Astrophysical Cosmology 4 2001–2002

J.A. Peacock

Room R5, Royal Observatory; jap@roe.ac.uk

http://www.roe.ac.uk/japwww/teaching/cos4.html

Synopsis

This course introduces the fundamental concepts of modern astrophysical cosmology. The meaning of time and space in an expanding universe is discussed, and the dynamics of the expansion are solved, yielding the tools needed to relate astronomical observations to the physical properties of objects seen at great distances. The time history of the expansion is studied, starting from the prediction of a hot big bang, and discussing the relics that remain from early times, especially light elements, dark matter and the microwave background. The initial conditions for the expansion are seen to require careful tuning, and the best modern explanation for this lies in the theory of inflation, which removes the idea of a big bang. Inflation can explain not only the existence of a uniform expanding universe, but can seed fluctuations via amplified quantum fluctuations, so that structures such as galaxies can form at late times. The basic elements of this theory are explained, and the course closes with a survey of open observational challenges.

Recommended books

Roos: Introduction to Cosmology (Wiley) The best short modern introductory text. Strong on the early universe, but a little brief on the observational side.

Bothun: Modern Cosmological Observations and Problems (Taylor & Francis) A good supplement to Roos for observational aspects of cosmology, and less mathematically demanding.

More advanced reference books for extra detail:

Peacock: Cosmological Physics (CUP)
Peebles: Principles of Physical Cosmology (Princeton)
Weinberg: Gravitation & Cosmology (Wiley)

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

Syllabus

(1) History and basic concepts of an expanding universe.

(2) Cosmological spacetime: the Robertson-Walker metric.

(3) Light propagation and redshift.

(4) Dynamics I: the Friedmann equation.

(5) Dynamics II: the expansion history.

(6) The hot big bang I: thermal history and relics.

(7) The hot big bang II: the microwave background.

(8) The hot big bang III: primordial nucleosynthesis.

(9) Observational cosmology: apparent ages, sizes and fluxes.

(10) Cosmological distance ladder and age scale.

(11) Observations of dark matter.

(12) Theories for dark matter.

(13) Large-scale structure.

(14) Structure formation I: gravitational collapse.

(15) Structure formation II: dark matter and clustering.

(16) Early universe I: initial conditions.

(17) Early universe II: inflation.

(18) Observing the evolution of the universe.

1 The 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 present 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 in fact the honour 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 $H \simeq 500\,\mathrm{km\,s^{-1}Mpc^{-1}}$, 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

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

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 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:

$$\nabla^2\Phi + \lambda\Phi = 4\pi G\rho, \tag{6}$$

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 $\rho_{\rm rep} = \lambda\Phi/4\pi G$:

$$\nabla^2\Phi = 4\pi G(\rho - \rho_{\rm rep}), \tag{7}$$

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.

When looked at in this way, 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.

2 Cosmological coordinates

These Newtonian arguments are useful for orientation, and contain part of the truth, because Newtonian physics does hold locally. However, a correct description of the totality of an expanding universe must be a relativistic one, since general relativity shows that space will be curved in general. Fortunately, by considering symmetry arguments, most of the complexities can be avoided.

2.1 Fundamental observers

Although spacetime in an expanding universe is indeed curved on a large scale, we need not worry about this locally. Newtonian physics works perfectly well over the distance to M31, at roughly 1 Mpc. This is not true when we look at a galaxy 1000 Mpc away, but conditions near to that galaxy will still seem Newtonian to an observer located there. This is just a consequence of the **equivalence principle**, which says that freely-falling observers in gravitational fields of any strength experience special relativity locally.

We therefore imagine the expanding universe filled with observers in different locations, all of whom are at rest with respect to the matter in their vicinity (these characters are usually termed **fundamental observers**). We can envisage them as each sitting on a different galaxy, and so receding from each other with the general expansion. 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. We can define 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.

2.2 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.

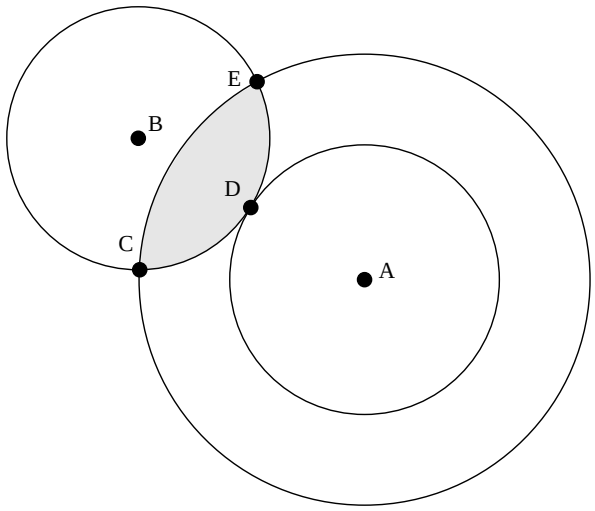

Figure 1. Isotropy about two points A and B shows that the universe is homogeneous. From isotropy about B, the density is the same at each of C,D,E. By constructing spheres of different radii about A, the shaded zone is swept out and shown to be homogeneous. By using large enough shells, this argument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

2.3 The metric

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.

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

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

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.

3 Spaces of constant curvature

3.1 Metrics on spheres

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

$$d\sigma^2 = R^2 \left(dr^2 + \sin^2 r \, d\phi^2 \right). \tag{9}$$

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

$$d\sigma^2 = R^2 \left(dr^2 + \sinh^2 r \, d\phi^2 \right). \tag{10}$$

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

$$d\sigma^2 = R^2 \left(\frac{dr^2}{1 - kr^2} + r^2 \, d\phi^2 \right), \tag{11}$$

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:

$$x^2 + y^2 + z^2 = R^2. \tag{12}$$

We can always satisfy this by defining some angles:

$$\begin{aligned} z &= R \cos\theta \\ y &= R \sin\theta \sin\phi \\ x &= R \sin\theta \cos\phi. \end{aligned} \tag{13}$$

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

$$d\sigma^2 = dx^2 + dy^2 + dz^2, \tag{14}$$

and we can express this in terms of the angles using

$$\begin{aligned} dx &= R(\cos\theta \cos\phi \, d\theta - \sin\theta \sin\phi \, d\phi) \\ dy &= R(\cos\theta \sin\phi \, d\theta + \sin\theta \cos\phi \, d\phi) \\ dz &= R(-\sin\theta \, d\theta). \end{aligned} \tag{15}$$

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. To evaluate this, start with the vectors generated by increments in $d\alpha$ *etc.*:

$$\begin{aligned}
\mathbf{e}_\alpha &= d\alpha\, R\,(\cos\alpha\sin\beta\sin\gamma,\ \cos\alpha\sin\beta\cos\gamma,\ \cos\alpha\cos\beta,\ -\sin\alpha\,) \\
\mathbf{e}_\beta &= d\beta\, R\,(\sin\alpha\cos\beta\sin\gamma,\ \sin\alpha\cos\beta\cos\gamma,\ -\sin\alpha\sin\beta,\ 0\,) \\
\mathbf{e}_\gamma &= d\gamma\, R\,(\sin\alpha\sin\beta\cos\gamma,\ -\sin\alpha\sin\beta\sin\gamma,\ 0,\ 0\,).
\end{aligned} \tag{16}$$

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

$$|(dx, dy, dz, dw)|^2 = R^2\left[d\alpha^2 + \sin^2\alpha\,(d\beta^2 + \sin^2\beta\; d\gamma^2)\right]. \tag{17}$$

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.

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.

3.2 The RW metric

We can now get the overall metric, since 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

$$S_k(r) = \begin{cases} \sin r & (k = 1) \\ \sinh r & (k = -1) \\ r & (k = 0). \end{cases} \tag{18}$$

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

$$c^2 d\tau^2 = c^2 dt^2 - R^2(t)\left[dr^2 + S_k^2(r)\, d\psi^2\right]. \tag{19}$$

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

$$c^2 d\tau^2 = c^2 dt^2 - R^2(t)\left(\frac{dr^2}{1 - kr^2} + r^2 d\psi^2\right). \tag{20}$$

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

$$a(t) \equiv \frac{R(t)}{R_0}, \tag{21}$$

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.

4 Light propagation

Now we have the RW metric, we can study the propagation of light in cosmology. This satisfies trajectories with zero proper time (**null geodesics**). The radial equation of motion is therefore

$$dr = c\,dt/R(t). \tag{22}$$

Many of the key equations in cosmology are obtained by starting with this simple equation.

4.1 Horizons

The first thing we can do with this is to see how far a photon can have travelled since the big bang. Integrating dr from 0 to t, we see that the behaviour of the scale factor at early times is rather important. Suppose it is a power law in time: $R \propto t^\alpha$. If $\alpha < 1$, then the integral converges to a finite value. We would then say that the universe possesses a **particle horizon**, meaning that light signals have only been able to propagate over a finite distance by the present. This may seem like common sense: surely at a time t, only points within a distance $< ct$ can have exchanged signals? This is not so, because the universe was small in the past; as we have seen, if it expanded slowly enough near to $t = 0$ light signals might have connected any two points.

However, when we study cosmological dynamics, we will see that the expected behaviour at early times is $R \propto t^{1/2}$, so there should be a horizon. This idea will be important later.

4.2 The redshift

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

$$H = \frac{\dot{R}}{R}. \tag{23}$$

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

$$\frac{\nu_{\rm emit}}{\nu_{\rm obs}} \equiv 1 + z \simeq 1 + \frac{v}{c}. \tag{24}$$

This defines the **redshift** z in terms of the shift of spectral lines. What is the equivalent of this relation at larger distances? We saw from the metric that the equation for a null geodesic is $r = \int c\,dt\,/\,R(t)$. The comoving distance is constant, whereas the domain of integration in time extends from $t_{\rm emit}$ to $t_{\rm obs}$; these are the times of emission and reception of a photon. Photons that are emitted at later times will be received at later times, but these changes in $t_{\rm emit}$ and $t_{\rm obs}$ cannot alter the integral, since r is a comoving quantity. This requires the condition $dt_{\rm emit}/dt_{\rm obs} = R(t_{\rm emit})/R(t_{\rm obs})$, which means that events on distant galaxies time-dilate according to how much the universe has expanded since the photons we see now were emitted. Clearly (think of events separated by one period), this dilation also applies to frequency, and we therefore get

$$\frac{\nu_{\rm emit}}{\nu_{\rm obs}} \equiv 1 + z = \frac{R(t_{\rm obs})}{R(t_{\rm emit})}. \tag{25}$$

In terms of the normalized scale factor $a(t)$ we have simply $a(t) = (1+z)^{-1}$.

Photon wavelengths therefore stretch with the universe, as may seem intuitively reasonable. We can prove this more directly, as follows. Suppose we send of 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:

$$\delta\nu \,/\, \nu = \delta v/c = -(Hd)/c = -H\delta t. \tag{26}$$

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

$$\delta\nu \,/\, \nu = -\delta R \,/\, R, \tag{27}$$

which proves the result. The redshift is 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

$$1 + z = \sqrt{\frac{1+v/c}{1-v/c}}. \tag{28}$$

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.

Finally, note that the law that frequency of photons scales as $1/R$ actually applies to the momentum of all particles – relativistic or not. 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).

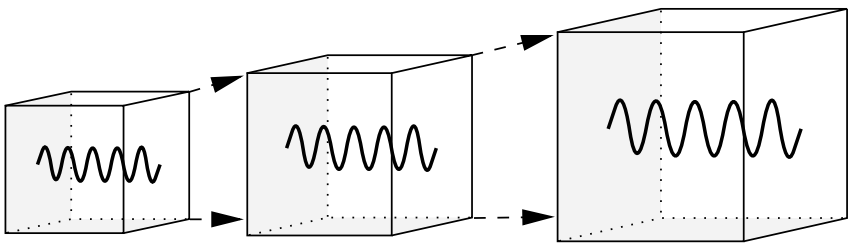

Figure 2. Suppose we trap some radiation inside a box with silvered sides, that expands with the universe. At least for an isotropic radiation field, the photons trapped in the box are statistically equivalent to any that would pass into this space from outside. Since the walls expand at $v \ll c$ for 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 $a(t)^{-1}$.

5 Dynamics of the expansion

5.1 The Friedmann equation

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

$$\dot{R}^2 - \frac{8\pi G}{3}\rho R^2 = -kc^2. \tag{29}$$

Note that this equation covers all contributions to ρ; it is independent of the equation of state.

This connection between the geometry of the universe and its density is one of the most profound results in cosmology. Since this course will not use general relativity, we can't prove it properly, which is a pity. Nevertheless, it is possible to give some partial justification, as follows. First note that any open model will evolve towards undecelerated expansion provided its equation of state is such that ρR^2 is a declining function of R – the potential energy becomes negligible by comparison with the total and $\dot{R}$ tends to c, so that $R = ct$. In this zero-density limit, there can be no spatial curvature and the open RW metric must be just a coordinate transformation of Minkowski spacetime. Later on, we will show how to make this transformation, proving that the Friedmann equation is correct for $k = -1$.

To prove the $k = 0$ case, rewrite the Friedmann equation in terms of the Hubble parameter:

$$H^2 - \frac{8\pi G}{3}\rho = \frac{\text{constant}}{R^2}. \tag{30}$$

Now consider holding the local observables H and ρ fixed but increasing R without limit. Clearly, in the RW metric this corresponds to going to the $k = 0$ form: the scale of spatial curvature goes to infinity and the comoving separation for any given proper separation goes to zero, so that the comoving geometry becomes indistinguishable from the Euclidean form. This case also has potential and kinetic energy much greater than total energy, so that the rhs of the Friedmann equation is effectively zero. This establishes the $k = 0$ case, leaving the closed universe as the only stubborn holdout against Newtonian arguments.

Accepting the Friedmann equation, there is thus always a **critical density** that will yield $k = 0$, making the comoving part of the metric look Euclidean:

$$\rho_c = \frac{3H^2}{8\pi G}. \tag{31}$$

A universe with density above this critical value will be **spatially closed**, whereas a lower-density universe will be **spatially open**. Note that the 'flat' universe with $k = 0$ is still curved spacetime. It is common to define a dimensionless **density parameter** as the ratio of density to critical density:

$$\Omega \equiv \frac{\rho}{\rho_c} = \frac{8\pi G\rho}{3H^2}. \tag{32}$$

In terms of this notation, the Friedmann equation is

$$\frac{kc^2}{H^2R^2} = \Omega - 1. \tag{33}$$

In this equation, R, H and Ω change with time; the current values of these parameters should be distinguished by a zero subscript. We can then use the Friedmann equation in the above form to deduce the present value of the scale factor:

$$R_0 = \frac{c}{H_0}\left[(\Omega_0 - 1)/k\right]^{-1/2}. \tag{34}$$

Another name for this is the **curvature length**; it becomes infinitely large as Ω_0 approaches unity from either direction. Models with Ω_0 very close to unity are thus practically indistinguishable from the $k = 0$ model in which the comoving part of the metric is exactly uncurved.

In practice, Ω_0 is such a common symbol in cosmological formulae, that it is normal to leave off the zero subscript. Henceforth, Ω means Ω_0; the density parameter at other epochs will be denoted by $\Omega(z)$. As a natural partner to Ω, we can also define a dimensionless (current) Hubble parameter as

$$h \equiv \frac{H_0}{100\,\mathrm{km\,s^{-1}Mpc^{-1}}}, \tag{35}$$

in terms of which the current density of the universe is

$$\begin{aligned}\rho_0 &= 1.88 \times 10^{-26}\,\Omega h^2\ \mathrm{kg\,m^{-3}} \\ &= 2.78 \times 10^{11}\,\Omega h^2\ M_\odot\,\mathrm{Mpc}^{-3}\end{aligned} \tag{36}$$

and the current curvature length is

$$R_0 = 3000\,|\Omega - 1|^{-1/2}\,h^{-1}\,\mathrm{Mpc}. \tag{37}$$

5.2 Solution with matter only

To see how the Friedmann equation works, it is convenient to start by thinking about universes in which only pressureless matter ('dust') exists. For this, conservation of particles requires that the density behaves as

$$\rho/\rho_0 = (R/R_0)^{-3} \tag{38}$$

This makes it easy to solve the equation if $k = 0$: we have

$$\dot{R}^2 = \frac{8\pi G}{3}\rho R^2 = \frac{8\pi G}{3}\rho_0 R_0^3\ R^{-1}, \tag{39}$$

so that $\dot{R} \propto R^{-1/2}$, which integrates to

$$R \propto t^{2/3}. \tag{40}$$

This $\Omega = 1$ matter-only universe is called the **Einstein–de Sitter model**. Notice that there is a finite time in the past at which $R \to 0$; the density diverges, and our assumption of cold pressureless material is probably wrong. Nevertheless, the basic conclusion still holds:

Friedmann's equation predicts a singular start to the expanding universe – the **big bang**. It is easy to work out H for this model, and deduce the time since the big bang: this is $2/3H_0$. If we lived in such a universe, and we use $H_0 = 70\,\mathrm{km\,s^{-1}Mpc^{-1}}$ as the best modern value, then the age of the universe would be 9.3 Gyr. As we will see later on, this is uncomfortably young.

If $k \neq 0$, the solution to the equation is a bit trickier. There is no equation for $R(t)$, but we can get a parametric solution where both R and t depend on some angle η:

$$\begin{aligned} R &= kR_*[1 - C_k(\eta)] \\ ct &= kR_*[\eta - S_k(\eta)]. \end{aligned} \tag{41}$$

Here, C_k means cos if $k = +1$ and cosh if $k = -1$. To check that this works, use

$$\dot{R} = \frac{dR/d\eta}{dt/d\eta} = \frac{ck\,S_k(\eta)}{1 - C_k(\eta)}. \tag{42}$$

If we use the trig-like identity $S_k^2 = k(1 - C_k^2) = k(1 - C_k)(1 + C_k)$, this becomes

$$\dot{R}^2 = -kc^2 + 2c^2R_*/R, \tag{43}$$

which is the Friedmann equation, with $R_* = (c/H_0)(\Omega[k/(\Omega - 1)]^{3/2}/2)$

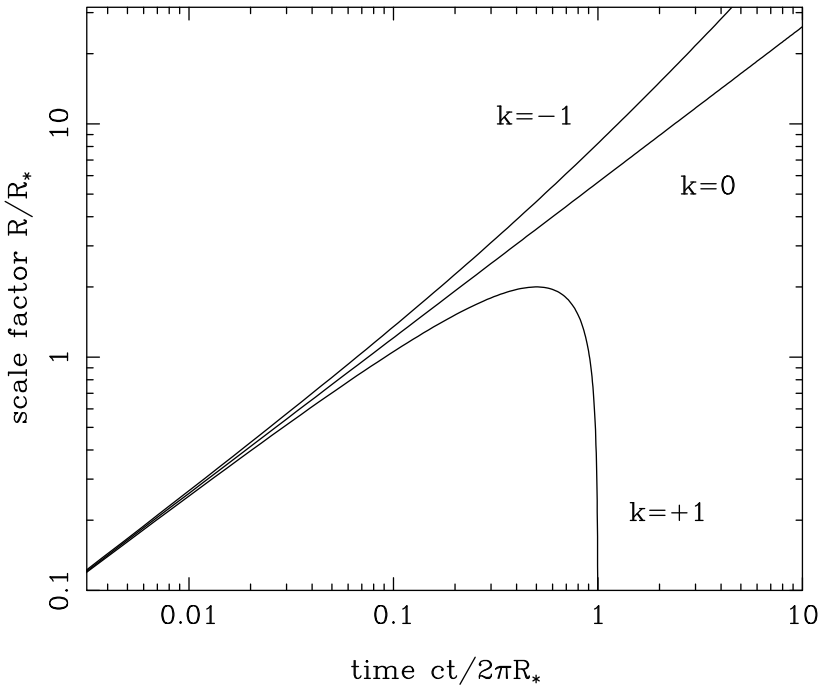

Figure 3. The time dependence of the scale factor for open, closed and critical matter-dominated cosmological models. The upper line corresponds to $k = -1$, the middle line to the flat $k = 0$ model, and the lowest line to the recollapsing closed $k = +1$ universe. The log scale is designed to bring out the early-time behaviour, although it obscures the fact that the closed model is a symmetric cycloid on a linear plot of R against t.

These solutions are plotted in figure 4. The $k = +1$ model has an interesting behaviour: after some time, it ceases to expand and falls back into a **big crunch**. From a Newtonian point of view, this is quite reasonable: the total 'energy' represented by $-kc^2$ is negative, so the universe is bound. The universe does not expand fast enough to have 'escape velocity' and must fall back on itself. The conclusion here is that there is a relation between the density of the universe, its geometry, and its eventual fate. If $\Omega > 1$, the universe must be closed and will recollapse; if $\Omega < 1$, the universe is open and infinite and will expand forever. Unfortunately, this simple story gets spoiled if the matter content is more complicated. The best current guess is that the universe is indeed very close to the critical density required for flatness, but that it will expand forever because some of that density is the peculiar vacuum energy introduced by Einstein.

6 Solutions to the Friedmann equation

6.1 Equation of state

In order to solve the Friedmann equation and learn the history of the scale factor, we need to know how the density changes as R changes. This can be achieved if we divide the contents of the universe into pressureless matter ($\rho \propto R^{-3}$), radiation ($\rho \propto R^{-4}$) and vacuum energy (ρ constant). The first two relations just say that the number density of particles is diluted by the expansion, with photons also having their energy reduced by the redshift; the third relation says that vacuum energy is just a constant property of empty space. In terms of observables, this means that the density is written as

$$\frac{8\pi G\rho}{3} = H_0^2(\Omega_v + \Omega_m a^{-3} + \Omega_r a^{-4}) \tag{44}$$

(using the normalized scale factor $a = R/R_0$).

It is important to be aware that this expression makes a very important assumption about the early universe: it was **radiation dominated**. In other words, even if radiation makes a relatively small contribution to the overall mass budget of the universe today, it would have been relatively more important in the past. There must have been a time at which the densities in matter and radiation were equal, with the radiation dominating at early times and the matter at late times (and the vacuum at very late times).

Friedmann's equation treats all contributions to the density parameter equally:

$$\frac{kc^2}{H^2R^2} = \Omega_m(a) + \Omega_r(a) + \Omega_v(a) - 1 \equiv \Omega(a) - 1, \tag{45}$$

so that a flat $k = 0$ universe requires $\sum \Omega_i = 1$ at all times, whatever the form of the contributions to the density.

6.2 Effects of pressure

Before going on, are we confident that the Friedmann equation will still apply for matter with a significant pressure? Suppose we differentiate it with respect to time to get an acceleration equation involving $\ddot{R}$. This requires a knowledge of $\dot{\rho}$, but this can be eliminated by means of conservation of energy: $d[\rho c^2 R^3] = -pd[R^3]$. We then obtain

$$\ddot{R} = -4\pi GR(\rho + 3p/c^2)/3. \tag{46}$$

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

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

$$\nabla^2 \Phi = 4\pi G(\rho + 3p/c^2). \tag{47}$$

It isn't easy to give a non-relativistic justification for why the pressure acts as an extra form of gravity. The trouble is that this extra term is unimportant in our everyday experience. However, it does matter in cosmology. 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. The main thing is to appreciate that some effect of the pressure has to appear, though conservation of energy. The simple Friedmann equations $\ddot{R} = -4\pi G R\rho/3$ and $\dot{R}^2 - 8\pi G\rho R^2/3 = \text{const}$ are inconsistent without a term in p: at least one of them has to change.

6.3 Energy density of the vacuum

A case where the gravitational effects of pressure are especially important is when considering the possibility of vacuum energy. As we saw earier, the possibility that empty space may have a non-zero density was first introduced by Einstein in an attempt to achieve a static universe (although he didn't phrase the argument in this way). 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:

$$p_{\rm vac} = -\rho_{\rm vac}\, c^2. \tag{48}$$

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 4). 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.

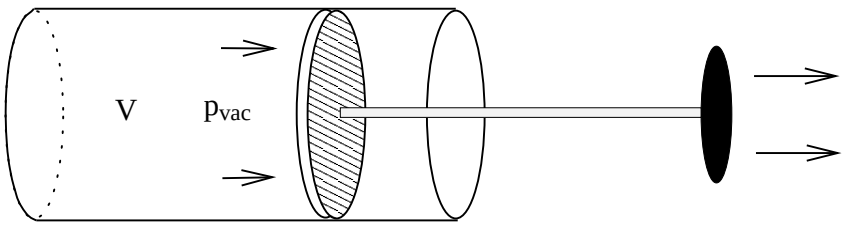

Figure 4. A thought experiment to illustrate the application of conservation of energy to the vacuum. If the vacuum density is $\rho_{\rm vac}$ then the energy created by withdrawing the piston by a volume dV is $\rho_{\rm vac} c^2\, dV$. This must be supplied by work done by the vacuum pressure $p_{\rm vac} dV$, and so $p_{\rm vac} = -\rho_{\rm vac} c^2$, as required.

6.4 No-vacuum universe

We already solved Friedmann's equation for no vacuum energy and no radiation. If we include radiation, we may as well stick to the simplest case, which is the $k = 0$ flat universe. This will always be a good approximation to the early phases of the universe, as can be seen by going back to the basic form of Friedmann's equation: $\dot{R}^2 = 8\pi G\rho R^2/3 - kc^2$. For matter and radiation, $\rho R^2 \propto R^{-1}$ and R^{-2} respectively. At small R, the density term therefore completely overwhelms the curvature term on the rhs; it is as good as zero. As long as this lasts, it is easy to solve the equation. A power-law in time clearly works, and we get

$$\begin{aligned} R \propto t^{2/3} \quad &\Rightarrow \quad t = \sqrt{\frac{1}{6\pi G\rho}} \qquad \text{(matter domination)} \\ R \propto t^{1/2} \quad &\Rightarrow \quad t = \sqrt{\frac{3}{32\pi G\rho}} \qquad \text{(radiation domination),} \end{aligned} \tag{49}$$

so that the age of the universe is always of order $1/\sqrt{G\rho}$.

Viewed in this way, it would be rather surprising if the universe was not flat today. The ρR^2 term in Friedmann's equation scales as $R^{-2} \propto t^{-1}$ in the radiation era. As we will see later, the earliest time in the big bang that we can sensibly discuss is about 10^{-43} s, at which time the curvature term must have been smaller that the density term by a factor of about 10^{60}. This means that the density at this time had to differ from the critical value by a **fine-tuned** factor of $1 + \mathrm{O}10^{-60}$. How could the universe have known to fix its density so precisely?

6.5 Vacuum-dominated universe

What happens if we go to the opposite extreme and have a universe where vacuum energy dominates? Consider again the Friedmann equation in its general form $\dot{R}^2 - 8\pi G\rho R^2/3 = -kc^2$. This is easy to solve for $k = 0$, since ρ is constant:

$$R \propto \exp Ht; \quad H = \sqrt{\frac{8\pi G\rho_v}{3}}. \tag{50}$$

The vacuum repulsion causes the expansion of the universe to increase without limit. If the curvature is non-zero, it is easy enough to see that the solutions still approach the exponential at large times: $R \propto \sinh Ht$ $(k = -1)$, or $R \propto \cosh Ht$ $(k = +1)$.

An interesting interpretation of this behaviour was promoted in the early days of cosmology by Eddington: the cosmological constant is what *caused* the expansion. In models without vacuum energy, the expansion is merely an initial condition: anyone who asks why the universe expands at a given epoch is given the unsatisfactory reply that it does so because it was expanding at some earlier time, and this chain of reasoning comes up against a barrier at $t = 0$. It would be more satisfying to have some mechanism that set the expansion into motion, and this is what is provided by vacuum repulsion. This tendency of vacuum-dominated models to end up undergoing an exponential phase of expansion exactly what is used in inflationary cosmology to generate the initial conditions for the big bang.

6.6 The general case

The solution of the Friedmann equation with matter, radiation, vacuum energy and curvature is not pretty, so we will be content here with just showing the results, which are summed up in figure 5. Since the radiation density is very small today, the main task of relativistic cosmology is to work out where on the $\Omega_{\rm matter} - \Omega_{\rm vacuum}$ plane the real universe lies.

The main things to note are that positive vacuum energy almost (but not quite) guarantees that the universe will expand forever. A universe with negative vacuum density will always recollapse. If the vacuum energy is positive and too large, it will prevent the existence of a big bang altogether. However, the existence of high-redshift objects rules out such 'bounce' models, so that the idea of a hot big bang cannot be evaded.

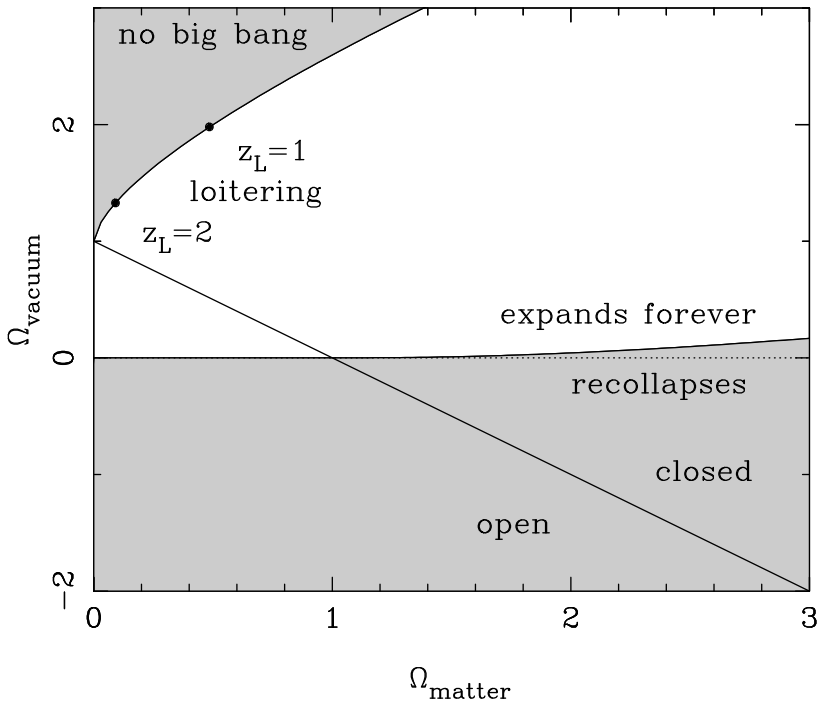

Figure 5. This plot shows the different possibilities for the cosmological expansion as a function of matter density and vacuum energy. Models with total $\Omega > 1$ are always spatially closed (open for $\Omega < 1$), although closed models can still expand to infinity if $\Omega_v \neq 0$. If the 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.

6.7 The age and size of the universe

Although the general solution of the Friedmann equation for $R(t)$ is difficult, it is easier to work out some of the quantities of particular interest from the point of view of observation. One of the most important of these is the age of the universe at a given redshift. Since $1+z = R_0/R(z)$, we have

$$\frac{dz}{dt} = -\frac{R_0}{R^2}\frac{dR}{dt} = -(1+z)H(z), \tag{51}$$

so we need to see how the expansion rate evolves – *i.e.* work out $H(z)$. 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

$$H^2(a) = H_0^2\left[\Omega_v + \Omega_m a^{-3} + \Omega_r a^{-4} - (\Omega - 1)a^{-2}\right]. \tag{52}$$

This shows that the universe expanded faster in the past (small a). The differential time–redshift relation is therefore

$$\begin{aligned} dt &= \frac{dz}{(1+z)\,H(z)} \\ &= \frac{1}{(1+z)\,H_0}\left[(1-\Omega)(1+z)^2 + \Omega_v + \Omega_m(1+z)^3 + \Omega_r(1+z)^4\right]^{-1/2} dz. \end{aligned} \tag{53}$$

To get the current age of the universe, we need to integrate this from 0 to ∞. This can be done analytically if we ignore radiation (valid at all times after about 10^5 years) and stick to flat models (so that $\Omega_v = 1 - \Omega_m$):

$$H_0 t_0 = \frac{2}{3}\,\frac{S_k^{-1}\left(\sqrt{|\Omega_m - 1|}/\Omega_m\right)}{\sqrt{|\Omega_m - 1|}}. \tag{54}$$

Here, k in S_k is used to mean sin if $\Omega_m > 1$, otherwise sinh; these are still $k = 0$ models. This is still not so easy to remember, but the following simple approximate formula is accurate to a few %, and also applies for non-flat universes for values of Ω_m and Ω_v of practical interest:

$$H_0 t_0 \simeq \frac{2}{3}\,(0.7\Omega_m - 0.3\Omega_v + 0.3)^{-0.3}. \tag{55}$$

For example, this gives $H_0 t_0 = 0.96$ for the empty universe, as against the exact answer, which we know to be 1. For a flat universe, the age is $H_0 t_0 \simeq (2/3)\Omega_m^{-0.3}$, so that a flat universe with $\Omega_m = 0.26$ has the same age as an empty model with the same H_0.

The time–redshift relation is easily converted to something else of key observational importance: the relation between redshift and comoving distance. The equation of motion for a photon is $R\,dr = c\,dt$, so $R_0 dr/dz = (1+z)c\,dt/dz$, or

$$\begin{aligned} R_0\,dr &= \frac{c}{H(z)}\,dz \\ &= \frac{c}{H_0}\left[(1-\Omega)(1+z)^2 + \Omega_v + \Omega_m(1+z)^3 + \Omega_r(1+z)^4\right]^{-1/2} dz. \end{aligned} \tag{56}$$

This equation can't be integrated in all cases, but it's simple to integrate numerically to get practical results. The simplest cases correspond to matter only, in which case

$$\begin{aligned} R_0 S_k(r) = R_0 r &= (2c/H_0)\left[\,1 - 1/(1+z)^{1/2}\,\right] \quad (\Omega = 1) \\ R_0 S_k(r) &= (c/2H_0)\left[\,1 + z - 1/(1+z)\,\right] \quad (\Omega = 0) \end{aligned} \tag{57}$$

Notice how both these expressions tend to a constant of order c/H_0 as $z \to \infty$. This is another example of the horizon: photons emitted just after the big bang and reaching us only now have travelled only a finite comoving distance.

7 The meaning of an expanding universe

Finally, having dealt with some of the formal apparatus of cosmology, it may be interesting to step back and ask what all this means. The idea of an expanding universe can easily lead to misconceptions, the worst of which is the 'expanding space' fallacy. 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. This is nonsense, and we can best prove this by considering the case of the empty universe.

The metric of uncurved Minkowski spacetime is

$$c^2 d\tau^2 = c^2 dt^2 - \left(dr^2 + r^2 d\psi^2\right), \tag{58}$$

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:

$$t' = t/\gamma = t\sqrt{1 - r^2/c^2t^2}. \tag{59}$$

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

$$c^2 d\tau^2 = c^2 dt'^2 - d\ell^2 - r^2 d\psi^2. \tag{60}$$

To complete the transition from Minkowski to fundamental-observer coordinates, we need to eliminate r. To do this, define the velocity variable ω:

$$v/c = \tanh\omega \quad \Rightarrow \quad \gamma = \cosh\omega. \tag{61}$$

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

$$r = c\sqrt{t^2 - t'^2} = ct' \sinh\omega, \tag{62}$$

and the radial increment of proper length is related to dr via length contraction (since $d\ell$ is at constant t'):

$$d\ell = dr/\gamma = ct'\, d\omega. \tag{63}$$

The metric therefore becomes

$$d\tau^2 = dt'^2 - t'^2\left(d\omega^2 + \sinh^2\omega\, d\psi^2\right). \tag{64}$$

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

8 The thermal history of the big bang

What was the state of matter in the early phases of the big bang? Since the present-day expansion will cause the density to decline in the future, conditions in the past must have corresponded to high density – and thus to high temperature. We can deal with this quantitatively by looking at the thermodynamics of the fluids that make up a uniform cosmological model. We have already used a simple model for the energy content of the universe in which we distinguish pressureless 'dust-like' matter (in the sense that $p \ll \rho c^2$) from relativistic 'radiation-like' matter (photons plus neutrinos). If these are assumed not to interact, then the energy densities scale as

$$\begin{aligned} \rho_m &\propto R^{-3} \\ \rho_r &\propto R^{-4} \end{aligned} \tag{65}$$

These are special cases of **adiabatic expansion**, where the entropy of a given comoving region does not change with time. Adiabatic changes satisfy $pV^\Gamma = \text{const}$, where Γ is the ratio of specific heats ($\Gamma = 5/3$ for a monatomic gas, for example). Since $p \propto \rho$ for radiation, this shows that radiation can be treated as a fluid with $\Gamma = 4/3$. Adiabatic change implies reversibility, but we know that irreversible changes happen in all astronomical systems. The idea of a reversible expansion is therefore only an approximation – but a very good one, as we will see.

The most important prediction of the adiabatic assumption is that the universe must have been **radiation dominated** at some time in the past, where the densities of matter and radiation cross over. To anticipate, we know that the current radiation density corresponds to thermal radiation with $T \simeq 2.73$K. We shall shortly show that one expects to find, in addition to this **cosmic microwave background** (CMB), a background in neutrinos that has an energy density 0.68 times that from the photons (if the neutrinos are massless and therefore relativistic). If there are no other contributions to the energy density from relativistic particles, then the total effective radiation density is $\Omega_r h^2 \simeq 4.2 \times 10^{-5}$ and the redshift of **matter–radiation equality** is

$$1 + z_{\rm eq} = 23\,900\,\Omega h^2\,(T/2.73{\rm K})^{-4}. \tag{66}$$

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.

8.1 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**:

$$\begin{aligned} m_{\rm P} &\equiv \sqrt{\frac{\hbar c}{G}} \simeq 10^{19}{\rm GeV} \\ \ell_{\rm P} &\equiv \sqrt{\frac{\hbar G}{c^3}} \simeq 10^{-35}{\rm m} \\ t_{\rm P} &\equiv \sqrt{\frac{\hbar G}{c^5}} \simeq 10^{-43}{\rm s}. \end{aligned} \tag{67}$$

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.

8.2 Thermal backgrounds

The study of matter under the extremes of pressure and temperature expected in the early phases of the expanding universe is easier that might be expected. Although the universe expands very fast in its early stages, it is also dense, and the interaction times for particles are often (but not always) shorter than the expansion timescale. We can therefore often consider thermal equilibrium. Also the fluids of interest are simple enough that we can treat them as perfect gases.

We therefore need to revise a few pieces of statistical thermodynamics. Consider some box of volume $V = L^3$, and expand the fields inside into periodic waves with **harmonic boundary conditions**. The density of states in k space is

$$dN = g\,\frac{V}{(2\pi)^3}\,d^3k \tag{68}$$

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

$$\langle f\rangle = \left[e^{(\epsilon-\mu)/kT} \pm 1\right]^{-1} \tag{69}$$

(+ 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

$$n = \frac{1}{V}\int f\,dN = g\,\frac{1}{(2\pi\hbar)^3}\int_0^\infty \frac{4\pi\,p^2dp}{e^{\epsilon(p)/kT}\pm 1}, \tag{70}$$

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

$$n = \left(\frac{kT}{c}\right)^3 \frac{4\pi g}{(2\pi\hbar)^3}\int_0^\infty \frac{y^2dy}{e^y \pm 1}. \tag{71}$$

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

$$n = e^{-mc^2/kT}(2mkT)^{3/2}\frac{4\pi g}{(2\pi\hbar)^3}\int_0^\infty e^{-y^2}y^2dy. \tag{72}$$

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

The fascinating thing about this idea is that it applies for any particle – and to its corresponding antiparticle. This means that, above about 10^{13} K, there was a thermal background of protons and antiprotons. As the temperature drops below this threshold, the protons and antiprotons annihilate, so that their number density drops exponentially, satisfying the above expression.

The problem with this reasoning is that it tells us that there should be no protons in the universe today. The thermodynamics doesn't distinguish particles from antiparticles, so they should cancel each other out at low T. However, the universe seems to be made of matter, and not antimatter. The inevitable conclusion is that the universe at early times must have had very slightly more protons than antiprotons (about 1 extra proton per 10^9 antiprotons, as we will see). The mechanism that causes this **matter–antimatter asymmetry** is still controversial, but it is clear that such an asymmetry persists once it it is set up – creating or destroying pairs cannot change it.

8.3 Energy, entropy and degrees of freedom

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:

$$u = \rho\, c^2 = g\, \frac{1}{(2\pi\hbar)^3} \int_0^\infty \frac{4\pi\ p^2\, dp}{e^{\epsilon(p)/kT} \pm 1}\ \epsilon(p). \tag{73}$$

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,

$$dE = TdS - PdV \quad \Rightarrow \quad \left(\frac{E}{V}dV + \frac{\partial E}{\partial T}dT\right) = \left(T\frac{S}{V}dV + T\frac{\partial S}{\partial T}dT\right) - PdV. \tag{74}$$

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

$$S = \frac{E + PV}{T} \tag{75}$$

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 $s \propto T^3$. 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. In the ultrarelativistic limit, the number density of protons and baryons would have been similar, which is why the fractional violation of matter–antimatter symmetry must have been at the 10^{-9} level. The fact that this ratio is so large also justifies the adiabatic assumption: pretty well all the entropy is in the photons.

We sometimes need to be a little more precise about the distinction between ultrarelativistic backgrounds of bosons and of fermions. These can be dealt with by the following trick:

$$\frac{1}{e^x + 1} = \frac{1}{e^x - 1} - \frac{2}{e^{2x} - 1}. \tag{76}$$

Thus, a gas of fermions looks like a mixture of bosons at two different temperatures. Knowing that boson number density and energy density scale as $n \propto T^3$ and $u \propto T^4$, we then get the corresponding fermionic results. The entropy requires just a little more care. Although we have said that entropy density is proportional to number density, in fact the entropy density for an ultrarelativistic gas was shown above to be $s = (4/3)u/T$, and so the fermionic factor is the same as for energy density:

$$\begin{aligned} n_{\rm F} &= \frac{3}{4}\frac{g_{\rm F}}{g_{\rm B}} n_{\rm B} \\ u_{\rm F} &= \frac{7}{8}\frac{g_{\rm F}}{g_{\rm B}} u_{\rm B}. \\ s_{\rm F} &= \frac{7}{8}\frac{g_{\rm F}}{g_{\rm B}} s_{\rm B}. \end{aligned} \tag{77}$$

Using these rules, we can count the effective number of relativistic bosons in the universe:

$$g_* \equiv \sum_{\rm bosons} g_i + \frac{7}{8} \sum_{\rm fermions} g_j. \tag{78}$$

This function falls as the temperature falls and more species of particles become nonrelativistic. At very high temperatures, it will asymptote to a number determined by the total number of distinct elementary particles that exist (of order 100, according to current theories).

8.4 Time and temperature

At early enough times, the typical photons become energetic enough that they interact strongly with matter – so the whole universe sits at a temperature dictated by the radiation. The behaviour of matter changes as a function of its temperature, and so a number of key events in the history of the universe happen according to a schedule dictated by the temperature – time relation. Using the expression we had earlier for the relation between time and density for a radiation–dominated universe, $t = (32\pi G\rho/3)^{-1/2}$, we can deduce the time–temperature relation:

$$t/\text{seconds} = g_*^{-1/2} \left(T/10^{10.26}\,\text{K}\right)^{-2}. \tag{79}$$

This is independent of the present-day temperature (which we will soon see to be very small: a mere 2.73 K). The present temperature does however set the redshift above which the universe is radiation-dominated: this is roughly $z = 10^4$.

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_*	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

9 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**:

$$\dot{n} + 3Hn = -\langle \sigma v \rangle n^2, \tag{80}$$

where σ is the reaction cross-section and v is the particle velocity. The $3Hn$ term just represents dilution by the expansion of the universe. What this shows is that the change in n involves two timescales:

$$\begin{aligned} \textbf{expansion timescale} &= H(z)^{-1} \\ \textbf{interaction timescale} &= (\langle \sigma v \rangle n)^{-1} \end{aligned} \tag{81}$$

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 expansion 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. Later on, we shall see that 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.

9.1 Neutrino decoupling

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

$$s(\gamma + e^+ + e^-) = \frac{11}{4} s(\gamma). \tag{82}$$

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

$$T_\nu = \left(\frac{4}{11}\right)^{1/3} T_\gamma = 1.95\,\mathrm{K}, \tag{83}$$

for $T_\gamma = 2.73$ K. Although it is hard to see how such low energy neutrinos could ever be detected directly, their gravitation is certainly not negligible: they contribute an energy density that is a factor $(7/8) \times (4/11)^{4/3}$ times that of the photons. 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).

9.2 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 consequences for the cosmological density are easily worked out to be

$$\Omega h^2 = \frac{\sum m_i}{93.5\,\mathrm{eV}}. \tag{84}$$

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

$$\nu_e \lesssim 15\,\mathrm{eV} \quad \nu_\mu \lesssim 0.17\,\mathrm{MeV} \quad \nu_\tau \lesssim 24\,\mathrm{MeV}. \tag{85}$$

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

10 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:

$$\begin{aligned} 1\mathrm{MeV} &= 10^{10.065}\,\mathrm{K} \\ m_e &= 0.511\,\mathrm{MeV} \\ m_p &= 939\,\mathrm{MeV} \\ m_n - m_p &= 1.3\,\mathrm{MeV} \end{aligned} \tag{86}$$

10.1 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:

$$\begin{aligned} p + e^- &\leftrightarrow n + \nu \\ n + e^+ &\leftrightarrow p + \bar{\nu}. \end{aligned} \tag{87}$$

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

$$\frac{n_n}{n_p} = e^{-\Delta mc^2/kT} \simeq e^{-1.5(10^{10}\,\mathrm{K}/T)}. \tag{88}$$

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

$$Y = \frac{4 \times n_n/2}{n_n + n_p} = \frac{2}{1 + n_p/n_n} \tag{89}$$

(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. These aren't part of this course, but 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

$$T(\text{n freezeout}) \simeq 10^{10.14}\,\text{K} \quad \Rightarrow \quad \frac{n_n}{n_p} \simeq 0.34. \tag{90}$$

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.

10.2 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 $n + p = \text{D} + \gamma$ 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:

$$T(\text{Deuterium formation}) \simeq 10^{8.9}\,\text{K}, \tag{91}$$

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 $10^{-3}\,\text{kg}\,\text{m}^{-3}$ 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.

10.3 Formation of Helium

The argument so far has produced a universe consisting of just Hydrogen and Deuterium, but this is not realistic, as one would expect ^{4}He to be preferred on thermodynamic grounds, owing to its greater binding energy per nucleon (7 MeV, as opposed to 1.1 MeV for Deuterium). In equilibrium, the first nuclei to come into existence in significant numbers should be ^{4}He: 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 is only $\sim 10^{-12}$.

Since the simplest way to synthesize Helium is by fusing Deuterium, it is no surprise that the equilibrium prediction fails miserably in the expanding universe: the production of Helium must await the synthesis of significant quantities of Deuterium, which we have seen happens at a temperature roughly one-third that at which Helium would be expected to dominate. What the thermodynamic argument does show, however, is that it is expected that the Deuterium will be rapidly converted to Helium once significant nucleosynthesis begins. This argument is what allows us to expect that the Helium abundance can be calculated from the final n/p ratio. The main reactions of importance are of course 2-body ones, rather than the improbable coincidence of 2 protons and 2 neutrons all arriving at the same point simultaneously to make ^{4}He in one go. 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):

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

$$\begin{aligned} \mathrm{D} + \mathrm{D} &= {}^4\mathrm{He} \\ \mathrm{T} + p &= {}^4\mathrm{He} \\ \mathrm{D} + {}^3\mathrm{He} &= {}^4\mathrm{He} + p \\ \mathrm{T} + \mathrm{D} &= {}^4\mathrm{He} + n. \end{aligned} \tag{92}$$

The same thermodynamic arguments that say that Helium should be favoured at temperatures around 0.1 MeV say that more massive nuclei would be preferred in equilibrium at lower temperatures still. 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.

10.4 The number of particle generations

An accurate fit for the final neutron-to-proton ratio is

$$\frac{n_n}{n_p} \simeq 0.163\,(\Omega_{\mathrm{B}} h^2)^{0.04}\ (N_\nu/3)^{0.2}. \tag{93}$$

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 neutrino 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 neutron freezeout happens earlier, again raising the abundance.

Increasing the number of neutrino species adds $\Delta Y \simeq 0.01$ for each additional neutrino species. It is therefore clear that strong limits can be set on the number of unobserved species, and thus on the number of possible additional families in particle physics. For many years, these nucleosynthesis limits were stronger than those that existed from particle physics experiments. This changed in 1990, with a critical series of experiments carried out in 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.

10.5 Weighing the baryons

Unlike Helium, the critical feature of the relic abundances of the other light elements is that they 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 by the COPERNICUS UV satellite in the early 1970s. This gave D/H $\simeq 2 \times 10^{-5}$, providing the first evidence for a low baryonic density, as follows. Figure 6 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:

$$\eta \equiv (n_p + n_n)/n_\gamma = 2.74 \times 10^{-8} (T/2.73\,\mathrm{K})^{-3} \Omega_{\mathrm{B}} h^2. \qquad (94)$$

Figure 6 shows that this Deuterium abundance favours a low density, $\Omega_{\mathrm{B}} h^2 \simeq 0.02$, and data on other elements give answers close to this. The constraint obtained from a comparison between nucleosynthesis predictions and observational data is rather tight:

$$\Omega_{\mathrm{B}} h^2 \simeq 0.02 \pm 0.002. \qquad (95)$$

(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.

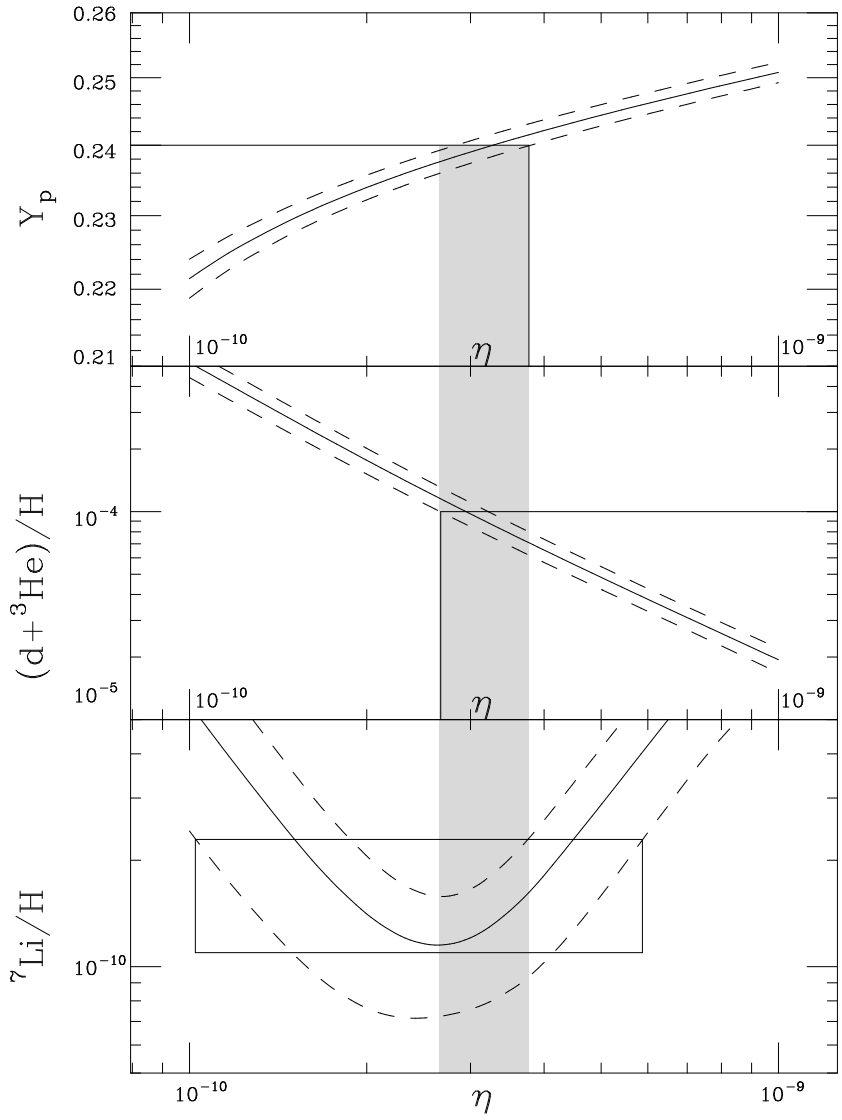

Figure 6. The predicted primordial abundances of the light elements, as a function of the baryon-to-photon ratio η (Smith, Kawano & Malaney 1993). For a microwave-background temperature of 2.73 K, this is related to the baryonic density parameter via $\Omega_{\rm B}h^2 = \eta/2.74 \times 10^{-8}$. Concordance with the data is found only for $\eta \simeq 3 \times 10^{-10}$, shown by the shaded strip.

11 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 first question to ask is how the ionization of a plasma would change as a function of time if it can be regarded as being in thermal equilibrium at the local temperature. This situation is described by the **Saha equation**:

$$\frac{x^2}{1-x} = \frac{(2\pi m_e kT)^{3/2}}{n(2\pi\hbar)^3} e^{-\chi/kT}, \tag{96}$$

where the fractional ionization, x, is the ratio of H ions to H ions + H atoms and χ is the ionization potential of Hydrogen. The Saha result is not simply a function of temperature, but the $\exp(-\chi/kT)$ term ensures that x in practice always diverges from 1 when $kT \sim \chi$.

The problem with the Saha approach is that the assumption of equilibrium rapidly ceases to be valid. This may seem paradoxical: processes with small cross-sections such as weak interactions may freeze out, but surely not the electromagnetic interaction? In fact, the problem is just the reverse: electromagnetic interactions 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).

There is a way out, however, using **two-photon emission**. The $2S \to 1S$ transition is strictly forbidden at first order and one can only conserve energy and angular momentum in the transition by emitting a *pair* of photons. Because of this slow bottleneck, the ionization at low redshift is far higher than would be suggested by the Saha prediction. Eventually, the process freezes out, and x gets stuck at $\sim 10^{-4}$ below redshifts of a few hundred.

11.1 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,

$$\tau = \int n_e x \sigma_{\rm T} d\ell_{\rm proper}; \qquad d\ell_{\rm proper} = R(z)\, dr = R_0\, dr/(1+z). \tag{97}$$

A good approximation for this quantity is

$$\tau(z) = 0.37 \left(\frac{z}{1000}\right)^{14.25}, \tag{98}$$

independent of cosmological parameters (although assuming the current temperature of 2.73 K). This is a very fortunate coincidence, but it would take too long to explain how it comes about.

Because τ changes rapidly with redshift, the distribution function for the redshift at which photons were last scattered, $e^{-\tau} d\tau/dz$, is sharply peaked, and is well fitted by a Gaussian of mean redshift 1065 and standard deviation in redshift 80. Thus, when we look at the sky, we can expect to see in all directions photons that originate from a **last-scattering surface** at

$z \simeq 1065$. It is worth noting that this redshift is very much lower than we would expect just from setting

$$k \times 2.73\ \mathrm{K} \times (1+z) = \chi, \tag{99}$$

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.

This argument does however emphasize that we can't know the redshift of last scattering without knowing the current radiation temperature. Historically, Gamow and collaborators used the nucleosynthesis argument to *predict* a background with a temperature of a few K – a prediction that unfortunately was not taken very seriously at the time.

12 The cosmic microwave background

12.1 The spectrum of the CMB

In a famous piece of serendipity, the CMB radiation was discovered in 1965, by Penzias & Wilson, who located an unaccounted-for source of noise in a radio telescope intended for studying our own galaxy. The timing of the discovery was especially ironic, given that experiments were under way at that time to test the theoretical prediction that such a background should exist – the progress of science is rarely a tidy business. Since the initial detection of the microwave background at $\lambda = 7.3$ cm, measurements of the spectrum have been made over an enormous range of wavelengths, from the depths of the Rayleigh–Jeans regime at 74 cm to well into the Wien tail at 0.5 mm. Prior to 1990, observations in the Rayleigh-Jeans portion of the spectrum had established the constancy of the temperature for $\lambda \gtrsim 1$ mm, buth there were suggestions of deviations from a Planck spectrum at short wavelengths in the Wien tail. However, **COBE** (the cosmic background explorer satellite, launched in 1989) showed that there is no deviation from a thermal spectrum at the 10^{-4} level (figure 7).

The COBE temperature measurement and 95% confidence range of

$$T = 2.728 \pm 0.004\,\mathrm{K} \tag{100}$$

improves significantly on the ground-based experiments, and indeed is limited in accuracy mainly by the systematics involved in making a good comparison black body. The lack of distortion in the shape of the spectrum is astonishing, and allows many competing cosmological models to be eliminated. Those with a distaste for the idea of a hot big bang suggested that the CMB could be due to starlight reprocessed by dust grains – but this would lead to a mixture of Planck functions because radiation would be received from a variety of redshifts. What the COBE result says is that the radiation must come from a shell within which temperature scales accurately as $T \propto 1/R(t)$ (so that the observed temperature of each shell is constant), which is something that only arises in the standard big-bang picture. The lack of any distortion from a Planck function moreover limits a number of possibilities for exotic processes in the later phases of the big bang. Although the CMB last scattered at $z \simeq 1100$, it was in one sense already a frozen-out background: Compton scattering conserves photon number, but full thermalization requires emission and absorption processes such as bremsstrahlung. These freeze out at a higher redshift – between 10^6 and 10^7.

Note that the CMB provides the answer to Olbers' paradox about why the night sky is dark. The answer is that it isn't, and only the expansion saves us from being cooked. Had we lived soon after recombination, the entire sky would have blazed like the surface of the sun.

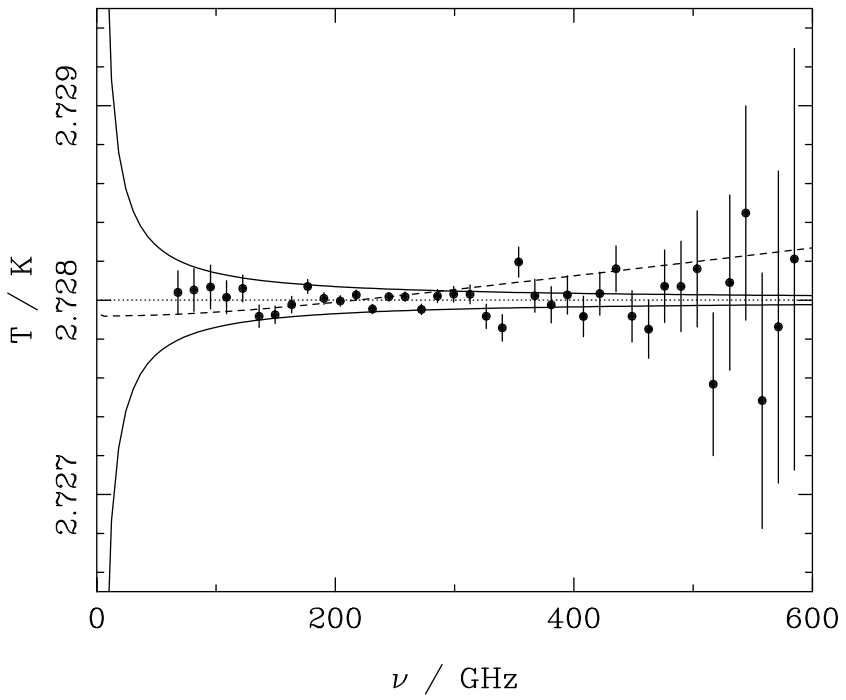

Figure 7. The spectrum of the microwave background as observed by the FIRAS experiment on the COBE satellite. The overall temperature scale has a systematic uncertainty of up to 0.004 K, but the deviations in shape are very much smaller even than this. The lines show the sort of distortions that might be expected from Compton scattering (broken) or non-zero chemical potential (solid).

12.2 The dipole anisotropy

The topic of deviations from isotropy in the microwave background is one of the dominant aspects of modern cosmology, and it is therefore treated separately below. However, this is a good place to study one important special case: the effects of the earth's absolute motion. The effect of the observer's motion on the specific intensity of the CMB can be calculated by using

$$\frac{I_\nu}{\nu^3} = \text{invariant.} \tag{101}$$

This is a general result that applies because the invariant is proportional to the **occupation number** of the radiation field. For example, consider blackbody radiation, where

$$\frac{I_\nu}{\nu^3} \propto \left(e^{h\nu/kT} - 1 \right)^{-1}, \tag{102}$$

which is the occupation number.

Now, we observe at some frequency ν, which has been Doppler-boosted from frequency ν_0 by our motion ($\nu = \mathcal{D}\,\nu_0$, where $\mathcal{D}$ denotes the Doppler factor). The invariant occupation number is

$$n = \frac{1}{\exp[h(\nu/\mathcal{D})/kT_0] - 1}, \tag{103}$$

where T_0 is the unboosted temperature. This tells us that boosted thermal radiation still appears *exactly thermal*, with a new temperature boosted in exactly the same way as frequency:

$$T = \mathcal{D}\, T_0. \tag{104}$$

If the velocity is small, then the Doppler factor is just $1 + \mathbf{v}\cdot\hat{\mathbf{r}}/c$, and

$$T_{\rm obs} \simeq T_0 \left[1 + \frac{v}{c}\cos\theta + O(v^2)\right]. \tag{105}$$

The dipole allows the absolute space velocity of the Earth to be measured, and this was first achieved by Smoot, Gorenstein & Muller in 1977. The COBE measurement of this motion is

$$v_\oplus = 371 \pm 1\,{\rm km\,s^{-1}} \quad \text{towards} \quad (\ell, b) = (264^\circ, 48^\circ). \tag{106}$$

This velocity measurement has been possible despite the invariant character of thermal radiation because of a leap of faith concerning the CMB: that it is very nearly isotropic. An observed dipole could of course be intrinsic to the universe (we cannot distinguish between this and any effect due to our motion). However, because the quadrupole term is only $\sim 1\%$ of the dipole, it is almost universally agreed that the observed dipole is entirely due to the motion of the Earth.

Of course, it was known long before the discovery of the CMB that the Earth is unlikely to be a sensible rest frame to adopt. There is the rotation of the Earth about the Milky Way; this is usually taken, according to the IAU convention, to be $300\,{\rm km\,s^{-1}}$ towards $(\ell, b) = (90^\circ, 0)$ – *i.e.* the galactic rotation has a left-hand screw. There are other corrections for the motion of the Milky Way relative to other members of the local group, but this is the dominant term. It was therefore predicted that a microwave dipole due to this motion would be seen. Imagine the surprise when the observed dipole turned out to be in a completely different direction! Correcting for the motion of the Earth with respect to the local group actually *increases* the velocity: the local group has the approximate motion

$$v_{\rm LG} \simeq 600\,{\rm km\,s^{-1}} \quad \text{towards} \quad (\ell, b) \simeq (270^\circ, 30^\circ). \tag{107}$$

The obvious interpretation of this motion is that it is caused by the perturbing gravity of large-scale structures beyond the local group.

13 Observations in cosmology

Having travelled from the extremes of quantum gravity, via nucleosynthesis, to a neutral universe at $T < 1000$ K, we reach the point where it is possible to do observational cosmology. Our observables are redshift, z, and angular difference between two points on the sky, $d\psi$. These can be converted into interesting intrinsic physical properties of the object under study, by using the RW metric and the Friedmann equation.

13.1 Sizes and volumes

We write the RW metric in the form

$$c^2 d\tau^2 = c^2 dt^2 - R^2(t) \left[dr^2 + S_k^2(r)\, d\psi^2 \right]. \tag{108}$$

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

$$d\ell_\perp = d\psi\ R(z) S_k(r) = d\psi\ R_0 S_k(r)/(1+z). \tag{109}$$

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

The element of proper distance in the radial direction is $R(z)\, dr$. This means that we can work out the cosmological volume element. If we survey a region of sky whose area is A steradians (think of a square $d\psi \times d\psi$), and which covers a range in comoving distance dr, the proper volume covered is

$$dV = [R(z)\, S_k(r)\, d\psi]^2 \times R(z)\, dr = A\, R(z)^3\, S_k(r)^2\, dr. \tag{110}$$

Often, we are more interested in the **comoving volume** (*i.e.* the volume that this expands to today). This would be given just by replacing $R(z)$ with R_0.

13.2 Luminosity and flux density

Probably the most important relation for observational cosmology is that between monochromatic flux density and luminosity. Start by assuming isotropic emission, so that the photons emitted by the source pass with a uniform flux density through any sphere surrounding the source. We can now make a shift of origin, and consider the RW metric as being centred on the source; however, because of homogeneity, the comoving distance between the source and the observer is the same as we would calculate when we place the origin at our location. The photons from the source are therefore passing through a sphere, on which we sit, of proper surface area $4\pi[R_0 S_k(r)]^2$. But redshift still affects the flux density in four further ways:

(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 $\nu_0(1+z)$.

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

$$S_\nu(\nu_0) = \frac{L_\nu([1+z]\nu_0)}{4\pi R_0^2 S_k^2(r)(1+z)} = \frac{L_\nu(\nu_0)}{4\pi R_0^2 S_k^2(r)(1+z)^{1+\alpha}}, \tag{111}$$

where the second expression assumes a power-law spectrum $L \propto \nu^{-\alpha}$. We can integrate over ν_0 to obtain the corresponding total or **bolometric** formulae, which are needed *e.g.* for spectral-line emission:

$$S_{\rm tot} = \frac{L_{\rm tot}}{4\pi R_0^2 S_k^2(r)(1+z)^2}; \tag{112}$$

13.3 Surface brightness

The flux density received by a given observer can be expressed by definition as the product of the **specific intensity** I_ν (the flux density received from unit solid angle of the sky) and the solid angle subtended by the source: $S_\nu = I_\nu \, d\Omega$. Combining the angular size and flux-density relations thus gives the relativistic version of surface-brightness conservation. This is independent of cosmology:

$$I_\nu(\nu_0) = \frac{B_\nu([1+z]\nu_0)}{(1+z)^3}, \tag{113}$$

where B_ν is **surface brightness** (luminosity emitted into unit solid angle per unit area of source). This works because I_ν/ν^3 is a relativistic invariant, which is just proportional to the photon occupation number (*cf.* the form of thermal radiation: $du/d\nu \propto \nu^3[\exp(h\nu/kT)-1]^{-1}$). This dimming makes it hard to detect extended objects at very high redshift.

13.4 Distance-redshift relation

The form of the above relations lead to the following definitions for particular kinds of distances, where we try to make things look Euclidean:

$$\begin{aligned} \textbf{angular}-\textbf{diameter distance}: &\quad D_{\rm A} = (1+z)^{-1} R_0 S_k(r) \\ \textbf{luminosity distance}: &\quad D_{\rm L} = (1+z)\, R_0 S_k(r). \end{aligned} \tag{114}$$

To complete the translation to observables, we need the relation between r and z derived from Friedmann's equation: $R_0 dr = [c/H(z)]\, dz$. For observational cosmology $z \lesssim 1000$, we can safely neglect radiation, so the distance-redshift relation is

$$R_0 dr = \frac{c}{H_0} \left[(1-\Omega_m-\Omega_v)(1+z)^2 + \Omega_v + \Omega_m(1+z)^3\right]^{-1/2} dz \tag{115}$$

(remember how this was derived: a null geodesic has $c\,dt = R\,dr$, and we write $dt = dR/\dot{R} = H^{-1} dR/R$ and use $dR/R = -dz/[1+z]$). Remembering also that we got $R_0 = (c/H_0)\,|1-\Omega|^{-1/2}$ from the Friedmann equation, we could write a complete equation for e.g. angular-diameter distance as a function of redshift. From an examination point of view, only proficiency in the $k=0$ case will be expected. This is easier, because $S_k(r)$ is then just r:

$$D_{\rm A}(z) = (1+z)^{-1} \frac{c}{H_0} \int_0^z \frac{dz'}{\sqrt{\Omega_v + \Omega_m(1+z')^3}} \tag{116}$$

For example, this gives

$$D_{\rm A}(z) = (1+z)^{-1}\frac{c}{H_0} \times 2\left(1-(1+z)^{-1/2}\right), \tag{117}$$

for the $\Omega_m = 1$ Einstein-de Sitter universe, and just

$$D_{\rm A}(z) = (1+z)^{-1}\frac{c}{H_0} \times z \tag{118}$$

for the $\Omega_v = 1$ de Sitter universe.

Some example distance-redshift relations are shown in figure 8. Notice how a high matter density tends to make distant objects brighter. It also tends to produce a maximum in the angular-diameter distance at $z \simeq 1$. Both these effects arise because gravitational deflection of light by mass inside light beams produces a focusing effect – as if we were observing distant objects through a colossal fish-eye lens.

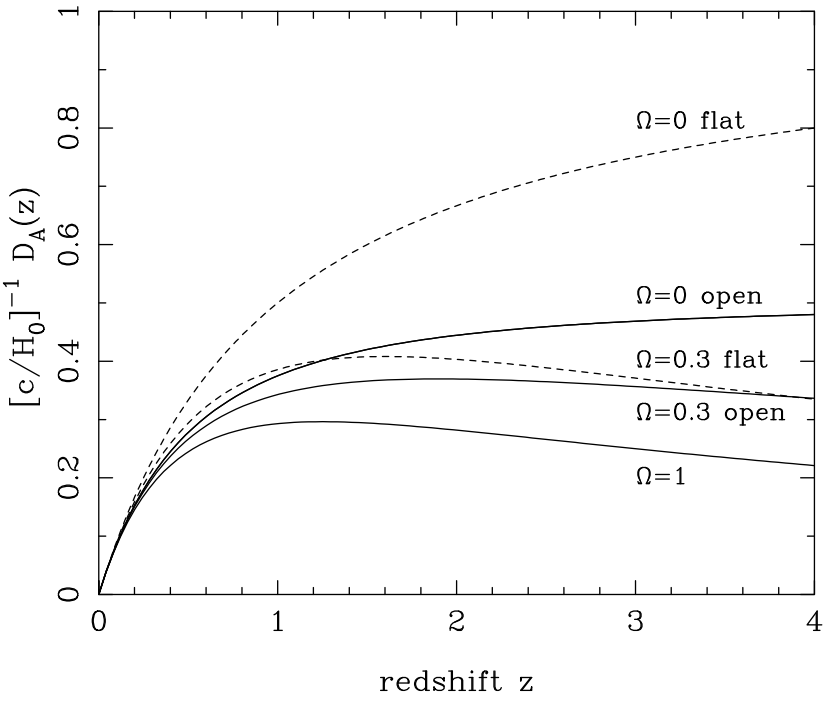

Figure 8. A plot of dimensionless angular-diameter distance versus redshift for various cosmologies. Solid lines show models with zero vacuum energy; dashed lines show flat models with $\Omega_m + \Omega_v = 1$. In both cases, results for $\Omega_m = 1, 0.3, 0$ are shown; higher density results in lower distance at high z, due to gravitational focusing of light rays.

13.5 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:

$$m = M + 5\log_{10}\left(\frac{D_{\rm L}}{10\ {\rm pc}}\right) + K(z), \tag{119}$$

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,

$$K(z) = 2.5(\alpha - 1)\log_{10}(1+z). \tag{120}$$

14 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.

14.1 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:

$$N(>S) = nV(S) = n(A/3)(L/4\pi S)^{3/2} \propto S^{-3/2}, \tag{121}$$

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:

$$I = \int S\, dN(S)/A. \tag{122}$$

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.

14.2 A relativistic count model

It should be helpful to contrast this analysis with the correct treatment for a specific case. We will choose the $\Omega_m = 1$ Einstein–de Sitter model. We have the following relations for this flat model. The comoving radius is

$$r = 2[1 - (1+z)^{-1/2}]. \tag{123}$$

The comoving volume is

$$V = \frac{4\pi A}{3} (R_0 r)^3 \propto [1 - (1+z)^{-1/2}]^3 \tag{124}$$

and the relation between flux density and distance is

$$S = \frac{L}{4\pi (R_0 r)^2 (1+z)^{1+\alpha}} \propto [1 - (1+z)^{-1/2}]^{-2} (1+z)^{-(1+\alpha)}. \tag{125}$$

For small z, we get $V \propto z^3$ and $S \propto z^{-2}$, implying $N \propto S^{-3/2}$ as before. However, for large z, V asymptotes to a constant (the horizon volume), whereas S continues to fall $\propto z^{-(1+\alpha)}$ for large z.

The number counts therefore inevitably depart from the Euclidean law, and tend to fall below the $S^{-3/2}$ scaling at about the point where a typical source reaches $z = 1$. The above example illustrates the general result, that it is the closing down of the volume elements, rather than the non-Euclidean dimming of high-z objects, which produces the low count slope.

Notice that we took the number of objects to be proportional to the *comoving* volume. The proper volume elements were higher at high z, but we would expect the proper number density of a conserved population of sources to scale as $n \propto (1+z)^3$. These effects cancel out if we combine comoving volume elements with a constant **comoving density** of sources.

A practical illustration of these effects is shown in figure 9. This displays observed galaxy counts in different wavebands. A range of different luminosities contributes to the counts, but we still see Euclidean counts until an object of typical luminosity reaches $z \simeq 1$, after which the counts rise more slowly. However, the data do not flatten as fast as predicted, and this is evidence for **cosmological evolution**: the galaxy population was more active at early times than at present.

14.3 Luminosity functions and galaxy evolution

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 $\ln L$:

$$dN = \phi(L, z)\; d\ln L\; dV(z). \tag{126}$$

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

$$\begin{aligned} d\phi &= \phi^*(L/L^*)^\alpha \exp(-L/L^*)\; dL/L^* \\ &= 0.921\, \phi^*(L/L^*)^{\alpha+1} \exp(-L/L^*)\; dM. \end{aligned} \tag{127}$$

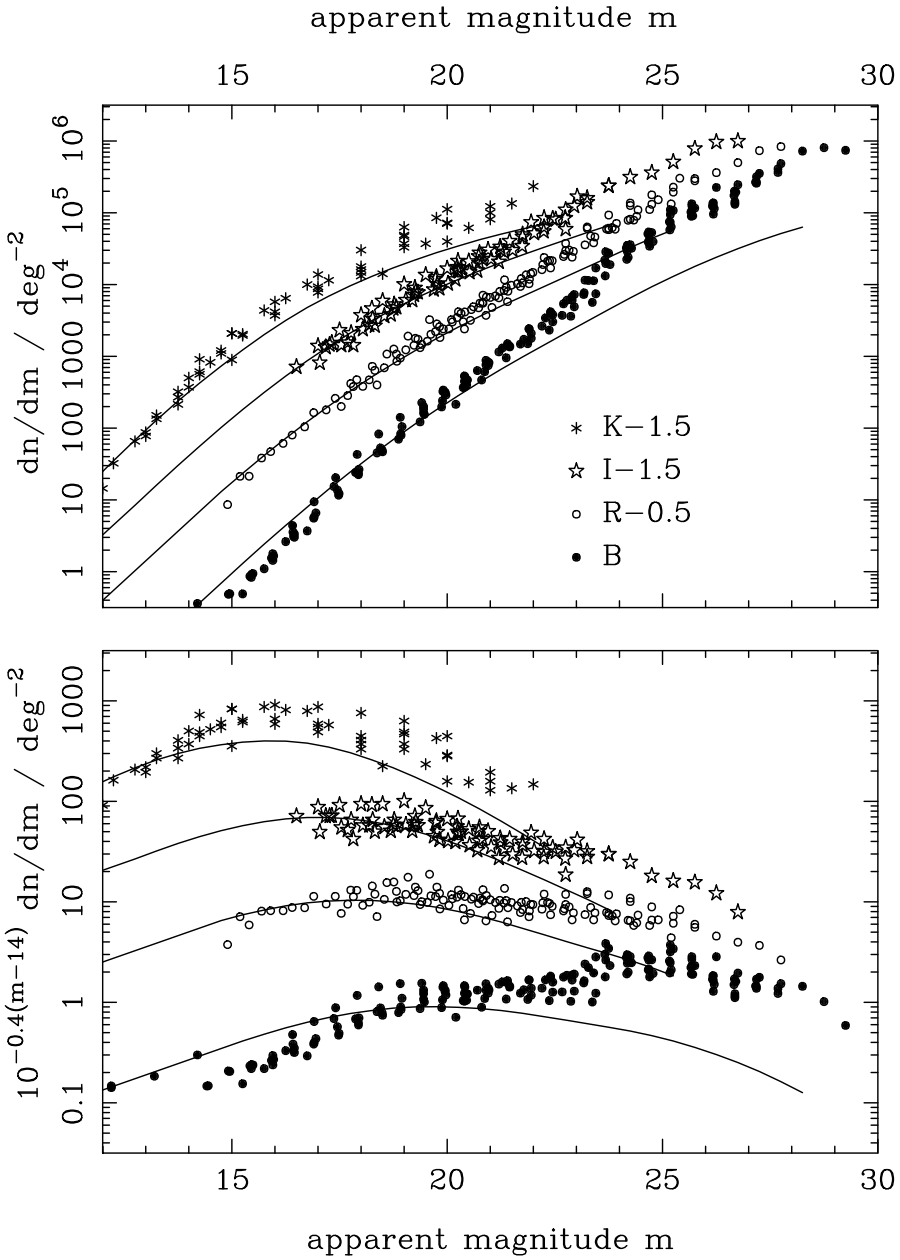

Figure 9. Differential counts of galaxies in the B, R, I and K bands. The upper panel shows the raw counts per unit magnitude interval, reaching up to a million galaxies per square degree at the faintest level. The bottom panel normalizes the counts by dividing by a count that scales as 1/(flux); this is the count for which the sky brightness diverges logarithmically. In this form, the significance of the deepest points (from the Hubble Deep Field) is revealed as indicating for the first time a convergence in the total amount of star-forming activity in the universe. The lines show no-evolution models, whereby the current galaxy population is transported unchanged to high redshifts and the resulting counts are predicted, assuming $\Omega = 1$. Although both sets of counts have very similar shapes, the different optical and infrared K-corrections mean that the blue counts were expected to cut off more rapidly, but this is not seen. The K counts are nearly consistent with the expectation for a non-evolving galaxy population, whereas there is a great excess of faint blue galaxies. The larger volume elements in low-density models raise the predicted faint counts somewhat, but do not change the qualitative point.

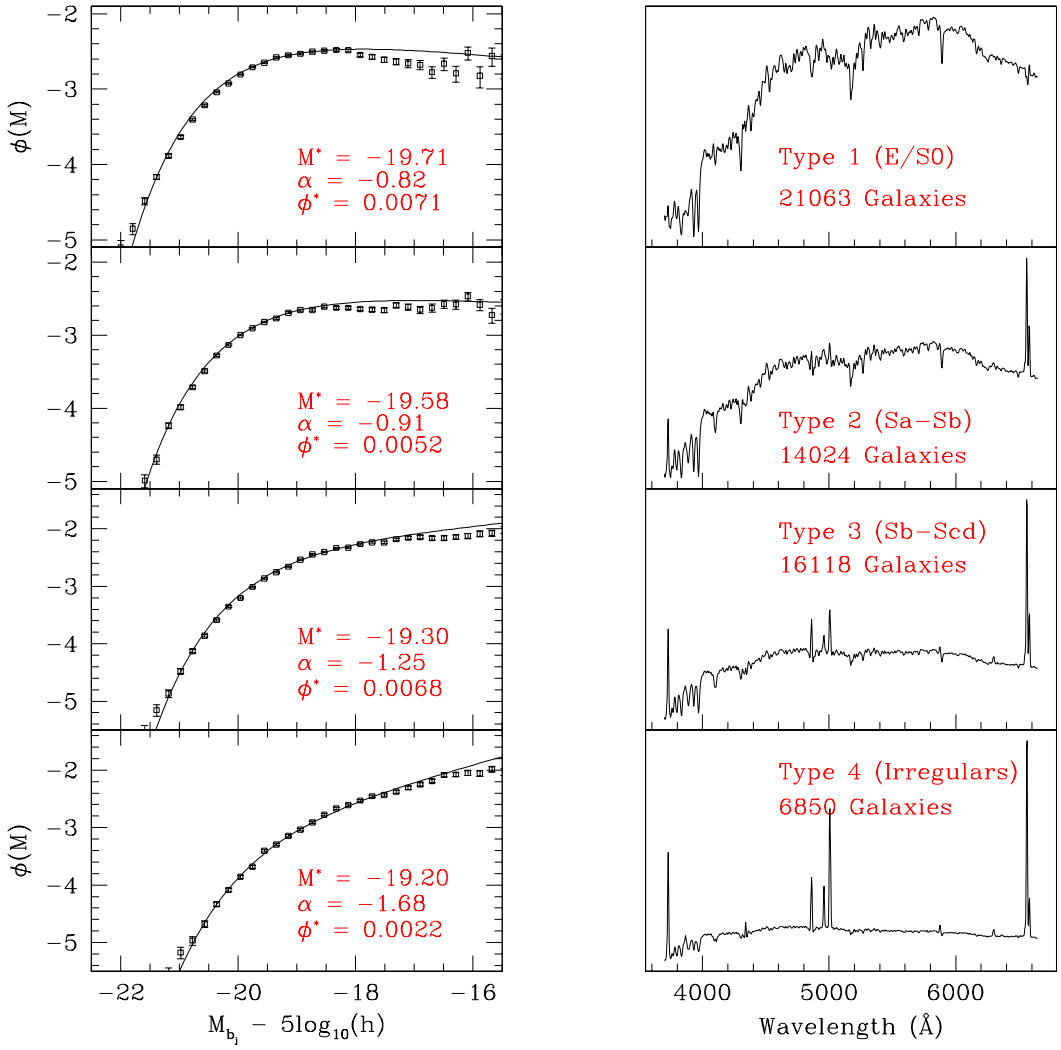

Figure 10. The galaxy luminosity function, with the population dissected into different types. At the top, we gave galaxies with old stellar populations (ellipticals); at the bottom, we move to galaxies dominated by younger stars (extreme spiral and irregular galaxies). The latter are systematically less luminous than the former (and less massive). In all cases, a Schechter function (power-law times exponential cutoff) gives a good fit.

Practical values of α for the galaxy population range between -1 and -1.5, depending on type (see figure 10). Note the factor 0.921, which is needed if we want numbers per magnitude rather than per $\ln L$.

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

$$L = S\,4\pi[R_0 S_k(r)]^2(1+z)^{1+\alpha}, \tag{128}$$

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

$$dN = \phi(L,z)\; d\ln S\; dV(z), \tag{129}$$

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.

Two simple extremes are usually discussed for describing the evolution of a luminosity function. The number of objects may be changed by scaling the whole luminosity function either vertically (**density evolution**) or horizontally (**luminosity evolution**):

$$\phi(L, z) = \begin{cases} f(z)\,\phi_0(L) & \text{(density evolution)} \\ \phi_0\big(L/g(z)\big) & \text{(luminosity evolution).} \end{cases} \tag{130}$$

Both these alternatives are clearly equivalent for a pure power-law luminosity function, but they can be distinguished if the luminosity function contains any curvature.

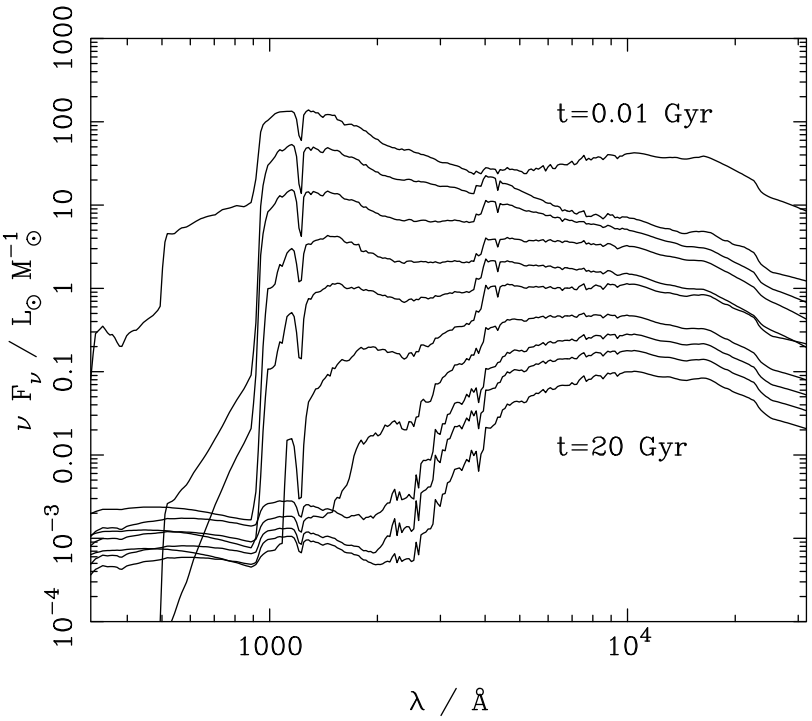

Figure 11. A plot of the expected evolution with time of the spectral energy distribution of a burst of star formation, assuming Solar metallicity. The time sampling is logarithmic, from an age of 0.1 Gyr at the top to 20 Gyr at the bottom. Note the ageing away of the blue O & B stars, leading to the establishment of the characteristic red spectrum shortwards of the '4000-Angstrom break' at an age of about 1 Gyr.

Luminosity evolution is certainly expected, because the stellar population in a galaxy ages. Once star formation has ceased, stars of progressively lower mass begin to use up their nuclear fuel and move off the main sequence. This removes luminous hot blue stars and replaces them with cooler red giants. Thus, the luminosity of a galaxy is expected to decline with time, with the decline being more rapid at the shorter wavelengths (as shown in figure 11). Redwards of the 4000Å break, the rate of evolution is similar at all wavelengths, and this is because the light in this region is dominated by the red giants. Their short lifetimes mean that the stellar

luminosity is determined by the rate at which stars evolve off the main sequence. In practice, this means that the luminosity of a galaxy decline as a power of time: approximately $L_{\rm G} \propto t^{-0.5}$.

Passively evolving galaxies with high formation redshifts are therefore expected to be between 0.4 magnitudes ($\Omega = 0$) and 0.6 magnitudes brighter ($\Omega = 1$) at $z = 1$ than at present. This evolution explains much of the behaviour seen in the number counts, but there is a limit to what luminosity evolution can achieve. This is because, even to the highest redshifts, there is only a finite amount of comoving volume to be sampled. This volume times the total comoving number density of galaxies integrated over all luminosities gives the maximum numer of galaxies that can be seen over the sky. In practice, this limit is exceeded if we assume an $\Omega = 1$ Einstein–de Sitter universe. The conclusion is either that there were more galaxies at high redshift (*i.e.* present-day galaxies were produced via **mergers** of smaller units), or that we need a cosmological model with more volume (a lower-density model). Current research still finds it difficult to distinguish these alternatives.

15 The distance scale and the age of the universe

We now consider the contribution that observational cosmology has made towards measuring one of the key cosmological parameters: the Hubble constant, H_0. As we have seen, this sets the basic length scale of the universe:

$$\textbf{Hubble length}: \quad c/H_0 \simeq 3000\, h^{-1}\ {\rm Mpc} \tag{131}$$

(recall the definition of the dimensionless Hubble parameter: $h \equiv H_0/100\ {\rm km\,s^{-1}Mpc^{-1}}$). It also fixes the characteristic age of the universe:

$$\textbf{Hubble time}: \quad H_0^{-1} = 9.78\, h^{-1}\ {\rm Gyr}, \tag{132}$$

If the expansion did not decelerate, the Hubble time would be exactly the age of the universe: $v = Hd$ and $d = vt$, so $t = 1/H$. In general, we must consider a photon null geodesic

$$c\, dt = R\, dr = R_0\, dr/(1+z), \tag{133}$$

and use Friedmann's equation to get the relation between comoving distance and redshift, and hence the relation between redshift and time. Ignoring radiation, this gives the age of the universe as

$$H_0 t_0 = \int_0^\infty \frac{dz}{(1+z)\sqrt{(1+z)^2(1+\Omega_m z) - z(2+z)\Omega_v}}. \tag{134}$$

Over the range of interest ($0.1 \lesssim \Omega_m \lesssim 1$, $|\Omega_v| \lesssim 1$), this exact answer may be approximated to a few % accuracy by

$$H_0 t_0 \simeq \tfrac{2}{3}\left(0.7\Omega_m + 0.3 - 0.3\Omega_v\right)^{-0.3}. \tag{135}$$

If the density content of the universe is known, then H_0 and t_0 are directly related. Alternatively, if both H_0 and t_0 are known, then we measure a combination of Ω_m and Ω_v.

15.1 Age estimates from nuclear decay

The most accurate means of obtaining ages for astronomical objects is **nuclear cosmochronology**, based on the natural clocks provided by radioactive decay. There exist a number of heavy nuclei with decay lifetimes of order 10 Gyr. The most useful decay clocks are based on thorium and uranium: $^{232}\mathrm{Th} \rightarrow {}^{208}\mathrm{Pb}$ (20.27 Gyr); $^{235}\mathrm{U} \rightarrow {}^{207}\mathrm{Pb}$ (1.02 Gyr); $^{238}\mathrm{U} \rightarrow {}^{206}\mathrm{Pb}$ (6.45 Gyr). The use of these clocks is complicated by a lack of knowledge of the initial conditions of the decay. Suppose (as is effectively the case in the above examples) that a given 'parent' (P) element decays only to a single 'daughter' (D), and that this daughter is produced by no other reaction. Once a sample of material is isolated from the nuclear reactions that produce the heavy elements (largely in supernovae), the abundances of the two elements concerned (by mass, say) satisfy

$$D = D_0 + P_0[1 - \exp(-t/\tau)] = D_0 + P[\exp(t/\tau) - 1]. \tag{136}$$

The quantities D and P are observed and τ is known; however, the age cannot be found unless the initial daughter abundance D_0 is known.

The way round this impasse is to exploit what at first sight seems to be a complication. Once solid bodies condense from a cloud of chemically uniform material, chemical fractionation produces spatial variations in P_0/D_0. Now pick a stable isotope S of D and express all abundances as a ratio to S; at the initial time, D_0/S will be a constant, whereas P_0/S will have a range of values. As a result, both D/S and P/S will have a range of values at some later time, but the scatter in these variables will be correlated:

$$D/S = D_0/S + (P/S)[\exp(t/\tau) - 1] \tag{137}$$

The age of a sample can thus be determined from the slope of a plot of P/S against D/S. This gives the time at which spatial homogeneity in P/S was first broken, and corresponds to the time of solidification of a given rock. It is only the introduction of a scatter in P/S with no scatter in D/S that makes the age measurement possible at all. When applied to meteorites, these methods give a highly precise age for the Solar system: this formed 4.57 Gyr ago, with an uncertainty of only about 1%. The oldest rocks on Earth are slightly younger: about 3.7 Gyr.

Such studies tell us not only the age of the Solar system, but also the abundances in the pre-Solar material: $^{235}\mathrm{U}/^{238}\mathrm{U} \simeq 0.33$; $^{232}\mathrm{Th}/^{238}\mathrm{U} \simeq 2.3$. Can we use these numbers to date the time of production of the heavy elements, and hence obtain an age for the galaxy itself? This can only be done if we know what abundance ratios are produced in the initial supernova explosion, which requires an input from nuclear physics theory. The resulting limit is somewhat model dependent, but gives an age for the local part of the Milky Way of about 9.5 Gyr.

15.2 Ages from stellar evolution

The other major means of obtaining cosmological age estimates is based on the theory of stellar evolution. Consider figure 12, which is a colour–magnitude diagram for a **globular cluster** – a system of roughly 10^5 stars. These systems are found in the outer parts of galaxy haloes, and are believed to be examples of a stellar population at a single age (**coeval**). The colour–magnitude diagram is the practical version of what can be predicted from theory: the **Hertzsprung-Russell diagram**, (HR diagram) or a plot of luminosity against effective temperature. On this plot, we can see very clearly the main elements of the life-cycle of a star, in particular the **main sequence**, which is the locus extending to low luminosities and red colours. This is a line parameterized by mass, where massive stars are bluer and more luminous. At the time of creation of the cluster, only the diagonal line of the main sequence would have been occupied.

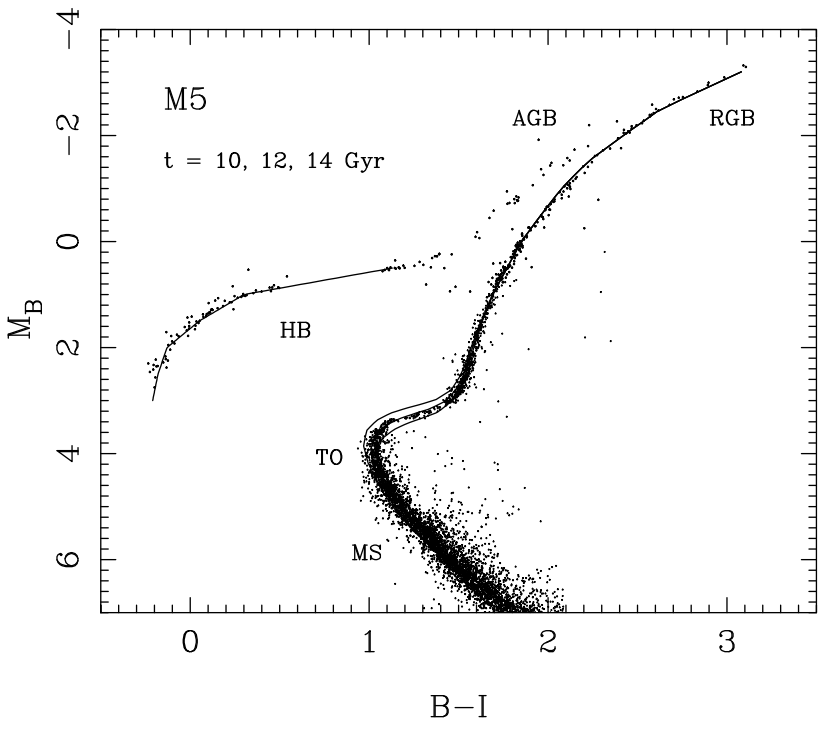

Figure 12. A colour–magnitude plot (the observational version of the Hertzsprung–Russell diagram) for stars in the globular cluster M5 (data courtesy of R. Jimenez). This illustrates well the main features of stellar evolution: the main sequence (MS) and its turn-off point (TO), followed by the red giant branch (RGB), horizontal branch (HB) and asymptotic giant branch (AGB). The main-sequence turn-off occurs at about $M_B = 4$, and is well-fitted by a theoretical isochrone of age 12 Gyr.

A star will shine on the main sequence until its fuel is exhausted, which corresponds to having fused a significant fraction of its total mass. Because the luminosity rises rapidly with mass (between $L \propto M^3$ and $L \propto M^5$), The lifetime is then given by $L\tau \propto Mc^2$, so the most luminous stars leave the main sequence soonest.

Once hydrogen is exhausted in the core, it continues to burn in a shell around the core. The helium-rich core contracts until it is supported by electron degeneracy pressure, as in white dwarfs. This phase is associated with an expansion in size of the outer parts of the star by roughly a factor 10 to form a red giant. There is thus a **turnoff point** on the main sequence, corresponding to stars that have just reached this point.

The age of a globular cluster may be obtained via a fit of a single-age evolution model (an **isochrone**) to the data in the colour–magnitude diagram, as illustrated in figure 12. The fits are impressively good in their detailed agreement with observation, although some significant difficulties are hidden when only the final plot is inspected. First of all, the distance to the cluster under study is not known, and the conversion from apparent magnitudes to luminosities is correspondingly uncertain. This might not seem a problem: the main sequence defines a colour–luminosity relation, and there will only be one choice of distance at which the model will match. However, both the metallicity of the cluster and the line-of-sight reddening will alter the main-sequence locus, and uncertainties in these parameters alter the best-fit distance.

The ages determined in this way for the low-metallicity clusters that are plausibly the oldest systems consistently come out in the range 13 – 17 Gyr. Given the random errors in the fitting procedure alone, it is plausible that the very largest figures may be upward fluctuations, but a figure of about 15 Gyr may be regarded as a good estimate of the typical age. However, because any uncertainty in the distance scale is systematic, this must carry a minimum error of ±1.5 Gyr. The 95% confidence limit for the age of the halo of the Milky Way is thus approximately $t > 12$ Gyr – consistent with the nuclear lower limit, but more restrictive.

15.3 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, and they have a positive correlation between luminosity and period (roughly $L \propto P^{1.3}$) that is very tightly defined. 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.

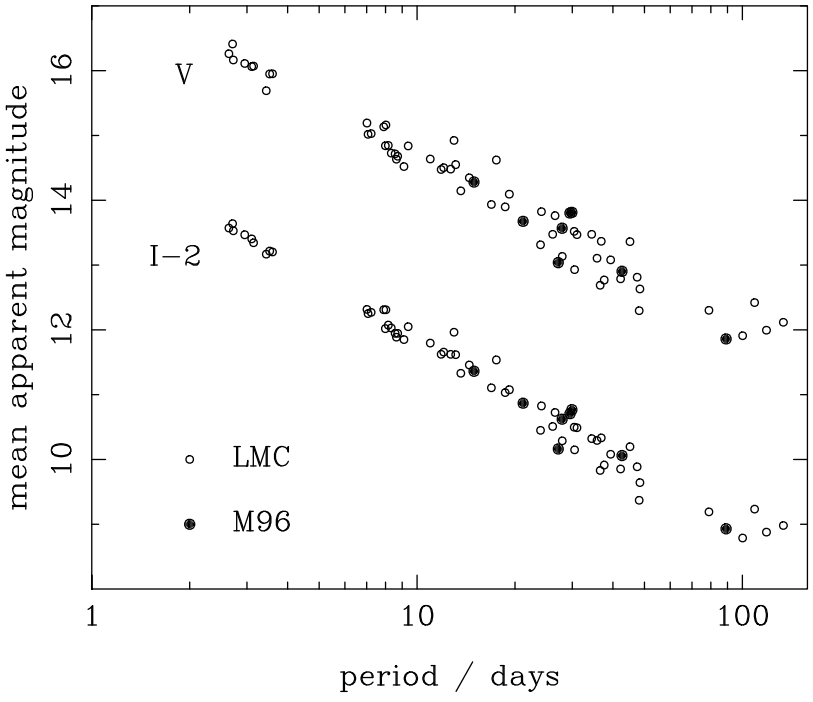

Figure 13. A plot of the Cepheid $P-L$ relation for stars in the Large Magellanic Cloud and in M96. The M96 stars have been shifted to overlay the LMC data, and the required shift gives the apparent difference in distance modulus. Examining this offset as a function of wavelength and fitting an extinction model allows the true relative distance to be established.

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. What is needed is in fact a stellar structure where the opacity is sensitive to small changes in density. This arises in Cepheids because of the existence of a transition zone between singly and doubly ionized helium, which is sensitive to the small changes in temperature associated with small radial perturbations.

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:

$$D_{\rm LMC} = 51 \text{ kpc } \pm 6\%. \tag{138}$$

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.

15.4 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 we learned from the CMB dipole that 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 **T**ully-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 10%, and hence giving relative distances to about 10% precision. The great discovery of the 1990s was that supernovae make much more accurate **standard candles**.

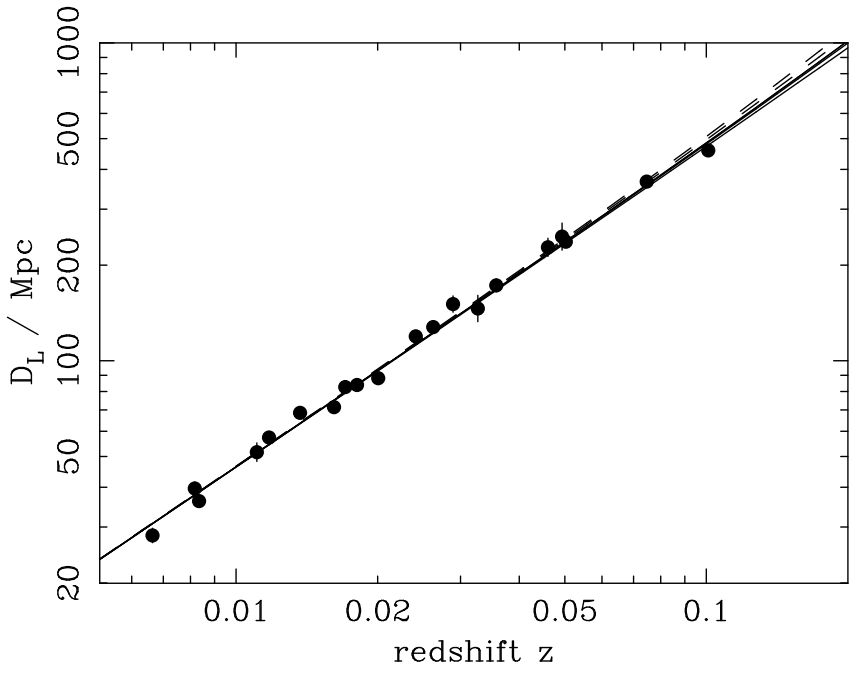

Figure 14. The type Ia supernova Hubble diagram. Using a measure of characteristic time as well as peak luminosity for the light curve, relative distances to individual SNe can be measured to 6% rms. Setting the absolute distance scale ($D_{\rm L}$ is luminosity distance) using local SNe in galaxies 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 14), and allows a very precise estimate of H_0, based on the HST Cepheid distances:

$$H_0 = 72 \pm 8\,\mathrm{km\,s^{-1}Mpc^{-1}}. \tag{139}$$

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

As discussed earlier, 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

$$0.55 < H_0 t_0 \simeq \tfrac{2}{3}\left(0.7\Omega_m + 0.3 - 0.3\Omega_v\right)^{-0.3} < 1.37, \tag{140}$$

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.

16 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. This is a big practical challenge, not only because accurate distance indicators are required, but especially because we must be sure that the 'standard candles' do not evolve. This rules out almost any indicator based on whole galaxies, but supernovae are suitable. Figure 15 shows the SNe Hubble diagram out to very large redshifts, emphasizing the curvature in the relation.

It is clear from figure 15 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 16, which can be used to set confidence limits on the cosmological parameters. The results very clearly require a low-density universe. For $\Omega_v = 0$, a very low density is just barely acceptable, with $\Omega_m \lesssim 0.1$. However, 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 $(\Omega_m, \Omega_v) = (0.3, 0.7)$.

These results add to the indications from the age of the universe that the universe contains non-zero vacuum energy. Although there were other lines of argument that yield the same conclusion, the SNe results are so direct and precise that they convinced the community very rapidly, yielding something of a scientific revolution in 1997/98.

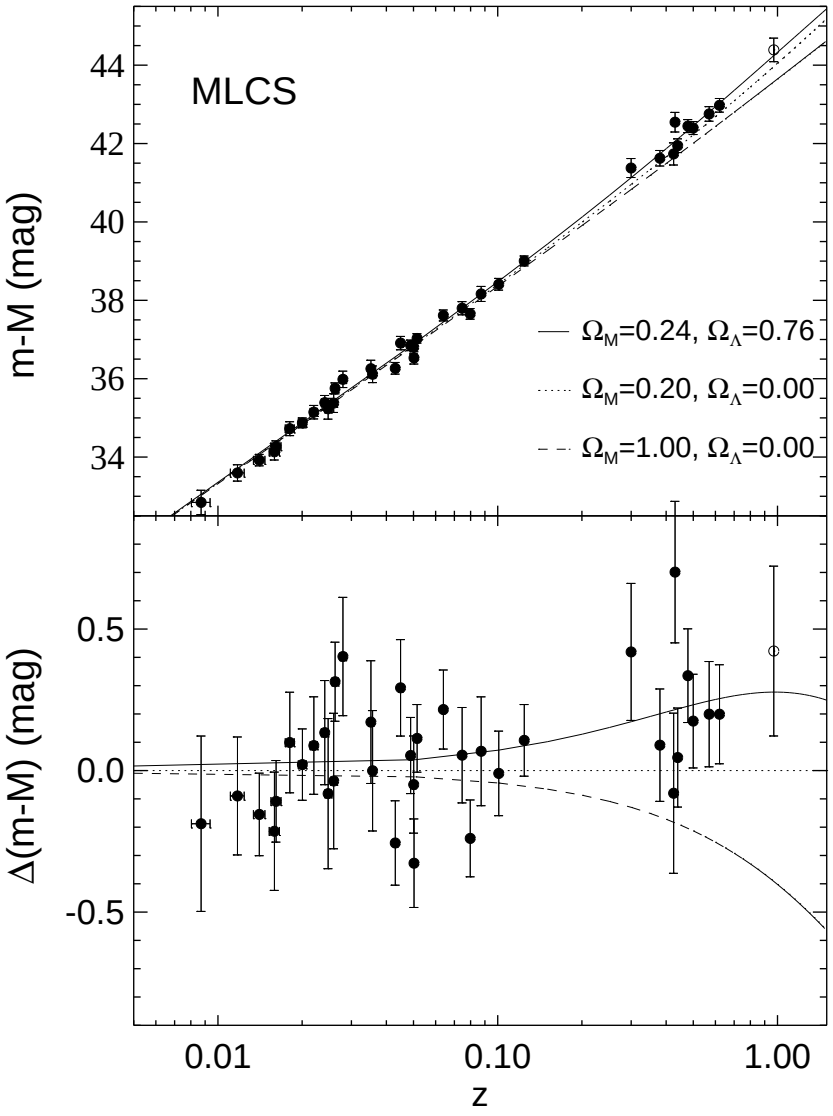

Figure 15. The Hubble diagram. The lower panel shows the data divided by a default model ($\Omega_m = 0.2$, $\Omega_v = 0$). The results lie clearly above this model, favouring a non-zero Ω_v. The lowest line is the Einstein-de Sitter model, which is in gross disagreement with observation.

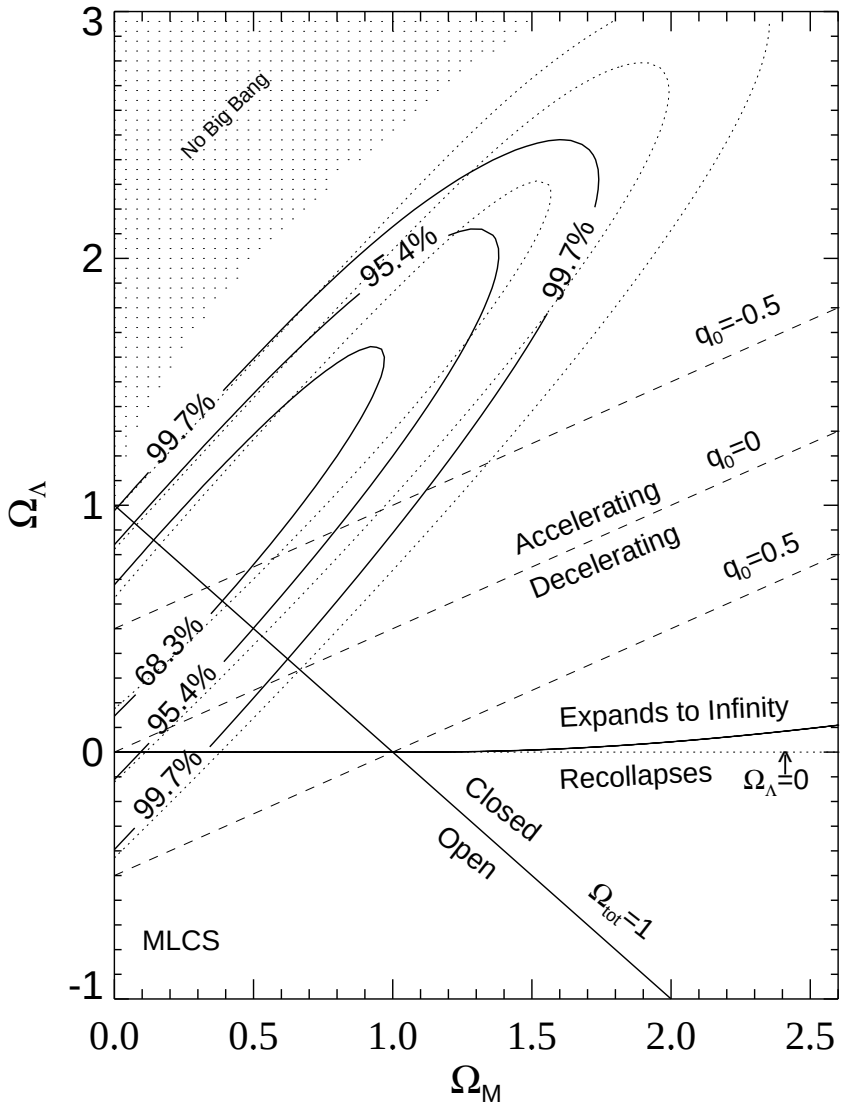

Figure 16. Confidence contours on the Ω_v–Ω_m plane. Open models of all but the lowest densities are apparently ruled out, and nonzero Λ is strongly preferred. If we restrict ourselves to $k = 0$, then $\Omega_m \simeq 0.3$ is required. The constraints perpendicular to the $k = 0$ line are not very tight, but CMB data can help here in limiting the allowed degree of curvature.

17 Dark matter

We have seen how the density of baryonic material in the universe was determined via nucleosynthesis: $\Omega_{\rm B} h^2 \simeq 0.02$. Together with the best guess for the Hubble constant of $h \simeq 0.7$, this suggests $\Omega_{\rm B} \simeq 0.04$, so the universe is very far from being closed by normal matter. No argument is ever watertight, however, so it is sensible to check this number. The best way to achieve this is to detect mass directly via gravity.

17.1 Mass-to-light ratios and Ω

The classical technique for weighing the whole universe is '$L \times M/L$'. The overall light density of the universe is reasonably well determined from redshift surveys of galaxies, so that a good determination of mass M and luminosity L for a single object suffices to determine Ω *if* the mass-to-light ratio is universal. Of course, different bodies will have different M/L values: if the Sun has a value of unity (these measurements are always quoted in Solar units), then low-mass stars have values of several tens while a comet can easily have $M/L \sim 10^{12}$. The stellar populations of galaxies typically produce M/L values between 1 and 10.

Galaxy redshift surveys allow us to deduce the galaxy luminosity function, ϕ, and hence the total luminosity density produced by integrating over all galaxies:

$$\rho_L = \int L \, d\phi(L). \tag{141}$$

In blue light, a canonical value for the total luminosity density is $\rho_{\rm L} = 2 \pm 0.7 \times 10^8 \; h L_\odot {\rm Mpc}^{-3}$. Note the h scaling: because cosmological distances scale as h^{-1}, luminosities go as h^{-2}, volumes as h^{-3} and number densities as h^3. Since the critical density is $2.78 \times 10^{11} \Omega h^2 M_\odot {\rm Mpc}^{-3}$, the critical M/L for closure is

$$\left(\frac{M}{L} \right)_{\rm crit,\, B} = 1390h \pm 35\%. \tag{142}$$

For reference, primordial nucleosynthesis implies Ωh^2 of 0.02, so we should certainly expect to find bodies with $M/L > 28/h$. We therefore know in advance that there must be more to the universe than normal stellar populations – **dark matter** of some form is inevitable.

17.2 Dark matter in galaxy haloes

One key piece of observational evidence for dark matter is that many galaxies show **flat rotation curves**. Rotation curves are readily analyzed by assuming a spherically symmetric model, so that the rotation speed for circular orbits, V, is just

$$V^2 = GM(< r)/r. \tag{143}$$

At large radii, one might expect that the total mass inside r would have converged, so that we expect the **Keplerian fall-off** $V \propto 1/r^{1/2}$. This is not seen in many cases. Rather, V is often nearly constant, as shown in figure 17.

It might seem that it would be very difficult to establish the behaviour of the mass at large radii, simply because one will run out of light. However, it is in fact possible to measure the rotation velocities of galaxies well beyond the point at which the emission from stars becomes undetectable, using 21-cm emission from neutral hydrogen (*e.g.* figure 17). This means that the best evidence for dark haloes exists for spiral galaxies.

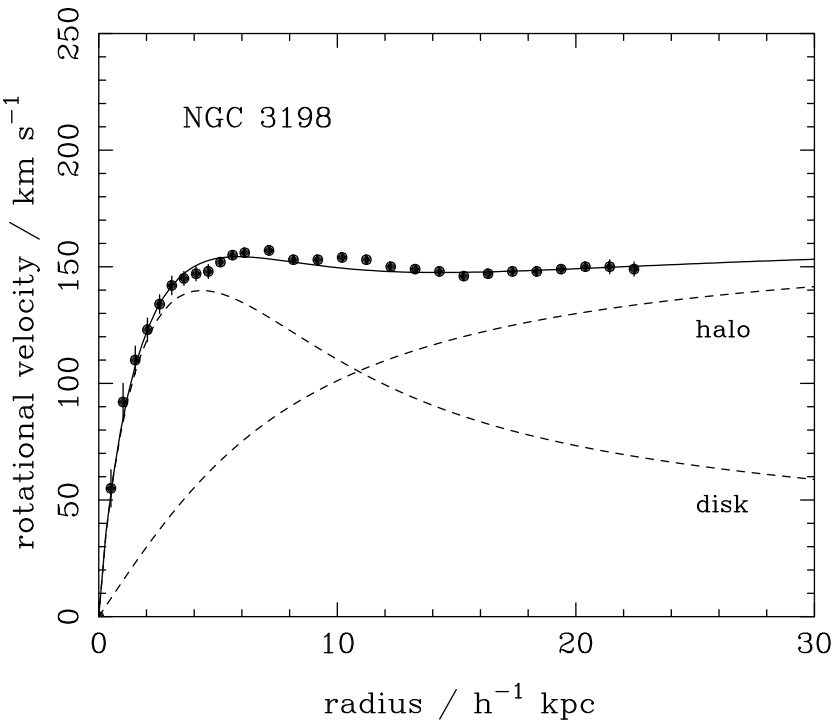

Figure 17. The brightness profile and rotation curve of the spiral galaxy NGC3198. Note the flat form of the curve at large radii by comparison with the contribution expected from the luminous disk.

For the above law, a constant V implies $M(< r) \propto r$, or a density law in the form of an **isothermal sphere**:

$$\rho \propto r^{-2}. \tag{144}$$

It might be objected that the mass may be not spherically symmetric. Consider the opposite extreme: a massive flat disk where the enclosed mass obeys the same $M(< r) \propto r$ law (*i.e.* a density per unit area $\propto 1/r$). Although it is not so easy to prove, the circular velocity of orbits in the disk is still $V^2 = GM(< r)/r$, which is identical to the spherical result. In fact, it is possible to probe whether or not the mass distribution is flattened, by looking for the occasional galaxies with material orbiting out of the plane (so-called **polar-ring galaxies**). Results here favour haloes that are mildly flattened (axial ratio perhaps 1.3:1).

Such rotation-curve studies have now demonstrated that the material in the outer parts of galaxies must have $M/L \gtrsim 1000$; within $30h^{-1}$ kpc, the typical values for the whole galaxy are $M/L \simeq 30h$ in the case of luminous galaxies similar to the Milky Way. Since we expect M/L of up to 10 for typical stellar populations, this shows that dark matter outweighs baryonic material by around 3 – 5 to 1. On the other hand, this figure is very much dependent on the mass of the galaxy under study. For dwarf galaxies with circular velocities below 100 km s^{-1}, only about 1% of the total dynamical mass can be accounted for in stars and gas.

17.3 Dark matter in groups and clusters of galaxies

Historically, the first detection of dark matter was made in 1933 by Zwicky. He looked at the dispersion of velocities in the Coma cluster and deduced $M/L \simeq 300h$ from application of the virial theorem. Modern techniques are able to probe the mass distribution in rather more detail, but end up with very much the same answer.

It is usually assumed that the cluster to be studied is spherically symmetric and in a relaxed equilibrium state. This means that the galaxies act similarly to a fluid in **hydrostatic equilibrium**, with their orbits in random directions playing the role of a pressure. Consider the radial equilibrium of a fluid element between the competing forces of gravity and 'pressure' of the galaxy orbits, $p = \rho\sigma_v^2$. This may look unfamiliar, but it is analogous to the familiar $p = mn\bar{v^2}/3$: the factor of 3 disappears because σ_v is defined as the rms dispersion of velocities in one direction. This gives the equilibrium equation

$$-\frac{\partial\Phi}{\partial r} = -\frac{GM(<r)}{r^2} = \frac{1}{\rho_g}\frac{\partial}{\partial r}\left(\rho_g\sigma_v^2\right), \tag{145}$$

where Φ is the gravitational potential and $M(<r)$ is the total mass integrated out to radius r. Remember that ρ_g is the mass density in the galaxies, which can differ from the total density. Indeed, it is clear from this equation that only the shape of $\rho_g(r)$ matters, rather than its absolute value.

We need to solve for $M(<r)$, given the observables of $\sigma_v^2(r)$ plus the projected galaxy surface density $\Sigma(r)$. The simplest approach assumes that light traces mass, so that the galaxy distribution gives the shape of the projected mass distribution. In this case, the only freedom is in M/L. The observed $\Sigma_{\rm g}(r)$ can be deprojected to yield $M(<r)$ up to an unknown factor, which predicts the shape of $\sigma_v^2(r)$. Scaling to the velocity-dispersion data gives an estimate of M/L.

If the constant M/L assumption is abandoned, the equations can still be solved, provided the observed projected velocity dispersion as a function of r can be deprojected to infer the true $\sigma_v^2(r)$. Reassuringly, these different methods all give rather similar answers for the mass in the central parts of the Coma cluster: about $6\times10^{14}h^{-1}M_\odot$ within a radius of $1h^{-1}$ Mpc. The light and mass can be mapped over a reasonable range of radii, and the result is that the two profiles follow each other quite closely. The implication is a nearly constant mass-to-light ratio, $M/L_B \simeq 300h - 400h$, implying $\Omega \simeq 0.2 - 0.3$ if clusters are large enough systems for their M/L ratios to be representative.

17.4 Masses from x-ray gas

A completely different tracer of mass is also available in clusters of galaxies in the form of hot gas that is visible in X-rays. For this material, it is reasonable to suppose that hydrodynamic equilibrium applies (with guaranteed isotropic pressure). The radial run of mass can then be deduced in terms of radial temperature and density gradients,

$$\frac{GM(<r)}{r} = -\frac{kT(r)}{\mu m_p}\left(\frac{d\ln T}{d\ln r} + \frac{d\ln\rho_{\rm gas}}{d\ln r}\right). \tag{146}$$

Unfortunately, the temperature run is often not very well constrained and the cluster is commonly assumed to be isothermal. This is something that will improve greatly with the new generation of X-ray satellites (Chandra and XMM); they have the spectroscopic capability to measure emission lines from the intracluster medium that will give accurate temperature profiles.

A common model used to fit cluster X-ray data is to assume that the total and gas density profiles satisfy a fitting formula with a core and a central density:

$$\begin{aligned} \rho_m &= \rho_m(0)\left[1+(r/r_c)^2\right]^{-3/2} \\ \rho_{\rm gas} &= \rho_{\rm gas}(0)\left[1+(r/r_c)^2\right]^{-3\beta/2}. \end{aligned} \tag{147}$$

This is a solution of the joint hydrostatic equations if the parameter β is a constant, satisfying

$$\beta = \frac{\mu m_p \sigma_v^2}{kT}, \tag{148}$$

where μ is the mean molecular weight in the gas. This parameter is just the 'temperature' of the dark-matter distribution divided by the temperature of the gas. The mean gas temperature can be measured from the overall shape of the X-ray spectrum, and the absolute value of the gas density follows from the X-ray bremsstrahlung luminosity. The emissivity scales as ρ^2, so the total X-ray luminosity is $L \propto \rho_{\rm gas}(0)^2 r_c^3$. Since the 'observed' luminosity scales as $L \propto h^{-2}$ and $r_c \propto h^{-1}$, the total inferred gas mass scales as $h^{-5/2}$. Note that both the gas mass and the total mass can be estimated from the X-ray data alone.

Again, Coma is a good example of the application of these methods: the gas mass is well constrained in the central $1h^{-1}$ Mpc:

$$M_{\rm gas} = 3\times 10^{13} h^{-5/2}\ M_\odot. \tag{149}$$

If we add this to the mass of stars in Coma ($\simeq 3\times 10^{13} h^{-1}\ M_\odot$), then we get a measure of the fraction of the mass in Coma that is in the form of baryons:

$$\frac{M_{\rm B}}{M_{\rm tot}} \simeq 0.01 + 0.05 h^{-3/2} \tag{150}$$

If the overall density is critical, this ratio is far above that predicted by the nucleosynthetic estimate, $\Omega_{\rm B} h^2 \simeq 0.02$. If we adopt $h = 0.7$, then a total density parameter $\Omega \simeq 0.4$ is implied. Thus, high Ω requires an abnormal concentration of baryons towards clusters, not just a greater star-formation efficiency there. This is a more robust argument in favour of low Ω, since it does not depend on assuming that the stellar populations in clusters are representative (which they are not). In other environments, some fraction of the intracluster gas might have formed stars, but so long as we count the totality of baryons this does not matter.

17.5 Cluster masses from gravitational lensing

In principle, the simplest method of measuring cluster masses is from gravitational deflection of light. The idea that massive bodies deflect light rays goes back several centuries, but was first correctly analyzed by Einstein, in his 1915 theory of general relativity. Nevertheless, we can get a good insight into the phenomenon using Newtonian ideas.

The deflection geometry for a single gravitational lens is shown in Figure 18. Source, lens and observer are shown as lying in the same plane, which is always possible even for a non-symmetric lens. The **bend angle**, α is given by the line integral of the gravitational acceleration perpendicular to the path, $a_\perp$:

$$\alpha = \frac{2}{c^2}\int a_\perp\, d\ell. \tag{151}$$

This simply says that photons are deflected twice as much as would be expected from a naive Newtonian argument.

Adding up distances at the left-hand side of Figure 18 in two different ways yields the fundamental lensing equation

$$\alpha(D_{\rm L}\theta_{\rm I}) = \frac{D_{\rm S}}{D_{\rm LS}}(\theta_{\rm I} - \theta_{\rm S}), \tag{152}$$

The different distances here are angular-diameter distances relative to different origins. For example, $D_{\rm LS}$ is the angular-diameter distance for points at the source as seen by the lens. These are simply related to the comoving radii of source and lens:

$$\begin{aligned} D_{\rm LS} &= \frac{R_0 S_k(r_{\rm S} - r_{\rm L})}{1 + z_{\rm S}} \\ D_{\rm L} &= \frac{R_0 S_k(r_{\rm L})}{1 + z_{\rm L}} \\ D_{\rm S} &= \frac{R_0 S_k(r_{\rm S})}{1 + z_{\rm S}}, \end{aligned} \tag{153}$$

so we know how to calculate these for any given cosmology.

In principle, the lens equation can be solved to find the mapping between the **object plane** and the **lens plane** or **image plane** – *i.e.* $\theta_{\rm I}(\theta_{\rm S})$ – and hence positions of the images. If the lensing deflection is small, this mapping is just a one-to-one and invertible distortion of the coordinates: this is **weak lensing**. For larger deflections, however, a unique inverse mapping from source to lens plane may not exist: this is the regime of **strong lensing**, associated with multiple imaging.

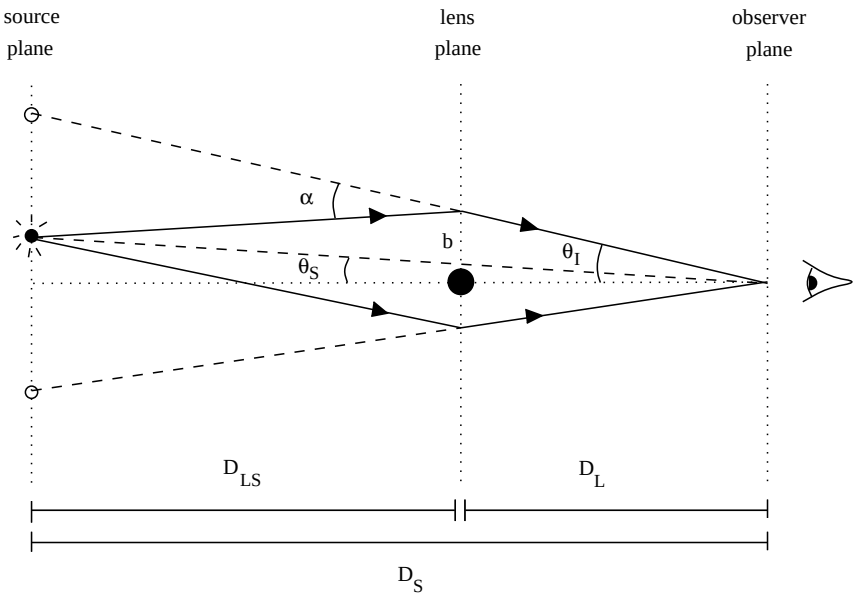

Figure 18. The geometry of a gravitational lensing event. For a thin lens, deflection through the small bend angle α may be taken as instantaneous. The angles $\theta_{\rm I}$ and $\theta_{\rm S}$ (two-dimensional vectors in general) specify respectively the observed and intrinsic positions of the source on the sky.

The only other thing we need in order to find the appearance of the images is the fact that gravitational lensing does not alter surface brightness. Formally, this comes about from the relativistic invariant I_ν/ν^3. Hence, the **amplification** of image flux densities is given simply by a ratio of image areas.

The lenses that have received the most attention are those where the mass distribution appears circularly symmetric on the sky. For these, there is the nice result (which we won't prove here) that

$$\alpha = \frac{4G}{c^2}\,\frac{M(<b)}{b}, \tag{154}$$

where $b = D_{\rm L}\theta_{\rm I}$ is the distance of closest approach (*not* the impact parameter) and $M(<b)$ is the mass seen *in projection* within a radius b.

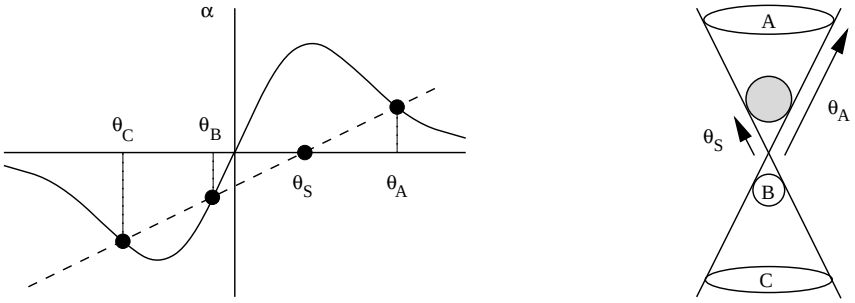

Figure 19. The graphical solution (left) and image geometry (right) for a spherical gravitational lens. The shaded circle on the right shows the image position in the absence of the lens. Images A, B and C are produced where there is an intersection between the bend-angle curve $\alpha(\theta)$ and a straight line that crosses the θ axis at the source radius $\theta_{\rm S}$. Because the acceleration is radially inwards, different portions of the images are displaced along different radial lines, often leading to crescent-shaped images.

The solution of the lensing equation for a symmetrical lens may be visualized graphically as the intersection of a straight line with the bend-angle curve. This is illustrated in figure 19, which also shows the two-dimensional image structure; different portions of the object are displaced along radial lines, leading to crescent-shaped images. For cases of close alignment (*i.e.* $\theta_{\rm S} \to 0$) the principal outer pair of images has an angular separation very close to the diameter of the **Einstein ring** formed by perfect alignment. The radius of this ring is the most important characteristic property of the lens, and is known as the **Einstein radius**:

$$\begin{aligned}\theta_{\rm E} &= \left(\frac{4GM}{c^2}\,\frac{D_{\rm LS}}{D_{\rm L}D_{\rm S}}\right)^{1/2} \\ &= \left(\frac{M}{10^{11.09}M_\odot}\right)^{1/2}\,\left(\frac{D_{\rm L}D_{\rm S}/D_{\rm LS}}{\rm Gpc}\right)^{-1/2} \quad \text{arcsec.}\end{aligned} \tag{155}$$

The Einstein ring radius is therefore a robust way of of measuring the mass internal to the ring. The best examples of Einstein rings have been found where the lens is a cluster core, producing spectacular **luminous arcs** several tens of arcsec in length. Such events are confined

to atypically compact clusters, since most clusters have central surface densities that are below critical. The cluster arcs are of especial interest, as the source is usually a low-luminosity galaxy of high redshift, and the lensing event also provides a 'gravitational telescope' to aid the study of objects that would otherwise be barely detectable.

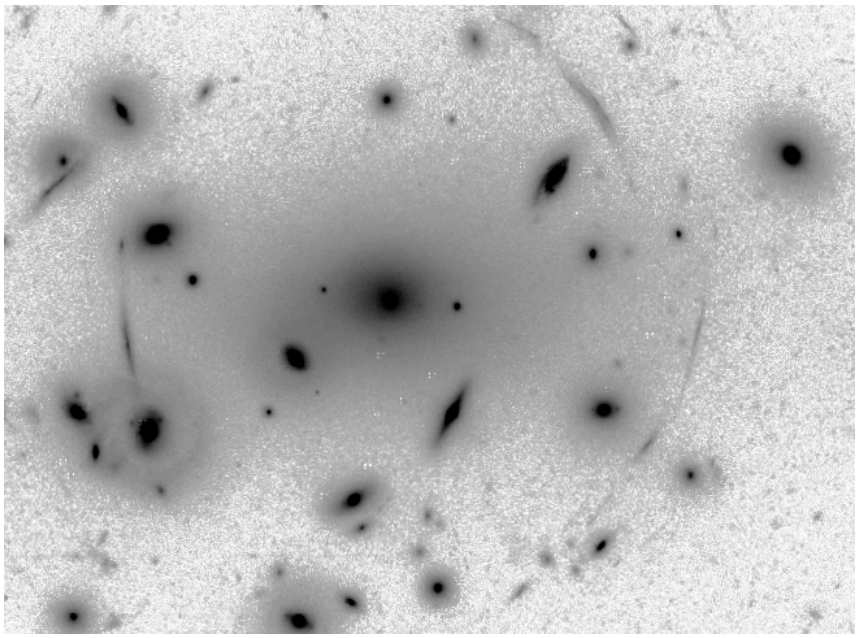

Figure 20. An image of the central 1.3×1.0 arcmin of the rich galaxy cluster A2218, taken with the Hubble Space Telescope. Note the spectacular arcs resulting from strong lensing of background galaxies.

How well do lensing-based mass estimates agree with other methods? A number of attempts have been made to carry out this comparison, with slightly inconsistent results. There have been claims that lensing results give masses that are larger than those from X-rays, by about a factor of 2. A plausible explanation for this is that the modelling of the X-ray gas may need to be more sophisticated: in a cluster with cooling flows, there will be a **multiphase IGM**: a mix of temperatures and densities at a given radius. With good spectral data, this can be allowed for, and consistent lensing and X-ray masses are found.

17.6 Summary

A variety of lines of evidence thus argue for a density of dark matter that is in the region of $\Omega_m \simeq 0.3$. The masses of clusters of galaxies seem to be robustly measured by a variety of methods. Applying their M/L ratios to galaxies in general is dangerous, because the stellar populations in clusters are atypical (mainly elliptical galaxies). The ratio to the baryon density is safer: because clusters are the largest collapsed systems we know of, it is hard to see how the mean ratio of baryons and total mass in these could differ greatly from the mean. Provided we trust the baryon density estimated from nucleosynthesis, the total mass density follows. If we compare this figure of $\Omega_m \simeq 0.3$ with the supernova data, we see that it indicates a model somewhere close to the $k = 0$ line, so that vacuum energy would have to dominate over dark matter by about 2:1.

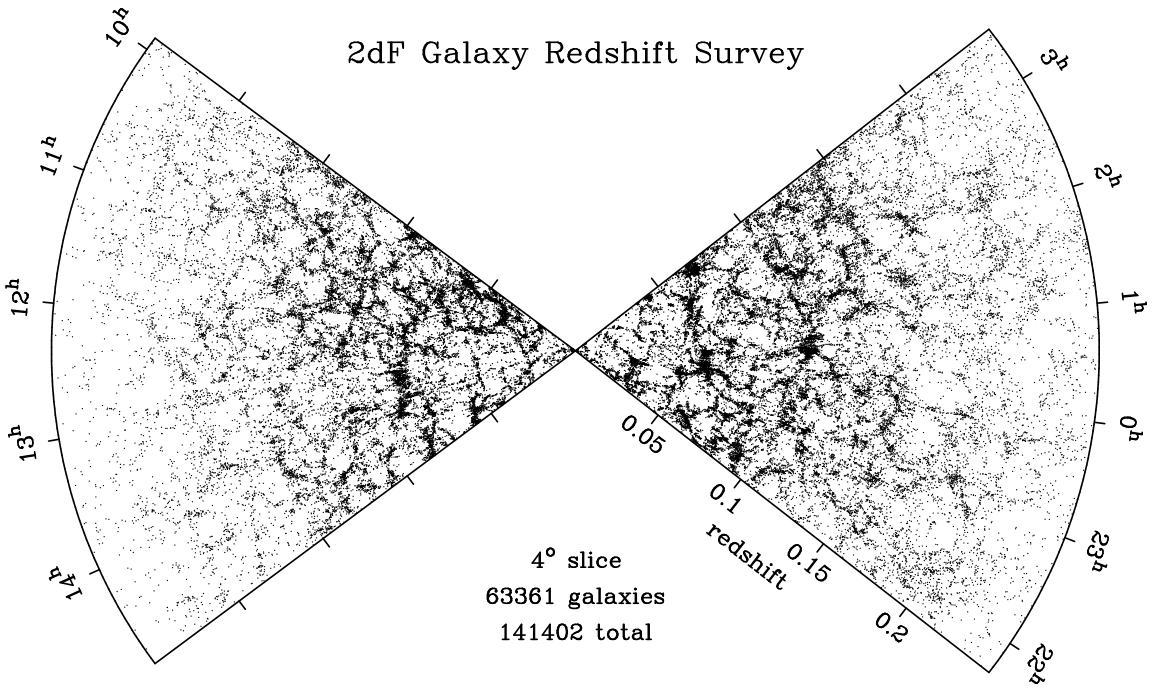

Figure 21. The spatial distribution of galaxies in two 4-degree strips on the sky, according to the 2dF Galaxy Redshift Survey. Note the 100-Mpc filamentary features and the prominent voids. One of the principal challenges in cosmology is to explain this pattern, which is most probably a relic of the very earliest stages of the expanding universe.

18 Large-scale structure

We now turn from questions of the global contents and properties of the universe to perhaps the major concern of modern cosmology: the initial conditions for the expanding universe. One of the main clues we possess about the processes that operated at this time comes from the inhomogeneity of the universe. Obviously, the universe does not locally conform to the ideal of the RW metric, but of greater significance is the fact that perturbations from uniformity exist on very large scales.

The most dramatic evidence for this come from **redshift surveys**, where a quasi-3D picture of the universe is built up by assuming that redshifts give exact radial distances via Hubble's law: $v = cz = H_0 d$, so $d = cz/H_0$ at small z. (the microwave dipole tells us that this distance will not be exact). The state of the art is shown in figure 21.

18.1 Statistics of density fluctuations

To quantify the patterns we see in large-scale structure, we use the dimensionless **density perturbation field**

$$\delta(\mathbf{x}) \equiv \frac{\rho(\mathbf{x}) - \langle\rho\rangle}{\langle\rho\rangle}. \tag{156}$$

This can be defined for any quantity, in particular the mass density, but we start by considering the density of galaxies.

The key issue of interest will be to dissect the pattern in figure 21as a function of scale, in order to see at what scale the universe starts to approach the RW ideal. The natural tool here is Fourier analysis, where the density fluctuation field is a sum over wave modes:

$$\delta(\mathbf{x}) = \sum \delta_{\mathbf{k}} e^{-i\mathbf{k}\cdot\mathbf{x}}. \tag{157}$$

Suppose we consider the distribution only in some periodic box of side L: the requirement of periodicity restricts the allowed wavenumbers to **harmonic boundary conditions**

$$k_x = n\,\frac{2\pi}{L}, \qquad n = 1, 2\cdots, \tag{158}$$

with similar expressions for k_y and k_z. Now, if we let the box become arbitrarily large, then the sum will go over to an integral that incorporates the density of states in k-space, exactly as in statistical mechanics. The Fourier expansion of density in 3D is thus

$$\delta(x) = \left(\frac{L}{2\pi}\right)^3 \int \delta_k(k)\exp\bigl(-i\mathbf{k}\cdot\mathbf{x}\bigr)\; d^3k. \tag{159}$$

Obviously, no theory is going to tell us whether the density is above or below the mean at a given point: the universe contains some random variations in density, and it is only sensible to try to consider their expectation properties. By definition of the mean density, $\langle\delta\rangle = 0$, but what about the variance, $\langle\delta^2\rangle$? If we average the Fourier expansion over the box,

$$\langle\delta^2\rangle = \frac{1}{V}\int \left|\sum \delta_{\mathbf{k}} e^{-i\mathbf{k}\cdot\mathbf{x}}\right|^2 \; dV, \tag{160}$$

all cross-terms in the expansion between different k values average to zero, so that

$$\langle\delta^2\rangle = \sum |\delta_{\mathbf{k}}|^2, \tag{161}$$

which justifies the term **power spectrum** for $|\delta_{\mathbf{k}}|^2$.

A similar analysis can be applied to the autocorrelation function of the density field – usually referred to simply as the **correlation function**:

$$\xi(\mathbf{r}) \equiv \langle\delta(\mathbf{x})\delta(\mathbf{x}+\mathbf{r})\rangle\,, \tag{162}$$

Applying the same reasoning shows that the correlation function is the Fourier transform of the power spectrum:

$$\xi(r) = \sum |\delta_{\mathbf{k}}|^2\, e^{-i\mathbf{k}\cdot\mathbf{r}}, \tag{163}$$

so these are alternative ways of characterizing the density fluctuations. An alternative definition of the autocorrelation function is as the **two-point correlation function**, which gives the excess probability for finding a neighbour a distance r from a given galaxy. By regarding this as the probability of finding a pair with one object in each of the volume elements dV_1 and dV_2,

$$dP = \rho_0^2\,[1 + \xi(r)]\,dV_1\,dV_2, \tag{164}$$

this is easily seen to be equivalent to the autocorrelation definition of ξ: the density fluctuation in each cell is $1+\delta$, so $\xi = \langle\delta(x_1)\delta(x_2)\rangle$.

18.2 Power-law spectra

The above shows that the power spectrum is a central quantity in cosmology, but how can we predict its functional form? For decades, this was thought to be impossible, and so a minimal set of assumptions was investigated. Consider the featureless power law:

$$\left\langle |\delta_k|^2 \right\rangle \propto k^n \tag{165}$$

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

$$\xi(r) = (r/r_0)^{-\gamma}; \quad \gamma = n + 3. \tag{166}$$

What general constraints can we set on the value of n? Asymptotic homogeneity clearly requires $n > -3$. The value $n = 0$ corresponds to **white noise**, the same power at all wavelengths. Most important of all is the **scale-invariant spectrum**, which corresponds to the value $n = 1$, *i.e.* $\Delta^2 \propto k^4$. To see how the name arises, consider a perturbation $\delta\Phi$ in the gravitational potential:

$$\nabla^2 \delta\Phi = 4\pi G \rho_0 \delta \quad \Rightarrow \quad \delta\Phi_k = -4\pi G \rho_0 \delta_k / k^2. \tag{167}$$

The two powers of k pulled down by ∇^2 mean that, if $n = 1$ for the power spectrum of density fluctuations, then $k^3 |\Phi_k|^2$ is a constant. The significance of this combination is that it is proportional to the contribution to the variance of the potential from waves in a given rance of $\ln k$. Since potential perturbations govern the flatness of spacetime, this says that the scale-invariant spectrum corresponds to a metric that is a **fractal**: spacetime has the same degree of 'wrinkliness' on each resolution scale. This spectrum is often known as the **Zeldovich spectrum**.

Practical spectra in cosmology actually turn out to have negative effective values of n over a large range of wavenumber. For many years, the data on the galaxy correlation function were consistent with a single power law:

$$\xi_g(r) \simeq \left(\frac{r}{5\, h^{-1}\,\mathrm{Mpc}} \right)^{-1.8} \qquad \left(1 \lesssim \xi \lesssim 10^4\right); \tag{168}$$

This corresponds to $n \simeq -1.2$. We now want to see how this can be understood.

19 Dynamics of structure formation

19.1 Spherical model and linear growth

The favoured mechanism for forming cosmological structure is gravitational instability. An easy way to analyze this is via the **spherical model**: consider an overdense sphere, which behaves in exactly the same way as a closed sub-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

$$\begin{aligned} r &= A(1 - \cos\theta) \\ t &= B(\theta - \sin\theta), \end{aligned} \tag{169}$$

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

$$r \simeq \frac{A}{2}\left(\frac{6t}{B}\right)^{2/3}\left[1 - \frac{1}{20}\left(\frac{6t}{B}\right)^{2/3}\right], \tag{170}$$

and we can identify the density perturbation within the sphere:

$$\delta \simeq \frac{3}{20}\left(\frac{6t}{B}\right)^{2/3}. \tag{171}$$

At early times, the density perturbations thus grow proportional to a. Although we have derived it via the spherical model, this is a general result that is valid for any linear perturbation with $\delta \ll 1$.

A simple way of summarizing this growth is that it keeps potential perturbations unchanged. If $\delta_k \propto a$ and we consider a perturbation of fixed *comoving* wavenumber,

$$\delta(a) = a\delta_k(z=0)\exp(i\mathrm{a}^{-1}\mathbf{k}_{\rm comoving}\cdot\mathbf{x}_{\rm proper}) \tag{172}$$

then

$$\nabla^2\Phi = -k^2a^{-2}\Phi = 4\pi G\rho(a)\delta(a). \tag{173}$$

Since $\rho \propto a^{-3}$, and $\delta \propto a$, we see that Φ is independent of a, as required. This makes intuitive sense: the fractional 'wrinkliness' of spacetime doesn't change with time. This result is also true in the radiation-dominated era, which requires $\delta \propto a^2$ in that case.

19.2 Formation of nonlinear structures

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 will never occur in practice; dissipative physics will eventually intervene and convert the kinetic energy of collapse into random motions. How dense will the resulting body be? Consider the time at which the sphere has collapsed by a factor 2 from maximum expansion. At this point, it has kinetic energy K related to potential energy V by $V = -2K$. This is the condition for equilibrium, according to the **virial theorem**. For this reason, many workers take this epoch as indicating the sort of density contrast to be expected as the endpoint of gravitational collapse. This occurs at $\theta = 3\pi/2$, and the corresponding density enhancement is $(9\pi + 6)^2/8 \simeq 147$, with $\delta_{\rm lin} \simeq 1.58$. Some authors prefer to assume that this virialized size is eventually achieved only at collapse, in which case the contrast becomes $(6\pi)^2/2 \simeq 178$ and $\delta_{\rm lin} \simeq 1.69$.

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 at a density $\simeq 200$ times the mean.

20 Dark matter and growth of structure

20.1 Types of non-baryonic dark matter

The above results show that the spectrum of any perturbations generated at early times evolves in such a way as to preserve its shape. However, this is not true on very small scales, in a way that depends on the matter content of the universe. We have already seen the evidence for dark matter beyond the contribution of baryons, and the most commonly considered explanation for this is in terms of exotic particles that are frozen-out **relics** from the early universe. A common collective term for these particles is **WIMP** – standing for weakly interacting massive particle. We have already seen one example of this in the massive neutrino, but there are really three generic types to consider, as follows.

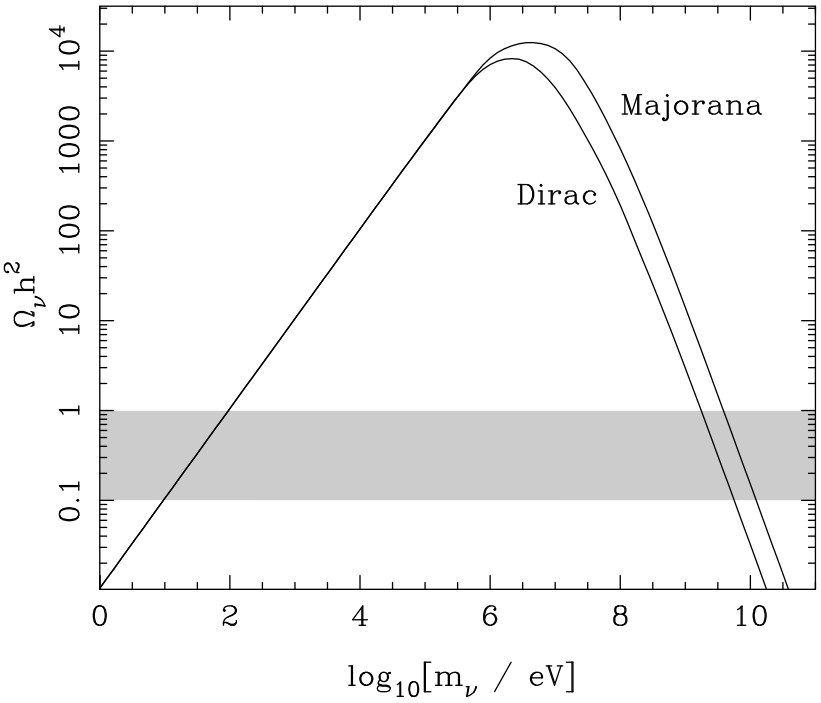

Figure 22. The contribution to the density parameter produced by relic neutrinos (or neutrino-like particles) as a function of their rest mass. At low masses, the neutrinos are highly relativistic when they decouple: their abundance takes the zero-mass value, and the density is just proportional to the mass. Above about 1 Mev, the neutrinos are non-relativistic at decoupling, and their relic density is reduced by annihilation.

(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.

(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, so $n \sim \exp(-M/\mathrm{MeV})$. The relic mass density is therefore $\propto M \exp(-M/\mathrm{MeV})$, so it falls with increasing mass (see figure 22). Interesting masses then lie in the $\gtrsim 10$ GeV range. This cannot correspond to the known neutrinos, but plausible candidates 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 (mainly in deep mines, to shield from cosmic rays). The chances of success are hard to estimate, but it would be a tremendous scientific achievement if dark matter particles were to be detected in this way.

20.2 Transfer functions for density fluctuations

In the meantime, we try to pin down the dark-matter particle via cosmology. One key property is that the dark matter influences the spectrum of density fluctuations in the universe. The dominant effect is common to all kinds of dark matter. It does not let us discriminate between the different types, but it does allow another means of weighing the universe. This is the **Mészáros effect**. The effect arises because the universe is radiation dominated at early times. This means that radiation density produces almost all the gravitational force: fluctuations in the dark matter can only grow if dark matter and radiation fall together. This does not happen for perturbations of small wavelength, because the photons can move out of the dark-matter potential wells at the speed of light. These fluctuations therefore do not grow. Growth only occurs for perturbations of very large wavelength, where there has been no time for the matter and radiation to separate. The length-scale that separates these two regimes will be the horizon distance.

This process continues, until the universe becomes matter dominated at $z_{\mathrm{eq}} = 23\,900\,\Omega h^2$. We therefore expect a characteristic 'break' in the fluctuation spectrum around the comoving horizon length at this time. If we use the comoving distance–redshift relation for matter plus radiation (neglecting vacuum energy at early times),

$$R_0 dr = \frac{c}{H_0} \frac{dz}{(1+z)\sqrt{1+\Omega_m z + (1+z)^2 \Omega_r}}, \tag{174}$$

the *comoving* horizon size at z_{eq} is easily derived:

$$D_{\mathrm{H}}(z_{\mathrm{eq}}) \equiv R_0 r_{\mathrm{H}}(z_{\mathrm{eq}}) = (\sqrt{2}-1)\frac{2c}{H_0}(\Omega_m z_{\mathrm{eq}})^{-1/2} = 16\,(\Omega_m h^2)^{-1}\mathrm{Mpc}. \tag{175}$$

Since distances in cosmology always scale as h^{-1}, this means that $\Omega_m h$ should be observable.

The result of the Mészáros effect is that modes of short wavelength have their amplitudes reduced relative to those of long wavelength. This effect is quantified via the **transfer function**:

$$T_k \equiv \frac{\delta_k(z=0)}{\delta_k(z)\ D(z)}, \tag{176}$$

where $D(z)$ is the linear growth factor between redshift z and the present. The normalization redshift is arbitrary, so long as it refers to a time before any scale of interest has entered the horizon. A plot of the transfer function for CDM and other models is shown in figure 23. The relative diminution in fluctuations at high k is the amount of growth missed out on between horizon entry and $z_{\rm eq}$. Although we will not prove it, this change is easily shown to be $\propto k^2$. The approximate limits of the CDM transfer function are therefore

$$T_k \simeq \begin{cases} 1 & (kD_{\rm H}(z_{\rm eq}) \ll 1) \\ [kD_{\rm H}(z_{\rm eq})]^{-2} & kD_{\rm H}(z_{\rm eq}) \gg 1). \end{cases} \tag{177}$$

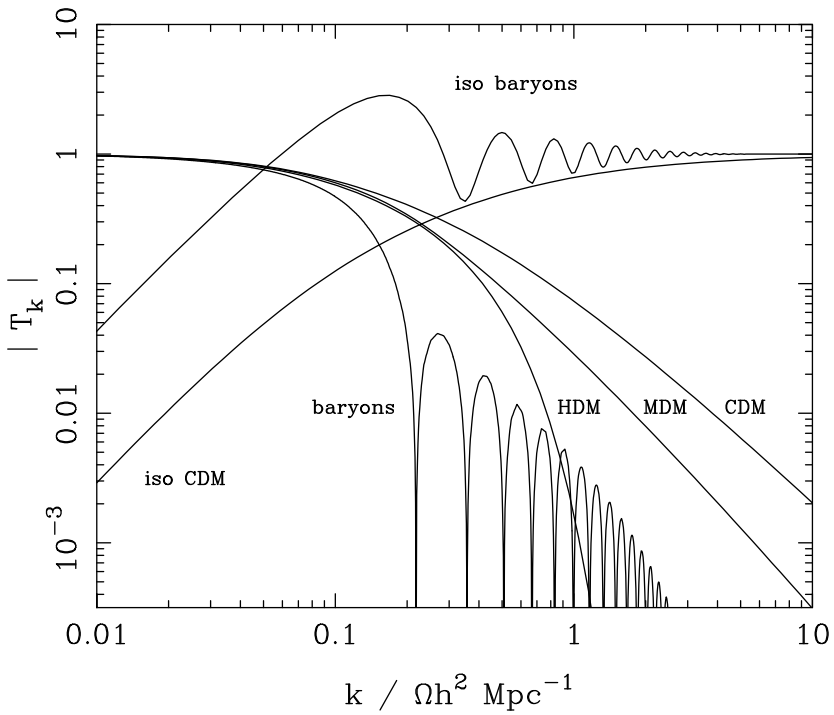

Figure 23. A plot of transfer functions for various models. For adiabatic models, $T_k \to 1$ at small k, whereas the opposite is true for isocurvature models. A number of possible matter contents are illustrated: pure baryons; pure CDM; pure HDM; MDM (30% HDM, 70% CDM). For dark-matter models, the characteristic wavenumber scales proportional to Ωh^2. The scaling for baryonic models does not obey this exactly; the plotted cases correspond to $\Omega = 1$, $h = 0.5$.

Figure 23 also shows a set of models labled **isocurvature**, which show a very different behaviour. This is because we have assumed that the character of the initial fluctuations was **adiabatic**: *i.e.* photon densities and matter densities were compressed equally. When discussing inflationary mechanisms for fluctuation generation later, we will see that this is the most natural type of perturbation. Nevertheless, it is possible to imagine more contrived types of initial conditions, in which the radiation is left unperturbed, and only the dark matter fluctuates (also known as **entropy perturbations**, since the ratio of photons to baryons fluctuates). These have the opposite behaviour: small-scale perturbations are preserved (they never grow), but large-scale perturbations diminish with time. These modes match the CMB anisotropy data poorly, so we will neglect them.

20.3 Damping scales

This relatively gentle filtering away of the initial fluctuations means that a CDM universe contains fluctuations in the dark matter on all scales. Structure formation in a CDM universe is then a **hierarchical** process in which nonlinear structures grow via the merger of very small initial units.

This is very different in the case of other kinds of dark matter, for which small-scale fluctuations are utterly damped away, creating a **coherence length** in the dark-matter distribution. For collisionless dark matter, perturbations can be erased simply by **free streaming**: random particle velocities cause blobs to disperse. At early times ($kT > mc^2$), the particles will travel at c, and so any perturbation that has entered the horizon will be damped. This process switches off when the particles become non-relativistic:

$$\text{proper} \quad L_{\rm damp} \simeq ct(kT = Mc^2). \tag{178}$$

Massive neutrinos would be just becoming non-relativistic at matter-radiation equality (since they have the same number density as the photons). The damping scale for HDM is therefore of order the horizon size then:

$$\text{comoving} \quad L_{\rm damp} \simeq 16(\Omega h^2)^{-1}\,\text{Mpc}. \tag{179}$$

Only perturbations on very large scales survive in the case of HDM, so it is difficult to make galaxies at high redshifts. Massive CDM particles go nonrelativistic long before $z_{\rm eq}$, so the damping is negligible for them. The existence of galaxies at $z \simeq 6$ tells us that the coherence scale must have been below about 100 kpc.

Similar processes operate in a purely baryonic universe. The corresponding process is called **Silk damping**: the mean free path of photons due to scattering by the plasma is non-zero, and so radiation can diffuse out of a perturbation, convecting the plasma with it.

20.4 Spectrum normalization

Assembling the ingredients so far, we have a model where the power spectrum of the density field is

$$|\delta_k|^2 \propto k^n T_k^2. \tag{180}$$

This contains three free parameters: Ωh (from T_k); n, and the normalization. The way to deal with this is to think about the variance in the density fluctuations:

$$\sigma^2 \propto \int |\delta_k|^2\, d^3k \propto \int |\delta_k|^2\, 4\pi k^2 dk \propto \int k^{3+n} T_k^2\, d\ln k. \tag{181}$$

In other words, $k^3|\delta_k|^2$ gives the power per log scale. Now suppose we smooth the density field by convolution with some filter, *e.g.* a sphere of radius R. This cuts off the integral at $k \sim 1/R$, so $\sigma^2(R)$ then measures roughly $k^3|\delta_k|^2$ at $k \sim 1/R$, giving the normalization. A common choice is $R = 8\,h^{-1}$ Mpc, since this contains the mass of a rich cluster of galaxies. These are the largest collapsed systems, so we know that $\sigma(R) \sim 1$ on these scales. Another way of fixing the normalization comes from the CMB, as we will see later. Given the normalization at two points, plus one of n or Ωh, we can fix the other. The evidence so far implies $\Omega h \simeq 0.2$, and this turns out to fit well with $n \simeq 1$.

21 Inflationary cosmology

We now return to some of the major problems with the whole framework of the hot big bang:

(1) The expansion problem. Why is the universe expanding at $t = 0$? This appears as an initial condition, but surely a mechanism is required to launch the expansion?

(2) The flatness problem. Furthermore, the expansion needs to be launched at just the correct rate, so that is is very close to the critical density, and can thus expand from perhaps near the Planck era to the present (a factor of over 10^{30}).

(3) The horizon problem. Models in which the universe is radiation dominated (with $a \propto t^{1/2}$ at early times) have a finite horizon. There is apparently no causal means for different parts of the universe to agree on the mean density or rate of expansion.

The list of problems with conventional cosmology provides a strong hint that the equation of state of the universe may have been very different at very early times. Consider the integral for the horizon length:

$$r_{\rm H} = \int \frac{c\,dt}{R(t)}. \tag{182}$$

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

$$\ddot{R} = -4\pi G R(\rho + 3p/c^2)/3, \tag{183}$$

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

$$\rho c^2 + 3p < 0. \tag{184}$$

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**:

$$R \propto \exp Ht, \tag{185}$$

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

$$\dot{R}^2 = \frac{8\pi G\rho R^2}{3} - kc^2. \tag{186}$$

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).

21.1 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 the Planck energy (10^{19} GeV) divided by the CMB energy ($kT \simeq 10^{-3.6}$ eV), or

$$z_{\rm P} \simeq 10^{31.6}. \tag{187}$$

This expands the Planck length to 0.4 mm today. This is far short of the present horizon ($6000\,h^{-1}$ 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:

$$\Delta t_{\rm inflation} > 60\, H^{-1}_{\rm inflation}. \tag{188}$$

The same criterion also solves the flatness problem. If $8\pi G\rho R^2/3 \sim kc^2$ today, then it was smaller than the curvature by roughly a factor $(1+z)^2$ at redshift z (assuming $\rho \propto (1+z)^{-4}$ for radiation, which is valid during most of the expansion history). At the GUT era, $z \sim 10^{28}$, so we need to make $8\pi G\rho R^2/3$ smaller than $-kc^2$ by a factor 10^{56}. A natural assumption about the initial conditions is instead that the density and curvature terms are of the same order. Exponential inflation for a time Δt will increase ρR^2 by $\exp(2H\Delta t)$, so again roughly 60 e-foldings of the expansion will do the job.

21.2 Inflationary dynamics

So far so good, but having used vacuum repulsion to launch the universe, we need to switch it off somehow, before the exponential expansion reduces the matter density effectively to zero. This requires some *dynamical* form of vacuum energy that is more sophisticated than a simple cosmological constant.

The models that achieve this come from particle physics. There are many variants, but the simplest concentrate on **scalar fields**. These are fields like the electromagnetic field, but differing in a number of respects. First, the field has only one degree of freedom: just a number that varies with position, not a vector like the EM field. The wave equation obeyed by such a field in flat space could be of a familiar form:

$$\frac{1}{c^2}\ddot{\phi} - \nabla^2\phi = 0, \tag{189}$$

but this applies only for the special case where the quanta associated with the ϕ field are massless. If they have mass, the equation becomes the **Klein–Gordon equation**:

$$\frac{1}{c^2}\ddot{\phi} - \nabla^2\phi + (m^2c^2/\hbar^2)\phi = 0, \tag{190}$$

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

$$\ddot{\phi} = -\frac{d}{d\phi}V(\phi); \quad V(\phi) = (m^2c^4/\hbar^2)\phi^2. \tag{191}$$

This is just a harmonic oscillator equation, and we can see that the field will oscillate in the potential. In classical dynamics of a ball in a potential, this motion will conserve energy: $\dot{\phi}^2/2+V(\phi) =$ constant. The energy transforms itself from all potential at the top of the motion, to all kinetic at the bottom. This behaviour is rather different to the familiar oscillations of the electromagnetic field: if the field is homogeneous, it does not oscillate. This is because the familiar energy density in electromagnetism ($\epsilon_0 E^2/2 + B^2/2\mu_0$) is entirely kinetic energy in this analogy (to see this, write the fields in terms of the potentials: $\mathbf{B} = \boldsymbol{\nabla} \wedge \mathbf{A}$ and $\mathbf{E} = -\boldsymbol{\nabla}\phi - \dot{\mathbf{A}}$. We don't see coherent oscillations in electromagnetism because the photon has no mass.

The ability of scalar-field oscillations to have a state of pure potential energy is what makes inflation possible. When the energy is potential-only, this is like adding to the vacuum energy. The gravitational effects of large $V(\phi)$ are therefore repulsive and can start exponential expansion.

In this simple model, the universe is started in a potential-dominated state, and inflates until the field falls enough that the kinetic energy becomes important. In practical models, this stage will be associated with **reheating**: although weakly interacting, the field does couple to other particles, and its oscillations can generate other particles – thus transforming the scalar-field energy into energy of a normal radiation-dominated universe. It may seem contrived to appeal to a potential-dominated state. However, if the scalar field fluctuates, there will always be some regions of large V, and these are the ones that inflate and 'invade' the rest of the universe. Only such uniform regions are suitable for life, so it is not unreasonable that we end up living in one.

22 Relic fluctuations from inflation

The idea of launching a flat and causally connected expanding universe, using only vacuum-energy antigravity, is attractive. What makes the package of inflationary ideas especially compelling is that there it is an inevitable outcome of this process that the post-inflation universe will be inhomogeneous to some extent.

The key idea is to appreciate that the inflation field cannot be a classical object, but must display quantum fluctuations. Well inside the horizon of de Sitter space, these must be calculable by normal flat-space quantum field theory. If we can calculate how these fluctuations evolve as the universe expands, we have a mechanism for seeding inhomogeneities in the expanding universe – which can then grow under gravity to make structure.

We will have to skip the details here, but assume that we can calculate the fluctuation in the field ϕ from the uncertainty principle (the actual answer is $\delta\phi = \hbar H/2\pi$ – so the uncertainty in ϕ goes up as the size of the de Sitter universe, which is c/H, goes down). Tha main effect of these fluctuations is to make different parts of the universe have fields that are perturbed by an amount $\delta\phi$. In other words, we are dealing with various copies of the same rolling behaviour $\phi(t)$, but viewed at different times

$$\delta t = \frac{\delta\phi}{\dot{\phi}}. \tag{192}$$

These universes will then finish inflation at different times, leading to a spread in energy densities (figure 24). The resulting density perturbation is determined by the different amounts that the universes have expanded following the end of inflation:

$$\frac{\delta\rho}{\rho} \sim \frac{\delta R}{R} = H\ \delta t = \frac{\hbar H^2}{2\pi\,\dot{\phi}}. \tag{193}$$

The dependence on $\dot{\phi}$ is especially interesting: the final density fluctuations depend on the dynamics of the field – *i.e.* on the form of $V(\phi)$. This can easily be more complicated than the simple $V \propto m^2\phi^2$ that we used to motivate the whole idea. There is therefore no model-independent prediction for the size of the density fluctuations.

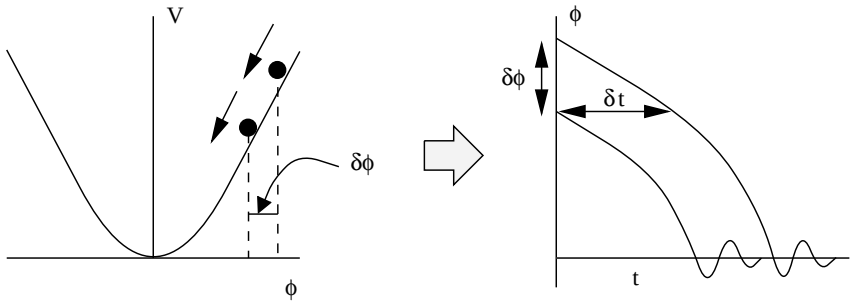

Figure 24. This plot shows how fluctuations in the scalar field transform themselves into density fluctuations at the end of inflation. Different points of the universe inflate from points on the potential perturbed by a fluctuation $\delta\phi$, like two balls rolling from different starting points. Inflation finishes at times separated by δt in time for these two points, inducing a density fluctuation $\delta = H\delta t$.

We can however predict the form of the fluctuations. Because the de Sitter expansion is invariant under time translation, the inflationary process must produce a universe that is fractal-like in the sense that scale-invariant fluctuations correspond to a metric that has the same 'wrinkliness' per log length-scale. The natural prediction of inflation is therefore an $n = 1$ fluctuation spectrum.

23 Anisotropies in the CMB

23.1 Mechanisms for primary fluctuations

How do we test this picture of fluctuation generation and growth? At the last-scattering redshift ($z \simeq 1000$), gravitational instability theory says that density perturbations must have existed in order for galaxies and clusters to have formed by the present. A long-standing challenge in cosmology has been to detect the corresponding fluctuations in brightness temperature of the cosmic microwave background (CMB) radiation, and the study of CMB fluctuations has now blossomed into a critical tool for pinning down cosmological models.

We distinguish **primary anisotropies** (those that arise due to effects at the time of recombination) from **secondary anisotropies**, which are generated by scattering along the line of sight. There are three basic primary effects, illustrated in figure 25, which are important on respectively large, intermediate and small angular scales:

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

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

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

To make quantitative progress, the next step is to see how to predict the size of these effects in terms of the spectrum of mass fluctuations.

23.2 Angular dependence of fluctuations

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

$$\frac{\delta T}{T} \sim \frac{\delta\Phi}{c^2} \quad \text{(gravity)}; \qquad \frac{\delta T}{T} \sim \frac{1}{3}\frac{\delta\rho}{\rho} \quad \text{(adiabatic)}. \tag{194}$$

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

$$k_{\rm crit} \sim (ct)^{-1}. \tag{195}$$

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

We have already seen how to calculate the horizon size. While the universe is matter dominated, the distance-redshift relation is

$$R_0\,dr = \frac{c}{H_0}\frac{dz}{(1+z)\sqrt{1+\Omega_m z}}, \tag{196}$$

so the *comoving* horizon size is

$$D_{\rm H}(z) \equiv R_0\int_z^\infty dr = \frac{2c}{H_0}\left[(1+z)\,\Omega_m\right]^{-1/2}, \tag{197}$$

which is $181\,\Omega_m^{-1/2}\,h^{-1}$ Mpc at last scattering (see the problem sheets). The relation between angle and comoving distance on the last-scattering sphere requires the comoving angular-diameter distance to the last-scattering sphere; because of its high redshift, this is effectively identical to the horizon size at the present epoch, $D_{\rm H}$:

$$\begin{aligned} D_{\rm H} &= \frac{2c}{\Omega_m H_0} \quad \text{(open)} \\ D_{\rm H} &\simeq \frac{2c}{\Omega_m^{0.4} H_0} \quad \text{(flat)}; \end{aligned} \tag{198}$$

the latter equation is a good approximation for models with $\Omega_m + \Omega_v = 1$. The change-over from scale-invariant Sachs–Wolfe fluctuations to fluctuations dominated by adiabatic perturbations thus occurs at a critical angle $\theta = D_{\rm LS}/D_{\rm H}$; for a matter-only model it takes the value

$$\theta = 1.8\,\Omega_m^{1/2}\ \text{degrees}. \tag{199}$$

For flat low-density models with significant vacuum density, conversely, θ is roughly independent of Ω.

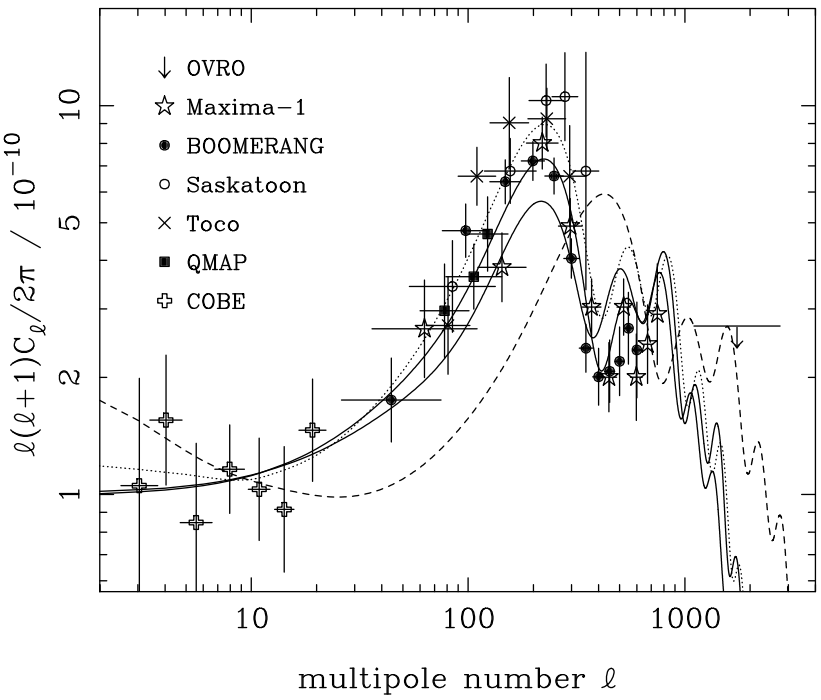

Figure 25. Angular power spectra $\mathcal{T}^2(\ell) = \ell(\ell+1)C_\ell/2\pi$ for the CMB, plotted against angular wavenumber ℓ in radians^{-1}. Various model predictions for adiabatic scale-invariant CDM fluctuations are shown. The 'acoustic oscillations' at high ℓ arise because the matter-radiation fluid oscillates as sound waves in the dark-matter potential wells. The two solid lines correspond to $(\Omega, \Omega_{\rm B}, h) = (1, 0.05, 0.5)$ and (1,0.1,0.5), with the higher $\Omega_{\rm B}$ increasing power by about 20% at the peak. The dotted line shows a flat Λ-dominated model with $(\Omega, \Omega_{\rm B}, h) = (0.3, 0.05, 0.65)$; the dashed line shows an open model with the same parameters. The main effects are that open models shift the peak to the right, and that the height of the peak increases with $\Omega_{\rm B}$ and h.

23.3 The CMB power spectrum

The important conclusion of the previous section was that there should be a feature in the pattern of CMB brightness fluctuations on an angular scale of about 1 degree, but moving to smaller scales if the universe is open and low-density. This gives us a nice way of measuring spatial curvature in cosmology.

To test for this, we want to look at the 2D power spectrum of the CMB brightness fluctuations. The temperature field can be decomposed into modes (actually spherical harmonics, because the sky is not flat), with angular wavenumber ℓ. As with the density field, it is convenient to define a dimensionless power spectrum of fractional temperature fluctuations, so that $\mathcal{T}^2$ is the fractional variance in temperature from modes in unit range of $\ln \ell$.

Our prediction is that $\mathcal{T}^2(\ell)$ will be a constant at small ℓ (scale-invariant fluctuations). We can also predict the amplitude empirically. What is required is the typical depth of large-scale potential wells in the universe, and many lines of argument point inevitably to numbers of

order $\delta\Phi/c^2 \sim 10^{-5}$. This is clear from the existence of massive clusters of galaxies with velocity dispersions of up to $1000\,\mathrm{km\,s^{-1}}$:

$$v^2 \sim \frac{GM}{r} \quad \Rightarrow \quad \frac{\Phi}{c^2} \sim \frac{v^2}{c^2}, \tag{200}$$

so the potential well of a cluster is of order 10^{-5} deep. We have previously shown that potential fluctuations do not evolve in gravitational collapse, so our prediction is

$$\left[\mathcal{T}^2(\ell)\right]^{1/2} \sim 10^{-5}. \tag{201}$$

As shown in figure 25, this prediction works pretty well, and we also see the expected sub-degree scale structure: the spectrum peaks at $\ell \simeq 200$, with a sharp fall by $\ell \simeq 1000$. This is at a large enough scale that it is very difficult to sustain the idea of significant spatial curvature. Combined with results from supernovae and large-scale structure, the simplest consistent model is thus the $k = 0$ ΛCDM universe.

24 The future: open questions

In conclusion, empirical cosmology seems to be in pretty good shape. Many lines of evidence point consistently to a universe which is close to being a $k = 0$ model, with $\Omega_m \simeq 0.3$ and $\Omega_v \simeq 0.7$. This model describes the growth of cosmic structure between $z \simeq 1100$ and the present in an accurate way.

Theoretically, there are also ground for optimism. Inflation provides a possible way of understanding how an expanding universe can be created that is so uniform, and yet which contains small seeds for the growth of structure. However, there are some key questions where our ignorance is almost total:

(1) What is the dark matter? Can a suitable elementary particle be detected by particle physicists?

(2) Why is the vacuum energy non-zero at such a low level today?

(3) Did inflation happen? Can we find a direct signature that proves that the vacuum-dominated phase actually happened? If not, is inflation science?

On the last question, the CMB provides the greatest hope. The mechanism of stretched quantum fluctuations implies that there should be relic fluctuations in all fields, not just the inflation field. In particular, there should be a background of gravity waves left over from inflation. These may eventually be detected directly by satellite interferometers, but the most immediate prospect is via looking for additional distortions in the CMB. So far, the data can be fit well by models with no gravity waves, and this is worrying. Because the predicted amplitude of inflationary fluctuations depends on the exact $V(\phi)$ model, a lack of detection would not rule out inflation – but then we will be left wondering whether or not the basic idea is right at all. The next generations of CMB anisotropy experiments will be able to detect a gravity-wave contribution to $\mathcal{T}^2$ at the 10% level, and seeing how these experiments work out will be perhaps the most closely-watched issue in cosmology over the next 5 years.