should be large and positive for a very red galaxy. For a $\nu^{-\alpha}$ spectrum,

$$K(z) = 2.5(\alpha - 1)\log_{10}(1 + z). \tag{96}$$

1.11 Counts and luminosity functions

We now have all the tools needed to understand how astronomical observations sample the population of objects in the universe. Historically, this began with almost no information on redshifts (which are still hard to obtain for the faintest objects). Nevertheless, there is useful information just in the number counts.

EUCLIDEAN COUNTS The number–flux relation assumes an important form if space is Euclidean. Consider first a universe populated with sources that all have the same luminosity, with number density n. The flux density is now just the normal inverse-square law $S = L/(4\pi D^2)$, so the distance to a given object is proportional to $(L/S)^{1/2}$. The number of objects brighter than S is just n times the volume of space within which they can be seen:

$$N(>S) = nV(S) = n(A/3)(L/4\pi S)^{3/2} \propto S^{-3/2}, \tag{97}$$

where A is the solid angle of the survey. This **Euclidean source count** is the baseline for all realistic surveys, and shows us that faint sources are likely to heavily outnumber bright ones. It obviously remains true if we now add in a more realistic population of sources with a wide range of luminosities. The relation is one form of **Olbers' paradox**: integration over S implies a divergent sky brightness:

$$I = \int S\, dN(S)/A. \tag{98}$$

Since the universe does not contain an infinite energy density, it is clear that relativistic effects in the distance–redshift and volume–redshift relations must cause the true counts to lie below the Euclidean prediction.

LUMINOSITY FUNCTIONS The evolution of the properties of a population of cosmological sources can be described via the luminosity function $\phi(L)$, which is the comoving number density of objects in some range of luminosity. Generally, the simplest results arise if we take ϕ to be the comoving density per interval of $\ln L$:

$$dN = \phi(L, z)\, d\ln L\, dV(z). \tag{99}$$

It is often convenient to describe the results analytically *e.g.* via a **Schechter function** fit at each redshift

$$d\phi = \phi^*(L/L^*)^\alpha \exp(-L/L^*)\, dL/L^*. \tag{100}$$

Practical values of α for the galaxy population range between -1 and -1.5, depending on type (see figure 6).

We saw earlier the relation between monochromatic luminosity, L, and flux density, S:

$$L = S\, 4\pi [R_0 S_k(r)]^2 (1+z)^{1+\alpha}, \tag{101}$$

where $S \propto \nu^{-\alpha}$ (unfortunately, spectral index and luminosity function slope are often both called α). So, denoting the luminosity function by $\phi(L,z)$, the expected number of objects seen in a range of flux and redshift is

$$dN = \phi(L,z)\; d\ln S\; dV(z), \tag{102}$$

because the Jacobian $\partial(\ln S, z)/\partial(\ln L, z)$ is unity. For an area of sky A sr, the differential volume element is $dV = A[R_0 S_k(r)]^2 R_0 dr$, where dr is the element of comoving radius.

1.12 The cosmological distance scale

CEPHEID VARIABLES AND LOCAL DISTANCES We now turn to the distance scale. Distances to the nearest few galaxies may be determined most accurately by the use of Cepheid variable stars. These are among the most luminous stars known: being objects between 3 and 9 $M_\odot$ that are in the process of evolving either off the main sequence onto the giant branch, or off the giant branch. They have a positive correlation between luminosity and period (roughly $L \propto P^{1.3}$) that is very tightly defined. All stars have a natural oscillation period, deriving from the crossing time for sound waves. This does not of course explain why some stars become unstable to radial oscillations with this period whereas others (such as the Sun) do not. Detailed study of stellar structure is needed in order to understand the instability. Data in different wavebands can be used to find the relative distance between two groups of Cepheids and also to determine the relative extinctions involved, so this is not a source of uncertainty in the method.

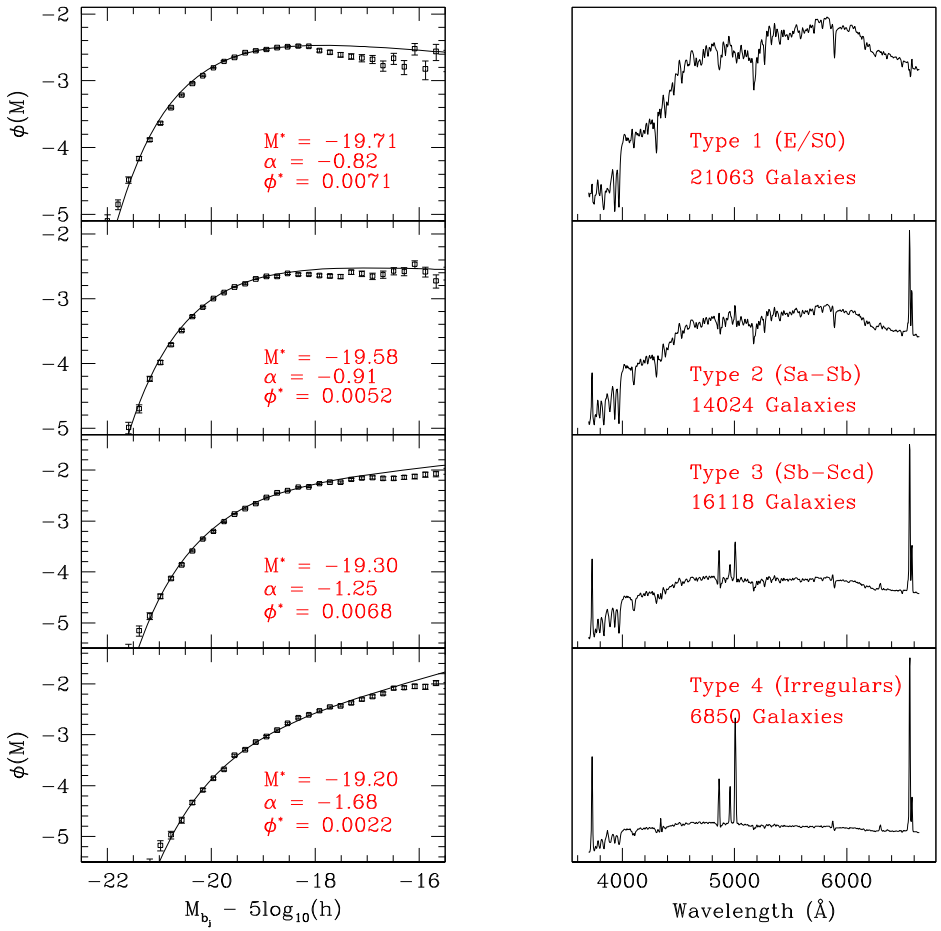

Figure 6. The galaxy luminosity function, with the population dissected into different types. At the top, we gave galaxies with old stellar populations (ellipticals); at the bottom, we move to galaxies dominated by younger stars (extreme spiral and irregular galaxies). The latter are systematically less luminous than the former (and less massive). In all cases, a Schechter function (power-law times exponential cutoff) gives a good fit.

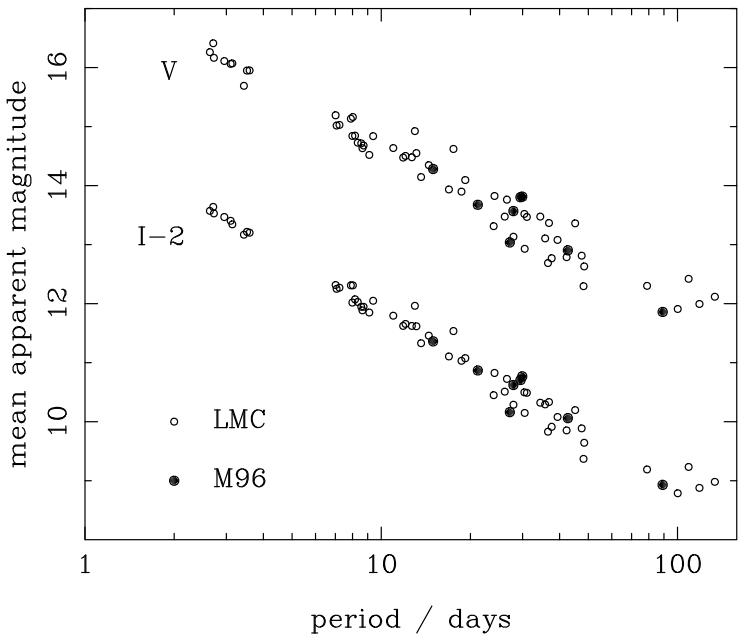

Figure 7. A plot of the Cepheid $P-L$ relation for stars in the Large Magellanic Cloud and in M96. The M96 stars have been shifted to overlay the LMC data, and the required shift gives the apparent difference in distance modulus. Examining this offset as a function of wavelength and fitting an extinction model allows the true relative distance to be established.

The Cepheid method is limited by the closest point at which there is a large group of Cepheids to calibrate the period-luminosity relation. Such a group is found in the **Large Magellanic Cloud** (LMC). We are therefore able to measure with some confidence relative distances between the LMC and nearby galaxies. The main concern with using the LMC as a zero point might be that this is a dwarf galaxy of low metal content relative to the Sun. However, no effect of metallicity on the cepheid distances has ever been detected.

This leaves the absolute distance to the LMC as one of the key numbers in cosmology, and we have a reasonably good idea what it is:

$$D_{\rm LMC} = 51 \text{ kpc } \pm 6\%. \tag{103}$$

This number has been established over the years by a number of methods. The simplest is to calibrate the luminosities of a few more nearby cepheids. This is done via main-sequence fitting (finding the offset in apparent magnitude at a given colour) of the HR diagrams of the star clusters that host Cepheids. For the most nearby star clusters, distances can be obtained via trigonometric parallax or related methods (the astrometric **HIPPARCOS** satellite has had a big impact here). A much more direct alternative came from observations of **SN1987A**: a supernova that took place in the LMC itself. This was observed to produce a ring of emission that was elliptical to high precision – and therefore almost certainly a circular ring seen inclined. Different parts of the ring were observed being illuminated at different times owing to finite light travel-time effects. Knowing the inclination, plus the observed angular size of the ring, the distance to the supernova follows. It agrees very well with the traditional figure.

1.13 Larger distances: the supernova Hubble diagram

Cepheid distances can thus be found for the more nearby galaxies. The **Hubble Space Telescope** has allowed this to be done for a few dozen galaxies out to distances of 10 to 20 Mpc. Unfortunately, this is not really far enough. At 10 Mpc, the recessional velocity is $1000\,h\,\mathrm{km\,s^{-1}}$, but peculiar velocities can reach $600\,\mathrm{km\,s^{-1}}$. In order to determine H_0 accurately, we need to attain recessional velocities of $> 10,000\,\mathrm{km\,s^{-1}}$.

This requires brighter objects than Cepheids, and traditional work concentrated on whole-galaxy luminosity indicators. These are variants of the **Tully-Fisher relation**, which says that the luminosity of a spiral galaxy scales with its rotational velocity roughly as $L \propto V^3$. This is reminiscent of $V^2 = GM/r$, but obviously raises questions about M/L ratios and sizes of galaxies. In any case, such methods are of limited accuracy, predicting L to about 40%, and hence giving relative distances to about 20% precision. The great discovery of the 1990s was that supernovae make much more accurate **standard candles**.

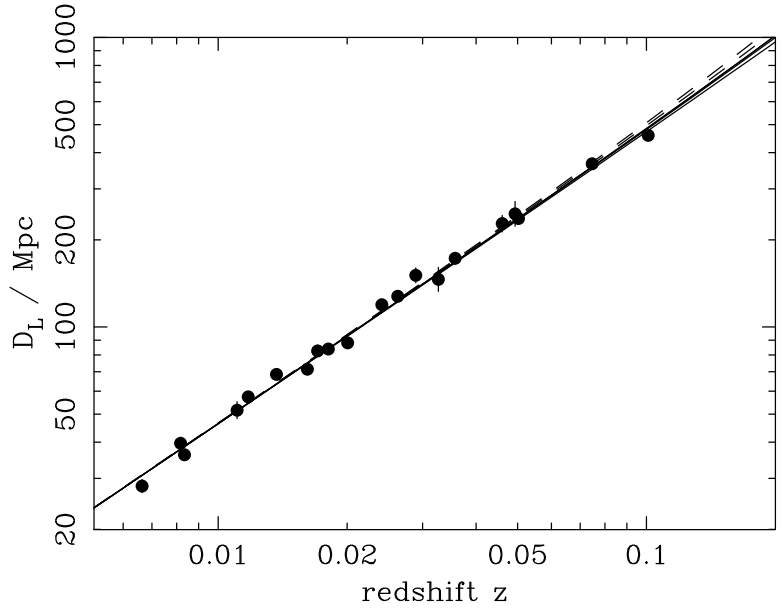

Figure 8. The type Ia supernova Hubble diagram. Using a measure of characteristic time as well as peak luminosity for the light curve, relative distances to individual SNe can be measured to 6% rms. Setting the absolute distance scale ($D_{\rm L}$ is luminosity distance) using local SNe in galaxies with Cepheid distances shows that the large-scale Hubble flow is indeed linear and uniform, and gives an estimate of $H_0 = 72 \pm 8$ km s^{-1}Mpc^{-1}.

Supernovae come in two-and-a-bit varieties, SNe Ia, Ib and II, distinguished according to whether or not they display absorption and emission lines of hydrogen. The SNe II do show hydrogen; they are associated with massive stars at the endpoint of their evolution, and are rather heterogeneous in their behaviour. The former, especially SNe Ia, are much more homogeneous in their properties, and can be used as standard candles. There is a characteristic rise to maximum, followed by a symmetric fall over roughly 30 days, after which the light decay becomes less rapid. Type Ib SNe are a complication to the scheme; they do not have the characteristic light curve, and also lack hydrogen lines.

The simplest use of these supernovae was to note that they empirically have a very small dispersion in luminosity at maximum light ($\lesssim 0.3$ magnitudes). However, one might legitimately ask why SNe Ia should be standard candles. After all, presumably the progenitor stars vary in mass, and this should affect the energy output. A more satisfactory treatment of the supernovae distance scale takes this possibility into account by measuring both the height of the light curve (apparent luminosity at maximum light) and the width (time taken to reach maximum light, or other equivalent measures). For SNe where relative distances are known by some other method, these parameters are seen to correlate: the maximum output of SNe scales as roughly the 1.7 power of the characteristic timescale. The physical interpretation of this relation is that both the measured parameters depend on *mass*: a more massive star has more fuel and so generates a more energetic explosion, but the resulting fireball has to expand for longer in order for its optical depth to reach unity, allowing the photons to escape.

It is therefore possible to turn SNe Ia into genuine standard candles, and the accuracy is astonishingly good: a corrected dispersion of 0.12 magnitudes, implying that relative distances to a single SN can be measured to 6% precision. The SN Hubble diagram is impressively linear (figure 8), and allows a very precise estimate of H_0, based on the HST Cepheid distances:

$$H_0 = 72 \pm 8\,\mathrm{km\,s^{-1}Mpc^{-1}}. \tag{104}$$

The uncertainty on this now comes largely from how accurately we know the distance to the LMC.

Values for t_0 in the range 12 – 16 Gyr are a reasonable summary of the present estimates from stellar evolution. If the globular-cluster ages are not trusted, however, nuclear decay ages do not compel us to believe that the universe is any older than 9 Gyr. If we take a conservative range from above of $0.6 < h < 0.84$, that allows an extreme range of

$$0.55 < H_0 t_0 \simeq \tfrac{2}{3}\left(0.7\Omega_m + 0.3 - 0.3\Omega_v\right)^{-0.3} < 1.37, \tag{105}$$

with a best guess of $H_0 t_0 \simeq 0.96$, for $h = 0.72$ and $t_0 = 13$ Gyr. If $\Omega_m > 0.1$ is accepted as a hard lower bound, then vacuum energy is required on the basis of this formula if $H_0 t_0 > 0.90$. The Einstein–de Sitter model requires $H_0 t_0 = 2/3$, and is very hard to reconcile with the data. The high apparent value of $H_0 t_0$ was historically one of the first indication that vacuum energy might be required in cosmology.

1.14 Measuring the cosmological geometry

Any method that can be used to estimate distances can be used not only to measure H_0, but also to look for curvature in $D(z)$ and measure the cosmological geometry. The accuracy of the SNe data makes this test a practical possibility, following decades of inconclusive efforts with low-accuracy distance indicators. Figure 9 shows the SNe Hubble diagram out to very large redshifts, emphasizing the curvature in the relation.

It is clear from figure 9 that the empirical distance-redshift relation is very different from the simplest model, which is the $\Omega = 1$ Einstein-de Sitter universe; by redshift 0.6, the SNe are fainter than expected in this model by about 0.5 magnitudes. If this model fails, we can try adjusting Ω_m and Ω_v in an attempt to do better. Comparing each such model to the data yields the likelihood contours shown in figure 10, which can be used to set confidence limits on the cosmological parameters. The preferred model has $\Omega_v \simeq 1$; if we restrict ourselves to models with $k = 0$ (as predicted by inflationary cosmology), then the required parameters are very close to $(\Omega_m, \Omega_v) = (0.3, 0.7)$: a low matter density, vacuum-dominated universe, in which the expansion is accelerating.

Although admirably direct, this route was not the first evidence for a vacuum-dominated universe. As will be seen in other lectures, the key argument is that the angular-diameter distance to high redshift is larger for an open universe than for a flat one of the same density. This implies that characteristic angular sizes of CMB anisotropies would be very small in low-density open models, and it was known even in the late 1980s that this would violate upper limits on 10-arcmin scales. The only alternatives were $\Omega_m = 1$ or a low matter density supplemented with vacuum energy. By about 1990, a range of arguments had dismissed $\Omega_m = 1$ as a plausible model, and our standard flat vacuum-dominated model took centre stage from then on (see e.g. Efstathiou, Sutherland & Maddox 1990).

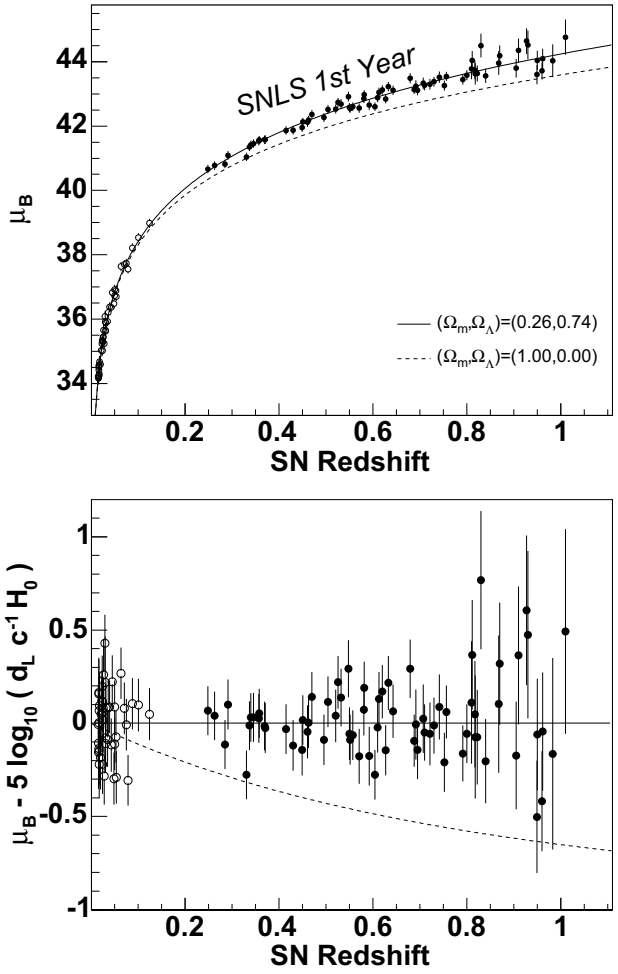

Figure 9. The Hubble diagram. The lower panel shows the data divided by a default model (flat $\Omega_m = 0.26$). The results lie clearly match this model, very precisely. The lowest line is the Einstein-de Sitter model, which is in gross disagreement with observation.

2 The perturbed universe

All discussion so far has applied to homogeneous models, but of course the real universe contains structure. The study of this structure via galaxy clustering and the CMB is the subject of separate courses here, but it may help to summarise some aspects of this in a way that should be complementary and fill in relevant background.

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

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

$$S'(x^\mu) = S(x^\mu) - \epsilon^\alpha \partial S/\partial x^\alpha. \tag{106}$$

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

$$\rho' = \rho - \epsilon^0 \dot{\rho}. \tag{107}$$

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

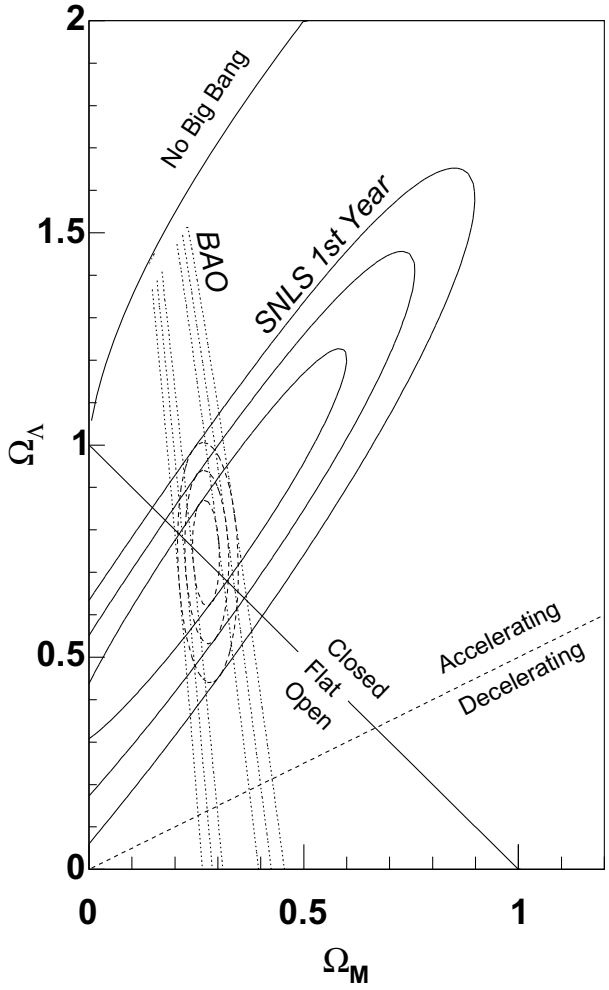

Figure 10. Confidence contours on the Ω_v–Ω_m plane. Open models of all but the lowest densities are apparently ruled out, and nonzero Λ is strongly preferred. If we restrict ourselves to $k = 0$, then $\Omega_m \simeq 0.3$ is required. Information from galaxy clustering ('BAO') prefers this density, yielding one argument for a flat vacuum-dominated universe.

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

$$d\tau^2 = a^2(\eta)\left\{(1+2\phi)d\eta^2 + 2w_i d\eta\, dx^i - [(1-2\psi)\gamma_{ij} + 2h_{ij}]\, dx^i\, dx^j\right\}. \tag{108}$$

In this equation, η is **conformal time**,

$$d\eta = dt/a(t), \tag{109}$$

and γ_{ij} is the comoving spatial part of the Robertson-Walker metric.

The total number of degrees of freedom here is apparently 2 (scalar fields ϕ and ψ) + 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:

$$\delta g_{\mu\nu} = a^2 \begin{pmatrix} 2\phi & -B_{,i} \\ -B_{,i} & 2[\psi\delta_{ij} - E_{,ij}] \end{pmatrix}, \tag{110}$$

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

$$\begin{aligned} \Phi &\equiv \phi + \frac{1}{a}\left[(B - E')a\right]' \\ \Psi &\equiv \psi - \frac{a'}{a}\,(B - E'), \end{aligned} \tag{111}$$

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 ϕ and ψ 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:

$$d\tau^2 = (1 + 2\Phi)dt^2 - (1 - 2\Phi)\gamma_{ij}\,dx^i\,dx^j, \tag{112}$$

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

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

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

$$\nabla^2\Phi/a^2 = 4\pi G(1+3w)\,\bar{\rho}\,\delta, \tag{113}$$

where we have defined a dimensionless fluctuation amplitude

$$\delta \equiv \frac{\rho-\bar{\rho}}{\bar{\rho}}. \tag{114}$$

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

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

$$\delta(\mathbf{x}) = \sum \delta_k e^{-i\mathbf{k}\cdot\mathbf{x}}, \tag{115}$$

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

$$k_x = n\,\frac{2\pi}{L}, \qquad n = 1, 2\cdots, \tag{116}$$

and the inverse Fourier relation would be

$$\delta_k(\mathbf{k}) = \left(\frac{1}{L}\right)^3 \int \delta(\mathbf{x}) \exp\bigl(i\mathbf{k}\cdot\mathbf{x}\bigr)\; d^3x. \tag{117}$$

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

$$\Phi_k = -4\pi G(1+3w)a^2\,\bar{\rho}\,\delta_k/k^2. \tag{118}$$

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

$$\langle \delta^2 \rangle = \sum |\delta_k|^2. \tag{119}$$

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

$$P(k) \equiv |\delta_k|^2 \tag{120}$$

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

$$\langle \delta^2 \rangle = \sum |\delta_k|^2 \to \frac{L^3}{(2\pi)^3} \int P(k)\, d^3k = \int \Delta^2(k)\, d\ln k. \tag{121}$$

In the last step, we have defined the combination

$$\Delta^2(k) \equiv \frac{L^3}{(2\pi)^3}\, 4\pi k^3\, P(k), \tag{122}$$

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).

2.1 Growth of linear perturbations

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

$$\begin{aligned} \mathbf{x}(t) &= a(t)\mathbf{r}(t) \\ \delta\mathbf{v}(t) &= a(t)\mathbf{u}(t), \end{aligned} \tag{123}$$

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 $\mathbf{r}$:

$$\dot{\mathbf{x}} = \dot{a}\mathbf{r} + a\dot{\mathbf{r}} = H\mathbf{x} + a\dot{\mathbf{r}}, \tag{124}$$

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

$$\ddot{\mathbf{x}} = a\dot{\mathbf{u}} + 2\dot{a}\mathbf{u} + \frac{\ddot{a}}{a}\,\mathbf{x} = \mathbf{g}_0 + \mathbf{g}. \tag{125}$$

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

$$\dot{\mathbf{u}} + 2(\dot{a}/a)\mathbf{u} = \mathbf{g}/a. \tag{126}$$

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

$$\frac{d}{dt}\rho_0(1+\delta) = -\rho_0(1+\delta)\,\boldsymbol{\nabla}\cdot\mathbf{u}. \tag{127}$$

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

$$\boldsymbol{\nabla} \equiv a\,\boldsymbol{\nabla}_{\rm proper}, \tag{128}$$

and the time derivative is a convective one:

$$\frac{d}{dt} = \frac{\partial}{\partial t} + \mathbf{u}\cdot\boldsymbol{\nabla}, \tag{129}$$

i.e. the time derivative measured by an observer who follows a particle's trajectory. Finally, when using a comoving frame, the background density ρ_0 is unaffected by d/dt, and so the full continuity equation can be written as

$$\frac{d}{dt}\delta = -(1+\delta)\boldsymbol{\nabla}\cdot\mathbf{u}. \tag{130}$$

LINEAR APPROXIMATION The equation for δ is not linear in the perturbations δ 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:

$$\begin{aligned} \dot{\mathbf{u}} + 2\frac{\dot{a}}{a}\,\mathbf{u} &= \frac{\mathbf{g}}{a} \\ \dot{\delta} &= -\boldsymbol{\nabla}\cdot\mathbf{u}, \end{aligned} \tag{131}$$

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 δ is

$$\ddot{\delta} + 2\frac{\dot{a}}{a}\dot{\delta} = 4\pi G\rho_0\,\delta. \tag{132}$$

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:

$$\delta(t) \propto t^{2/3} \quad \text{or} \quad t^{-1}. \tag{133}$$

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.

One further way of stating this result is that gravitational potential perturbations are independent of time (at least while $\Omega = 1$). Poisson's equation tells us that $-k^2\Phi/a^2 \propto \rho\,\delta$; since $\rho \propto a^{-3}$ for matter domination or a^{-4} for radiation, that gives $\Phi \propto \delta/a$ or δ/a^2 respectively, so that Φ 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 Λ, but an analytical solution does exist in this case. The perturbation equation is as before

$$\ddot{\delta} + 2\frac{\dot{a}}{a}\dot{\delta} = 4\pi G\rho_m\delta, \tag{134}$$

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

$$\delta'' + \frac{2+3y}{2y(1+y)}\,\delta' - \frac{3}{2y(1+y)}\,\delta = 0 \tag{135}$$

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

$$\delta \propto y + 2/3. \tag{136}$$

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

What this says is that, at early times, the dominant energy of radiation drives the universe to expand so fast that the matter has no time to respond, and δ is frozen at a constant value. At late times, the radiation becomes negligible, and the growth increases smoothly to the Einstein–de Sitter $\delta \propto a$ behaviour (Mészáros 1974). The overall behaviour is therefore 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.

2.2 Nonlinear structure formation

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

$$\frac{d}{dt}\mathbf{u} = -2\frac{\dot{a}}{a}\mathbf{u} - \frac{1}{a^2}\boldsymbol{\nabla}\Phi, \tag{137}$$

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

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

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

$$\mathbf{x}(t) = a(t)\left(\mathbf{q} + \mathbf{f}(\mathbf{q}, t)\right), \tag{138}$$

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:

$$\mathbf{u} = \frac{\partial \mathbf{f}}{\partial t} \tag{139}$$

(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:

$$\rho \,/\, \bar{\rho} = \left| \frac{\partial \mathbf{q}}{\partial \mathbf{x}/a} \right|. \tag{140}$$

When the displacement is small, this is just

$$\rho \,/\, \bar{\rho} = 1 - \boldsymbol{\nabla} \cdot \mathbf{f}(\mathbf{q}, t), \tag{141}$$

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

$$\mathbf{f} = -\boldsymbol{\nabla} \psi(\mathbf{q}), \tag{142}$$

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

$$\psi_k = -\delta_k / k^2, \tag{143}$$

and hence the displacement and velocity

$$\begin{aligned} \mathbf{f}_k &= i\mathbf{k}\, \psi_k \\ \mathbf{u}_k &= i\mathbf{k}\, \dot{\psi}_k. \end{aligned} \tag{144}$$

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 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

$$\begin{aligned} r &= A(1-\cos\theta) \\ t &= B(\theta-\sin\theta). \end{aligned} \tag{145}$$

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

$$r \simeq \frac{A}{2}\left(\frac{6t}{B}\right)^{2/3}\left[1-\frac{1}{20}\left(\frac{6t}{B}\right)^{2/3}\right], \tag{146}$$

and we can identify the density perturbation within the sphere:

$$\delta \simeq \frac{3}{20}\left(\frac{6t}{B}\right)^{2/3}. \tag{147}$$

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

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

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

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

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

These calculations are the basis for a common 'rule of thumb', whereby one assumes that linear theory applies until $\delta_{\rm lin}$ is equal to some δ_c a little greater than unity, at which point virialization is deemed to have occurred. Although the above only applies for $\Omega = 1$, analogous results can be worked out from the full $\delta_{\rm lin}(z, \Omega)$ and $t(z, \Omega)$ relations. These indicate that $\delta_{\rm lin} \simeq 1$ is a good criterion for collapse for any value of Ω likely to be of practical relevance. The density contrast at virialization tends to be higher in low-density universes, where the faster expansion means that, by the time a perturbation has turned round and collapsed to its final radius, a larger density contrast has been produced. For real non-spherical systems, it is not clear that this effect is meaningful, and in practice a fixed density contrast of around 200 is used to define the **virial radius** that marks the boundary of an object.

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

$$p(\delta > \delta_c \mid R) = \frac{1}{\sqrt{2\pi}\,\sigma(R)} \int_{\delta_c}^{\infty} \exp\left(-\delta^2/2\sigma^2(R)\right)\ d\delta, \tag{148}$$

where $\sigma(R)$ is the linear rms in the filtered version of δ. 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

$$F(> M) = \sqrt{\frac{2}{\pi}} \int_{\nu}^{\infty} \exp(-\nu^2/2)\ d\nu, \tag{149}$$

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

$$M = \frac{4\pi}{3}\,\bar{\rho}\,R^3. \tag{150}$$

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 virialized 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

$$Mf(M)/\rho_0 = |dF/dM|, \tag{151}$$

where ρ_0 is the total comoving density. Thus,

$$\frac{M^2 f(M)}{\rho_0} = \frac{dF}{d\ln M} = \left|\frac{d\,\ln\sigma}{d\,\ln M}\right| \sqrt{\frac{2}{\pi}}\,\nu\,\exp\left(-\frac{\nu^2}{2}\right). \tag{152}$$

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 ν seems to hold, and a good approximation is

$$F(>\nu) = (1 + a\,\nu^b)^{-1}\exp(-c\,\nu^2), \tag{153}$$

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 11, 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.

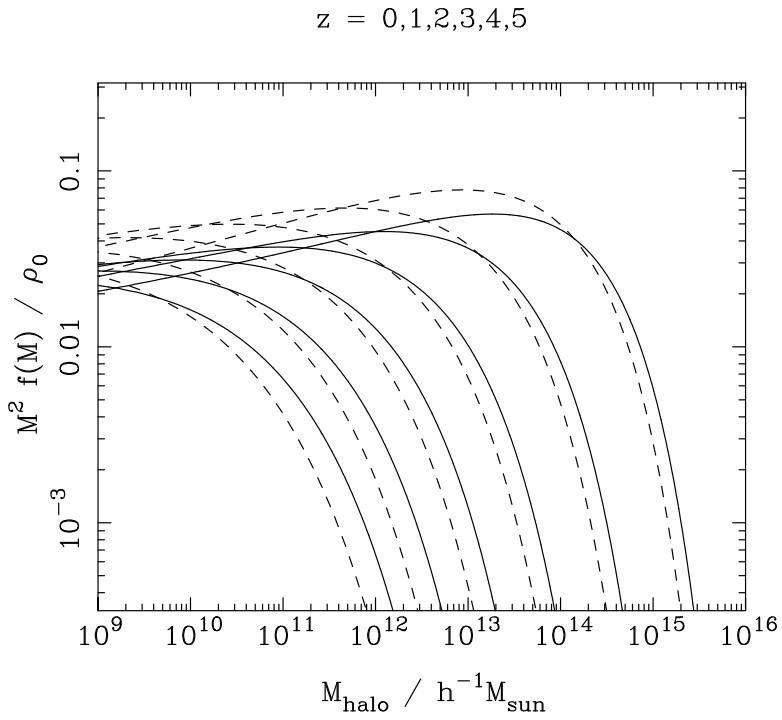

Figure 11. The mass function in the form of the **multiplicity function**: fraction of mass in the universe found in virialized haloes per unit range in $\ln M$. The solid lines show a fitting formula to N-body data and the dashed lines contrast the original Press-Schechter formula.

2.3 Biased clustering and haloes

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**:

$$\left(\frac{\delta\rho}{\rho}\right)_{\rm galaxies} = b\left(\frac{\delta\rho}{\rho}\right)_{\rm mass}. \tag{154}$$

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 12; 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

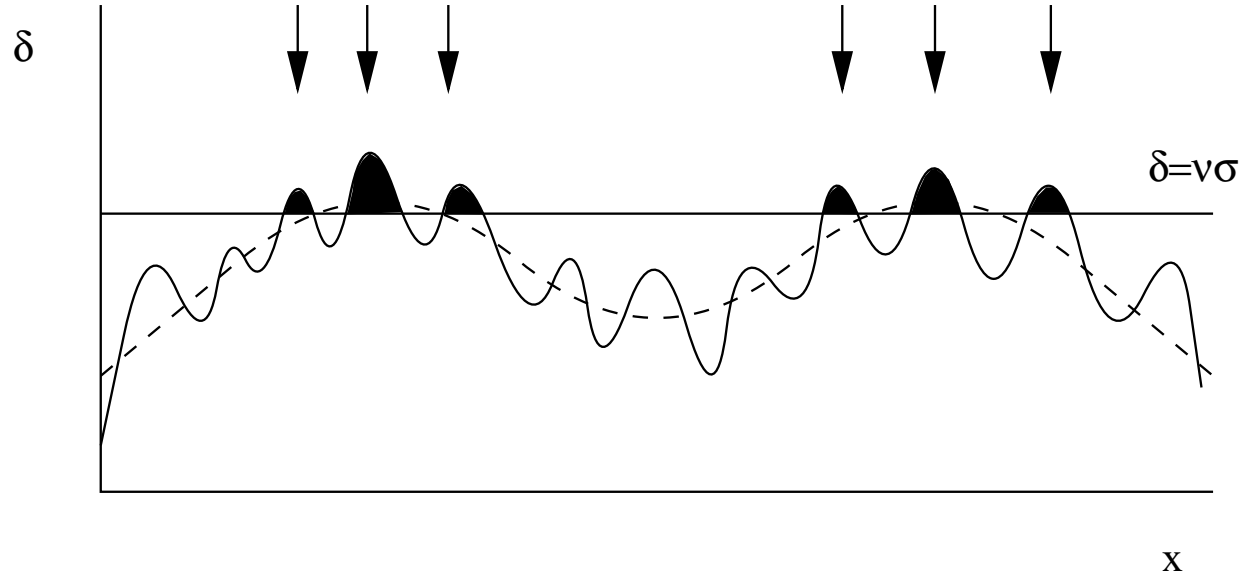

Figure 12. The high-peak bias model. If we decompose a density field into a fluctuating component on galaxy scales, together with a long-wavelength 'swell' (shown dashed), then those regions of density that lie above a threshold in density of ν times the rms will be strongly clustered. If proto-objects are presumed to form at the sites of these high peaks (shaded, and indicated by arrows), then this is a population with Lagrangian bias – i.e. a non-uniform spatial distribution even prior to dynamical evolution of the density field. 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.

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 δ averaged over a box containing mass M reaches some critical value δ_c. Generally, we shall assume the value $\delta_c = 1.686$ appropriate for spherical collapse in an Einstein–de Sitter universe. Now imagine that this situation is perturbed, by adding some constant shift ϵ to the density perturbations over some large region. The effect of this is to perturb the threshold: fluctuations now only need to reach $\delta = \delta_c - \epsilon$ in order to achieve collapse.

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

$$F \to F - (dF/d\delta_c)\,\epsilon. \tag{155}$$

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

$$F \to F - (dF/d\ln\nu)\,\epsilon/\delta_c. \tag{156}$$

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

$$M^2 f(M)/\rho_0 = (d\ln\nu/d\ln M)\,|dF/d\ln\nu|. \tag{157}$$

So the mass function is proportional to the positive quantity $G \equiv -dF/d\ln\nu$, which is perturbed as follows:

$$G \to G - (dG/d\ln\nu)\,\epsilon/\delta_c. \tag{158}$$

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

$$b_{\rm L} = -\frac{d\ln G}{d\ln\nu}/\delta_c. \tag{159}$$

In addition to this modulation of the halo properties, the large-scale disturbance will move haloes closer together where ϵ 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

$$b = 1 - \frac{d\ln G}{d\ln \nu}/\delta_c. \tag{160}$$

Of course, the field ϵ can hardly be imposed by hand; instead, we make the **peak-background split**, in which δ is mentally decomposed into a small-scale and a large-scale component – which we identify with ϵ. The scale above which the large-scale 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.

From this point of view, it is clear that the mean bias should come out as unity when we average over all haloes: in this limit, we recover the full fluctuating mass density field. Thus we wish to find $\langle b\rangle = \int b(M)\, M\, f(M)\, dM/\rho_0$, weighting each halo by mass. From the definition of the mass function, this is

$$\langle b\rangle = \int b(M)\, dF = 1 + \int (d\ln G/d\ln\nu)\, G\, d\ln\nu \;/\; \delta_c. \tag{161}$$

The latter integral is proportional to the difference in G between infinite mass and zero mass, which vanishes if the mass function vanishes in these limits.

To apply this apparatus, 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 $\nu = \delta_c/\sigma(M)$:

$$G(\nu) = \sqrt{\frac{2}{\pi}}\,\nu\,\exp\left(-\frac{\nu^2}{2}\right). \tag{162}$$

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

$$b(\nu) = 1 + \frac{\nu^2 - 1}{\delta_c}. \tag{163}$$

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):

$$G(\nu) = 0.21617[1 + (\sqrt{2}/\nu^2)^{0.3}]\,\nu\,\exp[-\nu^2/(2\sqrt{2})]. \tag{164}$$

As stated earlier, an even better approximation is

$$F(>\nu) = (1 + a\,\nu^b)^{-1}\exp(-c\,\nu^2), \tag{165}$$

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

$$b(\nu) = 1 + \frac{\nu^2 - 1}{(1 + z_f)\delta_c}. \tag{166}$$

(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.

3 The hot big bang

3.1 Thermal history

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

$$dN = g\,\frac{V}{(2\pi)^3}\,d^3k \tag{167}$$

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

$$\langle f\rangle = \left[e^{(\epsilon-\mu)/kT} \pm 1\right]^{-1} \tag{168}$$

(+ for fermions, − for bosons). Now, for a thermal radiation background, the **chemical potential**, μ is always zero. The reason for this is quite simple: μ appears in the first law of thermodynamics as the change in energy associated with a change in particle number, $dE = TdS - PdV + \mu dN$. So, as N adjusts to its equilibrium value, we expect that the system will be stationary with respect to small changes in N. The thermal equilibrium **background number density** of particles is

$$n = \frac{1}{V}\int f\,dN = g\,\frac{1}{(2\pi\hbar)^3}\int_0^\infty \frac{4\pi\,p^2dp}{e^{\epsilon(p)/kT}\pm 1}, \tag{169}$$

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

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

$$n = \left(\frac{kT}{c}\right)^3 \frac{4\pi g}{(2\pi\hbar)^3} \int_0^\infty \frac{y^2 dy}{e^y \pm 1}. \tag{170}$$

(2) Non-relativistic limit. Here we can neglect the ± 1 in the occupation number, in which case the number is suppressed by a dominant $\exp(-mc^2/kT)$ factor:

$$\frac{n}{n_{\rm ultrarel}} = (mc^2/kT)^{3/2} \exp(-mc^2/kT) \times [0.695\,({\rm F})\ 0.521\,({\rm B})]. \tag{171}$$

This shows us that the background 'switches on' at about $kT \sim mc^2$; at this energy, known as a **threshold**, photons and other species in equilibrium will have sufficient energy to create particle-antiparticle pairs.

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

$$u = \rho\, c^2 = g\, \frac{1}{(2\pi\hbar)^3} \int_0^\infty \frac{4\pi\ p^2\, dp}{e^{\epsilon(p)/kT} \pm 1}\ \epsilon(p). \tag{172}$$

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

$$u = \frac{\pi^2}{30(\hbar c)^3}\, g\, (kT)^4 \quad \text{(bosons)}, \tag{173}$$

and 7/8 of this for fermions.

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

$$dE = TdS - PdV \quad \Rightarrow \quad \left(\frac{E}{V}dV + \frac{\partial E}{\partial T}dT\right) = \left(T\frac{S}{V}dV + T\frac{\partial S}{\partial T}dT\right) - PdV. \tag{174}$$

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

$$S = \frac{E + PV}{T} \tag{175}$$

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

$$s = (4/3)u/T = \frac{2\pi^2 k}{45(\hbar c)^3}\, g\,(kT)^3 \quad \text{(bosons)}, \tag{176}$$

and 7/8 of this for fermions. Now, we saw earlier that the number density of an ultrarelativistic background also scales as T^3 – therefore we have the simple result that entropy just counts the number of particles. This justifies a common piece of terminology, in which the ratio of the number density of photons in the universe to the number density of **baryons** (protons plus neutrons) is called the **entropy per baryon**. As we will see later, this ratio is about 10^9. The fact that this ratio is so large justifies the adiabatic assumption: pretty well all the entropy is in the photons.

DEGREES OF FREEDOM Overall, the equilibrium relativistic density is

$$\rho c^2 = \frac{\pi^2}{30(\hbar c)^3}\, g_{\rm eff}\,(kT)^4; \quad g_{\rm eff} \equiv \sum_{\rm bosons} g_i + \frac{7}{8}\sum_{\rm fermions} g_j, \tag{177}$$

expressing the fermion contribution as an effective number of bosons. A similar relation holds for entropy density: $s = [2\pi^2 k/45(\hbar c)^3]\, h_{\rm eff}\ (kT)^3$. In equilibrium, $h_{\rm eff} = g_{\rm eff}$, but this ceases to be true at late times, when the neutrinos and photons have different temperatures. The $g_{\rm eff}$ functions are plotted against photon temperature in figure 13. They start at a number determined by the total number of distinct elementary particles that exist (of order 100, according to the standard model of particle physics), and fall as the temperature drops and more species of particles become nonrelativistic.

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

$$t/\text{seconds} = g_{\rm eff}^{-1/2}\ \left(T/10^{10.26}\,\text{K}\right)^{-2}. \tag{178}$$

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

$$T = 2.725 \pm 0.002\,\text{K}. \tag{179}$$

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

$$T(z) = 2.725(1+z), \tag{180}$$

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

The energy density in photons is supplemented by that of the neutrino background. Because they have a lower temperature, as shown below, they contribute an energy density 0.68 times that from the photons (if the neutrinos are massless and therefore relativistic). If there

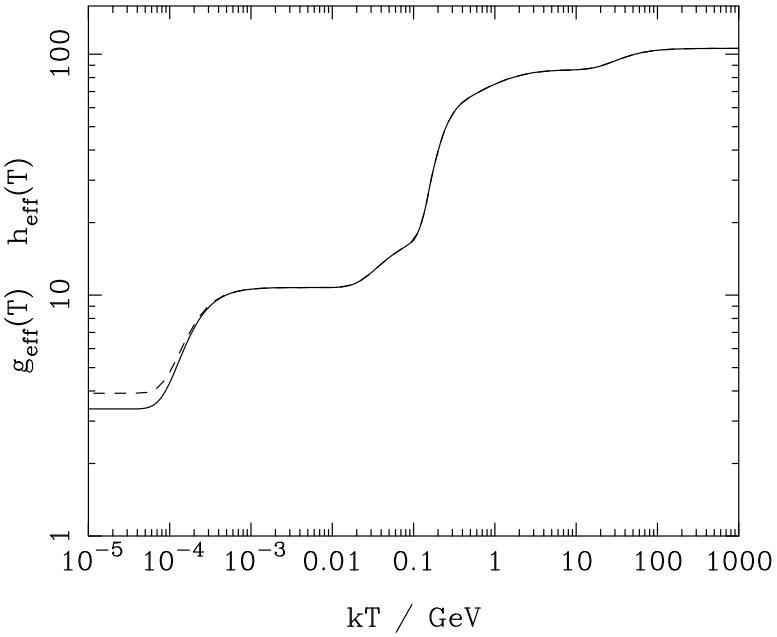

Figure 13. The number of relativistic degrees of freedom as a function of photon temperature. $g_{\rm eff}$ measures the energy density; $h_{\rm eff}$ the entropy (dashed line). The two depart significantly at low temperatures, when the neutrinos are cooler than the photons. For a universe consisting only of photons, we would expect $g = 2$. The main features visible are (1) The electroweak phase transition at 100 Gev; (2) The QCD phase transition at 200 MeV; (3) the $e^{\pm}$ annihilation at 0.3 MeV.

are no other contributions to the energy density from relativistic particles, then the total effective radiation density is $\Omega_r h^2 \simeq 4.2 \times 10^{-5}$ and the redshift of **matter–radiation equality** is

$$1 + z_{\rm eq} = 24\,074\,\Omega h^2\,(T/2.725\,{\rm K})^{-4}. \tag{181}$$

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

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

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

3.2 Freezeout and relics

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

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

$$\dot{n} + 3Hn = -\langle \sigma v \rangle n^2 + S, \tag{182}$$

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

$$\begin{aligned} \textbf{expansion timescale} &= H(z)^{-1} \\ \textbf{interaction timescale} &= (\langle \sigma v \rangle n)^{-1} \end{aligned} \tag{183}$$

Both these times increase as the universe expands, but the interaction time usually changes fastest. Two-body reaction rates scale proportional to density, times a cross-section that is often a declining function of energy, so that the interaction time changes at least as fast as R^3. In contrast, the Hubble time changes no faster than R^2 (in the radiation era), so that there is inevitably a crossover.

The situation therefore changes from one of thermal equilibrium at early times to a state of **freezeout** or **decoupling** at late times. Once the interaction timescale becomes much longer than the age of the universe, the particle has effectively ceased to interact. It thus preserves a 'snapshot' of the properties of the universe at the time the particle was last in thermal equilibrium. This phenomenon of freezeout is essential to the understanding of the present-day nature of the universe. It allows for a whole set of **relics** to exist from different stages of the hot big bang. The photons of the microwave background are one such relic, generated at redshift $z \simeq 1100$. A more exotic example is the case of neutrinos.

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

$$S = \langle \sigma v \rangle n_T^2. \tag{184}$$

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

$$\frac{d\ln N}{d\ln a} = -\frac{\Gamma}{H}\left[1 - \left(\frac{N_{_T}}{N}\right)^2\right], \tag{185}$$

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

Unfortunately, this equation must be solved numerically. The main features are easy enough to see, however. Suppose first that the universe is sustaining a population in approximate thermal equilibrium, $N \simeq N_{_T}$. If the population under study is relativistic, $N_{_T}$ does not change with time, because $n_{_T} \propto T^3$ and $T \propto a^{-1}$. This means that it is possible to keep $N = N_{_T}$ exactly, whatever Γ/H. It would however be grossly incorrect to conclude from this that the population stays in thermal equilibrium: if $\Gamma/H \ll 1$, a typical particle suffers no interactions even while the universe doubles in size, halving the temperature. A good example is the microwave background, whose photons last interacted with matter at $z \simeq 1100$. The CMB nevertheless still appears to be equilibrium black-body radiation because the number density of photons has fallen by the right amount to compensate for the redshifting of photon energy. This sounds like an incredible coincidence, but is in fact quite inevitable when looked at from the quantum-mechanical point of view. This says that the occupation number of a given mode, $= (\exp\hbar\omega/kT - 1)^{-1}$ for thermal radiation, is an adiabatic invariant that does not change as the universe expands – only the frequency alters, and thus the apparent temperature.

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

$$N(a \to \infty) = N_{_T}(\Gamma/H = 1). \tag{186}$$

Exact numerical solutions of the rate equation almost always turn out very close to this simple rule, as shown in figure 14.

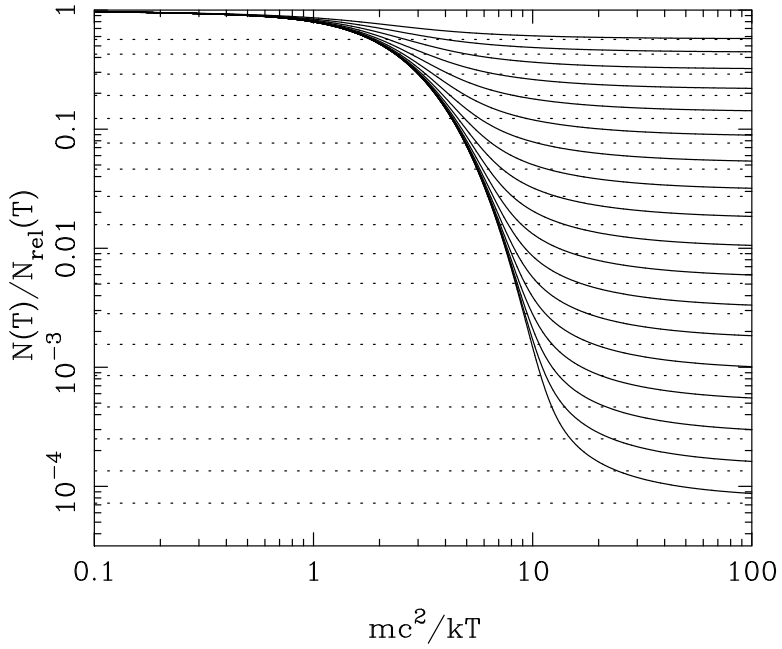

Figure 14. Solution of the Boltzmann equation for freezeout of a single massive fermion. We set $\Gamma/H = \epsilon(kT/mc^2)N/N_{\rm rel}$, as appropriate for a radiation-dominated universe in which $\langle\sigma v\rangle$ is assumed to be independent of temperature. The solid lines show the case $\epsilon = 1$ and increasing by powers of 2. A high value of ϵ leads to freezeout at increasingly low abundances. The dashed lines show the abundance predicted by the simple recipe of the thermal density for which $\Gamma/H = 1$.

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

$$\Omega_{\rm relic}h^2 \simeq 10^{-41.5}\ (\sigma/{\rm m}^2)^{-1}, \tag{187}$$

so that only a small range of annihilation cross-sections will be of observational interest. The steps needed to get this formula are as follows. (1) From $\Gamma/H = 1$, the number density of relics at freezeout is $n_f = H_f/\langle\sigma v\rangle$; (2) $H = (8\pi G\rho/3)^{1/2}$, where $\rho c^2 = (\pi^2/30\hbar^3c^3)g_{\rm eff}(kT)^4$; (3) $\Omega_{\rm relic} = 8\pi G m n_0/3H_0^2$. The only missing ingredient here is how to relate the present number density n_0 to the density n_f at temperature T_f. Since the relics are conserved, the number density must have fallen by the same factor as the entropy density:

$$n_f/n_0 = (h^f_{\rm eff}T_f^3)/(h^0_{\rm eff}T_0^3). \tag{188}$$

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

$$\Omega_{\rm relic}h^2 \simeq \frac{10^{-33.0}\ {\rm m}^2}{\langle\sigma v\rangle}\ \left(\frac{mc^2}{kT_f}\right)\ g_{\rm eff}^{-1/2}. \tag{189}$$

We see from figure 14 that $mc^2/kT_f \sim 10$ with only a logarithmic dependence on reaction rate, which roughly cancels the last factor on the rhs. Finally, since particles are nearly relativistic at freezeout, we set $\langle\sigma v\rangle = \sigma c$ to get our final estimate of the typical cross-section for an interesting relic abundance. The eventual conclusion makes sense: the higher the cross-section, the longer the particle can stay in equilibrium, and the more effective annihilations can be in suppressing the number density. Note that, in detail, we need to worry about whether the particle is a **Majorana particle** (i.e. its own antiparticle) or a **Dirac particle** where particles and antiparticles are distinct.

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

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

$$s(\gamma + e^+ + e^-) = \frac{11}{4}s(\gamma). \tag{190}$$

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

$$T_\nu = \left(\frac{4}{11}\right)^{1/3} T_\gamma = 1.945\,\mathrm{K}, \tag{191}$$

for $T_\gamma = 2.725$ K. Although it is hard to see how such low energy neutrinos could ever be detected directly, their gravitation is certainly not negligible: they contribute an energy density that is a factor $(7/8) \times (4/11)^{4/3}$ times that of the photons. For three neutrino species, this enhances the energy density in relativistic particles by a factor 1.68 (there are three different kinds of neutrinos, just as there are three **leptons**: the μ and τ particles are heavy analogues of the electron).

MASSIVE NEUTRINOS Although for many years the conventional wisdom was that neutrinos were massless, this assumption began to be increasingly challenged around the end of the 1970s. Theoretical progress in understanding the origin of masses in particle physics meant that it was no longer natural for the neutrino to be completely devoid of mass. Also, there is now some experimental evidence that neutrinos have a small non-zero mass. The consequences of this for cosmology could be quite profound, as relic neutrinos are expected to be very abundant. The above section showed that $n(\nu + \overline{\nu}) = (3/4)n(\gamma; T = 1.95\,\mathrm{K})$. That yields a total of 113 relic neutrinos in every cm^3 for each species. The density contributed by these particles is easily worked out provided the mass is small enough. If this is the case, then the neutrinos were ultrarelativistic at decoupling and their statistics were those of massless particles. As the universe expands to $kT < m_\nu c^2$, the total number of neutrinos is preserved. We therefore obtain the present-day mass density in neutrinos just by multiplying the zero-mass number density by m_ν, and the consequence for the cosmological density in light neutrinos is easily worked out to be

$$\Omega_\nu h^2 = \frac{\sum m_i}{94.1\,\mathrm{eV}}. \tag{192}$$

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

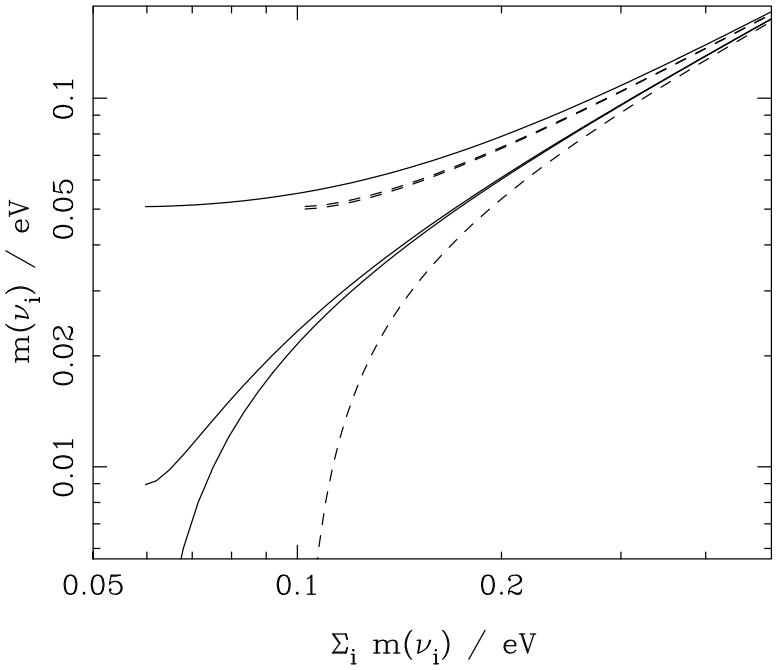

Figure 15. The masses of the individual neutrino mass eigenstates, plotted against the total neutrino mass for a normal hierarchy (solid lines) and an inverted hierarchy (dashed lines). Current cosmological data set an upper limit on the total mass of light neutrinos of around 0.5 eV.

The current direct laboratory limits to the neutrino masses are

$$\nu_e \lesssim 2.2\,\mathrm{eV} \quad \nu_\mu \lesssim 0.17\,\mathrm{MeV} \quad \nu_\tau \lesssim 15\,\mathrm{MeV}. \tag{193}$$

Based on this, even the electron neutrino could be of great cosmological significance. But in practice, we will see later that studies of cosmological large-scale structure limit the sum of the masses to a maximum of about 0.5 eV. This is becoming interesting, since it is

known that neutrino masses must be non-zero. In brief, this comes from studies of **neutrino mixing**, in which each neutrino type is a mixture of energy eigenstates. The energy differences can be measured, which yields a measure of the difference in the square of the masses (consider the relativistic relation $E^2 = m^2 + p^2$, and expand to get $E \simeq m + m^2/2p$). These mixings are known from wonderfully precise experiments detecting neutrinos generated in the sun and the Earth's atmosphere:

$$\begin{aligned} \Delta(m_{21})^2 &= 8.0 \times 10^{-5}\ \mathrm{eV}^2 \\ \Delta(m_{32})^2 &= 2.5 \times 10^{-3}\ \mathrm{eV}^2, \end{aligned} \tag{194}$$

where m_1, m_2 and m_3 are the three mass eigenstates. This information does not give the absolute mass scale, nor does it tell us whether there is a **normal hierarchy** with $m_3 \gg m_2 \gg m_1$, or an **inverted hierarchy** in which states 1 & 2 are a close doublet lying well above state 3. Cosmology can settle both these issues by measuring the total density in neutrinos. The absolute minimum situation is a normal hierarchy with m_1 negligibly small, in which case the mass is dominated by m_3, which is around 0.05 eV. The cosmological limits are within a power of 10 of this interesting point.

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

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

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

(3) **Cold Dark Matter (CDM)** If the relic particles decouple while they are nonrelativistic, the number density can be exponentially suppressed. If the interactions are like those of neutrinos, then the freezeout temperature is about 1 MeV, and the relic mass density then falls with increasing mass (see figure 16). For weak interactions, cross-sections scale as (energy)^2, so that the relic density falls as $1/m^2$. Interesting masses then lie in the $\simeq 10$ GeV range, this cannot correspond to the known neutrinos, since such particles

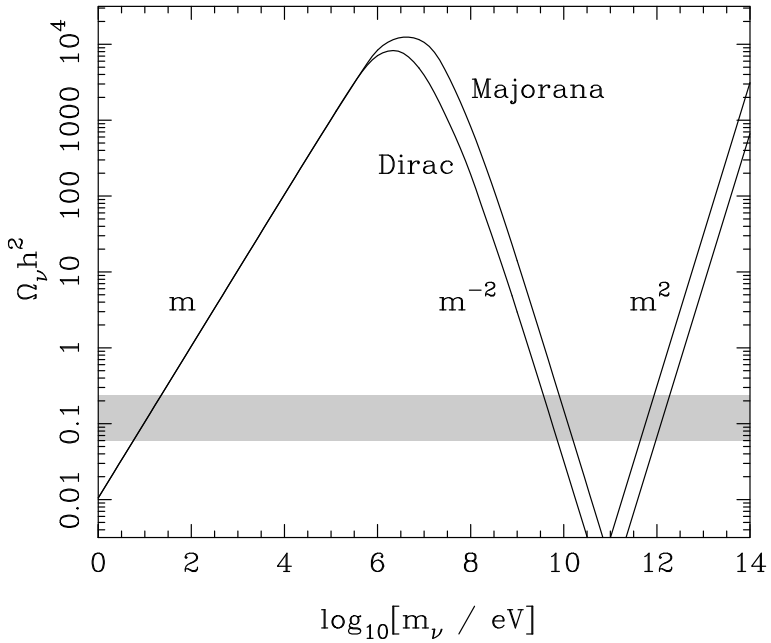

Figure 16. The contribution to the density parameter produced by relic neutrinos (or neutrino-like particles) as a function of their rest mass. The shaded band shows a factor of 2 either side of the observed CDM density. At low masses, the neutrinos are highly relativistic when they decouple: their abundance takes the zero-mass value, and the density is just proportional to the mass. Above about 1 MeV, the neutrinos are non-relativistic at decoupling, and their relic density is reduced by annihilation. Above the mass of the Z boson, the cross-section falls, so that annihilation is less effective and the relic density rises again.

would have been seen in accelerators. But beyond about 90 GeV (the mass of the Z boson), the strength of the weak interaction is reduced, with cross-section going as (energy)$^{-2}$. The relic density now rises as m^2, so that the observed dark matter density is

attained at $m \simeq 1$ TeV. Plausible candidates of this sort are found among so-called **supersymmetric** theories, which predict many new weakly-interacting particles. The favoured particle for a CDM relic is called the **neutralino**.

Since these particles exist to explain galaxy rotation curves, they must be passing through us right now. There is therefore a huge effort in the direct laboratory detection of dark matter, mainly via cryogenic detectors that look for the recoil of a single nucleon when hit by a DM particle (in deep mines, to shield from cosmic rays). So far, there are no detections, but well-constructed experiments with low backgrounds are starting to set interesting limits, as shown in figure 17. There is no unique target to aim for, since even the simplest examples of supersymmetric models contain a variety of free parameters. These allow models that are optimistically close to current limits, but also some that will be hard to verify. The public-domain package **DarkSUSY** is available at `www.physto.se/~edsjo/darksusy` to make these detailed abundance calculations.

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

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

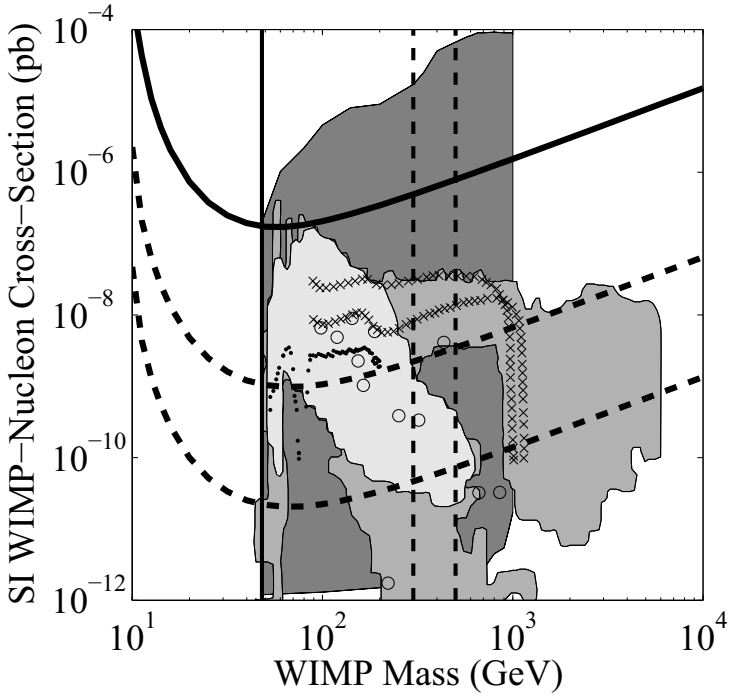

Figure 17. A plot of the dark-matter experimentalists' space: cross-section for scattering off nucleons (in wonderfully baroque units: the 'picobarn' is $10^{-40}\,\mathrm{m}^2$) against WIMP mass. The shaded areas and points indicate various supersymmetric models that match particle-physics constraints and have the correct relic density. The upper curve indicates current direct (non)detection limits, and dashed curves are where we might be in about a decade. Vertical lines are current collider limits, and predictions for the LHC and a future linear collider.

3.3 Primordial nucleosynthesis

At sufficiently early times, the temperature of the universe will be that of the centre of the Sun (1.55×10^7 K), where we know that nuclear reactions occur. Starting in the 1940s, Gamow considered the fascinating question of whether nuclear reactions were possible in the early universe. He noted that the abundances of some elements in stars showed great regularities, especially a universal proportion of about 25% Helium by mass. This led to the vision that a chain of nuclear reactions in the early universe could generate not only Helium, but all elements. In 1957, the Burbidges, Fowler & Hoyle showed that almost all elements could in fact be generated in stars, but the problem of Helium remained. Gamow showed that its existence could be used to predict the present radiation temperature, as argued below.

For this part, it will be convenient to refer to particle masses and temperatures in nuclear-physics units, which are MeV. Some useful conversions are:

$$\begin{aligned} 1\text{MeV} &= 10^{10.065}\,\text{K} \\ m_e &= 0.511\ \text{MeV} \\ m_p &= 939\ \text{MeV} \\ m_n - m_p &= 1.3\ \text{MeV} \end{aligned} \tag{195}$$

NEUTRON FREEZEOUT We have shown that, at temperatures below the nucleon mass threshold (about 10^{13} K), nucleon pairs will annihilate, leaving behind the residual matter imbalance over antimatter. For a while, the balance between neutrons and protons will be maintained in equilibrium by weak interactions:

$$\begin{aligned} p + e^- &\leftrightarrow n + \nu \\ n + e^+ &\leftrightarrow p + \bar{\nu}. \end{aligned} \tag{196}$$

While this persists, the relative number densities of neutrons and protons should vary according to a Boltzmann factor based on their mass difference:

$$\frac{n_n}{n_p} = e^{-\Delta mc^2/kT} \simeq e^{-1.5(10^{10}\,\text{K}/T)}. \tag{197}$$

The reason that neutrons exist today is that the timescale for the weak interactions needed to keep this equilibrium set up eventually becomes longer than the expansion timescale. The reactions thus rapidly cease, and the neutron–proton ratio undergoes **freezeout** at some characteristic value, which determines the He abundance. Since most He is ^{4}He, with 2 nucleons out of 4 being neutrons, the He fraction by mass (denoted by Y) is

$$Y = \frac{4 \times n_n/2}{n_n + n_p} = \frac{2}{1 + n_p/n_n} \tag{198}$$

(neglecting neutrons in other elements). So, $Y = 0.25$ requires freezeout at $n_n/n_p \simeq 1/7$.

To calculate when the neutron-to-proton ratio undergoes freezeout, we need to know the rates of weak nuclear reactions; Fermi discovered how to calculate the relevant cross-sections in the 1930s. Remember that, at $T \sim 10^{10}$ K, we are above the e^+e^- threshold, so there exist thermal populations of both neutrinos and electrons, to make the reaction $p + e^- \leftrightarrow n + \nu$ go equally well in either direction. All that is needed is therefore to consider either the reaction timescale for one proton immersed in a thermal bath of electrons or of one neutron immersed in a bath of neutrinos (the rates are the same). When this timescale equals the local Hubble time, $R(t)/\dot{R}(t)$, we get freezeout of the neutron-to-proton ratio. Taking the known weak reaction rates, this happens at

$$T(\text{n freezeout}) \simeq 10^{10.14}\,\text{K} \quad \Rightarrow \quad \frac{n_n}{n_p} \simeq 0.34. \tag{199}$$

This number is not a precisely correct result, because nucleosynthesis is a process that contains a number of interesting (but potentially confusing) coincidences:

(1) The freezeout condition was calculated assuming a temperature well above the electron mass threshold, but freezeout actually happens only a very little above this critical temperature.

(2) Neutrons are not stable: they decay spontaneously with the e-folding lifetime of $\tau_n = 887 \pm 2$ s. Unless the frozen-out neutrons can be locked away in nuclei before $t = 887$ s, the relic abundance will decay freely to zero. The freezeout point occurs at an age of a few seconds, so there are only a few e-foldings of expansion available in which to salvage some neutrons.

LOCKING UP THE NEUTRONS It may seem implausible that we can add one more coincidence – *i.e.* that nuclear reactions will become important at about the same time – but this is just what does happen. The Deuteron binding energy of 2.225 MeV is only 4.3 times larger than $m_e c^2$ and only 1.7 times larger than the neutron–proton mass difference. At higher temperatures, the strong interaction $n + p = \mathrm{D} + \gamma$ is fast enough to produce Deuterium, but thermal equilibrium favours a small Deuterium fraction – *i.e.* typical photons are energetic enough to disrupt Deuterium nuclei very easily. The second key temperature in nucleosynthesis is therefore where the universe has cooled sufficiently for the equilibrium to swing in favour of Deuterium. In practice, this happens at a temperature a little below the Deuteron binding energy. This is because of the large photon-to-baryon ratio: even if most photons lack sufficient energy to disintegrate Deuterons, the rare ones in the tail of the distribution can still do the job.

Nevertheless, the temperature at which Deuterium switches from being rare to dominating the equilibrium is still at kT of order the Deuterium binding energy:

$$T(\text{Deuterium formation}) \simeq 10^{8.9}\,\mathrm{K}, \tag{200}$$

or a time of about 3 minutes.

Notice that we have not needed to know the nuclear reaction rates that form Deuterium, since the argument is an equilibrium one. However, if the matter density is too low, the nuclear reactions will freeze out before much Deuterium is formed. Gamow took the known nuclear cross-sections, and argued that the typical reaction time for Deuterium formation had to be the cosmological age at that temperature (3 minutes). This let him conclude that the matter density must have been about $10^{-3}\,\mathrm{kg\,m^{-3}}$ at that time. This gives a ratio of number densities of photons to nucleons, which is preserved as the universe expands. Therefore, the present-day matter density allows a prediction of the present photon density, and hence its temperature. Alpher & Herman used Gamow's argument to predict a current temperature of 4 K to 5 K, which is impressively accurate. On the other hand, this prediction was based on a figure for the $z = 0$ matter density that is probably too low by at least a factor 100, raising the temperature estimate by a factor 5. Actually, Gamow's argument is an inequality: there is a minimum matter density at 10^9 K, but it could have been higher. The prediction for the current temperature is therefore really an upper limit. It works because the nuclear reactions are not too far from freezeout when Deuterium forms.

FORMATION OF HELIUM A universe consisting of just Hydrogen and Deuterium is not realistic: ^{4}He should be preferred on thermodynamic grounds, owing to its greater binding energy per nucleon (7 MeV, as opposed to 1.1 MeV for Deuterium). In equilibrium,

the abundance of ^{4}He relative to protons should reach unity at an energy of about 0.3 MeV, at which point the relative abundance of Deuterium should be only $\sim 10^{-12}$.

Since the simplest way to synthesize Helium is by fusing Deuterium, the production of Helium must in practice await the synthesis of significant quantities of Deuterium. Nevertheless, the Deuterium will be rapidly converted to Helium once significant nucleosynthesis begins. This argument is what allows us to expect that the Helium abundance can be calculated from the final n/p ratio. The process starts by fusing Deuterium to make either Tritium and ^{3}He, following which there are four main ways of reaching ^{4}He (leaving aside rarer reactions involving residual free neutrons):

$$\begin{aligned} \mathrm{D} + \mathrm{D} &= {}^3\mathrm{He} + n \\ \mathrm{D} + \mathrm{D} &= \mathrm{T} + p \end{aligned}$$

$$\begin{aligned} \mathrm{D} + \mathrm{D} &= {}^4\mathrm{He} \\ \mathrm{T} + p &= {}^4\mathrm{He} \\ \mathrm{D} + {}^3\mathrm{He} &= {}^4\mathrm{He} + p \\ \mathrm{T} + \mathrm{D} &= {}^4\mathrm{He} + n. \end{aligned} \tag{201}$$

A universe that stayed in nuclear equilibrium as it cooled would eventually consist entirely of Iron, since this has the highest binding energy per nucleon. However, by the time Helium synthesis is accomplished, the density and temperature are too low for significant synthesis of heavier nuclei to proceed. Apart from Helium, the main nuclear residue of the big bang is therefore those Deuterium nuclei that escape being mopped up into Helium, plus a trace of ^{3}He. The other intermediate product, Tritium, is not so strongly bound and thus leaves no significant relic. There also exist extremely small fractions of other elements: ^{7}Li ($\sim 10^{-9}$ by mass) and ^{7}Be ($\sim 10^{-11}$).

In summary, nucleosynthesis starts at about 10^{10} K, when the universe was about 1 s old, and effectively ends when it has cooled by a factor of 10, and is about 100 times older.

THE NUMBER OF PARTICLE GENERATIONS An accurate fit for the final neutron-to-proton ratio is

$$\frac{n_n}{n_p} \simeq 0.163\ (\Omega_{\rm B} h^2)^{0.04}\ (N_\nu/3)^{0.2}. \tag{202}$$

The signs of the dependences on the baryon density and on the number of neutrino species are easily understood. A high baryon density increases the temperature at which nuclei form and gives a higher neutron abundance because fewer of them have decayed. This is a weak effect, because the neutron fraction is largely set by weak-interaction freeze-out, which is independent of the baryon density. The effect of extra neutrino species is to boost the total relativistic density. This increase the overall rate of expansion, so that weak-interaction freezeout happens earlier and thus at higher T, again raising the neutron abundance.

It is therefore clear that strong limits can be set on the number of unobserved neutrino species, and thus on the number of possible additional families in particle physics. For many years, these nucleosynthesis limits were stronger than those that existed from particle physics experiments. This changed in 1990, with a critical series of experiments carried out in the **LEP** (large electron-positron) collider at CERN, which was the first experiment to produce Z^0 particles in large numbers. The Z^0 can decay to pairs of neutrinos so long as their rest mass sums to less than 91.2 GeV; more species increase the decay rate, and decrease the Z^0 lifetime. Since 1990, these arguments have required N to be very close to 3; it is a matter of detailed argument over the Helium data as to whether $N = 4$ was ruled out from cosmology prior to this.

WEIGHING THE BARYONS Unlike Helium, the relic abundances of the other light elements are rather sensitive to density. We have seen that Helium formation occurs at very nearly a fixed temperature, depending only weakly on density or neutrino species. The residual Deuterium will therefore freeze out at about this temperature, leaving a number density fixed at whatever sets the reaction rate low enough to survive for a Hubble time. Since this density is a fixed quantity, the *proportion* of the baryonic density that survives as Deuterium (or ^{3}He) should thus decline roughly as 1/(density).

This provides a relatively sensitive means of weighing the baryonic content of the universe. A key event in the development of cosmology was thus the determination of the D/H ratio in the interstellar medium, carried out in the early 1970s. This gave D/H $\simeq 2\times10^{-5}$, providing the first evidence for a low baryonic density, as follows. Figure 18 shows how the abundances of light elements vary with the

cosmological density, according to detailed calculations. The baryonic density in these calculations is traditionally quoted in the field as the reciprocal of the entropy per baryon:

$$\eta \equiv (n_p + n_n)/n_\gamma = 2.74 \times 10^{-8} (T/2.73\,\mathrm{K})^{-3} \Omega_{\mathrm{B}} h^2. \tag{203}$$

Figure 18 shows that this Deuterium abundance favours a low density:

$$\Omega_{\mathrm{B}} h^2 \simeq 0.02 \pm 0.002. \tag{204}$$

(although lower values are favoured if a higher weight is given to the Helium abundance). Baryons therefore cannot close the universe. If $\Omega = 1$, the dark matter must be non-baryonic.

3.4 Recombination

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

THE RATE EQUATION A natural first thought is that the ionization of the plasma may be treated by a thermal-equilibrium approach, using the Saha equation, which applies in stellar atmospheres. In fact, such an approach is almost always invalid. This is not because electromagnetic interactions are too slow to maintain equilibrium: rather, they are too fast. Consider a single recombination; if this were to occur directly to the ground state, a photon with $\hbar\omega > \chi$ would be produced. Such photons are almost immediately destroyed by ionizing another neutral atom. Similarly, reaching the ground state requires the production of photons at least as energetic as the $2P \to 1S$ spacing (Lyman α, with $\lambda = 1216$Å), and these also are re-absorbed very efficiently. This is a common phenomenon in astrophysics: the Lyman α photons undergo **resonant scattering** and are very hard to get rid of (unlike a finite HII region, where the Lyα photons can escape).

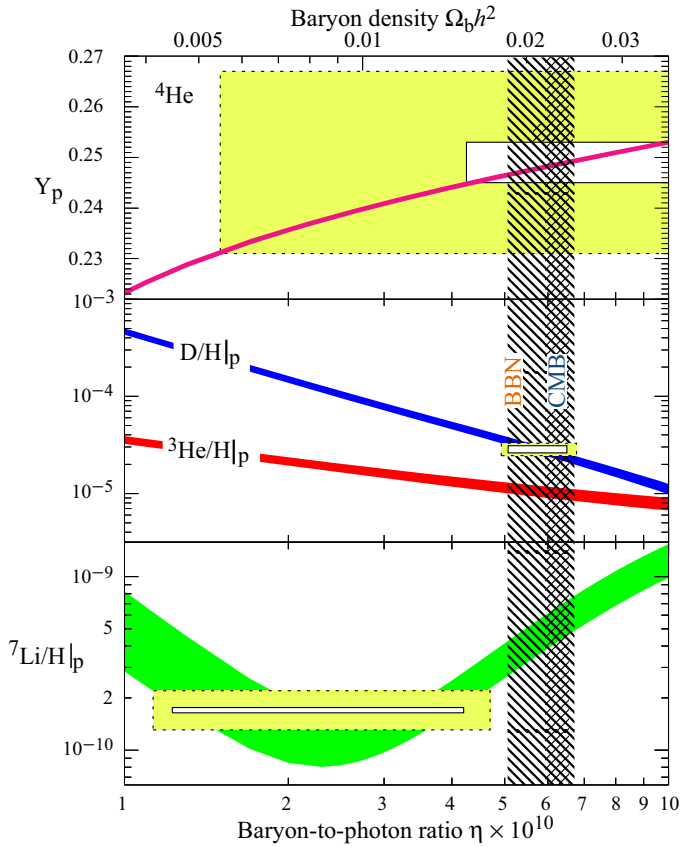

Figure 18. The predicted primordial abundances of the light elements, as a function of the baryon density and the baryon-to-photon ratio η; for a microwave-background temperature of 2.73 K, this is related to the baryonic density parameter via $\Omega_B h^2 = \eta/2.74\times10^{-8}$. Rectangular bars show the 95% confidence regions for observations of various species, with and without allowance for systematics. Concordance with the data is found only for $\eta \simeq 3\times10^{-10}$, driven mainly by the Deuterium data. This is consistent with the rather more precise determination that arises from CMB data. Taken from the PDG (Amsler et al. 2008).

There is a way out, however, using **two-photon emission**. The $2S \rightarrow 1S$ transition is strictly forbidden at first order and one can only conserve energy and angular momentum in the transition by emitting a *pair* of photons. Because of this slow bottleneck, the ionization at low redshift is far higher than would be suggested by the Saha prediction.

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

$$\frac{d(nx)}{dt} = -R\,(nx)^2\,\frac{\Lambda_{2\gamma}}{\Lambda_{2\gamma} + \Lambda_{\rm U}(T)}, \tag{205}$$

where n is the number density of protons, x is the fractional ionization, R is the recombination coefficient ($R \simeq 3 \times 10^{-17} T^{-1/2}\,{\rm m^3 s^{-1}}$), $\Lambda_{2\gamma}$ is the two-photon decay rate, and $\Lambda_{\rm U}(T)$ is the stimulated transition rate upwards from the $2S$ state. This equation just says that recombinations are a two-body process, which create excited states that cascade down to the $2S$ level, from whence a competition between the upward and downward transition rates determines the fraction that make the downward transition. A fuller discussion (see chapter 6 of Peebles 1993 or section 3.6 of Mukhanov 2005) would include a number of other processes: depopulation of the ground state by inverse 2-photon absorption; redshifting of Ly alpha photons due to the universal expansion, which can prevent them being re-absorbed. At the redshifts of practical interest (1000 to 10), the simplified equation captures the main effect, although detailed calculations have to include the recombination of He as well as H.

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

$$\frac{dt}{dz} \simeq -3.09 \times 10^{17} (\Omega_m h^2)^{-1/2}\, z^{-5/2}\ {\rm s}, \tag{206}$$

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

$$x(z) \propto \frac{(\Omega_m h^2)^{1/2}}{\Omega_b h^2}. \tag{207}$$

This is a very different scaling from the prediction of the Saha equation.

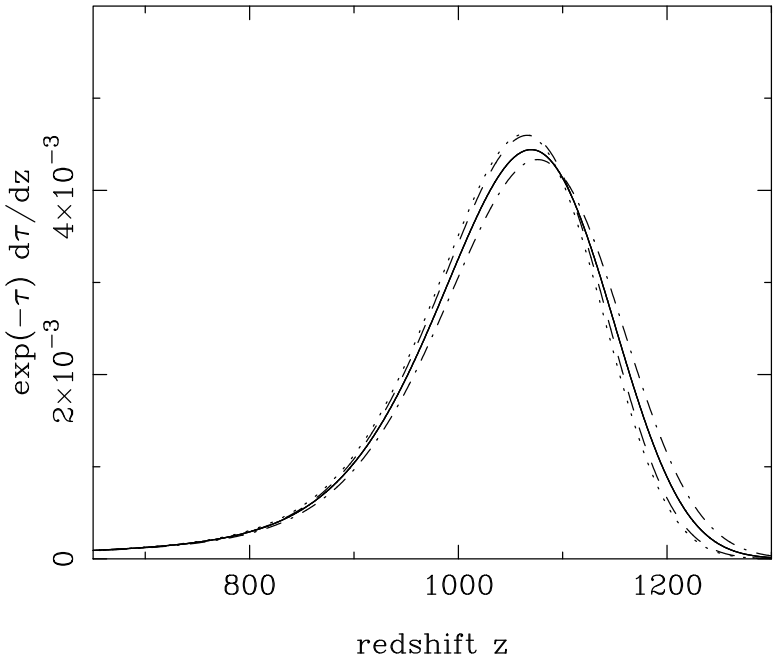

Figure 19. The 'visibility function' governing where photons in the CMB undergo their final scattering. This is very nearly independent of cosmological parameters, as illustrated by the effect of a 50% increase in Ω_b (dotted line), Ω_m (dot-dashed line) and h (dashed line), relative to the standard model (solid line).

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

$$\tau = \int n_e x \sigma_{\rm T} d\ell_{\rm proper}; \qquad d\ell_{\rm proper} = R(z)\, dr = R_0\, dr/(1+z). \tag{208}$$

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

$$n_e(z) = 9.83\, \Omega_b h^2\, (1+z)^3 \ {\rm m}^{-3}. \tag{209}$$

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

$$\tau(z) \simeq \left(\frac{1+z}{1080} \right)^{13} \tag{210}$$

(*cf.* Jones & Wyse 1985).

This approximation is not perfect, however, and very accurate work needs detailed numerical solution of the evolution equations, including the omitted processes. See Seager, Sasselov & Scott (2000; ApJS, 128, 407). Because τ changes rapidly with redshift, the **visibility function** for the redshift at which photons were last scattered, $e^{-\tau} d\tau/dz$, is sharply peaked, and is well fitted by a Gaussian of mean redshift 1070 and standard deviation in redshift 80. As illustrated in figure 19, these properties are in practice insensitive to the cosmological parameters. Thus, when we look at the sky, we can expect to see in all directions photons that originate from a **last-scattering surface** at $z \simeq 1100$. It is worth noting that this redshift is very much lower than we would expect just from setting

$$k \times 2.725 \ {\rm K} \times (1+z) = \chi, \tag{211}$$

which gives $z \simeq 10^{4.8}$. The difference is partly because the ionization falls slower than Saha, but also because even a tiny ionization easily causes scattering. The fact that the properties of the last-scattering surface are almost independent of all the unknowns in cosmology is immensely satisfying, and gives us at least one relatively solid piece of ground to act as a base in exploring the trackless swamp of cosmology.

4 Frontiers

In the remaining time, we survey some of the big questions raised by the standard model of cosmology, in the hope that subsequent lecturers will be able to answer them satisfactorily.

4.1 Initial condition problems

The expanding universe of the big-bang model is surprising in many ways: (1) What caused the expansion? (2) Why is the expansion so close to flat – i.e. $\Omega \sim 1$ today? (3) Why is the universe close to isotropic (the same in all directions)? (4) Why does it contain structure? Some of these problems may seem larger than others, but when examined in detail all point to something missing in our description of the early stages of cosmological history. It is normally assumed that the solution to this will be some piece of additional physics that is important at high energies. Before discussing possibilities, there are a few important general ideas that will be needed.

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

$$\begin{aligned} m_{\rm P} &\equiv \sqrt{\frac{\hbar c}{G}} \simeq 10^{19}\text{GeV} \\ \ell_{\rm P} &\equiv \sqrt{\frac{\hbar G}{c^3}} \simeq 10^{-35}\text{m} \\ t_{\rm P} &\equiv \sqrt{\frac{\hbar G}{c^5}} \simeq 10^{-43}\text{s}. \end{aligned} \tag{212}$$

The Planck time therefore sets the origin of time for the classical phase of the big bang. It is incorrect to extend the classical solution to $R = 0$ and conclude that the universe began in a singularity of infinite density. A common question about the big bang is 'what happened at $t < 0$?', but in fact it is not even possible to get to zero time without adding new physical laws.

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

$$k = \hbar = c = \mu_0 = \epsilon_0 = 1. \tag{213}$$

This convention makes the meaning of equations clearer by reducing the algebraic clutter, and is also useful in the construction of intuitive arguments for the order of magnitude of quantities of interest. In the professional world of cosmology, natural units are frequently adopted without comment, so this is something to watch out for. Hereafter, natural units will frequently be adopted, although it will occasionally be convenient to re-insert explicit powers of $\hbar$ *etc.*

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

$$\begin{aligned} [E] = [T] &= [m] \\ [L] &= [m]^{-1} \end{aligned} \tag{214}$$

Hence, the dimensions of energy density are

$$[u] = [m]^4, \tag{215}$$

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

$$G = m_{\mathrm{P}}^{-2}. \tag{216}$$

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

$$\Omega(a) = \frac{8\pi G\rho(a)}{H^2(a)} = \frac{\Omega_v + \Omega_m a^{-3} + \Omega_r a^{-4}}{\Omega_v + \Omega_m a^{-3} + \Omega_r a^{-4} - (\Omega - 1)a^{-2}} \tag{217}$$

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

$$\Omega(a_{\rm init}) \simeq 1 + \frac{(\Omega - 1)}{\Omega_r}\, a_{\rm init}^2. \tag{218}$$

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

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

$$r_{\rm H} = \int \frac{c\,dt}{R(t)}. \tag{219}$$

The standard radiation-dominated $R \propto t^{1/2}$ law makes this integral converge near $t = 0$. To solve the horizon problem and allow causal contact over the whole of the region observed at last scattering requires a universe that expands 'faster than light' near $t = 0$: $R \propto t^{\alpha}$, with

$\alpha > 1$. It is tempting to assert that the observed homogeneity *proves* that such causal contact must once have occurred, but this means that the equation of state at early times must have been different. Indeed, if we look at Friedmann's equation in its second form,

$$\ddot{R} = -4\pi GR(\rho + 3p/c^2)/3, \tag{220}$$

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

$$\rho c^2 + 3p < 0. \tag{221}$$

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

$$R \propto \exp Ht, \tag{222}$$

where $H = \sqrt{8\pi G\rho_{\rm vac}/3}$. This is the basic idea of the **inflationary universe**: vacuum repulsion can cause the universe to expand at an ever-increasing rate. This launches the Hubble expansion, and solves the horizon problem by stretching a small causally-connected patch to a size large enough to cover the whole presently-observable universe. This is illustrated by in figure 20, where we assume that the universe can be made to change its equation of state abruptly from vacuum dominated to radiation dominated at some time t_c. Before t_c, we have $R \propto \exp Ht$; after t_c, $R \propto t^{1/2}$. We have to match R and $\dot{R}$ at the join (otherwise the acceleration form of Friedmann's equation would be singular); it is then easy to show that $t_c = 1/2H$. When we observe the universe at $t > t_c$, we predict that there was a singularity at $t = 0$, but the real universe existed far earlier than this. In principle, the question 'what happened before the big bang?' is now answered: there was no big bang. There might have still been a singularity at large negative time, but one could imagine the de Sitter phase being of indefinite duration, so that the true origin of everything can be pushed back to $t = -\infty$. In a sense, then, an inflationary start to the expansion would in reality be a very slow one – as compared to the common popular description of 'an extraordinarily rapid phase of expansion'.

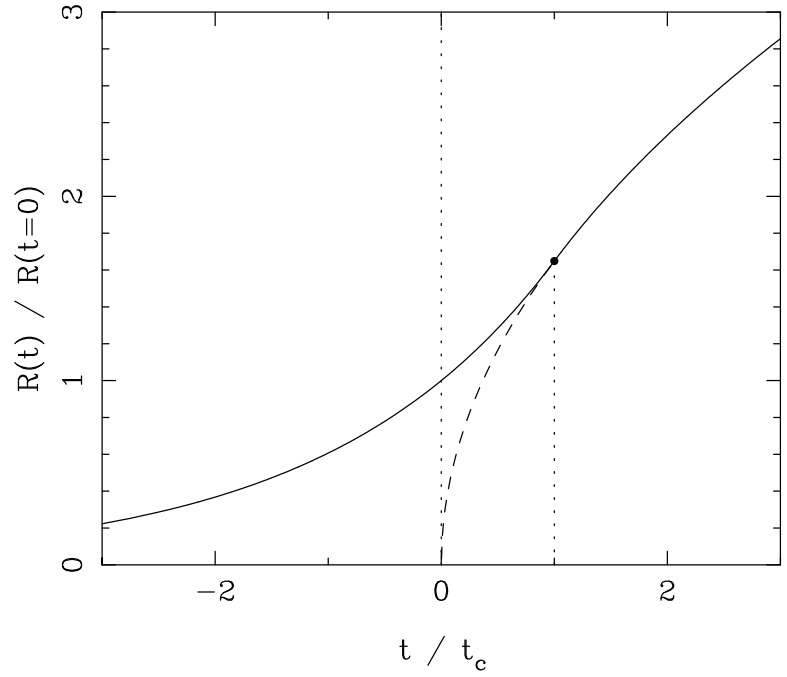

Figure 20. Illustrating the true history of the scale factor in the simplest possible inflationary model. Here, the universe stays in an exponential de Sitter phase for an indefinite time until its equation of state abruptly changes from vacuum dominated to radiation dominated at time t_c. This must occur in such a way as to match R and $\dot{R}$, leading to the solid curve, where the plotted point indicates the join. For $0 < t < t_c$, the dashed curve indicates the time dependence we would infer if vacuum energy was ignored. This reaches $R = 0$ at $t = 0$: the classical 'big bang'. The inflationary solution clearly removes this feature, placing any singularity at large negative time. The universe is much older than we would expect from observations at $t > t_c$, which is one way of seeing how the horizon problem can be evaded.

This idea of a non-singular origin to the universe was first proposed by the Soviet cosmologist E.B. Gliner, in 1969. He suggested no

mechanism by which the vacuum energy could change its level, however. Before trying to plug this critical gap, we can note that an early phase of vacuum-dominated expansion can also solve the flatness problem. Consider the Friedmann equation,

$$\dot{R}^2 = \frac{8\pi G\rho R^2}{3} - kc^2. \tag{223}$$

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

$$R \propto \begin{cases} \sinh Ht & (k=-1) \\ \cosh Ht & (k=+1) \\ \exp Ht & (k=0), \end{cases} \tag{224}$$

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

$$\Omega = 1 + O(f^{-2}). \tag{225}$$

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

$$\ln f \gtrsim 60 \tag{226}$$

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

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

$$z_{\rm P} \simeq 10^{31.6}. \tag{227}$$

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

$$\Delta t_{\rm inflation} > 60\,H^{-1}_{\rm inflation}. \tag{228}$$

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

4.2 The puzzle of dark energy

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

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

$$\dot{R}^2 - \frac{8\pi G}{3}\,\rho\,R^2 = -kc^2. \tag{229}$$

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

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

$$w \equiv p/\rho\, c^2 \tag{230}$$

If this is constant, adiabatic expansion of the vacuum gives

$$\frac{8\pi G\rho}{3H_0^2} = \Omega_v a^{-3(w+1)}. \tag{231}$$

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

$$\frac{8\pi G\rho}{3H_0^2} = \Omega_v \exp\left(\int -3(w(a)+1)\; d\ln a\right). \tag{232}$$

In general, we therefore need

$$H^2(a) = H_0^2\left[\Omega_v e^{\int -3(w(a)+1)\; d\ln a} + \Omega_m a^{-3} + \Omega_r a^{-4} - (\Omega-1)a^{-2}\right]. \tag{233}$$

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

$$w(a) = w_0 + w_a(1-a). \tag{234}$$

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

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

$$t_{\rm rip} - t_0 \simeq \frac{2}{3} H_0^{-1} |1+w|^{-1} (1-\Omega_m)^{-1/2}. \tag{235}$$

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

$$D \equiv R_0 r = \int_0^z \frac{c}{H(z)}\, dz. \tag{236}$$

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

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

$$\ddot{\delta} + 2\frac{\dot{a}}{a}\dot{\delta} = 4\pi G \rho_0 \delta. \tag{237}$$

The vacuum energy manifests itself in the factor of H in the 'Hubble drag' term $2(\dot{a}/a)\dot{\delta}$. For flat models with $w = -1$, we have seen that the growing mode for density perturbations is approximately as $g(a) \propto a\Omega(a)^{0.23}$. If w is made more negative, this makes the growth law closer to the Einstein–de Sitter $g(a) \propto a$ (for very large negative w, the vacuum was unimportant until very recently). Therefore, increasing

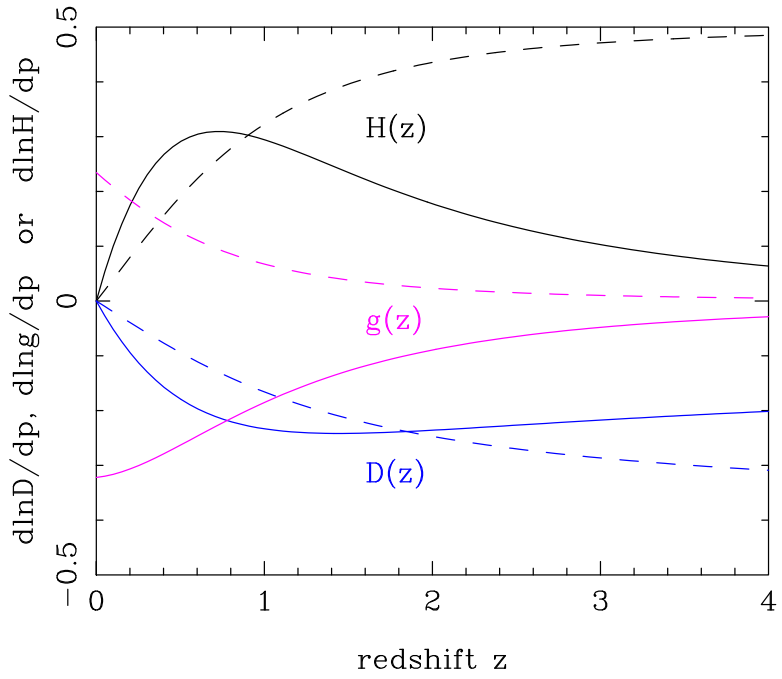

Figure 21. Perturbation around $\Omega_m = 0.25$ of distance-redshift and growth-redshift relations. Solid line shows the effect of increase in w; dashed line the effect of increase in Ω_m

w (making it less negative) has an effect in the same sense as *decreasing* Ω_m. As shown in figure 21, the degeneracy between variations in Ω_m and w thus has the opposite sign to the degeneracy in $D(z)$. Ideally, one would therefore try to observe both effects.

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

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

$$\theta_{\rm H} = D(z_{\rm LS})/D(z=0). \tag{238}$$

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

$$\theta_{\rm H} \propto (\Omega_m h^{3.3})^{0.15}\Omega_m^{\alpha-0.4}; \quad \alpha(w) = -2w/(1-3.8w). \tag{239}$$

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

$$D_0 = 2\frac{c}{H_0}\Omega_m^{-\alpha(w)}. \tag{240}$$

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

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

$$w = -0.926^{+0.054}_{-0.053}, \tag{241}$$

for a spatially flat model – see Spergel et al. (2006). The confidence contours are plotted in detail in figure 22, and it is clear that so far there is very good consistency with a simple cosmological constant. But as we will see, plenty of models exist in which some deviation is predicted. The next goal of the global cosmology community is therefore to push the errors on w down substantially – to about 1%. There is no guarantee that this will yield any signal, but certainly it will cut down the range of viable models for dark energy.

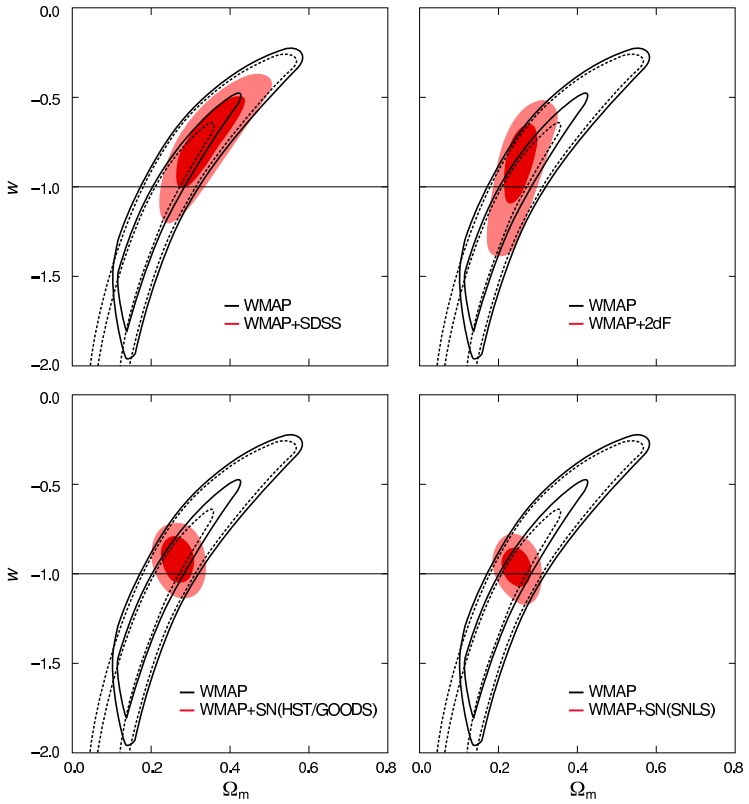

Figure 22. The marginalized WMAP3 confidence contours on the plane of dark-energy equation of state (w) vs Ω_m (from Spergel et al. 2006). A flat universe is assumed, although this is not critical to the conclusions.

4.3 The anthropic approach

Whether or not one finds the 'essence' approach compelling, there remains one big problem. All the models are constructed using Lagrangians with a particular zero level. All quintessence potentials tend to zero for large fields, and k-essence models lack a potential altogether. They are therefore subject to the classical dilemma of the cosmological constant: adding a pure constant to the Lagrangian has no affect on field dynamics, but mimics a cosmological constant. With so many possible contributions to this vacuum energy from the zero-point energies of different fields (if nothing else), it seems contrived to force $V(\phi)$ to asymptote to zero without a reason.

This leads us in the direction of anthropic arguments, which are able to limit Λ to some extent: if the universe had become vacuum-dominated at $z > 1000$, gravitational instability would have been impossible – so that galaxies, stars and observers would not have been possible (Weinberg 1989). Indeed, Weinberg made the astonishingly prescient prediction on this basis that a non-zero vacuum density would be detected at Ω_v of order unity, since there was no reason for it to be much smaller.

MANY UNIVERSES At first sight, this argument seems quite appealing, but it rapidly leads us into deep waters. How can we talk about changing Λ? It has the value that it has. We are implicitly invoking an **ensemble picture** in which there are many universes with differing properties. This is a big step (although exciting, if this turns out to be the only way to explain the vacuum level we see). In fact, the idea of an ensemble emerges inevitably from the framework of inflationary cosmology, since the fluctuations in the scalar field can affect the progress of inflation itself. We have used this idea to look at the changes in when inflation ends – but fluctuations can affect the field at all stages of its evolution. They can be thought of as adding a random-walk element to the classical rolling of the scalar field down the trough defined by $V(\phi)$. In cases where ϕ is too close to the origin for inflation to persist for sufficiently long, it is possible for the quantum fluctuations to push ϕ further out – creating further inflation in a self-sustaining process. This is the concept of **stochastic eternal inflation** due to Linde. Sufficiently far from the origin, the random walk effect of fluctuations becomes more marked and can overwhelm the classical downhill rolling. This means that some regions of space can inflate for an indefinite time, and a single inflating universe automatically breaks up into different bubbles with their own histories. Some random subset of these eventually random-walk close enough to the origin that the classical end of inflation can occur, thus creating a set of 'universes' each of which can potentially host observers.

With this as a starting point, the question now becomes whether we can arrange for the different members of this ensemble to have different values of Λ. This is easily achieved. Let there be some quintessence field with a very flat potential, so that it is capable of

simulating Λ effectively. Quantum fluctuations during inflation can also displace this field, so that each member of the **multiverse** would have a different Λ.

THE DISTRIBUTION OF Λ We are now almost in a position to calculate a probability distribution for Λ. First, we have to set some ground rules: what will vary and what will be held fixed? We should try to change as little as possible, so we assume that all universes have the same values for

(1) The Baryon fraction $f_b = \rho_b/\rho_m$.

(2) The entropy per particle $S = (T/2.73)^3/\Omega_m h^2$

(3) The horizon-scale inhomogeneity $\delta_{\rm H} \simeq 10^{-5}$.

It is far from clear that these minimal assumptions are correct. For example, string theorists have evolved the notion of the **landscape**, in which there is no unique form for low-energy particle physics, but instead a large number of possibilities in which numbers such as the fine-structure constant, neutrino masses etc. are different. From the point of view of understanding Λ, we need there to be at least 10^{100} possible states so that at least some have Λ smaller than the natural $m_{\rm p}^4$ density by a sufficient factor. The landscape hypothesis really took off in 2001, when this number was first shown to be about 10^{500}. But to start with, the simplest approach makes sense: if the simplest forms of anthropic variation can be ruled out, this might be taken as evidence in favour of the landscape picture.

We then take a Bayesian viewpoint to the distribution of Λ given the existence of observers:

$$P(\Lambda \mid \text{Observer}) \propto P_{\rm prior}(\Lambda) P(\text{Observer} \mid \Lambda), \tag{242}$$

where we need both the prior distribution of Λ between different members of the ensemble and how the chance of getting an observer is modified by Λ. The latter factor should be proportional to the number of stars, which is generally take to be proportional to the fraction of the baryons that are incorporated into nonlinear structures. We can estimate this using the Press-Schechter apparatus to get the collapse fraction into systems of a galaxy-scale mass. The exact definition of this is not very important, since the CDM power spectrum is very flat on small scales: any mass at all close to $10^{12}\,M_\odot$ gives similar answers.

The more difficult part is the prior distribution of Λ, and a common argument is to say that it has a uniform distribution – which seems reasonable enough if we are to allow it to have either sign, but know that we will be interested in a very small range near zero. This is the startling proposition of the anthropic model: the vacuum density takes large ranges, and in almost all realizations, the values are comparable in magnitude to the natural scale m_{P}^4; such models are stupendously inimical to life.

We therefore have the simple model

$$dP(\rho_v) \propto f_c \, d\rho_v, \tag{243}$$

where f_c is the collapse fraction into galaxy-scale objects. For large values of Λ, growth ceases at high redshift, and f_c is exponentially suppressed. But things are less clear-cut if $\Lambda < 0$. Here the universe eventually recollapses, and the high density means that the collapse fraction always tends to unity. So why do we not observe $\Lambda < 0$? The answer is that we have to cut off the calculation at late stages of recollapse: once the universe becomes too hot, star-formation may be affected and in any case there is little time for life to form.

With this proviso, figure 23 shows the posterior distribution of Λ conditional on the existence of observers in the multiverse. We express things in natural units: if we adopt the values $\Omega_v = 0.75$ and $h = 0.73$ for the key cosmological parameters, then

$$\rho_v = 7.51 \times 10^{-27}\,\mathrm{kg\,m^{-3}} = \frac{\hbar}{c}\left(\frac{E_v}{\hbar c}\right)^4, \tag{244}$$

where $E_v = 2.39$ meV is known to a tolerance of about 1 %. Provided we consider recollapse only to a maximum temperature of about 10 K, the observed figure is matched well by the anthropic prediction: with this cutoff, most observers will see a positive Λ, and something of order 10% of observers will see Λ as big as we do, or smaller.

So is the anthropic explanation the correct one? Many people find the hypothesis too radical: why postulate an infinity of universes in order to explain a detail of one of them? Certainly, if an alternative explanation for the 'why now' problem existed in the form of e.g. a naturally successful quintessence model, one might tend to prefer that. But so far, there is no such alternative. The longer this situation persists, the more we will be forced to accept that the universe we see can only be understood by making proper allowance for our role as observers.

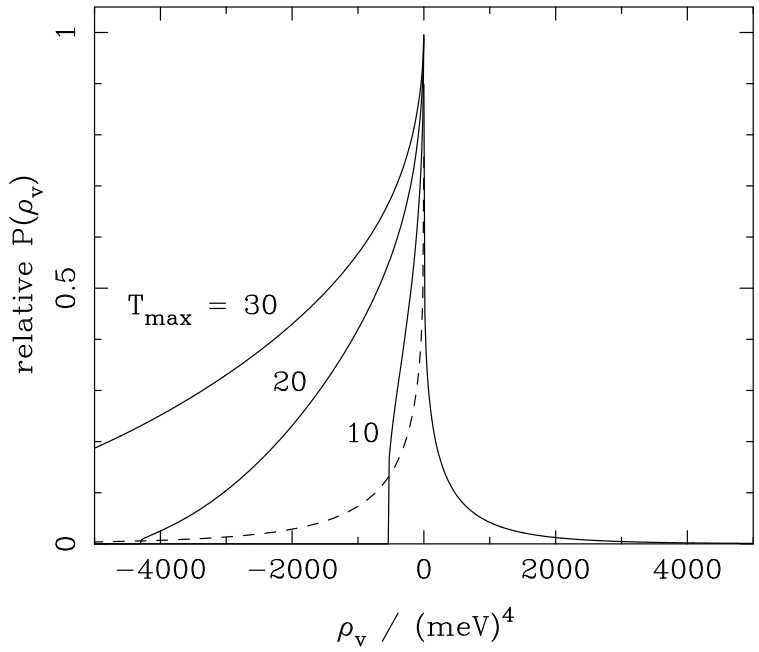

Figure 23. The collapse fraction as a function of the vacuum density, which is assumed to give the relative weighting of different models. The dashed line for negative density corresponds to the expanding phase only, whereas the solid lines for negative density include the recollapse phase, up to maximum temperatures of 10 K, 20 K, 30 K.

Bibliography

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`

Exercises

Here are some problems that may be useful in testing comprehension of the material in these lectures.

Problem (1)

(a) Starting from the Friedmann equation, show that the Hubble parameter as a function of epoch can be written as

$$H^2(a) = H_0^2 \left(\Omega_m a^{-3} + \Omega_r a^{-4} + \Omega_v + \Omega_k a^{-2} \right), \tag{245}$$

defining carefully all terms that appear. Explain why the Ω_k term does not represent a contribution to the matter density. How does the equation chance if the vacuum energy is given a constant equation of state $w \neq -1$? Show that, provided $w < -1/3$, the universe will become vacuum dominated in the far future.

(b) If $w < -1$, show from the Friedmann equation that the scale factor diverges at a finite time in the future. If the current matter density is neglected in comparison with the vacuum density, show that the time to this event is approximately

$$t - t_0 \simeq \frac{2}{3} H_0^{-1} |1 + w|^{-1} \Omega_v^{-1/2}. \tag{246}$$

(c) If the vacuum density is negative, prove that the expansion of the universe will always result in a maximum for the scale factor, followed by collapse to a big crunch, provided $w < -1/3$. For the case of a flat universe containing only matter and vacuum with $w = -1$, show that the Friedmann equation may be written as $\dot{a}^2 = -a^2 + 1/a$ with a suitable choice of time unit. Thus derive the exact expression for $a(t)$ and hence the time of maximum expansion and the time of the big crunch. The substitution $y = a^{3/2}$ should be useful.

(d) Write down the integral for the relation between comoving distance and redshift. Discuss the use of this relation at the time of maximum expansion in the above recollapsing universe, and show that the leading dependence of redshift on distance is quadratic in distance.

Problem (2)

(a) Write down the integrals for the relation between cosmological time and redshift, and for the relation between the particle horizon and redshift. Show that the proper size of the particle horizon during the early radiation-dominated era is $2ct$, and explain how this distance can be larger than ct.

(b) Write down the integral for the relation between comoving distance and redshift, and explain the meaning of the terms 'particle horizon' and 'event horizon'. Give the relation between the current distance to an object at redshift z, the current particle horizon, and the particle horizon at the time when the light we now receive was emitted.

(c) The North and South Hubble Deep Fields are two small patches that lie in opposite directions on the sky, and which contain statistically identical galaxy populations. Show that, according to the above model for $H(a)$, there are critical redshifts beyond which points in the two Hubble Deep Fields have not established causal contact (a) by the present day; (b) by the time at which the light we now see was emitted. Considering the following table of comoving distances for a flat $\Omega_m = 0.25$ model, estimate these redshifts.

z	$D(z)/\,h^{-1}\,\mathrm{Mpc}$
0.5	1345
1	2385
1.5	3178
2	3795
3	4690
5	5775
10	7051
∞	10666

Problem (3)

The equation describing the growth of density fluctuations in a matter-dominated expanding universe is

$$\ddot{\delta} + 2\frac{\dot{a}}{a}\dot{\delta} = \delta\big(4\pi G\bar{\rho}_m - c_s^2 k^2/a^2\big), \tag{247}$$

where δ is the fractional density fluctuation, $\rho = \bar{\rho}_m(1+\delta)$, $\bar{\rho}_m$ is the mean matter density, c_s is the speed of sound, and k is comoving wavenumber.

(a) Show that, for a static model, density fluctuations grow exponentially as long as the wavelength is sufficiently large, and explain physically why this is so.

(b) Derive the solutions to the perturbation equation for a universe of critical density, in the limit of infinitely long wavelength.

(c) At time $t = t_c$, a homogeneous critical-density universe is given a velocity perturbation, such that $\dot{\delta} = A$. Evaluate the density perturbation as a function of time following this event.

(d) If the universe contains a homogeneous component in addition to matter that can clump, the perturbation equation still applies – but $\bar{\rho}_m$ does not include the homogeneous component. Consider a flat universe that contains a mixture of hot and cold dark matter: for sufficiently small wavelengths, the hot component can be assumed to be uniform, with density parameter Ω_h. Show that fluctuations in the cold component grow as $\delta \propto t^\alpha$, where $\alpha = (\sqrt{25 - 24\Omega_h} - 1)/6$.