Chapter 2

An introduction to the physics of cosmology

John A Peacock
Institute for Astronomy, University of Edinburgh, United Kingdom

In asking me to write on ‘The Physics of Cosmology’, the editors of this book have placed no restrictions on the material, since the wonderful thing about modern cosmology is that it draws on just about every branch of physics. In practice, this chapter attempts to set the scene for some of the later more specialized topics by discussing the following subjects:

(1) some cosmological aspects of general relativity,
(2) basics of the Friedmann models,
(3) quantum fields and physics of the vacuum and
(4) dynamics of cosmological perturbations.

2.1 Aspects of general relativity

The aim of general relativity is to write down laws of physics that are valid descriptions of nature as seen from any viewpoint. Special relativity shares the same philosophy, but is restricted to inertial frames. The mathematical tool for the job is the 4-vector; this allows us to write equations that are valid for all observers because the quantities on either side of the equation will transform in the same way. We ensure that this is so by constructing physical 4-vectors out of the fundamental interval

$$\mathrm{d}x^\mu = (c\,\mathrm{d}t, \mathrm{d}x, \mathrm{d}y, \mathrm{d}z) \qquad \mu = 0, 1, 2, 3,$$

using relativistic invariants such as the the rest mass m and proper time $\mathrm{d}\tau$.

For example, defining the 4-momentum $P^\mu = m\,\mathrm{d}x^\mu/\mathrm{d}\tau$ allows an immediate relativistic generalization of conservation of mass and momentum, since the equation $\Delta P^\mu = 0$ reduces to these laws for an observer who sees a set of slowly-moving particles.

None of this seems to depend on whether or not observers move at constant velocity. We have in fact already dealt with the main principle of general relativity, which states that the only valid physical laws are those that equate two quantities that transform in the same way under any arbitrary change of coordinates. We may distinguish equations that are *covariant*—i.e. relate two tensors of the same rank—and *invariants*, where contraction of a tensor yields a number that is the same for all observers:

$$\begin{aligned} \Delta P^\mu &= 0 \qquad \text{covariant} \\ P^\mu P_\mu &= m^2c^2 \qquad \text{invariant.} \end{aligned}$$

The constancy of the speed of light is an example of this: with $\mathrm{d}x_\mu = (c\,\mathrm{d}t, -\mathrm{d}x, -\mathrm{d}y, -\mathrm{d}z)$, we have $\mathrm{d}x^\mu\,\mathrm{d}x_\mu = 0$.

Before getting too pleased with ourselves, we should ask how we are going to construct general analogues of 4-vectors. We want general 4-vectors V^μ to transform like $\mathrm{d}x^\mu$ under the adoption of a new set of coordinates x'^μ:

$$V'^\mu = \frac{\partial x'^\mu}{\partial x^\nu} V^\nu .$$

This relation applies for 4-velocity $U^\mu = \mathrm{d}x^\mu/\tau$, but fails when we try to differentiate this equation to form the 4-acceleration $A^\mu = \mathrm{d}U^\mu/\mathrm{d}\tau$:

$$A'^\mu = \frac{\partial x'^\mu}{\partial x^\nu} A^\nu + \frac{\partial^2 x'^\mu}{\partial\tau\,\partial x^\nu} U^\nu .$$

The second term on the right-hand side is zero only when the transformation coefficients are constants. This is so for the Lorentz transformation, but not in general.

The need is therefore to be able to remove the effects of such *local* coordinate transformations from the laws of physics. Technically, we say that physics should be invariant under *Lorentz group symmetry*.

One difficulty with this programme is that general relativity makes no distinction between coordinate transformations associated with the motion of the observer and a simple change of variable. For example, we might decide that henceforth we will write down coordinates in the order (x, y, z, ct) rather than (ct, x, y, z). General relativity can cope with these changes automatically. Indeed, this flexibility of the theory is something of a problem: it can sometimes be hard to see when some feature of a problem is 'real', or just an artifact of the coordinates adopted. People attempt to distinguish this second type of coordinate change by distinguishing between 'active' and 'passive' Lorentz transformations; a more common term for the latter class is *gauge transformations*.

2.1.1 The equivalence principle

The problem of how to generalize the laboratory laws of special relativity is solved by using the equivalence principle, in which the physics in the vicinity of freely falling observers is assumed to be equivalent to special relativity. We can in fact obtain the full equations of general relativity in this way, in an approach pioneered by Weinberg (1972). In what follows, Greek indices run from 0 to 3 (spacetime), Roman from 1 to 3 (spatial). The summation convention on repeated indices of either type is assumed.

Consider freely falling observers, who erect a special-relativity coordinate frame ξ^μ in their neighbourhood. The equation of motion for nearby particles is simple:

$$\frac{\mathrm{d}^2\xi^\mu}{\mathrm{d}\tau^2} = 0; \qquad \xi^\mu = (ct, x, y, z),$$

i.e. they have zero acceleration, and we have Minkowski spacetime

$$c^2\,\mathrm{d}\tau^2 = \eta_{\alpha\beta}\,\mathrm{d}\xi^\alpha\,\mathrm{d}\xi^\beta,$$

where $\eta_{\alpha\beta}$ is just a diagonal matrix $\eta_{\alpha\beta} = \mathrm{diag}(1, -1, -1, -1)$. Now suppose the observers make a transformation to some other set of coordinates x^μ. What results is the perfectly general relation

$$\mathrm{d}\xi^\mu = \frac{\partial\xi^\mu}{\partial x^\nu}\,\mathrm{d}x^\nu,$$

which, on substitution, leads to the two principal equations of dynamics in general relativity:

$$\begin{aligned}\frac{\mathrm{d}^2x^\mu}{\mathrm{d}\tau^2} + \Gamma^\mu_{\alpha\beta}\frac{\mathrm{d}x^\alpha}{\mathrm{d}\tau}\frac{\mathrm{d}x^\beta}{\mathrm{d}\tau} &= 0\\ c^2\,\mathrm{d}\tau^2 &= g_{\alpha\beta}\,\mathrm{d}x^\alpha\,\mathrm{d}x^\beta.\end{aligned}$$

At this stage, the new quantities appearing in these equations are defined only in terms of our transformation coefficients:

$$\begin{aligned}\Gamma^\mu_{\alpha\beta} &= \frac{\partial x^\mu}{\partial\xi^\nu}\frac{\partial^2\xi^\nu}{\partial x^\alpha\partial x^\beta}\\ g_{\mu\nu} &= \frac{\partial\xi^\alpha}{\partial x^\mu}\frac{\partial\xi^\beta}{\partial x^\nu}\eta_{\alpha\beta}.\end{aligned}$$

This tremendously neat argument effectively uses the equivalence principle to prove what is often merely assumed as a starting point in discussions of relativity: that spacetime is governed by Riemannian geometry. There is a metric tensor, and the gravitational force is to be interpreted as arising from non-zero derivatives of this tensor.

The most well-known example of the power of the equivalence principle is the thought experiment that leads to gravitational time dilation. Consider an accelerating frame, which is conventionally a rocket of height h, with a clock mounted on the roof that regularly disgorges photons towards the floor. If the rocket accelerates upwards at g, the floor acquires a speed $v = gh/c$ in the time taken for a photon to travel from roof to floor. There will thus be a blueshift in the frequency of received photons, given by $\Delta\nu/\nu = gh/c^2$, and it is easy to see that the rate of reception of photons will increase by the same factor.

Now, since the rocket can be kept accelerating for as long as we like, and since photons cannot be stockpiled anywhere, the conclusion of an observer on the floor of the rocket is that in a real sense the clock on the roof is running fast. When the rocket stops accelerating, the clock on the roof will have gained a time Δt by comparison with an identical clock kept on the floor. Finally, the equivalence principle can be brought in to conclude that gravity must cause the same effect. Noting that $\Delta\phi = gh$ is the difference in potential between roof and floor, it is simple to generalize this to

$$\frac{\Delta t}{t} = \frac{\Delta\phi}{c^2}.$$

The same thought experiment can also be used to show that light must be deflected in a gravitational field: consider a ray that crosses the rocket cabin horizontally when stationary. This track will appear curved when the rocket accelerates.

2.1.2 Applications of gravitational time dilation

For many purposes, the effects of weak gravitational fields can be dealt with by bolting gravitational time dilation onto Newtonian physics. One good example is in resolving the twin paradox (see p 8 of Peacock 1999).

Another nice paradox is the following: Why do distant stars suffer no time dilation due to their apparently high transverse velocities as viewed from the frame of the rotating Earth? At cylindrical radius r, a star appears to move at $v = r\omega$, implying time dilation by a factor $\Gamma \simeq 1+r^2\omega^2/2c^2$; this is not observed. However, in order to maintain the stars in circular orbits, a centripetal acceleration $a = v^2/r$ is needed. This is supplied by an apparent gravitational acceleration in the rotating frame (a 'non-inertial' force). The necessary potential is $\Phi = r^2\omega^2/2$, so gravitational blueshift of the radiation cancels the kinematic redshift (at least to order r^2). This example captures very well the main philosophy of general relativity: correct laws of physics should allow us to explain what we see, whatever our viewpoint.

For a more important practical application of gravitational time dilation, consider the *Sachs–Wolfe effect*. This is the dominant source of large-scale anisotropies in the cosmic microwave background (CMB), which arise from potential perturbations at last scattering. These have two effects:

(i) they redshift the photons we see, so that an overdensity *cools* the background as the photons climb out, $\delta T/T = \delta\Phi/c^2$;

(ii) they cause time dilation at the last-scattering surface, so that we seem to be looking at a younger (and hence *hotter*) universe where there is an overdensity.

The time dilation is $\delta t/t = \delta\Phi/c^2$; since the time dependence of the scale factor is $a \propto t^{2/3}$ and $T \propto 1/a$, this produces the counterterm $\delta T/T = -(2/3)\delta\Phi/c^2$. The net effect is thus one-third of the gravitational redshift:

$$\frac{\delta T}{T} = \frac{\delta\Phi}{3c^2}.$$

This effect was originally derived by Sachs and Wolfe (1967) and bears their name. It is common to see the first argument alone, with the factor 1/3 attributed to some additional complicated effect of general relativity. However, in weak fields, general relativistic effects should already be incorporated within the concept of gravitational time dilation; the previous argument shows that this is indeed all that is required to explain the full result.

2.2 The energy–momentum tensor

The only ingredient now missing from a classical theory of relativistic gravitation is a field equation: the presence of mass must determine the gravitational field. To obtain some insight into how this can be achieved, it is helpful to consider first the weak-field limit and the analogy with electromagnetism. Suppose we guess that the weak-field form of gravitation will look like electromagnetism, i.e. that we will end up working with both a scalar potential ϕ and a vector potential $\boldsymbol{A}$ that together give a velocity-dependent acceleration $\boldsymbol{a} = -\boldsymbol{\nabla}\phi - \dot{\boldsymbol{A}} + \boldsymbol{v} \wedge (\boldsymbol{\nabla} \wedge \boldsymbol{A})$. Making the usual $e/4\pi\epsilon_0 \rightarrow Gm$ substitution would suggest the field equation

$$\partial^\nu \partial_\nu A^\mu \equiv \Box A^\mu = \frac{4\pi G}{c^2} J^\mu,$$

where $\Box$ is the d'Alembertian wave operator, $A^\mu = (\phi/c, \boldsymbol{A})$ is the 4-potential and $J^\mu = (\rho c, \boldsymbol{j})$ is a quantity that resembles a 4-current, whose components are a mass density and mass flux density. The solution to this equation is well known:

$$A^\mu(\boldsymbol{r}) = \frac{G}{c^2} \int \frac{[J^\mu(\boldsymbol{x})]}{|\boldsymbol{r} - \boldsymbol{x}|} \, d^3x,$$

where the square brackets denote retarded values.

Now, in fact this analogy can be discarded immediately as a theory of gravitation in the weak-field limit. The problem lies in the vector J^μ: what would the meaning of such a quantity be? In electromagnetism, it describes conservation of charge via

$$\partial_\mu J^\mu = \dot{\rho} + \boldsymbol{\nabla} \cdot \boldsymbol{j} = 0$$

(notice how neatly such a conservation law can be expressed in 4-vector form). When dealing with mechanics, on the other hand, we have not one conserved quantity, but *four*: energy and vector momentum.

The electromagnetic analogy is nevertheless useful, as it suggests that the source of gravitation might still be mass and momentum: what we need first is to find the object that will correctly express conservation of 4-momentum. Informally, what is needed is a way of writing four conservation laws for each component of P^μ. We can clearly write four equations of the previous type in matrix form:

$$\partial_\nu T^{\mu\nu} = 0.$$

Now, if this equation is to be covariant, $T^{\mu\nu}$ must be a tensor and is known as the *energy–momentum tensor* (or sometimes as the stress–energy tensor). The meanings of its components in words are: $T^{00} = c^2 \times$ (mass density) $=$ energy density; $T^{12} = x$-component of current of y-momentum etc. From these definitions, the tensor is readily seen to be symmetric. Both momentum density and energy flux density are the product of a mass density and a net velocity, so $T^{0\mu} = T^{\mu 0}$. The spatial stress tensor T^{ij} is also symmetric because any small volume element would otherwise suffer infinite angular acceleration: any asymmetric stress acting on a cube of side L gives a couple $\propto L^3$, whereas the moment of inertia is $\propto L^5$.

An important special case is the energy–momentum tensor for a perfect fluid. In matrix form, the rest-frame $T^{\mu\nu}$ is given by just diag$(c^2\rho, p, p, p)$ (using the fact that the meaning of the pressure p is just the flux density of x-momentum in the x direction etc.). We can bypass the step of carrying out an explicit Lorentz transformation (which would be rather cumbersome in this case) by the powerful technique of manifest covariance. The following expression is clearly a tensor and reduces to the previous rest-frame answer in special relativity:

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

Thus it must be the general expression for the energy–momentum tensor of a perfect fluid.

2.2.1 Relativistic fluid mechanics

A nice application of the energy–momentum tensor is to show how it generates the equations of relativistic fluid mechanics. Given $T^{\mu\nu}$ for a perfect fluid, all that needs to be done is to insert the specific components $U^\mu = \gamma(c, \boldsymbol{v})$ into the fundamental conservation laws: $\partial T^{\mu\nu}/\partial x^\nu = 0$. The manipulation of the resulting equations is a straightforward exercise. Note that it is immediately clear that the results will involve the total or *convective derivative*:

$$\frac{\mathrm{d}}{\mathrm{d}t} \equiv \frac{\partial}{\partial t} + \boldsymbol{v} \cdot \nabla = \gamma^{-1} U^\mu \partial_\mu .$$

The idea here is that the changes experienced by an observer moving with the fluid are inevitably a mixture of temporal and spatial changes. This two-part derivative arises automatically in the relativistic formulation through the 4-vector dot product $U^\mu \partial_\mu$, which arises from the 4-divergence of an energy–momentum tensor containing a term $\propto U^\mu U^\nu$.

The equations that result from unpacking $T^{\mu\nu}{}_{,\nu} = 0$ in this way have a familiar physical interpretation. The $\mu = 1, 2, 3$ components of $T^{\mu\nu}_{,\nu} = 0$ give the relativistic generalization of *Euler's equation* for momentum conservation in fluid mechanics (not to be confused with Euler's equation in variational calculus):

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{v} = -\frac{1}{\gamma^2(\rho + p/c^2)}(\boldsymbol{\nabla} p + \dot{p}\boldsymbol{v}/c^2),$$

and the $\mu = 0$ component gives a generalization of conservation of energy:

$$\frac{\mathrm{d}}{\mathrm{d}t}[\gamma^2(\rho + p/c^2)] = \dot{p}/c^2 - \gamma^2(\rho + p/c^2)\boldsymbol{\nabla}\cdot\boldsymbol{v},$$

where $\dot{p} \equiv \partial p/\partial t$. The meaning of this equation may be made clearer by introducing one further conservation law: particle number. This is governed by a 4-current having zero 4-divergence:

$$\frac{\mathrm{d}}{\mathrm{d}x^\mu}J^\mu = 0, \qquad J^\mu \equiv nU^\mu = \gamma n(c, \boldsymbol{v}).$$

If we now introduce the *relativistic enthalpy* $w = \rho + p/c^2$, then energy conservation becomes

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\gamma w}{n}\right) = \frac{\dot{p}}{\gamma n c^2}.$$

Thus, in steady flow, γ × (enthalpy per particle) is constant.

A very useful general procedure can be illustrated by *linearizing* the fluid equations. Consider a small perturbation about each quantity ($\rho \to \rho + \delta\rho$ etc) and subtract the unperturbed equations to yield equations for the perturbations valid to first order. This means that any higher-order term such as $\delta\boldsymbol{v}\cdot\boldsymbol{\nabla}\delta\rho$ is set equal to zero. If we take the initial state to have constant density and pressure and zero velocity, then the resulting equations are simple:

$$\begin{aligned}\frac{\partial}{\partial t}\delta\boldsymbol{v} &= -\frac{1}{\rho + p/c^2}\boldsymbol{\nabla}\delta p\\ \frac{\partial}{\partial t}\delta\rho &= -(\rho + p/c^2)\boldsymbol{\nabla}\cdot\delta\boldsymbol{v}.\end{aligned}$$

Now eliminate the perturbed velocity (via the divergence of the first of these equations minus the time derivative of the second) to yield the wave equation:

$$\nabla^2\delta\rho - \left(\frac{\partial\rho}{\partial p}\right)\frac{\partial^2\delta\rho}{\partial t^2} = 0.$$

This defines the speed of sound to be $c_{\mathrm{S}}^2 = \partial p/\partial\rho$. Notice that, by a fortunate coincidence, this is exactly the same as is derived from the non-relativistic equations, although we could not have relied upon this in advance. Thus, the speed of sound in a radiation-dominated fluid is just $c/\sqrt{3}$.

2.3 The field equations

The energy–momentum tensor plausibly plays the role that the charge 4-current J^μ plays in the electromagnetic field equations, $\Box A^\mu = \mu_0 J^\mu$. The tensor on the left-hand side of the gravitational field equations is rather more complicated. Weinberg (1972) showed that it is only possible to make one tensor that is linear in second derivatives of the metric, which is the *Riemann tensor*:

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

This tensor gives a covariant description of spacetime curvature. For the field equations, we need a second-rank tensor to match $T^{\mu\nu}$, and the Riemann tensor may be contracted to the Ricci tensor $R^{\mu\nu}$, or further to the *curvature scalar* R:

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

Unfortunately, these definitions are not universally agreed, All authors, however, agree on the definition of the Einstein tensor $G^{\mu\nu}$:

$$G^{\mu\nu} = R^{\mu\nu} - \tfrac{1}{2}g^{\mu\nu}R.$$

This tensor is what is needed, because it has zero covariant divergence. Since $T^{\mu\nu}$ also has zero covariant divergence by virtue of the conservation laws it expresses, it therefore seems reasonable to guess that the two are proportional:

$$G^{\mu\nu} = -\frac{8\pi G}{c^4}T^{\mu\nu}.$$

These are Einstein's gravitational field equations, where the correct constant of proportionality has been inserted. This is obtained by considering the weak-field limit.

2.3.1 Newtonian limit

The relation between Einstein's and Newton's descriptions of gravity involves taking the limit of weak gravitational fields ($\phi/c^2 \ll 1$). We also need to consider a classical source of gravity, with $p \ll \rho c^2$, so that the only non-zero component of $T^{\mu\nu}$ is $T^{00} = c^2\rho$. Thus, the spatial parts of $R^{\mu\nu}$ must be given by

$$R^{ij} = \tfrac{1}{2}g^{ij}R.$$

Converting this to an equation for R^i_j, it follows that $R = R^{00} + \frac{3}{2}R$ and hence that

$$G^{00} = G_{00} = 2R_{00}.$$

Discarding nonlinear terms in the definition of the Riemann tensor leaves

$$R_{\alpha\beta} = \frac{\partial \Gamma^\mu_{\alpha\mu}}{\partial x^\beta} - \frac{\partial \Gamma^\mu_{\alpha\beta}}{\partial x^\mu} \Rightarrow R_{00} = -\Gamma^i_{00,i}$$

for the case of a stationary field. We have already seen that $c^2\Gamma^i_{00}$ plays the role of the Newtonian acceleration, so the required limiting expression for G^{00} is

$$G^{00} = -\frac{2}{c^2}\nabla^2\phi,$$

and comparison with Poisson's equation gives us the constant of proportionality in the field equations.

2.3.2 Pressure as a source of gravity

Newtonian gravitation is modified in the case of a relativistic fluid (i.e. where we cannot assume $p \ll \rho c^2$). It helps to begin by recasting the field equations (this would also have simplified the previous discussion). Contract the equation using $g^\mu_\mu = 4$ to obtain $R = (8\pi G/c^4)T$. This allows us to write an equation for $R^{\mu\nu}$ directly:

$$R^{\mu\nu} = -\frac{8\pi G}{c^4}(T^{\mu\nu} - \tfrac{1}{2}g^{\mu\nu}T).$$

Since $T = c^2\rho - 3p$, we get a modified Poisson equation:

$$\nabla^2\phi = 4\pi G(\rho + 3p/c^2).$$

What does this mean? For a gas of particles all moving at the same speed u, the effective gravitational mass density is $\rho(1 + u^2/c^2)$; thus a radiation-dominated fluid generates a gravitational attraction twice as strong as one would expect from Newtonian arguments. In fact, this factor applies also to individual particles and leads to an interesting consequence. One can turn the argument round by going to the rest frame of the gravitating mass. We will then conclude that a passing test particle will exhibit an acceleration transverse to its path greater by a factor $(1 + u^2/c^2)$ than that of a slowly moving particle. This gives an extra factor of two deflection in the trajectories of photons, which is of critical importance in gravitational lensing.

2.3.3 Energy density of the vacuum

One consequence of the gravitational effects of pressure that may seem of mathematical interest only is that a negative-pressure equation of state that achieved $\rho c^2 + 3p < 0$ would produce *antigravity*. Although such a possibility may seem physically nonsensical, it is in fact one of the most important concepts in contemporary cosmology. The origin of the idea goes back to the time when Einstein was first thinking about the cosmological consequences of general relativity. At that time, the universe was believed to be static—although this was simply a prejudice, rather than being founded on any observational facts. The problem of how a uniform distribution of matter could remain static was one that had faced Newton, and Einstein gave a very simple Newtonian solution. He reasoned that a static homogeneous universe required both the density, ρ, and the gravitational potential, Φ, to be constants. This does not solve Poisson's equation, $\nabla^2\Phi = 4\pi G\rho$, so he suggested that the equation should be changed to $(\nabla^2 + \lambda)\Phi = 4\pi G\rho$, where λ is a new constant of nature: the *cosmological constant*. Almost as an afterthought, Einstein pointed out that this equation has the natural relativistic generalization of

$$G^{\mu\nu} + \Lambda g^{\mu\nu} = -\frac{8\pi G}{c^4} T^{\mu\nu}.$$

What is the physical meaning of Λ? In the current form, it represents the curvature of empty space. The modern approach is to move the Λ term to the right-hand side of the field equations. It now looks like the energy–momentum tensor of the vacuum:

$$T^{\mu\nu}_{\text{vac}} = \frac{\Lambda c^4}{8\pi G} g^{\mu\nu}.$$

How can a vacuum have a non-zero energy density and pressure? Surely these are zero by definition in a vacuum? What we can be sure of is that the absence of a preferred frame means that $T^{\mu\nu}_{\text{vac}}$ must be the same for all observers in special relativity . Now, apart from zero, there is only one isotropic tensor of rank 2: $\eta^{\mu\nu}$. Thus, in order for $T^{\mu\nu}_{\text{vac}}$ to be unaltered by Lorentz transformations, the only requirement we can have is that it must be proportional to the metric tensor. Therefore, it is inevitable that the vacuum (at least in special relativity) will have a negative-pressure equation of state:

$$p_{\text{vac}} = -\rho_{\text{vac}} c^2.$$

In this case, $\rho c^2 + 3p$ is indeed negative: a positive Λ will act to cause a large-scale repulsion. The vacuum energy density can thus play a crucial part in the dynamics of the early universe.

It may seem odd to have an energy density that does not change as the universe expands. What saves us is the peculiar equation of state of the vacuum: the work done by the pressure is just sufficient to maintain the energy density constant (see figure 2.1). 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.

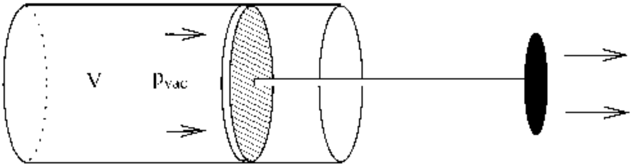

Figure 2.1. A thought experiment to illustrate the application of conservation of energy to the vacuum. If the vacuum density is $\rho_{\rm vac}$ then the energy created by withdrawing the piston by a volume $\mathrm{d}V$ is $\rho_{\rm vac}c^2\,\mathrm{d}V$. This must be supplied by work done by the vacuum pressure $p_{\rm vac}\,\mathrm{d}V$, and so $p_{\rm vac} = -\rho_{\rm vac}c^2$, as required.

2.4 The Friedmann models

Many of the chapters in this book discuss observational cosmology, assuming a body of material on standard homogeneous cosmological models. This section attempts to set the scene by summarizing the key basic features of relativistic cosmology.

2.4.1 Cosmological coordinates

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

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

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

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

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

$$\mathrm{d}\sigma^2 = R^2(\mathrm{d}r^2 + \sin^2 r\,\mathrm{d}\phi^2).$$

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

$$\mathrm{d}\sigma^2 = R^2(\mathrm{d}r^2 + \sinh^2 r\,\mathrm{d}\phi^2).$$

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

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

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

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

$$|(\mathrm{d}x, \mathrm{d}y, \mathrm{d}z, \mathrm{d}w)|^2 = R^2[\mathrm{d}\alpha^2 + \sin^2\alpha(\mathrm{d}\beta^2 + \sin^2\beta\,\mathrm{d}\gamma^2)],$$

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

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

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

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

$$c^2\,\mathrm{d}\tau^2 = c^2\,\mathrm{d}t^2 - R^2(t)[\mathrm{d}r^2 + S_k^2(r)\,\mathrm{d}\psi^2].$$

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

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

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

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

so that $a = 1$ at the present.

2.4.2 The redshift

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

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

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

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

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

$$r = \int \frac{c\,\mathrm{d}t}{R(t)}.$$

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 $\mathrm{d}t_{\rm emit}/\mathrm{d}t_{\rm obs} = R(t_{\rm emit})/R(t_{\rm obs})$, 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

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

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

2.4.3 Dynamics of the expansion

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: $(\dot{R}r)^2/2 - GM/(Rr) = \text{constant}$. 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:

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

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

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

$$\rho_{\rm c} = \frac{3H^2}{8\pi G}.$$

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

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

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

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

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

then the current density of the universe may be expressed as

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

A powerful approximate model for the energy content of the universe is to divide it into pressureless matter ($\rho \propto R^{-3}$), radiation ($\rho \propto R^{-4}$) and vacuum energy (ρ constant). The first two relations just say that the number density of particles is diluted by the expansion, with photons also having their energy reduced by the redshift; the third relation applies for Einstein's cosmological constant. In terms of observables, this means that the density is written as

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

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

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

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

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

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

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

2.4.4 Solutions to the Friedmann equation

The Friedmann equation may be solved most simply in 'parametric' form, by recasting it in terms of the conformal time $\mathrm{d}\eta = c\,\mathrm{d}t/R$ (denoting derivatives with respect to η by primes):

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

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

$$a'^2 = \frac{k}{(\Omega - 1)}[\Omega_\mathrm{r} + \Omega_\mathrm{m} a - (\Omega - 1)a^2 + \Omega_\mathrm{v} a^4],$$

which is straightforward to integrate provided $\Omega_\mathrm{v} = 0$.

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

$$H^2(a) = H_0^2[\Omega_\mathrm{v} + \Omega_\mathrm{m} a^{-3} + \Omega_\mathrm{r} a^{-4} - (\Omega - 1)a^{-2}].$$

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

$$R_0\,\mathrm{d}r = \frac{c}{H(z)}\,\mathrm{d}z = \frac{c}{H_0}\,\mathrm{d}z[(1-\Omega)(1+z)^2 + \Omega_\mathrm{v} + \Omega_\mathrm{m}(1+z)^3 + \Omega_\mathrm{r}(1+z)^4]^{-1/2}.$$

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

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

$$\Omega(z) - 1 = \frac{\Omega - 1}{1 - \Omega + \Omega_\mathrm{v} a^2 + \Omega_\mathrm{m} a^{-1} + \Omega_\mathrm{r} a^{-2}}.$$

This last equation is very important. It tells us that, at high redshift, all model universes apart from those with only vacuum energy will tend to look like the $\Omega = 1$ model. If $\Omega \neq 1$, then in the distant past $\Omega(z)$ must have differed from unity by a tiny amount: the density and rate of expansion needed to have been finely balanced for the universe to expand to the present. This tuning of the initial conditions is called the *flatness problem*.

The solution of the Friedmann equation becomes more complicated if we allow a significant contribution from vacuum energy—i.e. a non-zero cosmological constant. Detailed discussions of the problem are given by Felten and Isaacman (1986) and Carroll *et al* (1992); the most important features are outlined later.

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

$$\Omega_{\rm v} = \frac{8\pi G\rho_{\rm v}}{3H^2} = \frac{\Lambda c^2}{3H^2}.$$

With the addition of Λ, the Friedmann equation can only in general be solved numerically. However, we can find the conditions for the different behaviours described earlier analytically, at least if we simplify things by ignoring radiation. The equation in the form of the time-dependent Hubble parameter looks like

$$\frac{H^2}{H_0^2} = \Omega_{\rm v}(1 - a^{-2}) + \Omega_{\rm m}(a^{-3} - a^{-2}) + a^{-2}.$$

This equation allows the left-hand side to vanish, defining a turning point in the expansion. Vacuum energy can thus remove the possibility of a big bang in which the scale factor goes to zero. Setting the right-hand side to zero yields a cubic equation, and it is possible to give the conditions under which this has a solution (see Felten and Isaacman 1986). The main results of this analysis are summed up in figure 2.2. Since the radiation density is very small today, the main task of relativistic cosmology is to work out where on the $\Omega_{\rm matter}$–$\Omega_{\rm vacuum}$ plane the real universe lies. The existence of high-redshift objects rules out the bounce models, so that the idea of a hot big bang cannot be evaded.

The most important model in cosmological research is that with $k = 0 \Rightarrow \Omega_{\rm total} = 1$; when dominated by matter, this is often termed the *Einstein–de Sitter* model. Paradoxically, this importance arises because it is an unstable state: as we have seen earlier, the universe will evolve away from $\Omega = 1$, given a slight perturbation. For the universe to have expanded by so many *e-foldings* (factors of e expansion) and yet still have $\Omega \sim 1$ implies that it was very close to being spatially flat at early times.

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

$$\dot{a}^2 = H_0^2(\Omega_{\rm m}a^{-1} + \Omega_{\rm r}a^{-2}),$$

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

$$H_0t = \frac{2}{3\Omega_{\rm m}^2}\left[\sqrt{\Omega_{\rm r} + \Omega_{\rm m}a}(\Omega_{\rm m}a - 2\Omega_{\rm r}) + 2\Omega_{\rm r}^{3/2}\right];$$

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

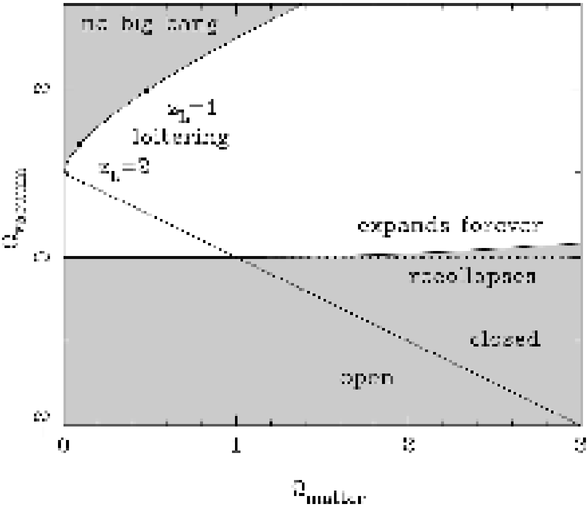

Figure 2.2. This plot shows the different possibilities for the cosmological expansion as a function of matter density and vacuum energy. Models with total $\Omega > 1$ are always spatially closed (open for $\Omega < 1$), although closed models can still expand to infinity if $\Omega_{\rm v} \neq 0$. If the cosmological constant is negative, recollapse always occurs; recollapse is also possible with a positive $\Omega_{\rm v}$ if $\Omega_{\rm m} \gg \Omega_{\rm v}$. If $\Omega_{\rm v} > 1$ and $\Omega_{\rm m}$ 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.

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

$$t = \sqrt{\frac{1}{6\pi G\rho}} \qquad \text{(matter domination)}$$

$$t = \sqrt{\frac{3}{32\pi G\rho}} \qquad \text{(radiation domination).}$$

An alternative $k = 0$ model of greater observational interest has a significant cosmological constant, so that $\Omega_{\rm m} + \Omega_{\rm v} = 1$ (radiation being neglected for simplicity). The advantage of this model is that it is the only way of retaining the theoretical attractiveness of $k = 0$ while changing the age of the universe from the relation $H_0 t_0 = 2/3$, which characterizes the Einstein–de Sitter model. Since much observational evidence indicates that $H_0 t_0 \simeq 1$, this model has received a good deal of interest in recent years. To keep things simple we shall neglect radiation, so that the Friedmann equation is

$$\dot{a}^2 = H_0^2[\Omega_{\rm m} a^{-1} + (1 - \Omega_{\rm m})a^2],$$

and the $t(a)$ relation is

$$H_0 t(a) = \int_0^a \frac{x\,\mathrm{d}x}{\sqrt{\Omega_\mathrm{m} x + (1-\Omega_\mathrm{m})x^4}}.$$

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

$$H_0 t(a) = \frac{2}{3}\frac{S_k^{-1}\left(\sqrt{a^3|\Omega_\mathrm{m} - 1|/\Omega_\mathrm{m}}\right)}{\sqrt{|\Omega_\mathrm{m} - 1|}}.$$

Here, k in S_k is used to mean sin if $\Omega_\mathrm{m} > 1$, otherwise sinh; these are still $k = 0$ models. Since there is nothing special about the current era, we can clearly also rewrite this expression as

$$H(a)t(a) = \frac{2}{3}\frac{S_k^{-1}\left(\sqrt{|\Omega_\mathrm{m}(a) - 1|/\Omega_\mathrm{m}(a)}\right)}{\sqrt{|\Omega_\mathrm{m}(a) - 1|}} \simeq \frac{2}{3}\Omega_\mathrm{m}(a)^{-0.3},$$

where we include a simple approximation that is accurate to a few per cent over the region of interest ($\Omega_\mathrm{m} \gtrsim 0.1$). In the general case of significant Λ but $k \neq 0$, this expression still gives a very good approximation to the exact result, provided Ω_m is replaced by $0.7\Omega_\mathrm{m} - 0.3\Omega_\mathrm{v} + 0.3$ (Carroll *et al* 1992).

2.4.5 Horizons

For photons, the radial equation of motion is just $c\,\mathrm{d}t = R\,\mathrm{d}r$. How far can a photon get in a given time? The answer is clearly

$$\Delta r = \int_{t_0}^{t_1} \frac{c\,\mathrm{d}t}{R(t)} = \Delta\eta,$$

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

2.4.6 Observations in cosmology

We can now assemble some essential formulae for interpreting cosmological observations. Our observables are the redshift, z, and the angular difference between two points on the sky, $\mathrm{d}\psi$. We write the metric in the form

$$c^2\,\mathrm{d}\tau^2 = c^2\,\mathrm{d}t^2 - R^2(t)[\mathrm{d}r^2 + S_k^2(r)\,\mathrm{d}\psi^2],$$

so that the *comoving* volume element is

$$\mathrm{d}V = 4\pi [R_0 S_k(r)]^2 R_0 \, \mathrm{d}r.$$

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

$$\mathrm{d}\ell = \mathrm{d}\psi \; R_0 S_k(r)/(1+z).$$

Probably the most important relation for observational cosmology is that between monochromatic flux density and luminosity. Start by assuming isotropic emission, so that the photons emitted by the source pass with a uniform flux density through any sphere surrounding the source. We can now make a shift of origin, and consider the RW metric as being centred on the source; however, because of homogeneity, the comoving distance between the source and the observer is the same as we would calculate when we place the origin at our location. The photons from the source are therefore passing through a sphere, on which we sit, of proper surface area $4\pi[R_0 S_k(r)]^2$. But redshift still affects the flux density in four further ways: photon energies and arrival rates are redshifted, reducing the flux density by a factor $(1+z)^2$; opposing this, the bandwidth $\mathrm{d}\nu$ is reduced by a factor $1+z$, so the energy flux per unit bandwidth goes down by one power of $1+z$; finally, the observed photons at frequency ν_0 were emitted at frequency $\nu_0(1+z)$, so the flux density is the luminosity at this frequency, divided by the total area, divided by $1+z$:

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

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: $S_\nu = I_\nu \, \mathrm{d}\Omega$. Combining the angular size and flux–density relations thus gives the relativistic version of surface-brightness conservation. This is independent of cosmology:

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

where B_ν is *surface brightness* (luminosity emitted into unit solid angle per unit area of source). We can integrate over ν_0 to obtain the corresponding total or *bolometric* formulae, which are needed, for example, for spectral-line emission:

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

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

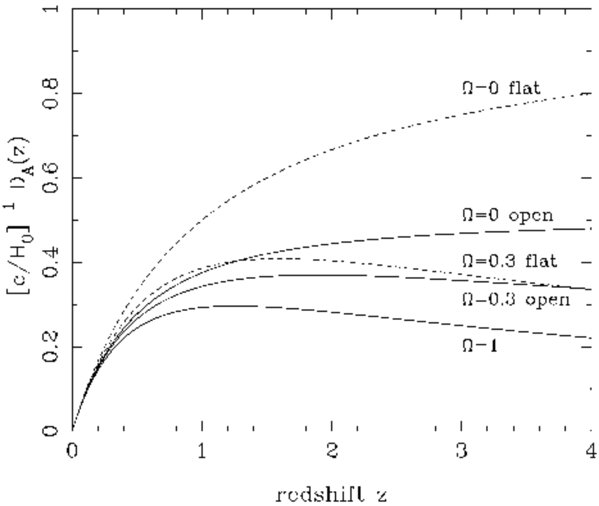

Figure 2.3. A plot of dimensionless angular-diameter distance versus redshift for various cosmologies. Full curves show models with zero vacuum energy; broken curves show flat models with $\Omega_m + \Omega_v = 1$. In both cases, results for $\Omega_m = 1, 0.3, 0$ are shown; higher density results in lower distance at high z, due to gravitational focusing of light rays.

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

$$\textit{angular-diameter distance}: \qquad D_A = (1+z)^{-1} R_0 S_k(r)$$
$$\textit{luminosity distance}: \qquad D_L = (1+z) R_0 S_k(r).$$

The angular-diameter distance is plotted against redshift for various models in figure 2.3.

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

$$c\,\mathrm{d}t = R_0\,\mathrm{d}r/(1+z).$$

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

$$R_0\,\mathrm{d}r = \frac{c}{H(z)}\,\mathrm{d}z = \frac{c}{H_0}\,\mathrm{d}z[(1-\Omega)(1+z)^2+\Omega_v+\Omega_m(1+z)^3+\Omega_r(1+z)^4]^{-1/2}.$$

2.4.7 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 confusion, and this section tries to counter some of the more tenacious misconceptions.

The worst of these 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 suggests some completely new physical effect that is not covered by Newtonian concepts. However, on scales much smaller than the current horizon, we should be able to ignore curvature and treat galaxy dynamics as occurring in Minkowski spacetime; this approach works in deriving the Friedmann equation. How do we relate this to 'expanding space'? It should be clear that Minkowski spacetime does not expand – indeed, the very idea that the motion of distant galaxies could affect local dynamics is profoundly anti-relativistic: the equivalence principle says that we can always find a tangent frame in which physics is locally special relativity.

To clarify the issues here, it should help to consider an explicit example, which makes quite a neat paradox. Suppose we take a nearby low-redshift galaxy and give it a velocity boost such that its redshift becomes zero. At a later time, will the expansion of the universe have cause the galaxy to recede from us, so that it once again acquires a positive redshift? To idealize the problem, imagine that the galaxy is a massless test particle in a homogeneous universe.

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

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

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

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

$$\ddot{r} = \frac{-GM(\langle r|t)}{r^2},$$

and the mass internal to r is just

$$M(\langle r|t) = \frac{4\pi}{3}\rho r^3 = \frac{4\pi}{3}\rho_0 a^{-3} r^3,$$

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

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

Adding vacuum energy is easy enough:

$$\ddot{r} = -\frac{H_0^2}{2} r(\Omega_{\rm m} a^{-3} - 2\Omega_{\rm v}).$$

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

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

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

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

$$H^2(t) = H_0^2[\Omega_{\rm v} + \Omega_{\rm m} a^{-3} + (1 - \Omega_{\rm m} - \Omega_{\rm v})a^{-2}].$$

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

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

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

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

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

At large t, this becomes $r = -r_0 t^{2/3}$. This is indeed the equation of motion of a particle moving with the Hubble flow, but it arises because the particle has fallen right through the origin and emerged on the other side. In no sense, therefore, can 'expanding space' be said to have operated: in an Einstein–de Sitter model, a particle initially at rest with respect to the origin falls towards the origin, passes through it, and asymptotically regains its initial comoving radius on the opposite side of the sky. This behaviour can be understood quantitatively using only Newtonian dynamics.

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

$$r = r_0 \cosh H_0(t - t_0),$$

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

2.5 Inflationary cosmology

We now turn from classical cosmology to aspects of cosmology in which quantum processes are important. This is necessary in order to solve the major problems of the simple 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 lauch 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. To solve the horizon problem and allow causal contact over the whole of the region observed at last scattering requires a universe that expands 'faster than light' near $t = 0$: $R \propto t^{\alpha}$, with $\alpha > 1$. If such a phase had existed, the integral for the comoving horizon would have diverged, and there would be no difficulty in understanding the overall homogeneity of the universe—this could then be established by causal processes. Indeed, it is tempting to assert that the observed homogeneity *proves* that such causal contact must once have occurred.

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

$$\Gamma < 2/3 \Rightarrow \rho c^2 + 3p < 0.$$

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

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

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

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

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

where $H = \sqrt{\Lambda c^2/3} = \sqrt{8\pi G\rho_{\mathrm{vac}}/3}$; all solutions evolve towards the exponential $k = 0$ solution, known as *de Sitter spacetime*. Note that H is not the Hubble parameter at an arbitrary time (unless $k = 0$), but it becomes so exponentially fast as the hyperbolic trigonometric functions tend to the exponential.

Because de Sitter space clearly has H^2 and ρ in the right ratio for $\Omega = 1$ (obvious, since $k = 0$), the density parameter in all models tends to unity as the Hubble parameter tends to H. If we assume that the initial conditions are not fine tuned (i.e. $\Omega = O(1)$ initially), then maintaining the expansion for a factor f produces

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

This can solve the flatness problem, provided f is large enough. To obtain Ω of order unity today requires $|\Omega - 1| \lesssim 10^{-52}$ at the Grand Unified Theory (GUT) epoch, and so $\ln f \gtrsim 60$ e-foldings of expansion are needed; it will be proved later that this is also exactly the number needed to solve the horizon problem. It then seems almost inevitable that the process should go to completion and yield $\Omega = 1$ to measurable accuracy today.

2.5.1 Inflation field dynamics

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

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

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

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

In familiar examples of quantum fields, the potential would be

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

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

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

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

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

If the field is constant both spatially and temporally, the equation of state is then $p = -\rho$, as required if the scalar field is to act as a cosmological constant; note that derivatives of the field spoil this identification.

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

$$\ddot{\phi} + 3H\dot{\phi} - \nabla^2\phi + \mathrm{d}V/\mathrm{d}\phi = 0.$$

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

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

$$V \gg \phi^2/T^2, \qquad V \gg \phi^2/X^2,$$

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

$$3H\dot{\phi} = -\mathrm{d}V/\mathrm{d}\phi.$$

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

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

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

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

(recall that for de Sitter space $H = \sqrt{8\pi G V(\phi)/3} \sim \sqrt{V}/m_\mathrm{P}$ in natural units). These two criteria make perfect intuitive sense: the potential must be flat in the sense of having small derivatives if the field is to roll slowly enough for inflation to be possible.

Similar arguments can be made for the spatial parts. However, they are less critical: what matters is the value of $\nabla\phi = \nabla_{\rm comoving}\phi/R$. Since R increases exponentially, these perturbations are damped away: assuming V is large enough for inflation to start in the first place, inhomogeneities rapidly become negligible. This 'stretching' of field gradients as we increase the cosmological horizon beyond the value predicted in classical cosmology also solves a related problem that was historically important in motivating the invention of inflation—the *monopole problem*. Monopoles are point-like topological defects that would be expected to arise in any phase transition at around the GUT scale ($t \sim 10^{-35}$ s). If they form at approximately one per horizon volume at this time, then it follows that the present universe would contain $\Omega \gg 1$ in monopoles. This unpleasant conclusion is avoided if the horizon can be made much larger than the classical one at the end of inflation; the GUT fields have then been aligned over a vast scale, so that topological-defect formation becomes extremely rare.

2.5.2 Ending inflation

Although spatial derivatives of the scalar field can thus be neglected, the same is not always true for time derivatives. Although they may be negligible initially, the relative importance of time derivatives increases as ϕ rolls down the potential and V approaches zero (leaving aside the subtle question of how we know that the minimum is indeed at zero energy). Even if the potential does not steepen, sooner or later we will have $\epsilon \simeq 1$ or $|\eta| \simeq 1$ and the inflationary phase will cease. Instead of rolling slowly 'downhill', the field will oscillate about the bottom of the potential, with the oscillations becoming damped by the $3H\dot{\phi}$ friction term. Eventually, we will be left with a stationary field that either continues to inflate without end, if $V(\phi = 0) > 0$, or which simply has zero density. This would be a most boring universe to inhabit, but fortunately there is a more realistic way in which inflation can end. We have neglected so far the couplings of the scalar field to matter fields. Such couplings will cause the rapid oscillatory phase to produce particles, leading to *reheating*. Thus, even if the minimum of $V(\phi)$ is at $V = 0$, the universe is left containing roughly the same energy density as it started with, but now in the form of normal matter and radiation—-which starts the usual FRW phase, albeit with the desired special 'initial' conditions.

As well as being of interest for completing the picture of inflation, it is essential to realize that these closing stages of inflation are the *only* ones of observational relevance. Inflation might well continue for a huge number of *e*-foldings, all but the last few satisfying $\epsilon, \eta \ll 1$. However, the scales that left the de Sitter horizon at these early times are now vastly greater than our observable horizon, c/H_0, which exceeds the de Sitter horizon by only a finite factor. If inflation was terminated by reheating to the GUT temperature, then the expansion factor required to reach the present epoch is

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

The comoving horizon size at the end of inflation was therefore

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

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

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

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

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

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

Successful inflation in any of these models requires > 60 e-foldings of the expansion. The implications of this are easily calculated using the slow-roll equation, which gives the number of e-foldings between ϕ_1 and ϕ_2 as

$$N = \int H\ {\rm d}t = -\frac{8\pi}{m_{\rm P}^2}\int_{\phi_1}^{\phi_2}\frac{V}{V'}\ {\rm d}\phi.$$

For any potential that is relatively smooth, $V' \sim V/\phi$, and so we get $N \sim (\phi_{\rm start}/m_{\rm P})^2$, assuming that inflation terminates at a value of ϕ rather smaller than at the start. The criterion for successful inflation is thus that the initial value of the field exceeds the Planck scale:

$$\phi_{\rm start} \gg m_{\rm P}.$$

By the same argument, it is easily seen that this is also the criterion needed to make the slow-roll parameters ϵ and $\eta \ll 1$. To summarize, any model in which the potential is sufficiently flat that slow-roll inflation can commence will probably achieve the critical 60 e-foldings. Counterexamples can of course be constructed, but they have to be somewhat special cases.

It is interesting to review this conclusion for some of the specific inflation models listed earlier. Consider a mass-like potential $V = m^2\phi^2$. If inflation starts near the Planck scale, the fluctuations in V are $\sim m_{\rm P}^4$ and these will drive $\phi_{\rm start}$ to $\phi_{\rm start} \gg m_{\rm P}$ provided $m \ll m_{\rm P}$; similarly, for $V = \lambda\phi^4$, the condition is weak coupling: $\lambda \ll 1$. Any field with a rather flat potential will thus tend to inflate, just because typical fluctuations leave it a long way from home in the form of the potential minimum. In a sense, inflation is realized by means of 'inertial confinement': there is nothing to prevent the scalar field from reaching the minimum of the potential—-but it takes a long time to do so, and the universe has meanwhile inflated by a large factor.

2.5.3 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. There is not time to go into much detail on this here, but we summarize some of the key aspects, in order to make a bridge to the following material on structure formation.

The key idea is to appreciate that the inflaton 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.

To anticipate the detailed treatment, the inflationary prediction is of a horizon-scale fractional perturbation to the density

$$\delta_{\rm H} = \frac{H^2}{2\pi\dot{\phi}}$$

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

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

These universes will then finish inflation at different times, leading to a spread in energy densities (figure 2.4). The horizon-scale density amplitude is given by the different amounts that the universes have expanded following the end of inflation:

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

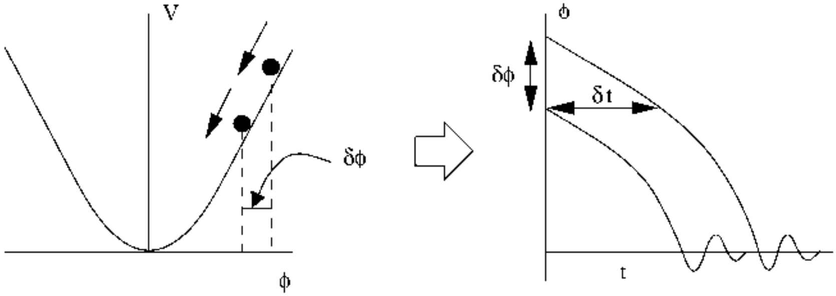

Figure 2.4. This plot shows how fluctuations in the scalar field transform themselves into density fluctuations at the end of inflation. Different points of the universe inflate from points on the potential perturbed by a fluctuation $\delta\phi$, like two balls rolling from different starting points. Inflation finishes at times separated by δt in time for these two points, inducing a density fluctuation $\delta = H\delta t$.

where the last step uses the crucial input of quantum field theory, which says that the rms $\delta\phi$ is given by $H/2\pi$. This is the classical amplitude that results from the stretching of sub-horizon flat-space quantum fluctuations. We will not attempt to prove this key result here (see chapter 12 of Peacock 1999, or Liddle and Lyth 1993, 2000).

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

In summary, we have the following three key equations for basic inflationary model building. The fluctuation amplitude can be thought of as supplying the variance per $\ln k$ in potential perturbations, which we show later does not evolve with time:

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

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

2.5.4 Gravity waves and tilt

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

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

$$h_{\mathrm{rms}} \sim H/m_{\mathrm{P}}.$$

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

This redshifting produces a break in the spectrum of waves. Prior to horizon entry, the gravity waves produce a scale-invariant spectrum of metric distortions, with amplitude h_{rms} per $\ln k$. These distortions are observable via the large-scale CMB anisotropies, where the tensor modes produce a spectrum with the same scale dependence as the Sachs–Wolfe gravitational redshift from scalar metric perturbations. In the scalar case, we have $\delta T/T \sim \phi/3c^2$, i.e. of order the Newtonian metric perturbation; similarly, the tensor effect is

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

where the second step follows because the tensor modes can constitute no more than 100% of the observed CMB anisotropy.

A detailed estimate of the ratio between the tensor effect of gravity waves and the normal scalar Sachs–Wolfe effect was first analysed in a prescient paper by Starobinsky (1985). Denote the fractional temperature variance per natural logarithm of angular wavenumber by Δ^2 (constant for a scale-invariant spectrum). The tensor and scalar contributions are, respectively,

$$\begin{aligned}\Delta_{\mathrm{T}}^2 &\sim h_{\mathrm{rms}}^2 \sim (H^2/m_{\mathrm{P}}^2) \sim V/m_{\mathrm{P}}^4 \\ \Delta_{\mathrm{S}}^2 &\sim \delta_{\mathrm{H}}^2 \sim \frac{H^2}{\dot{\phi}} \sim \frac{H^6}{(V')^2} \sim \frac{V^3}{m_{\mathrm{P}}^6 V'^2}.\end{aligned}$$

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

$$\frac{\Delta_{\mathrm{T}}^2}{\Delta_{\mathrm{S}}^2} \simeq 12.4\epsilon,$$

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

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

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

We then want to express the tilt in terms of parameters of the inflationary potential, ϵ and η. These are of order unity when inflation terminates; ϵ and η must therefore be evaluated when the observed universe left the horizon, recalling that we only observe the last 60-odd e-foldings of inflation. The way to introduce scale dependence is to write the condition for a mode of given comoving wavenumber to cross the de Sitter horizon,

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

Since H is nearly constant during the inflationary evolution, we can replace $\mathrm{d}/\mathrm{d}\ln k$ by $\mathrm{d}\ln a$, and use the slow-roll condition to obtain

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

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

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

and derivatives of V can be expressed in terms of the dimensionless parameters ϵ and η. The tilt of the density perturbation spectrum is thus predicted to be

$$1 - n = 6\epsilon - 2\eta.$$

In section 2.8.5 on CMB anisotropies, we discuss whether this relation is observationally testable.