1

The Expansion of the Universe

The visible universe seems the same in all directions around us, at least if we look out to distances larger than about 300 million light years.[1] The isotropy is much more precise (to about one part in 10^{-5}) in the cosmic microwave background, to be discussed in Chapters 2 and 7. As we will see there, this radiation has been traveling to us for about 14 billion years, supporting the conclusion that the universe at sufficiently large distances is nearly the same in all directions.

It is difficult to imagine that we are in any special position in the universe, so we are led to conclude that the universe should appear isotropic to observers throughout the universe. But not to all observers. The universe does not seem at all isotropic to observers in a spacecraft whizzing through our galaxy at half the speed of light. Such observers will see starlight and the cosmic microwave radiation background coming toward them from the direction toward which they are moving with much higher intensity than from behind. In formulating the assumption of isotropy, one should specify that the universe seems the same in all directions to a family of "typical" freely falling observers: those that move with the average velocity of typical galaxies in their respective neighborhoods. That is, conditions must be the same at the same time (with a suitable definition of time) at any points that can be carried into each other by a rotation about any typical galaxy. But any point can be carried into any other by a sequence of such rotations about various typical galaxies, so the universe is then also homogeneous — observers in all typical galaxies at the same time see conditions pretty much the same.[2]

The assumption that the universe is isotropic and homogeneous will lead us in Section 1.1 to choose the spacetime coordinate system so that the metric takes a simple form, first worked out by Friedmann[3] as a solution of the Einstein field equations, and then derived on the basis of isotropy and homogeneity alone by Robertson[4] and Walker.[5] Almost all of modern cosmology is based on this Robertson–Walker metric, at least as a first

[1] K. K. S. Wu, O. Lahav, and M. J. Rees, *Nature* **397**, 225 (January 21, 1999). For a contrary view, see P. H. Coleman, L. Pietronero, and R. H. Sanders, *Astron. Astrophys.* **200**, L32 (1988): L. Pietronero, M. Montuori, and F. Sylos-Labini, in *Critical Dialogues in Cosmology*, (World Scientific, Singapore, 1997): 24; F. Sylos-Labini, F. Montuori, and L. Pietronero, *Phys. Rep.* **293**, 61 (1998).

[2] The Sloan Digital Sky Survey provides evidence that the distribution of galaxies is homogeneous on scales larger than about 300 light years; see J. Yadav, S. Bharadwaj, B. Pandey, and T. R. Seshadri, *Mon. Not. Roy. Astron. Soc.* **364**, 601 (2005) [astro-ph/0504315].

[3] A. Friedmann, *Z. Phys.* **10**, 377 (1922); *ibid.* **21**, 326 (1924).

[4] H. P. Robertson, *Astrophys. J.* **82**, 284 (1935); *ibid.*, **83**, 187, 257 (1936).

[5] A. G. Walker, *Proc. Lond. Math. Soc.* (2) **42**, 90 (1936).

approximation. The observational implications of these assumptions are discussed in Sections 1.2–1.4, without reference to any dynamical assumptions. The Einstein field equations are applied to the Robertson–Walker metric in Section 1.5, and their consequences are then explored in Sections 1.6–1.13.

1.1 Spacetime geometry

As preparation for working out the spacetime metric, we first consider the geometry of a three-dimensional homogeneous and isotropic space. As discussed in Appendix B, geometry is encoded in a *metric* $g_{ij}(\mathbf{x})$ (with i and j running over the three coordinate directions), or equivalently in a *line element* $ds^2 \equiv g_{ij}\, dx^i\, dx^j$, with summation over repeated indices understood. (We say that ds is the *proper distance* between $\mathbf{x}$ and $\mathbf{x} + d\mathbf{x}$, meaning that it is the distance measured by a surveyor who uses a coordinate system that is Cartesian in a small neighborhood of the point $\mathbf{x}$.) One obvious homogeneous isotropic three-dimensional space with positive definite lengths is flat space, with line element

$$ds^2 = d\mathbf{x}^2 \,. \tag{1.1.1}$$

The coordinate transformations that leave this invariant are here simply ordinary three-dimensional rotations and translations. Another fairly obvious possibility is a spherical surface in four-dimensional Euclidean space with some radius a, with line element

$$ds^2 = d\mathbf{x}^2 + dz^2 \,, \qquad z^2 + \mathbf{x}^2 = a^2 \,. \tag{1.1.2}$$

Here the transformations that leave the line element invariant are *four*-dimensional rotations; the direction of $\mathbf{x}$ can be changed to any other direction by a four-dimensional rotation that leaves z unchanged (that is, an ordinary three-dimensional rotation), while $\mathbf{x}$ can be carried into any other point by a four-dimensional rotation that does change z. It can be proved[6] that the only other possibility (up to a coordinate transformation) is a hyperspherical surface in four-dimensional pseudo-Euclidean space, with line element

$$ds^2 = d\mathbf{x}^2 - dz^2 \,, \qquad z^2 - \mathbf{x}^2 = a^2 \,, \tag{1.1.3}$$

where a^2 is (so far) an arbitrary positive constant. The coordinate transformations that leave this invariant are four-dimensional pseudo-rotations, just like Lorentz transformations, but with z instead of time.

[6] See S. Weinberg, *Gravitation and Cosmology* (John Wiley & Sons, New York, 1972) [quoted below as G&C], Sec. 13.2.

We can rescale coordinates

$$\mathbf{x}' \equiv a\mathbf{x} \,, \qquad z' \equiv az \,. \tag{1.1.4}$$

Dropping primes, the line elements in the spherical and hyperspherical cases are

$$ds^2 = a^2\Big[d\mathbf{x}^2 \pm dz^2\Big] \,, \qquad z^2 \pm \mathbf{x}^2 = 1 \,. \tag{1.1.5}$$

The differential of the equation $z^2 \pm \mathbf{x}^2 = 1$ gives $zdz = \mp\mathbf{x}\cdot d\mathbf{x}$ so

$$ds^2 = a^2\Big[d\mathbf{x}^2 \pm \frac{(\mathbf{x}\cdot d\mathbf{x})^2}{1 \mp \mathbf{x}^2}\Big] \,. \tag{1.1.6}$$

We can extend this to the case of Euclidean space by writing it as

$$ds^2 = a^2\left[d\mathbf{x}^2 + K\frac{(\mathbf{x}\cdot d\mathbf{x})^2}{1 - K\mathbf{x}^2}\right] \,, \tag{1.1.7}$$

where

$$K = \begin{cases} +1 & \text{spherical} \\ -1 & \text{hyperspherical} \\ 0 & \text{Euclidean} \end{cases} \,. \tag{1.1.8}$$

(The constant K is often written as k, but we will use upper case for this constant throughout this book to avoid confusion with the symbols for wave number or for a running spatial coordinate index.) Note that we must take $a^2 > 0$ in order to have ds^2 positive at $\mathbf{x} = 0$, and hence everywhere.

There is an obvious way to extend this to the geometry of spacetime: just include a term (1.1.7) in the spacetime line element, with a now an arbitrary function of time (known as the *Robertson–Walker scale factor*):

$$d\tau^2 \equiv -g_{\mu\nu}(x)dx^\mu dx^\nu = dt^2 - a^2(t)\left[d\mathbf{x}^2 + K\frac{(\mathbf{x}\cdot d\mathbf{x})^2}{1 - K\mathbf{x}^2}\right] \,. \tag{1.1.9}$$

Another theorem[7] tells us that this is the unique metric (up to a coordinate transformation) if the universe appears spherically symmetric and isotropic to a set of freely falling observers, such as astronomers in typical galaxies. The components of the metric in these coordinates are:

$$g_{ij} = a^2(t)\left(\delta_{ij} + K\frac{x^i x^j}{1 - K\mathbf{x}^2}\right) \,, \quad g_{i0} = 0 \,, \quad g_{00} = -1 \,, \tag{1.1.10}$$

[7] G&C, Sec. 13.5.

with i and j running over the values 1, 2, and 3, and with $x^0 \equiv t$ the time coordinate in our units, with the speed of light equal to unity. Instead of the quasi-Cartesian coordinates x^i, we can use spherical polar coordinates, for which

$$d\mathbf{x}^2 = dr^2 + r^2 d\Omega \ , \quad d\Omega \equiv d\theta^2 + \sin^2\theta d\phi^2 \ .$$

so

$$d\tau^2 = dt^2 - a^2(t)\left[\frac{dr^2}{1-Kr^2} + r^2 d\Omega\right] \ . \tag{1.1.11}$$

in which case the metric becomes diagonal, with

$$g_{rr} = \frac{a^2(t)}{1-Kr^2} \ , \quad g_{\theta\theta} = a^2(t) r^2 \ , \quad g_{\phi\phi} = a^2(t) r^2 \sin^2\theta \ , \quad g_{00} = -1 \ . \tag{1.1.12}$$

We will see in Section 1.5 that the dynamical equations of cosmology depend on the overall normalization of the function $a(t)$ only through a term $K/a^2(t)$, so for $K = 0$ this normalization has no significance; all that matters are the ratios of the values of $a(t)$ at different times.

The equation of motion of freely falling particles is given in Appendix B by Eq. (B.12):

$$\frac{d^2 x^\mu}{du^2} + \Gamma^\mu_{\nu\kappa} \frac{dx^\nu}{du} \frac{dx^\kappa}{du} = 0 \ , \tag{1.1.13}$$

where $\Gamma^\mu_{\nu\kappa}$ is the affine connection, given in Appendix B by Eq. (B.13),

$$\Gamma^\mu_{\nu\kappa} = \frac{1}{2} g^{\mu\lambda}\left[\frac{\partial g_{\lambda\nu}}{\partial x^\kappa} + \frac{\partial g_{\lambda\kappa}}{\partial x^\nu} - \frac{\partial g_{\nu\kappa}}{\partial x^\lambda}\right] \ . \tag{1.1.14}$$

and u is a suitable variable parameterizing positions along the spacetime curve, proportional to τ for massive particles. (A spacetime path $x^\mu = x^\mu(u)$ satisfying Eq. (1.1.13) is said to be a *geodesic*, meaning that the integral $\int d\tau$ is stationary under any infinitesimal variation of the path that leaves the endpoints fixed.) Note in particular that the derivatives $\partial_i g_{00}$ and $\dot{g}_{0i}$ vanish, so $\Gamma^i_{00} = 0$. A particle at rest in these coordinates will therefore stay at rest, so *these are co-moving coordinates,* which follow the motion of typical observers. Because $g_{00} = -1$, the proper time interval $(-g_{\mu\nu} dx^\mu dx^\nu)^{1/2}$ for a co-moving clock is just dt, so t is the time measured in the rest frame of a co-moving clock.

The meaning of the Robertson–Walker scale factor $a(t)$ can be clarified by calculating the proper distance at time t from the origin to a co-moving

object at radial coordinate r:

$$d(r,t) = a(t)\int_0^r \frac{dr}{\sqrt{1-Kr^2}} = a(t)\times\begin{cases}\sin^{-1} r & K=+1\\ \sinh^{-1} r & K=-1\\ r & K=0\end{cases} \quad (1.1.15)$$

In this coordinate system a co-moving object has r time-independent, so the proper distance from us to a co-moving object increases (or decreases) with $a(t)$. Since there is nothing special about our own position, the proper distance between any two co-moving observers anywhere in the universe must also be proportional to $a(t)$. The rate of change of any such proper distance $d(t)$ is just

$$\dot{d} = d\,\dot{a}/a\ . \quad (1.1.16)$$

We will see in the following section that in fact $a(t)$ is increasing.

We also need the non-zero components of the affine connection, given by Eq. (1.1.14) as:

$$\Gamma^0_{ij} = -\frac{1}{2}\Big(g_{0i,j}+g_{0j,i}-g_{ij,0}\Big) = a\dot{a}\left(\delta_{ij}+K\frac{x^i x^j}{1-K\mathbf{x}^2}\right)$$
$$= a\dot{a}\tilde{g}_{ij}\ , \quad (1.1.17)$$

$$\Gamma^i_{0j} = \frac{1}{2}g^{il}\Big(g_{l0,j}+g_{lj,0}-g_{0j,l}\Big) = \frac{\dot{a}}{a}\delta_{ij}\ , \quad (1.1.18)$$

$$\Gamma^i_{jl} = \frac{1}{2}\tilde{g}^{im}\left(\frac{\partial\tilde{g}_{jm}}{\partial x^l}+\frac{\partial\tilde{g}_{lm}}{\partial x^j}-\frac{\partial\tilde{g}_{jl}}{\partial x^m}\right)\equiv\tilde{\Gamma}^i_{jl}\ . \quad (1.1.19)$$

Here $\tilde{g}_{ij}$ and $\tilde{\Gamma}^i_{jl}$ are the purely spatial metric and affine connection, and $\tilde{g}^{ij}$ is the reciprocal of the 3×3 matrix $\tilde{g}_{ij}$, which in general is different from the ij component of the reciprocal of the 4×4 matrix $g_{\mu\nu}$. In quasi-Cartesian coordinates,

$$\tilde{g}_{ij} = \delta_{ij}+K\frac{x^i x^j}{1-K\mathbf{x}^2}\ , \qquad \tilde{\Gamma}^i_{jl} = K\,\tilde{g}_{jl}x^i\ . \quad (1.1.20)$$

We can use these components of the affine connection to find the motion of a particle that is not at rest in the co-moving coordinate system. First, let's calculate the rate of change of the momentum of a particle of non-zero mass m_0. Consider the quantity

$$P \equiv m_0\sqrt{g_{ij}\frac{dx^i}{d\tau}\frac{dx^j}{d\tau}} \quad (1.1.21)$$

where $d\tau^2 = dt^2 - g_{ij}dx^i dx^j$. In a locally inertial Cartesian coordinate system, for which $g_{ij}=\delta_{ij}$, we have $d\tau = dt\sqrt{1-\mathbf{v}^2}$ where $v^i = dx^i/dt$,

so Eq. (1.1.21) is the formula given by special relativity for the magnitude of the momentum. On the other hand, the quantity (1.1.21) is evidently invariant under arbitrary changes in the *spatial* coordinates, so we can evaluate it just as well in co-moving Robertson–Walker coordinates. This can be done directly, using Eq. (1.1.13), but to save work, suppose we adopt a spatial coordinate system in which the particle position is near the origin $x^i = 0$, where $\tilde{g}_{ij} = \delta_{ij} + O(\mathbf{x}^2)$, and we can therefore ignore the purely spatial components Γ^i_{jk} of the affine connection. General relativity gives the equation of motion

$$\frac{d^2x^i}{d\tau^2} = -\Gamma^i_{\mu\nu}\frac{dx^\mu}{d\tau}\frac{dx^\nu}{d\tau} = -\frac{2}{a}\frac{da}{dt}\frac{dx^i}{d\tau}\frac{dt}{d\tau} \ .$$

Multiplying with $d\tau/dt$ gives

$$\frac{d}{dt}\left(\frac{dx^i}{d\tau}\right) = -\frac{2}{a}\frac{da}{dt}\frac{dx^i}{d\tau} \ ,$$

whose solution is

$$\frac{dx^i}{d\tau} \propto \frac{1}{a^2(t)} \ . \tag{1.1.22}$$

Using this in Eq. (1.1.21) with a metric $g_{ij} = a^2(t)\delta_{ij}$, we see that

$$P(t) \propto 1/a(t) \ . \tag{1.1.23}$$

This holds for any non-zero mass, however small it may be compared to the momentum. Hence, although for photons both m_0 and $d\tau$ vanish, Eq. (1.1.23) is still valid.

It is important to characterize the paths of photons and material particles in interpreting astronomical observations (especially of gravitational lenses, in Chapter 9). Photons and particles passing through the origin of our spatial coordinate system obviously travel on straight lines in this coordinate system, which are spatial geodesics, curves that satisfy the condition

$$\frac{d^2x^i}{ds^2} + \tilde{\Gamma}^i_{jl}\frac{dx^j}{ds}\frac{dx^l}{ds} = 0 \ , \tag{1.1.24}$$

where ds is the three-dimensional proper length

$$ds^2 \equiv \tilde{g}_{ij}\, dx^i\, dx^j \ . \tag{1.1.25}$$

But the property of being a geodesic is invariant under coordinate transformations (since it states the vanishing of a vector), so the path of the photon

or particle will also be a spatial geodesic in any spatial coordinate system, including those in which the photon or particle's path does *not* pass through the origin. (This can be seen in detail as follows. Using Eqs. (1.1.17) and (1.1.18), the equations of motion (1.1.13) of a photon or material particle are

$$0 = \frac{d^2x^i}{du^2} + \Gamma^i_{jl}\frac{dx^j}{du}\frac{dx^l}{du} + \frac{2\dot{a}}{a}\frac{dx^i}{du}\frac{dt}{du} \tag{1.1.26}$$

$$0 = \frac{d^2t}{du^2} + a\dot{a}\tilde{g}_{ij}\frac{dx^i}{du}\frac{dx^j}{du} \ . \tag{1.1.27}$$

Eq. (1.1.26) can be written

$$0 = \left(\frac{ds}{du}\right)^2\left[\frac{d^2x^i}{ds^2} + \Gamma^i_{jl}\frac{dx^j}{ds}\frac{dx^l}{ds}\right] + \left[\frac{d^2s}{du^2} + \frac{2\dot{a}}{a}\frac{dt}{du}\frac{ds}{du}\right]\frac{dx^i}{ds} \ , \tag{1.1.28}$$

where s is so far arbitrary. If we take s to be the proper length (1.1.25) in the spatial geometry, then as we have seen

$$du^2 \propto d\tau^2 \propto dt^2 - a^2\, ds^2$$

Dividing by du^2, differentiating with respect to u, and using Eq. (1.1.27) shows that

$$\frac{d^2s}{du^2} + \frac{2\dot{a}}{a}\frac{dt}{du}\frac{ds}{du} = 0 \ ,$$

so that Eq. (1.1.28) gives Eq. (1.1.24).)

There are various smoothed-out vector and tensor fields, like the current of galaxies and the energy-momentum tensor, whose mean values satisfy the requirements of isotropy and homogeneity. Isotropy requires that the mean value of any three-vector v^i must vanish, and homogeneity requires the mean value of any three-scalar (that is, a quantity invariant under purely spatial coordinate transformations) to be a function only of time, so the current of galaxies, baryons, etc. has components

$$J^i = 0 \ , \quad J^0 = n(t) \ , \tag{1.1.29}$$

with $n(t)$ the number of galaxies, baryons, etc. per proper volume in a comoving frame of reference. If this is conserved, in the sense of Eq. (B.38), then

$$0 = J^\mu{}_{;\mu} = \frac{\partial J^\mu}{\partial x^\mu} + \Gamma^\mu_{\mu\nu}J^\nu = \frac{dn}{dt} + \Gamma^i_{i0}n = \frac{dn}{dt} + 3\frac{da}{dt}\frac{n}{a}$$

so

$$n(t) = \frac{\text{constant}}{a^3(t)} \ . \tag{1.1.30}$$

This shows the decrease of number densities due to the expansion of the co-moving coordinate mesh for increasing $a(t)$.

Likewise, isotropy requires the mean value of any three-tensor t^{ij} at $\mathbf{x} = 0$ to be proportional to δ_{ij} and hence to g^{ij}, which equals $a^{-2}\delta_{ij}$ at $\mathbf{x} = 0$. Homogeneity requires the proportionality coefficient to be some function only of time. Since this is a proportionality between two three-tensors t^{ij} and g^{ij} it must remain unaffected by an arbitrary transformation of space coordinates, including those transformations that preserve the form of g^{ij} while taking the origin into any other point. Hence homogeneity and isotropy require the components of the energy-momentum tensor *everywhere* to take the form

$$T^{00} = \rho(t)\ , \quad T^{0i} = 0\ , \quad T^{ij} = \tilde{g}^{ij}(\mathbf{x})\, a^{-2}(t)\, p(t)\ . \tag{1.1.31}$$

(These are the conventional definitions of proper energy density ρ and pressure p, as given by Eq. (B.43) in the case of a velocity four-vector with $u^i = 0$, $u^0 = 1$.) The momentum conservation law $T^{i\mu}{}_{;\mu} = 0$ is automatically satisfied for the Robertson–Walker metric and the energy-momentum tensor (1.1.31), but the energy conservation law gives the useful information

$$\begin{aligned} 0 = T^{0\mu}{}_{;\mu} &= \frac{\partial T^{0\mu}}{\partial x^\mu} + \Gamma^0_{\mu\nu} T^{\nu\mu} + \Gamma^\mu_{\mu\nu} T^{0\nu} \\ &= \frac{\partial T^{00}}{\partial t} + \Gamma^0_{ij} T^{ij} + \Gamma^i_{i0} T^{00} = \frac{d\rho}{dt} + \frac{3\dot{a}}{a}\Big(p + \rho\Big)\ , \end{aligned}$$

so that

$$\frac{d\rho}{dt} + \frac{3\dot{a}}{a}\Big(p + \rho\Big) = 0\ . \tag{1.1.32}$$

This can easily be solved for an equation of state of the form

$$p = w\rho \tag{1.1.33}$$

with w time-independent. In this case, Eq. (1.1.32) gives

$$\rho \propto a^{-3-3w}\ . \tag{1.1.34}$$

In particular, this applies in three frequently encountered extreme cases:

- **Cold Matter** (e.g. dust): $p = 0$

$$\rho \propto a^{-3} \tag{1.1.35}$$

- **Hot Matter** (e.g. radiation): $p = \rho/3$

$$\rho \propto a^{-4} \tag{1.1.36}$$

- **Vacuum energy**: As we will see in Section 1.5, there is another kind of energy-momentum tensor, for which $T^{\mu\nu} \propto g^{\mu\nu}$, so that $p = -\rho$, in which case the solution of Eq. (1.1.32) is that ρ is a constant, known (up to conventional numerical factors) either as the *cosmological constant* or the *vacuum energy*.

These results apply separately for coexisting cold matter, hot matter, and a cosmological constant, provided that there is no interchange of energy between the different components. They will be used together with the Einstein field equations to work out the dynamics of the cosmic expansion in Section 1.5.

So far, we have considered only local properties of the spacetime. Now let us look at it in the large. For $K = +1$ space is finite, though like any spherical surface it has no boundary. The coordinate system used to derive Eq. (1.1.7) with $K = +1$ only covers half the space, with $z > 0$, in the same way that a polar projection map of the earth can show only one hemisphere. Taking account of the fact that z can have either sign, the circumference of the space is $2\pi a$, and its volume is $2\pi^2 a^3$.

The spaces with $K = 0$ or $K = -1$ are usually taken to be infinite, but there are other possibilities. It is also possible to have finite spaces with the same local geometry, constructed by imposing suitable conditions of periodicity. For instance, in the case $K = 0$ we might identify the points $\mathbf{x}$ and $\mathbf{x} + n_1\mathbf{L}_1 + n_2\mathbf{L}_2 + n_3\mathbf{L}_3$, where n_1, n_2, n_3 run over all integers, and $\mathbf{L}_1$, $\mathbf{L}_2$, and $\mathbf{L}_3$ are fixed non-coplanar three-vectors that characterize the space. This space is then finite, with volume $a^3\mathbf{L}_1 \cdot (\mathbf{L}_2 \times \mathbf{L}_3)$. Looking out far enough, we should see the same patterns of the distribution of matter and radiation in opposite directions. There is no sign of this in the observed distribution of galaxies or cosmic microwave background fluctuations, so any periodicity lengths such as $|\mathbf{L}_i|$ must be larger than about 10^{10} light years.[8]

There are an infinite number of possible periodicity conditions for $K = -1$ as well as for $K = +1$ and $K = 0$.[9] We will not consider these possibilities further here, because they seem ill-motivated. In imposing conditions of periodicity we give up the rotational (though not translational) symmetry that led to the Robertson–Walker metric in the first place, so there seems little reason to impose these periodicity conditions while limiting the local spacetime geometry to that described by the Robertson–Walker metric.

[8] N. J. Cornish *et al.*, *Phys. Rev. Lett.* **92**, 201302 (2004); N. G. Phillips & A. Kogut, *Astrophys. J.* **545**, 820 (2006) [astro-ph/0404400].

[9] For reviews of this subject, see G. F. R. Ellis, *Gen. Rel. & Grav.* **2**, 7 (1971); M. Lachièze-Rey and J.-P. Luminet, Phys. Rept. **254**, 135 (1995); M. J. Rebouças, in *Proceedings of the Xth Brazilian School of Cosmology and Gravitation*, eds. M. Novello and S. E. Perez Bergliaffa (American Institute of Physics Conference Proceedings, Vol. 782, New York, 2005): 188 [astro-ph/0504365].

1.2 The cosmological redshift

The general arguments of the previous section gave no indication whether the scale factor $a(t)$ in the Robertson–Walker metric (1.1.9) is increasing, decreasing, or constant. This information comes to us from the observation of a shift in the frequencies of spectral lines from distant galaxies as compared with their values observed in terrestrial laboratories.

To calculate these frequency shifts, let us adopt a Robertson–Walker coordinate system in which we are at the center of coordinates, and consider a light ray coming to us along the radial direction. A ray of light obeys the equation $d\tau^2 = 0$, so for such a light ray Eq. (1.1.11) gives

$$dt = \pm a(t) \frac{dr}{\sqrt{1 - Kr^2}} \tag{1.2.1}$$

For a light ray coming toward the origin from a distant source, r decreases as t increases, so we must choose the minus sign in Eq. (1.2.1). Hence if light leaves a source at co-moving coordinate r_1 at time t_1, it arrives at the origin $r = 0$ at a later time t_0, given by

$$\int_{t_1}^{t_0} \frac{dt}{a(t)} = \int_0^{r_1} \frac{dr}{\sqrt{1 - Kr^2}} \ . \tag{1.2.2}$$

Taking the differential of this relation, and recalling that the radial coordinate r_1 of co-moving sources is time-independent, we see that the interval δt_1 between departure of subsequent light signals is related to the interval δt_0 between arrivals of these light signals by

$$\frac{\delta t_1}{a(t_1)} = \frac{\delta t_0}{a(t_0)} \tag{1.2.3}$$

If the "signals" are subsequent wave crests, the emitted frequency is $\nu_1 = 1/\delta t_1$, and the observed frequency is $\nu_0 = 1/\delta t_0$, so

$$\nu_0/\nu_1 = a(t_1)/a(t_0) \ . \tag{1.2.4}$$

If $a(t)$ is increasing, then this is a redshift, a decrease in frequency by a factor $a(t_1)/a(t_0)$, equivalent to an increase in wavelength by a factor conventionally called $1 + z$:

$$1 + z = a(t_0)/a(t_1) \ . \tag{1.2.5}$$

Alternatively, if $a(t)$ is decreasing then we have a blueshift, a decrease in wavelength given by the factor Eq. (1.2.5) with z negative. These results are frequently interpreted in terms of the familiar Doppler effect; Eq. (1.1.15)

shows that for an increasing or decreasing $a(t)$, the proper distance to any comoving source of light like a typical galaxy increases or decreases with time, so that such sources are receding from us or approaching us, which naturally produces a redshift or blueshift. For this reason, galaxies with redshift (or blueshift) z are often said to have a cosmological radial velocity cz. (The meaning of relative velocity is clear only for $z \ll 1$, so the existence of distant sources with $z > 1$ does not imply any violation of special relativity.) However, the interpretation of the cosmological redshift as a Doppler shift can only take us so far. In particular, the increase of wavelength from emission to absorption of light does not depend on the rate of change of $a(t)$ at the times of emission or absorption, but on the increase of $a(t)$ in the whole period from emission to absorption.

We can also understand the frequency shift (1.2.4) by reference to the quantum theory of light: The momentum of a photon of frequency ν is $h\nu/c$ (where h is Planck's constant), and we saw in the previous section that this momentum varies as $1/a(t)$.

For nearby sources, we may expand $a(t)$ in a power series, so

$$a(t) \simeq a(t_0)\,[1 + (t - t_0)H_0 + \dots] \tag{1.2.6}$$

where H_0 is a coefficient known as the *Hubble constant*:

$$H_0 \equiv \dot{a}(t_0)/a(t_0)\ . \tag{1.2.7}$$

Eq. (1.2.5) then gives the fractional increase in wavelength as

$$z = H_0\ (t_0 - t_1) + \dots\ . \tag{1.2.8}$$

Note that for close objects, $t_0 - t_1$ is the proper distance d (in units with $c = 1$). We therefore expect a redshift (for $H_0 > 0$) or blueshift (for $H_0 < 0$) that increases linearly with the proper distance d for galaxies close enough to use the approximation (1.2.6):

$$z = H_0 d + \dots\ . \tag{1.2.9}$$

The redshift of light from other galaxies was first observed in the 1910s by Vesto Melvin Slipher at the Lowell Observatory in Flagstaff, Arizona. In 1922, he listed 41 spiral nebulae, of which 36 had positive z up to 0.006, and only 5 had negative z, the most negative being the Andromeda nebula M31, with $z = -0.001$.[1] From 1918 to 1925 C. Wirtz and K. Lundmark[2]

[1] V. M. Slipher, table prepared for A. S. Eddington, *The Mathematical Theory of Relativity*, 2nd ed. (Cambridge University Press, London, 1924): 162.

[2] C. Wirtz, *Astr. Nachr.* **206**, 109 (1918); *ibid.* **215**, 349 (1921); *ibid.* **216**, 451 (1922); *ibid.* **222**, 21 (1924); *Scientia* **38**, 303 (1925); K. Lundmark, *Stock. Hand.* **50**, No. 8 (1920); *Mon. Not. Roy. Astron. Soc.* **84**, 747 (1924); *ibid.* **85**, 865 (1925).

discovered a number of spiral nebulae with redshifts that seemed to increase with distance. But until 1923 it was only possible to infer the *relative* distances of the spiral nebulae, using observations of their apparent luminosity or angular diameter. With the absolute luminosity and physical dimensions unknown, it was even possible that the spiral nebulae were outlying parts of our own galaxy, as was in fact believed by many astronomers. Edwin Hubble's 1923 discovery of Cepheid variable stars in the Andromeda nebula M31 (discussed in the next section) allowed him to estimate its distance and size, and made it clear that the spiral nebulae are galaxies like our own, rather than objects in our own galaxy.

No clear linear relation between redshift and distance could be seen in the early data of Slipher, Wirtz, and Lundmark, because of a problem that has continued to bedevil measurements of the Hubble constant down to the present. Real galaxies generally do not move only with the general expansion or contraction of the universe; they typically have additional "peculiar" velocities of hundreds of kilometers per second, caused by gravitational fields of neighboring galaxies and intergalactic matter. To see a linear relation between redshift and distance, it is necessary to study galaxies with $|z| \gg 10^{-3}$, whose cosmological velocities zc are thousands of kilometers per second.

In 1929 Hubble[3] announced that he had found a "roughly linear" relation between redshift and distance. But at that time redshifts and distances had been measured only for galaxies out to the large cluster of galaxies in the constellation Virgo, whose redshift indicates a radial velocity of about 1,000 km/sec, not much larger than typical peculiar velocities. His data points were therefore spread out widely in a plot of redshift versus distance, and did not really support a linear relation. But by the early 1930s he had measured redshifts and distances out to the Coma cluster, with redshift $z \simeq 0.02$, corresponding to a recessional velocity of about 7,000 km/sec, and a linear relation between redshift and distance was evident. The conclusion was clear (at least, to some cosmologists): the universe really is expanding. The correctness of this interpretation of the redshift is supported by observations to be discussed in Section 1.7.

From Hubble's time to the present galaxies have been discovered with ever larger redshifts. Galaxies were found with redshifts of order unity, for which expansions such as Eq. (1.2.9) are useless, and we need formulas that take relativistic effects into account, as discussed in Sections 1.4 and 1.5. At the time of writing, the largest accurately measured redshift is for a galaxy observed with the Subaru telescope.[4] The Lyman alpha line from

[3] E. P. Hubble, *Proc. Nat. Acad. Sci.* **15**, 168 (1929).

[4] M. Iye *et al.*, *Nature* **443**, 186 (2006) [astro-ph/0609393].

this galaxy (emitted in the transition from the $2p$ to $1s$ levels of hydrogen), which is normally at an ultraviolet wavelength of 1,215 Å, is observed at the infrared wavelength of 9,682 Å, indicating a redshift $1 + z = 9682/1215$, or $z = 6.96$.

It may eventually become possible to measure the expansion rate $H(t) \equiv \dot{a}(t)/a(t)$ at times t earlier than the present, by observing the change in very accurately measured redshifts of individual galaxies over times as short as a decade.[5] By differentiating Eq. (1.2.5) we see that the rate of change of redshift with the time of observation is

$$\frac{dz}{dt_0} = \frac{\dot{a}(t_0)}{a(t_1)} - \frac{a(t_0)\,\dot{a}(t_1)}{a^2(t_1)}\frac{dt_1}{dt_0} = \left[H_0 - H(t_1)\frac{dt_1}{dt_0}\right](1+z) \ .$$

From the same argument that led to Eq. (1.2.3) we have $dt_1/dt_0 = 1/(1+z)$, so if we measure dz/dt_0 we can find the expansion rate at the time of light emission from the formula

$$H(t_1) = H_0(1+z) - \frac{dz}{dt_0} \ . \tag{1.2.10}$$

1.3 Distances at small redshift: The Hubble constant

We must now think about how astronomical distances are measured. In this section we will be considering objects that are relatively close, say with z not much greater than 0.1, so that effects of the spacetime curvature and cosmic expansion on distance determinations can be neglected. These measurements are of cosmological importance in themselves, as they are used to learn the value of the Hubble constant H_0. Also, distance measurements at larger redshift, which are used to find the shape of the function $a(t)$, rely on the observations of "standard candles," objects of known intrinsic luminosity, that must be identified and calibrated by studies at these relatively small redshifts. Distance determinations at larger redshift will be discussed in Section 1.6, after we have had a chance to lay a foundation in Sections 1.4 and 1.5 for an analysis of the effects of expansion and spacetime geometry on measurements of distances of very distant objects.

It is conventional these days to separate the objects used to measure distances in cosmology into primary and secondary distance indicators. The absolute luminosities of the primary distance indicators in our local group

[5] A. Loeb, *Astrophys. J.* **499**, L111 (1998) [astro-ph/9802122]; P-S. Corasaniti, D. Huterer, and A. Melchiorri, *Phys. Rev. D* **75**, 062001 (2007) [astro-ph/0701433]. For an earlier suggestion along this line, see A. Sandage, *Astrophys. J.* **139**, 319 (1962).

of galaxies are measured either directly, by kinematic methods that do not depend on an *a priori* knowledge of absolute luminosities, or indirectly, by observation of primary distance indicators in association with other primary distance indicators whose distance is measured by kinematic methods. The sample of these relatively close primary distance indicators is large enough to make it possible to work out empirical rules that give their absolute luminosities as functions of various observable properties. Unfortunately, the primary distance indicators are not bright enough for them to be studied at distances at which z is greater than about 0.01, redshifts at which cosmological velocities cz would be greater than typical random departures of galactic velocities from the cosmological expansion, a few hundred kilometers per second. Thus they cannot be used directly to learn about $a(t)$. For this purpose it is necessary to use secondary distance indicators, which are bright enough to be studied at these large distances, and whose absolute luminosities are known through the association of the closer ones with primary distance indicators.

A. Primary distance indicators[1]

Almost all distance measurements in astronomy are ultimately based on measurements of the distance of objects within our own galaxy, using one or the other of two classic kinematic methods.

1. Trigonometric parallax

The motion of the earth around the sun produces an annual motion of the apparent position of any star around an ellipse, whose maximum angular radius π is given in radians (for $\pi \ll 1$, which is the case for all stars) by

$$\pi = \frac{d_E}{d} \tag{1.3.1}$$

where d is the star's distance from the solar system, and d_E is the mean distance of the earth from the sun,[2] defined as the *astronomical unit*,

[1] For a survey, see M. Feast, in *Nearby Large-Scale Structures and the Zone of Avoidance*, eds. A. P. Fairall and P. Woudt (ASP Conference Series, San Francisco, 2005) [astro-ph/0405440].

[2] The history of measurements of distances in the solar system goes back to Aristarchus of Samos (circa 310 BC–230 BC). From the ratio of the breadth of the earth's shadow during a lunar eclipse to the angular diameter of the moon he estimated the ratio of the diameters of the moon and earth; from the angular diameter of the moon he estimated the ratio of the diameter of the moon to its distance from the earth; and from the angle between the lines of sight to the sun and moon when the moon is half full he estimated the ratio of the distances to the sun and moon; and in this way he was able to measure the distance to the sun in units of the diameter of the earth. Although the method of Aristarchus was correct, his observations were poor, and his result for the distance to the sun was far too low. [For an account of Greek astronomy before Aristarchus and a translation of his work, see T. L. Heath, *Aristarchus of*

$1\,\mathrm{AU} = 1.496 \times 10^8$ km. A parsec (pc) is defined as the distance at which $\pi = 1''$; there are 206,264.8 seconds of arc per radian so

$$1\ \mathrm{pc} = 206{,}264.8\ \mathrm{AU} = 3.0856 \times 10^{13}\ \mathrm{km} = 3.2616\ \text{light years}.$$

The parallax in seconds of arc is the reciprocal of the distance in parsecs.

The first stars to have their distances found by measurement of their trigonometric parallax were α Centauri, by Thomas Henderson in 1832, and 61 Cygni, by Friedrich Wilhelm Bessel in 1838. These stars are at distances 1.35 pc and 3.48 pc, respectively. The earth's atmosphere makes it very difficult to measure trigonometric parallaxes less than about $0.03''$ from ground-based telescopes, so that for many years this method could be used to find the distances of stars only out to about 30 pc, and at these distances only for a few stars and with poor accuracy.

This situation has been improved by the launching of a European Space Agency satellite known as Hipparcos, used to measure the apparent positions and luminosities of large numbers of stars in our galaxy.[3] For stars of sufficient brightness, parallaxes could be measured with an accuracy (standard deviation) in the range of 7 to 9 $\times 10^{-4}$ arc seconds. Of the 118,000 stars in the Hipparcos Catalog, it was possible in this way to find distances with a claimed uncertainty of no more than 10% for about 20,000 stars, some at distances over 100 pc.

2. Proper motions

A light source at a distance d with velocity $v_\perp$ perpendicular to the line of sight will appear to move across the sky at a rate μ in radians/time given by

$$\mu = v_\perp / d\ . \tag{1.3.2}$$

This is known as its *proper motion*. Of course, astronomers generally have no way of directly measuring the transverse velocity $v_\perp$, but they can measure the component v_r of velocity along the line of sight from the Doppler shift of the source's spectral lines. The problem is to infer $v_\perp$ from the measured value of v_r. This can be done in a variety of special cases:

- Moving clusters are clusters of stars that were formed together and hence move on parallel tracks with equal speed. (These are *open*

Samos (Oxford University Press, Oxford, 1913).] The first reasonably accurate determination of the distance of the earth to the sun was made by the measurement of a parallax. In 1672 Jean Richer and Giovanni Domenico Cassini measured the distance from the earth to Mars, from which it was possible to infer the distance from the earth to the sun, by observing the difference in the apparent direction to Mars as seen from Paris and Cayenne, which are separated by a known distance of 6,000 miles. Today distances within the solar system are measured very accurately by measurement of the timing of radar echoes from planets and of radio signals from transponders carried by spacecraft.

[3] M. A. C. Perryman *et al.*, *Astron. Astrophys.* **323**, L49 (1997).

clusters, in the sense that they are not held together by gravitational attraction, in distinction to the much larger globular clusters whose spherical shape indicates a gravitationally bound system.) The most important such cluster is the Hyades (called by Tennyson's Ulysses the "rainy Hyades"), which contains over 100 stars. The velocities of these stars along the line of sight are measured from their Doppler shifts, and if we knew the distance to the cluster then the velocities of its stars at right angles to the line of sight could be measured from their proper motions. The distance to the cluster was determined long ago to be about 40 pc by imposing the constraint that all these velocities are parallel. Distances measured in this way are often expressed as *moving cluster parallaxes*. Since the advent of the Hipparcos satellite, the moving cluster method has been supplemented with a direct measurement of the trigonometric parallax of some of these clusters, including the Hyades.

- A second method is based on the statistical analysis of the Doppler shifts and proper motions of stars in a sample whose *relative* distances are all known, for instance because they all have the same (unknown) absolute luminosity, or because they all at the same (unknown) distance. The Doppler shifts give the velocities along the line of sight, and the proper motions and the relative distances give the velocities transverse to the line of sight, up to a single overall factor related to the unknown absolute luminosity or distance. This factor can be determined by requiring that the distribution of velocities transverse to the line of sight is the same as the distribution of velocities along the line of sight. Distances measured in this way are often called *statistical parallaxes*, or *dynamical distances*.

- The distance to the Cepheid variable star ζ Geminorum has been measured[4] by comparing the rate of change of its physical diameter, as found from the Doppler effect, with the rate of change of its angular diameter, measured using an optical interferometer. (About Cepheids, more below.) The distance was found to be 336 ± 44 pc, much greater than could have been found from a trigonometric parallax. This method has subsequently been extended to eight other Cepheids.[5]

- It is becoming possible to measure distances by measuring the proper motion of the material produced by supernovae, assuming a

[4] B. F. Lane, M. J. Kuchner, A. F. Boden, M. Creech-Eakman, and S. B. Kulkarni, *Nature* **407**, 485 (September 28, 2000).

[5] P. Kervella *et al.*, *Astron. Astrophys.* **423**, 327 (2004) [astro-ph/0404179].

more-or-less cylindrically symmetric explosion, so that the transverse velocity $v_\perp$ can be inferred from the radial velocity v_r measured by Doppler shifts. This method has been applied[6] to the ring around the supernova SN1987A, observed in 1987 in the Large Magellanic Cloud, with the result that its distance is 52 kpc (thousand parsecs).

- The measurement of the time-varying Doppler shift and proper motion of an object in orbit around a central mass can be used to find the distance to the object. For instance, if the line of sight happens to be in the plane of the orbit, and if the orbit is circular, then the Doppler shift is a maximum when the object is moving along the line of sight, and hence gives the orbital velocity v, while the proper motion μ is a maximum when the object is moving with the same velocity at right angles to the line of sight, and gives the distance as v/μ. This method can also be used for orbits that are inclined to the line of sight and not circular, by studying the time-variation of the Doppler shift and proper motion. The application of this method to the star S2, which orbits the massive black hole in the galactic center, gives what is now the best value for the distance of the solar system from the galactic center,[7] as 8.0 ± 0.4 kpc. This method also allows the measurement of some distances outside our galaxy, by using the motion of masers — point microwave sources — in the accretion disks of gas and dust in orbit around black holes at the centers of galaxies. The orbital velocity can be judged from the Doppler shifts of masers at the edge of the accretion disk, which are moving directly toward us or away from us, and if this is the same as the orbital velocities of masers moving transversely to the line of sight, then the ratio of this orbital velocity to their observed proper motion gives the distance to the galaxy. So far, this method has been used to measure the distance to the galaxy NGC 4258,[8] as 7.2 ± 0.5 Mpc (million parsecs), and to the galaxy M33,[9] as 0.730 ± 0.168 Mpc.

These kinematic methods have limited utility outside the solar neighborhood. We need a different method to measure larger distances.

3. Apparent luminosity

The most common method of determining distances in cosmology is based on the measurement of the apparent luminosity of objects of known (or

[6] N. Panagia, *Mem. Soc. Astron. Italiana* **69**, 225 (1998).

[7] F. Eisenhauer *et al.*, *Astrophys. J. Lett.* **597**, L121 (2003) [astro-ph/0306220].

[8] J. Herrnstein *et al.*, *Nature* **400**, 539 (3 August 1999).

[9] A. Brunthaler, M. J. Reid, H. Falcke, L. J. Greenhill, and C. Henkel, *Science* **307**, 1440 (2005) [astro-ph/0503058].

supposedly known) absolute luminosity. The absolute luminosity L is the energy emitted per second, and the apparent luminosity ℓ is the energy received per second per square centimeter of receiving area. If the energy is emitted isotropically, then we can find the relation between the absolute and apparent luminosity in Euclidean geometry by imagining the luminous object to be surrounded with a sphere whose radius is equal to the distance d between the object and the earth. The total energy per second passing through the sphere is $4\pi d^2 \ell$, so

$$\ell = \frac{L}{4\pi d^2} \ . \tag{1.3.3}$$

This relation is subject to corrections due to interstellar and/or intergalactic absorption, as well as possible anisotropy of the source, which though important in practice involve too many technicalities to go into here.

Astronomers unfortunately use a traditional notation for apparent and absolute luminosity in terms of apparent and absolute *magnitude*.[10] In the second century A.D., the Alexandrian astronomer Claudius Ptolemy published a list of 1,022 stars, labeled by categories of apparent brightness, with bright stars classed as being of first magnitude, and stars just barely visible being of sixth magnitude.[11] This traditional brightness scale was made quantitative in 1856 by Norman Pogson, who decreed that a difference of five magnitudes should correspond to a ratio of a factor 100 in apparent luminosities, so that $\ell \propto 10^{-2m/5}$. With the advent of photocells at the beginning of the twentieth century, it became possible to fix the constant of proportionality: the apparent bolometric luminosity (that is, including all wavelengths) is given in terms of the apparent bolometric magnitude m by

$$\ell = 10^{-2m/5} \times 2.52 \times 10^{-5} \,\text{erg cm}^{-2}\ \text{s}^{-1} \ . \tag{1.3.4}$$

For orientation, Sirius has a visual magnitude $m_{\rm vis} = -1.44$, the Andromeda nebula M31 has $m_{\rm vis} = 0.1$, and the large galaxy M87 in the nearest large cluster of galaxies has $m_{\rm vis} = 8.9$. The absolute magnitude in any wavelength band is defined as the apparent magnitude an object would have at a distance of 10 pc, so that the absolute bolometric luminosity is given in terms of the absolute bolometric magnitude M by

$$L = 10^{-2M/5} \times 3.02 \times 10^{35} \text{erg}\ \text{s}^{-1} \ . \tag{1.3.5}$$

[10] For the history of the apparent magnitude scale, see J. B. Hearnshaw, *The Measurement of Starlight: Two centuries of astronomical photometry* (Cambridge University Press, Cambridge, 1996); K. Krisciunas, astro-ph/0106313.

[11] For the star catalog of Ptolemy, see M. R. Cohen and I. E. Drabkin, *A Source Book in Greek Science* (Harvard University Press, Cambridge, MA, 1948): p. 131.

For comparison, in the visual wavelength band the absolute magnitude $M_{\rm vis}$ is +4.82 for the sun, +1.45 for Sirius, and −20.3 for our galaxy. Eq. (1.3.3) may be written as a formula for the distance in terms of the *distance-modulus* $m - M$:

$$d = 10^{1+(m-M)/5}\text{pc} \ . \tag{1.3.6}$$

There are several different kinds of star that have been used in measurements of distance through the observation of apparent luminosity:

- **Main Sequence**: Stars that are still burning hydrogen at their cores obey a characteristic relation between absolute luminosity and color, both depending on mass. This is known as the main sequence, discovered in the decade before the First World War by Ejnar Hertzsprung and Henry Norris Russell. The luminosity is greatest for blue-white stars, and then steadily decreases for colors tending toward yellow and red. The *shape* of the main sequence is found by observing the apparent luminosities and colors of large numbers of stars in clusters, all of which in each cluster may be assumed to be at the same distance from us, but we need to know the distances to the clusters to calibrate absolute luminosities on the main sequence. For many years the calibration of the main sequence absolute luminosities was based on observation of a hundred or so main sequence stars in the Hyades cluster, whose distance was measured by the moving cluster method described above. With the advent of the Hipparcos satellite, the calibration of the main sequence has been greatly improved through the observation of colors and apparent luminosities of nearly 100,000 main sequence stars whose distance is known through measurement of their trigonometric parallax. Including in this sample are stars in open clusters such as the Hyades, Praesepe, the Pleiades, and NGC 2516; these clusters yield consistent main sequence absolute magnitudes if care is taken to take proper account of the varying chemical compositions of the stars in different clusters.[12] With the main sequence calibrated in this way, we can use Eq. (1.3.3) to measure the distance of any star cluster or galaxy in which it is possible to observe stars exhibiting the main sequence relation between apparent luminosity and color. Distances measured in this way are sometimes known as *photometric parallaxes.*

  The analysis of the Hipparcos parallax measurements revealed a discrepancy between the distances to the Pleiades star cluster measured by observations of main sequence stars and by measurements of

[12] S. M. Percival, M. Salaris, and D. Kilkenny, *Astron. Astrophys.* **400**, 541 (2003) [astro-ph/0301219].

trigonometric parallax.[13] The traditional method, using a main sequence calibration based on the application of the moving cluster method to the closer Hyades cluster, gave a distance to the Pleiades[14] of 132 ± 4 pc. Then trigonometric parallaxes of a number of stars in the Pleiades measured by the Hipparcos satellite gave a distance[15] of 118 ± 4 pc, in contradiction with the results of main sequence fitting. More recently, these Hipparcos parallaxes have been contradicted by more accurate measurements of the parallaxes of three stars in the Pleiades with the Fine Guidance Sensor of the Hubble Space Telescope,[16] which gave a distance of 133.5 ± 1.2 pc, in good agreement with the main sequence results. At the time of writing, the balance of astronomical opinion seems to be favoring the distances given by main sequence photometry.[17]

- **Red Clump Stars**: The color–magnitude diagram of clusters in metal-rich[18] parts of the galaxy reveals distinct clumps of red giant stars in a small region of the diagram, with a spread of only about 0.2 in visual magnitude. These are stars that have exhausted the hydrogen at their cores, with helium taking the place of hydrogen as the fuel for nuclear reactions at the stars' cores. The absolute magnitude of the red clump stars in the infrared band (wavelengths around 800 nm) has been determined[19] to be $M_I = -0.28 \pm 0.2$ mag, using the distances and apparent magnitudes measured with the Hipparcos satellite and in an earlier survey.[20] In this band there is little dependence of absolute magnitude on color, but it has been argued that even the infrared magnitude may depend significantly on metallicity.[21]

- **RR Lyrae Stars**: These are variable stars that have been used as distance indicators for many decades.[22] They can be recognized by their periods, typically 0.2 to 0.8 days. The use of the statistical parallax, trigonometric parallax and moving cluster methods (with data

[13] B. Paczynski, *Nature* **227**, 299 (22 January, 2004).

[14] G. Meynet, J.-C. Mermilliod, and A. Maeder, *Astron. Astrophys. Suppl. Ser.* **98**, 477 (1993).

[15] J.-C. Mermilliod, C. Turon, N. Robichon, F. Arenouo, and Y. Lebreton, in *ESA SP-402 Hipparcos–Venice '97*, eds. M.A.C. Perryman and P. L. Bernacca (European Space Agency, Paris, 1997), 643; F. van Leeuwen and C. S. Hansen Ruiz, *ibid*, 689; F. van Leeuwen, *Astron. Astrophys.* **341**, L71 (1999).

[16] D. R. Soderblom *et al.*, *Astron. J.* **129**, 1616 (2005) [astro-ph/0412093].

[17] A new reduction of the raw Hipparcos data is given by F. van Leeuwen and E. Fantino, *Astron. Astrophys.* **439**, 791 (2005) [astro-ph/0505432].

[18] Astronomers use the word "metal" to refer to all elements heavier than helium.

[19] B. Paczyński and K. Z. Stanek, *Astrophys. J.* **494**, L219 (1998).

[20] A. Udalski *et al.*, *Acta. Astron.* **42**, 253 (1992).

[21] L. Girardi, M. A. T. Groenewegen, A. Weiss, and M. Salaris, astro-ph/9805127.

[22] For a review, see G. Bono, *Lect. Notes Phys.* **635**, 85 (2003) [astro-ph/0305102].

from both ground-based observatories and the Hipparcos satellite) give respectively[23] an absolute visual magnitude for RR Lyrae stars in our galaxy's halo of 0.77 ± 0.13, 0.71 ± 0.15, and 0.67 ± 0.10, in good agreement with an earlier result[24] $M_{\rm vis} = 0.71 \pm 0.12$ for halo RR Lyrae stars and 0.79 ± 0.30 for RR Lyrae stars in the thick disk of the galaxy. RR Lyrae stars are mostly too far for a measurement of their trigonometric parallax, but recently measurements[25] with the Hubble Space Telescope have given a value of 3.82×10^{-3} arcsec for the trigonometric parallax of the eponymous star RR Lyr itself, implying an absolute visual magnitude of $0.61^{-0.11}_{+0.10}$.

- **Eclipsing Binaries**: In favorable cases it is possible to estimate the intrinsic luminosity of a star that is periodically partially eclipsed by a smaller companion, without the use of any intermediate distance indicators. The velocity of the companion can be inferred from the Doppler shift of its spectral lines (with the ellipticity of the orbit inferred from the variation of the Doppler shift with time), and the radius of the primary star can then be calculated from the duration of the eclipse. The temperature of the primary can be found from measurement of its spectrum, typically from its apparent luminosity in various wavelength bands. Knowing the radius, and hence the area, and the temperature of the primary, its absolute luminosity can then be calculated from the Stefan–Boltzmann law for black body radiation. This method has been applied to measure distances to two neighboring dwarf galaxies, the Large Magellanic Cloud (LMC)[26] and the Small Magellanic Cloud (SMC),[27] and to the Andromeda galaxy M31[28] and its satellite M33.[29]

- **Cepheid variables**: Because they are so bright, these are by far the most important stars used to measure distances outside our galaxy. Named after the first such star observed, δ Cephei, they can be

[23] P. Popowski and A. Gould, Astrophys. J. **506**, 259, 271 (1998); also astro-ph/9703140, astro-ph/9802168; and in *Post-Hipparcos Cosmic Candles*, eds. A. Heck and F. Caputo (Kluwer Academic Publisher, Dordrecht) [astro-ph/9808006]; A. Gould and P. Popowski, *Astrophys. J.* **568**, 544 (1998) [astro-ph/9805176]; and references cited therein.

[24] A. Layden, R. B. Hanson, S. L. Hawley, A. R. Klemola, and C. J. Hanley, *Astron. J.* **112**, 2110 (1996).

[25] G. F. Benedict *et al.*, *Astrophys. J.* **123**, 473 (2001) [astro-ph/0110271]

[26] E. F. Guinan *et al.*, *Astrophys. J.* **509**, L21 (1998); E. L. Fitzpatrick *et al. Astrophys. J.* **587**, 685 (2003).

[27] T. J. Harries, R. W. Hilditch, and I. D. Howarth, *Mon. Not. Roy. Astron. Soc.* **339**, 157 (2003); R. W. Hilditch, I. D. Howarth, and T. J. Harries, *Mon. Not. Roy. Astron. Soc.* **357**, 304 (2005).

[28] I. Ribas *et al.*, *Astrophys. J.* **635**, L37 (2005).

[29] A. Z. Bonanos *et al.*, *Astrophys. Space Sci.* **304**, 207 (2006) [astro-ph/0606279].

recognized from the characteristic time dependence of their luminosity, with periods ranging from 2 to 45 days. (Cepheids in other galaxies have been observed with periods extending up to 100 days.) In 1912 Henrietta Swan Leavitt discovered that the Cepheid variables in the Small Magellanic Cloud (SMC) have apparent luminosities given by a smooth function of the period of the variation in luminosity, but the distance to the SMC was not known. Having measured the distances and apparent luminosities of several Cepheids in open clusters, and hence their absolute luminosities, it became possible to calibrate the relation between period and luminosity. Cepheid variables thus became a "standard candle" that could be used to measure the distance to any galaxy close enough for Cepheids to be seen. It was the discovery of Cepheids in M31, together with Leavitt's calibration of the Cepheid period–luminosity relation, that allowed Edwin Hubble in 1923 to measure the distance of M31, and show that it was far outside our own galaxy, and hence a galaxy in its own right.

Today the form of the Cepheid period–luminosity relation is derived more from the Large Magellanic Cloud (LMC), where there are many Cepheids, and the dependence of the absolute luminosity on color is also taken into account. The calibration of Cepheid absolute luminosities can therefore be expressed as (and often in fact amounts to) a measurement of the distance to the LMC. Main sequence photometry and other methods gave what for some years was a generally accepted LMC distance modulus of 18.5 mag, corresponding according to Eq. (1.3.6) to a distance of 5.0×10^4 pc. The use of red clump stars[30] has given a distance modulus of 18.47, with a random error ± 0.01, and a systematic error $^{+0.05}_{-0.06}$. A large catalog[31] of Cepheids in the LMC has been interpreted by the members of the Hubble Space Telescope Key Project[32] to give the Cepheid visual and infrared absolute magnitudes as functions of the period P in days:

$$M_V = -2.760 \log_{10} P - 1.458 \,, \qquad M_I = -2.962 \log_{10} P - 1.942 \,, \tag{1.3.7}$$

under the assumption that the LMC distance modulus is 18.5.

This result was challenged in two distinct ways, which illustrate the difficulty of this sort of distance measurement:

First, there have been discordant measurements of the distance to the LMC. Under the assumption that red clump stars in the LMC

[30] M. Salaris, S. Percival, and L. Girardi, *Mon. Not. Roy. Astron. Soc.* **345**, 1030 (2003) [astro-ph/0307329].

[31] A. Udalski *et al.*, *Acta Astr.* **49**, 201 (1999): Table 1.

[32] W. L. Freedman *et al.*, *Astrophys. J.* **553**, 47 (2001).

have the same infrared luminosity as those in the local galactic disk, a distance modulus was found[33] that was 0.45 magnitudes smaller, giving a distance to the LMC that is smaller by a factor 0.8. This has in turn been challenged on the grounds that the stars in the LMC have distinctly lower metallicity than in the local disk; two groups taking this into account[34] have given LMC distance moduli of 18.36 ± 0.17 mag and 18.28 ± 0.18 mag, in fair agreement with the previously accepted value. This also agrees with the measurement of the distance to the LMC inferred[35] from observations of RR Lyrae stars, which gives a distance modulus of 18.33 ± 0.06 mag. This distance modulus for the LMC is further confirmed by the measurement of the distance of the eclipsing binary HV2274; taking account of its distance from the center of the LMC gives[36] a distance modulus for the LMC of 18.30 ± 0.07.

Second, there have been new calibrations of the Cepheid period–luminosity relation, that do not rely on Cepheids in the LMC, which together with observations of Cepheids in the LMC can be used to give an independent estimate of the LMC distance.[37] In recent years the satellite Hipparcos[38] has measured trigonometric parallaxes for 223 Cepheid variables in our galaxy, of which almost 200 can be used to calibrate the period–luminosity relation, without relying on main sequence photometry, red clump stars, or RR Lyrae stars. The nearest Cepheids are more than 100 pc away from us (the distance to Polaris is about 130 pc), so the parallaxes are just a few milliarcseconds, and individual measurements are not very accurate, but with about 200 Cepheids measured it has been possible to get pretty good accuracy. One early result[39] gave the relation between the absolute visual magnitude M_V and the period P (in days) as

$$M_V = -2.81 \log_{10} P - 1.43 \pm 0.10 \; .$$

This was a decrease of about 0.2 magnitudes from previous results, i.e., an increase of the intrinsic luminosity of Cepheids by a factor $10^{0.2 \times 2/5} = 1.20$ leading to a 10% increase in all cosmic distances based

[33] K. Z. Stanek, D. Zaritsky, and J. Harris, *Astrophys. J.* **500**, L141 (1998) [astro-ph/9803181].

[34] A. A. Cole, Astrophys. J. **500**, L137 (1998) [astro-ph/9804110]; L. Girardi *et al., op. cit.*.

[35] P. Popowski and A. Gould, *op. cit.*.

[36] E. F. Guinan *et al., op. cit.*.

[37] For a review, see M. Feast, *Odessa Astron. Publ.* **14** [astro-ph/0110360].

[38] M. A. C. Perryman, *Astron. Astrophys.* **323**, L49 (1997).

[39] M. W. Feast and R. M. Catchpole, *Mon. Not. Roy. Astron. Soc.* **286**, L1 (1997); also see F. Pont, in *Harmonizing Cosmic Distances in a Post-Hipparcos Era*, eds. D. Egret and A. Heck (ASP Conference Series, San Francisco, 1998) [astro-ph/9812074]; H. Baumgardt, C. Dettbarn, B. Fuchs, J. Rockmann, and R. Wielen, in *Harmonizing Cosmic Distance Scales in a Post-Hipparcos Era*, *ibid* [astro-ph/9812437].

directly or indirectly on the Cepheid period–luminosity relation. With this value of Cepheid absolute luminosity, the LMC distance modulus would be 18.66, or slightly less with corrections for the metallicity of the LMC (though with the absolute luminosity of Cepheids calibrated by Hipparcos observations, the only relevance of the LMC for the Cepheid period–luminosity relation is to determine its shape.) This result for the Cepheid absolute luminosities has in turn been contradicted.[40]

These uncertainties may now have been resolved by measurements of the trigonometric parallax of Cepheids in our galaxy with the Fine Guidance Sensor of the Hubble Space Telescope. First, the trigonometric parallax of δ Cephei[41] gave a distance of 273 ± 11 pc, corresponding to an LMC distance modulus of 18.50 ± 0.13. More recently, trigonometric parallaxes have been measured for nine Galactic Cepheids, giving an LMC distance modulus of 18.50 ± 0.03, or with metallicity corrections, 18.40 ± 0.05.[42]

There has also been an independent calibration of the Cepheid period–luminosity relation through observations[43] of Cepheids in the galaxy NGC 4258, whose distance 7.2 ± 0.5 Mpc has been measured using the observations of proper motions of masers in this galaxy mentioned above. This distance is in satisfactory agreement with the distance 7.6 ± 0.3 Mpc obtained from the Cepheids in NGC 4258 under the assumption that these Cepheids have the period–luminosity relation (1.3.7) obtained under the assumption that the LMC distance modulus is 18.5, which tends to confirm this period–luminosity relation. But there are differences in the metallicity of the Cepheids in NGC 4258 and in the LMC, which makes this conclusion somewhat controversial.[44] A 2006 calibration of the Cepheid period–luminosity relation based on the study of 281 Cepheids in NGC 4258[45] (whose distance, as we have seen, is known from observations of maser Doppler shifts and proper motions) gave an LMC distance modulus 18.41 ± 0.10 (stat.) ± 0.13 (syst.). This study includes both a field that is metal rich, like our Galaxy, and a field that is metal poor, like the LMC, so

[40] See, e.g., B. F. Madore and W. L. Freedman, *Astrophys. J.* **492**, 110 (1998) For a recent survey of the theory underlying the Cepheid period–luminosity relation, see A. Gautschy, in *Recent Results on H_0–19th Texas Symposium on Relativistic Astrophysics* [astro-ph/9901021].

[41] G. F. Benedict *et al.*, *Astrophys. J.* **124**, 1695 (2002).

[42] G. F. Benedict *et al.*, *Astron. J.* **133**, 1810 (2007), [astro-ph/0612465].

[43] J. A. Newman *et al.*, *Astrophys. J.* **553**, 562 (2001) [astro-ph-0012377].

[44] For instance, see B. Paczynski, *Nature* **401**, 331 (1999); F. Caputo, M. Marconi, and I. Musella, *Astrophys. J.* [astro-ph/0110526].

[45] L. M. Macri *et al.*, *Astrophys. J.* **652**, 1133 (2006) [astro-ph/0608211].

it provides a calibration of the metallicity dependence of the Cepheid period–luminosity relation.

In a 2003 survey[46] the LMC distance modulus measured using a variety of distance indicators *other* than Cepheid variables (including RR Lyrae stars, red clump stars, etc.) was found to be 18.48 ± 0.04, in very good agreement with the earlier value 18.52 ± 0.05 found by observation of Cepheids, with the corrections adopted by the Hubble Space Telescope group.

B. Secondary distance indicators

None of the above distance indicators are bright enough to be used to measure distances at redshifts large enough so that peculiar velocities can be neglected compared with the expansion velocity, say, $z > 0.03$. For this we need what are called *secondary distance indicators* that are brighter than Cepheids, such as whole galaxies, or supernovae, which can be as bright as whole galaxies.

For many years Cepheids could be used as distance indicators only out to a few million parsecs (Mpc), which limited their use to the Local Group (which consists of our galaxy and the Andromeda nebula M31, and a dozen or so smaller galaxies like M33 and the LMC and SMC) and some other nearby groups (the M81, M101, and Sculptor groups). This was not enough to calibrate distances to an adequate population of galaxies or supernovae, and so it was necessary to use a variety of intermediate distance indicators: globular clusters, HII regions, brightest stars in galaxies, etc. Now the Hubble Space Telescope allows us to observe Cepheids in a great many galaxies at much greater distances, out to about 30 Mpc, and so the secondary distance indicators can now be calibrated directly, without the use of intermediate distance indicators. Four chief secondary distance indicators have been developed:

1. The Tully–Fisher relation
Although whole galaxies can be seen out to very large distances, it has not been possible to identify any class of galaxies with the same absolute luminosity. However, in 1977 Tully and Fisher[47] developed a method for estimating the absolute luminosity of suitable spiral galaxies. The 21 cm absorption line in these galaxies (arising in transitions of hydrogen atoms from lower to the higher of their two hyperfine states) is widened by the

[46] M. Feast, *Lect. Notes Phys.* **635**, 45 (2003) [astro-ph/0301100].
[47] R. B. Tully and J. R. Fisher, *Astron. Astrophys.* **54**, 661 (1977)

Doppler effect, caused by the rotation of the galaxy. The line width W gives an indication of the maximum speed of rotation of the galaxy, which is correlated with the mass of the galaxy, which in turn is correlated with the galaxy's absolute luminosity.[48] (It is also possible to apply the Tully–Fisher relation using the width of other lines, such as a radio frequency transition in the carbon monoxide molecule.[49])

In one application of this approach[50] the *shape* of the function $L_I(W)$ that gives the infrared band absolute luminosity as a function of 21 cm line width (that is, the absolute luminosity up to a common constant factor) was found from a sample of 555 spiral galaxies in 24 clusters, many with redshifts less than 0.01. (The relative distances to these galaxies were found from the ratios of their redshifts, using Eq. (1.2.9), so that the peculiar velocities of these galaxies introduced considerable errors into the estimated ratios of absolute luminosities of individual pairs of galaxies, but with 555 galaxies in the sample it could be assumed that these errors would cancel in a least-squares fit of the measured relative values of absolute luminosity to a smooth curve.) Roughly speaking, $L_I(W)$ turned out to be proportional to W^3. The overall scale of the function $L_I(W)$ was then found by fitting it to the absolute luminosities of 15 spiral galaxies whose distances were accurately known from observations of Cepheid variables they contain. (These 15 galaxies extended out only to 25 Mpc, not far enough for them to be used to measure the Hubble constant directly.) The Hubble constant could then be found by using the function $L_I(W)$ calibrated in this way to find the distances to galaxies in 14 clusters with redshifts ranging from 0.013 to 0.03, and comparing the results obtained with Eq. (1.2.8). (These redshifts may not be large enough to ignore peculiar velocities altogether, but again, this problem is mitigated by the use of a fairly large number of galaxies.) The Hubble constant found in this way was 70 ± 5 km s^{-1} Mpc^{-1}. More recently, the Hubble Space Telescope Key Project to Measure the Hubble Constant has used Cepheid variables to recalibrate the Tully–Fisher relation (assuming an LMC distance of 50 kpc) and then found H_0 by plotting distances found from the Tully–Fisher relation against redshift for a sample of 19 clusters with redshifts from 0.007 to 0.03,[51] taken from the G97 survey of Giovanelli *et al.*[52] The result was $H_0 = 71 \pm 3 \pm 7$ km s^{-1} Mpc^{-1}, with the first quoted uncertainty statistical and the second systematic.

[48] M. Aaronson, J. R. Mould, and J. Huchra, *Astrophys. J.* **229**, 1 (1979).

[49] Y. Tutui *et al.*, *Publ. Astron. Soc. Japan* **53**, 701 (2001) [astro-ph/0108462].

[50] R. Giovanelli, in *The Extragalactic Distance Scale - Proceedings of the Space Telescope Science Institute Symposium held in Baltimore, MD, May, 1996* (Cambridge University Press, Cambridge, 1997): 113; R. Giovanelli *et al.*, *Astron. J.* **113**, 22 (1997).

[51] S. Sakai *et al.*, *Astrophys. J.* **529**, 698 (2000); W. L. Freedman *et al.*, *Astrophys. J.* **553**, 47 (2001); and references cited therein.

[52] Giovanelli *et al., op. cit.*

2. Faber–Jackson relation
Just as the Tully–Fisher method is based on a correlation of orbital velocities with absolute luminosities in spiral galaxies, the Faber–Jackson method is based on a correlation of random velocities with absolute luminosities in elliptical galaxies.[53] An advantage of this method over the Tully–Fisher method is that it has a firmer theoretical foundation, provided by the virial theorem to be discussed in Section 1.9, which directly relates the mean square random velocity to the galaxy mass.

3. Fundamental plane
The Faber–Jackson method was improved by the recognition that the correlation between orbital velocity and absolute luminosity depends also on the surface brightness of the cluster, and hence on its area.[54] (The term "fundamental plane" refers to the way that data on elliptical galaxies are displayed graphically.) This method has been used[55] to estimate that $H_0 = 78 \pm 5$ (stat.) ± 9 (syst.) km sec^{-1} Mpc^{-1}.

4. Type Ia supernovae
Supernovae of Type Ia are believed to occur when a white dwarf star in a binary system accretes sufficient matter from its partner to push its mass close to the Chandrasekhar limit, the maximum possible mass that can be supported by electron degeneracy pressure.[56] When this happens the white dwarf becomes unstable, and the increase in temperature and density allows the conversion of carbon and oxygen into ^{56}Ni, triggering a thermonuclear explosion that can be seen at distances of several thousand megaparsecs. The exploding star always has a mass close to the Chandrasekhar limit, so there is little variation in the absolute luminosity of these explosions, making them nearly ideal distance indicators.[57] What variation there is seems

[53] S. M. Faber and R. E. Jackson, *Astrophys. J.* **204**, 668 (1976).

[54] A. Dressler *et al.*, Astrophys. J. **313**, 42 (1987).

[55] D. D. Kelson *et al.*, *Astrophys. J.* **529**, 768 (2000) [astro-ph/9909222]; J. P. Blakeslee, J. R. Lucey, J. L. Tonry, M. J. Hudson, V. K. Narayan, and B. J. Barris, *Mon. Not. Roy. Astron. Soc.* **330**, 443 (2002) [astro-ph/011183].

[56] W. A. Fowler and F. Hoyle, *Astrophys. J.* **132**, 565 (1960). Calling a supernova Type I simply means that hydrogen lines are not observed in its spectrum. In addition to Type Ia supernovae, there are other Type I supernovae that occur in the collapse of the cores of stars much more massive than white dwarfs, whose outer layer of hydrogen has been lost in stellar winds, as well as Type II supernovae, produced by core collapse in massive stars that have not lost their outer layer of hydrogen. For a discussion of the Chandrasekhar limit, see G&C, Section 11.3.

[57] The use of Type Ia supernovae as distance indicators was pioneered by A. Sandage and G. A. Tammann, *Astrophys. J.* **256**, 339 (1982), following an earlier observation that they had fairly uniform luminosity by C. T. Kowal, *Astron. J.* **73**, 1021 (1968). In 1982 it was necessary to use brightest supergiant stars as intermediate distance indicators, to bridge the gap between the distances that could then be measured using Cepheids and the distances at which the Type Ia supernova could be found. For reviews of the use of type Ia supernovae as standard candles, see D. Branch, *Ann. Rev. Astron. & Astrophys.* **36**, 17 (1998); P. Höflich, C. Gerardy, E. Linder, and H. Marion, in *Stellar Candles*, eds. W. Gieren *et al.* (*Lecture Notes in Physics*) [astro-ph/0301334].

to be well correlated with the rise time and decline time of the supernova light: the slower the decline, the higher the absolute luminosity.[58]

This relation has been calibrated by measurements of Type Ia supernovae in several galaxies of known distance. From 1937 to 1999 there were ten supernovae in galaxies whose distance had been measured by observation of Cepheid variables they contain.[59] Of these, six Type Ia supernovae were used by the HST Key H_0 Group[60] to calibrate the relation between absolute luminosity and decline time. This relation was then used to calculate distances to a sample of 29 Type Ia supernovae in galaxies with redshifts extending from 0.01 to 0.1, observed at the Cerro Tololo Inter-American Observatory.[61] Plotting these distances against measured redshifts gave a Hubble constant[62] of 71 ± 2(statistical) $\pm$ 6(systematic) km s^{-1} Mpc^{-1}. This agrees well with an older determination using Type Ia supernovae by a Harvard group,[63] which found $H_0 = 67 \pm 7$ km s^{-1} Mpc^{-1}. Members of this group have superceded this result,[64] now giving a Hubble constant $H_0 = 73 \pm 4$(stat.) $\pm$ 5(syst.) km s^{-1} Mpc^{-1}. On the other hand, a group headed by Sandage using Type Ia supernovae and the Tully–Fisher relation has consistently found lower values of H_0.[65] The gap seems to be narrowing; in 2006, this group quoted[66] a value $H_0 = 62.3 \pm 1.3$(stat.) $\pm$ 5.0(syst.) km s^{-1} Mpc^{-1}. (According to Sandage *et al.*, the difference between these results is due to a difference in the Cepheid period–luminosity relation used to measure distances to the galaxies that host the supernovae that are used to calibrate the relation between supernova absolute luminosity and decline time. Sandage *et al.* use a metallicity-dependent period–luminosity relation. However Macri *et al.*[45] subsequently reported no difference in the period-luminosity relation for Cepheids in a metal-rich and a metal-poor region of NGC 4258.)

It is an old hope that with a sufficient theoretical understanding of supernova explosions, it might be possible to measure their distance

[58] M. Phillips, *Astrophys. J.* **413**, L105 (1993); M. Hamuy *et al.*, *Astron. J.* **109**, 1 (1995); A. Reiss, W. Press, and R. Kirshner, *Astrophys. J.* **438**, L17 (1996); S. Jha, A. Riess, & R. P. Kirshner, Astrophys. J. **659**, 122 (20007). A dependence of absolute luminosity on color as well as decline time has been considered by R. Tripp and D. Branch, *Astrophys. J.* [astro-ph/9904347].

[59] For a list, see Tripp and Branch, *op. cit.*.

[60] B. Gibson *et al.*, *Astrophys. J.* **529**, 723 (2000) [astro-ph/9908149].

[61] M. Hamuy *et al.*, *Astron. J.* **112**, 2398 (1996).

[62] L. Ferrarese *et al.*, in *Proceedings of the Cosmic Flows Workshop*, eds. S. Courteau *et al.* (ASP Conference Series) [astro-ph/9909134]; W. L. Freedman *et al.*, *Astrophys. J.* **553**, 47 (2001).

[63] A. G. Riess, W. H. Press, And R. P. Kirshner, *Astrophys. J.* **438**, L17 (1995)

[64] A. Riess *et al.*, *Astrophys. J.* **627**, 579 (2005) [astro-ph/0503159].

[65] For a 1996 summary, see G. A. Tammann and M. Federspeil, in *The Extragalactic Distance Scale*, eds. M. Livio, M. Donahue, and N. Panagia (Cambridge University Press, 1997): 137.

[66] A. Sandage *et al.*, *Astrophys. J.* **653**, 843 (2006) [astro-ph/0603647].

without use of primary distance indicators. A 2003 comparison[67] of observed light curves (apparent magnitude as a function of time) and spectra with theory for 26 Type Ia supernovae with redshifts extending up to 0.05, plus one with redshift 0.38, gave $H_0 = 67$ km sec^{-1}Mpc^{-1}, with a two standard deviation uncertainty of 8 km sec^{-1}Mpc^{-1}. It is too soon for this method to replace the older method based on the use of primary distance indicators to calibrate the supernova absolute luminosities, but the agreement between the values of H_0 found in these two ways provides some reassurance that no large error is being made with the older method.

It is instructive to consider a fifth secondary distance indicator that is also used to measure the Hubble constant:

5. Surface brightness fluctuations
In 1988 Tonry and Schneider[68] suggested using the fluctuations in the observed surface brightness of a galaxy from one part of the image to another as a measure of the galaxy's distance. Suppose that the stars in a galaxy can be classified in luminosity classes, all the stars in a luminosity class i having the same absolute luminosity L_i. The rate of receiving energy per unit area of telescope aperture in a small part of the galactic image (as for instance, a single pixel in a charge-coupled device) is

$$\ell = \sum_i \frac{N_i L_i}{4\pi d^2} \tag{1.3.8}$$

where N_i the number of stars of class i in this part of the galaxy's image, and d is the distance of the galaxy. Usually only the brightest stars can be resolved, so it is not possible to measure all the N_i directly, but one can measure the fluctuations in ℓ from one part of the image to another due to the finite values of the N_i. Suppose that the different N_i fluctuate independently from one small part of the galaxy's image to another, and obey the rules of Poisson statistics, so that

$$\Big\langle (N_i - \langle N_i \rangle)\,(N_j - \langle N_j \rangle) \Big\rangle = \delta_{ij} \langle N_i \rangle \ , \tag{1.3.9}$$

with brackets denoting an average over small parts of the central portion of the galaxy's image. It follows then that

$$\frac{\langle (\ell - \langle \ell \rangle)^2 \rangle}{\langle \ell \rangle} = \frac{\bar{L}}{4\pi d^2} \ , \tag{1.3.10}$$

[67] P. Höflich, C. Gerardy, E. Linder, and H. Marion, *op. cit.*
[68] J. Tonry and D. P. Schneider, *Astron. J.* **96**, 807 (1988).

where $\bar{L}$ is a luminosity-weighted mean *stellar* luminosity

$$\bar{L} \equiv \frac{\sum_i \langle N_i \rangle L_i^2}{\sum_i \langle N_i \rangle L_i} \tag{1.3.11}$$

which is expected to vary much less from one galaxy to another than the luminosities of the galaxies themselves. Eq. (1.3.10) can be used to measure distances once this relation is calibrated by measuring $\bar{L}$. By studying surface brightness fluctuations in a survey of galaxies whose distances were found by observations of Cepheids they contain, Tonry *et al.*[69] found an absolute magnitude $\bar{M}_I$ that in the infrared band is equivalent to the absolute luminosity $\bar{L}$:

$$\bar{M}_I = (-1.74 \pm 0.07) + (4.5 \pm 0.25)\,[m_V - m_I - 1.15] \tag{1.3.12}$$

where $m_V - m_I$ is a parameter characterizing the color of the galaxy, equal to the difference of its apparent magnitudes in the infrared and visual bands, assumed here to lie between 1.0 and 1.5. Using Eq. (1.3.10) to find distances of galaxies of higher redshift, they obtained a Hubble constant $81 \pm 6\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$.

There are other phenomena that are used to measure the Hubble constant, including the comparison of apparent and absolute luminosity of supernovae of other types, novae, globular clusters, and planetary nebulae, the diameter–velocity dispersion relation for elliptical galaxies, gravitational lenses (discussed in Section 1.12), the Sunyaev–Zel'dovich effect (discussed in Section 2.3), etc.[70] The HST Key H_0 Group have put together their results of measurements of the Hubble constant using the Tully–Fisher relation, Type Ia supernovae, and several of these other secondary distance indicators, and conclude that[71]

$$H_0 = 71 \pm 6\ \mathrm{km\ s^{-1}\ Mpc^{-1}}\ .$$

As we will see in Section 7.2, the study of anisotropies in the cosmic microwave background has given a value $H_0 = 73 \pm 3\ \mathrm{km\ s^{-1}\ Mpc^{-1}}$. This does not depend on any of the tools discussed in this section, but it does depend on some far-reaching cosmological assumptions: including flat spatial geometry, time-independent vacuum energy, and cold dark matter. For this reason, the increasingly precise measurement of H_0 provided by the

[69] J. L. Tonry, J. P. Blakeslee, E. A. Ajhar, and A. Dressler, *Astrophys. J.* **473**, 399 (1997). For a more recent survey, see J. L. Tonry *et al.*, *Astrophys. J.* **546**, 681 (2001) [astro-ph/0011223].

[70] For a survey of most of these methods, with references, see G. H. Jacoby, D. Branch, R. Ciardullo, R. L. Davies, W. E. Harris, M. J. Pierce, C. J. Pritchet, J. L. Tonry, and D. L. Welch, *Publ. Astron. Soc. Pacific* **104**, 599 (1992).

[71] L. Ferrarese *et al.*, *op. cit.*; W. L. Freedman *et al.*, *Astrophys. J.* **553**, 47 (2001).

cosmic microwave background will not supplant the older measurements discussed in this section — rather, the agreement (or possible future disagreement) between the values of H_0 provided by these very different methods will serve to validate (or possibly invalidate) the cosmological assumptions made in the analysis of the cosmic microwave background.

To take account of the remaining uncertainty in the Hubble constant, it is usual these days to take

$$H_0 = 100\,h \text{ km s}^{-1} \text{ Mpc}^{-1} , \tag{1.3.13}$$

with the dimensionless parameter h assumed to be in the neighborhood of 0.7. This corresponds to a Hubble time

$$1/H_0 = 9.778 \times 10^9 \; h^{-1} \text{ years} . \tag{1.3.14}$$

1.4 Luminosity distances and angular diameter distances

We must now consider the measurement of distances at large redshifts, say $z > 0.1$, where the effects of cosmological expansion on the determination of distance can no longer be neglected. It is these measurements that can tell us whether the expansion of the universe is accelerating or decelerating, and how fast. Before we can interpret these measurements, we will need to consider in this section how to define distance at large redshifts, and we will have to apply Einstein's field equations to the Robertson–Walker metric in the following section. After that, we will return in Section 1.6 to the measurements of distances for large redshift, and their interpretation.

In the previous section we derived the familiar relation $\ell = L/4\pi d^2$ for the apparent luminosity ℓ of a source of absolute luminosity L at a distance d. At large distances this derivation needs modification for three reasons:

1. At the time t_0 that the light reaches earth, the proper area of a sphere drawn around the luminous object and passing through the earth is given by the metric (1.1.10) as $4\pi r_1^2 a^2(t_0)$, where r_1 is the coordinate distance of the earth as seen from the luminous object, which is just the same as the coordinate distance of the luminous object as seen from the earth. The fraction of the light received in a telescope of aperture A on earth is therefore $A/4\pi r_1^2 a^2(t_0)$, and so the factor $1/d^2$ in the formula for ℓ must be replaced with $1/r_1^2 a^2(t_0)$.

2. The rate of arrival of individual photons is lower than the rate at which they are emitted by the redshift factor $a(t_1)/a(t_0) = 1/(1+z)$.

3. The energy $h\nu_0$ of the individual photons received on earth is less than the energy $h\nu_1$ with which they were emitted by the same redshift factor $1/(1+z)$.

Putting this together gives the correct formula for apparent luminosity of a source at radial coordinate r_1 with a redshift z of any size:

$$\ell = \frac{L}{4\pi r_1^2 a^2(t_0)(1+z)^2} \ . \tag{1.4.1}$$

It is convenient to introduce a "luminosity distance" d_L, which is defined so that the relation between apparent and absolute luminosity and luminosity distance is the same as Eq. (1.3.3):

$$\ell = \frac{L}{4\pi d_L^2} \ . \tag{1.4.2}$$

Eq. (1.4.1) can then be expressed as

$$d_L = a(t_0) r_1 (1+z) \ . \tag{1.4.3}$$

For objects with $z \ll 1$, we can usefully write the relation between luminosity distance and redshift as a power series. The redshift $1+z \equiv a(t_0)/a(t_1)$ is related to the "look-back time" $t_0 - t_1$ by

$$z = H_0(t_0 - t_1) + \frac{1}{2}(q_0+2)H_0^2(t_0-t_1)^2 + \ldots \tag{1.4.4}$$

where H_0 is the Hubble constant (1.2.7) and q_0 is the *deceleration parameter*

$$q_0 \equiv \frac{-1}{H_0^2\, a(t_0)} \left.\frac{d^2a(t)}{dt^2}\right|_{t=t_0} \ . \tag{1.4.5}$$

This can be inverted, to give the look-back time as a power series in the redshift

$$H_0(t_0 - t_1) = z - \frac{1}{2}(q_0+2)z^2 + \ldots \ . \tag{1.4.6}$$

The coordinate distance r_1 of the luminous object is given by Eq. (1.2.2) as

$$\frac{t_0 - t_1}{a(t_0)} + \frac{H_0(t_0-t_1)^2}{2a(t_0)} + \cdots = r_1 + \ldots \ , \tag{1.4.7}$$

with the dots on the right-hand side denoting terms of *third* and higher order in r_1. Using Eq. (1.4.6), the solution is

$$r_1\, a(t_0)\, H_0 = z - \frac{1}{2}(1+q_0)z^2 + \cdots \ . \tag{1.4.8}$$

This gives the luminosity distance (1.4.3) as a power series

$$d_L = H_0^{-1}\Big[z + \frac{1}{2}(1-q_0)z^2 + \cdots\Big]\,. \tag{1.4.9}$$

We can therefore measure q_0 as well as H_0 by measuring the luminosity distance as a function of redshift to terms of order z^2. The same reasoning has been used to extend the expression (1.4.9) to fourth order in z:[1]

$$\begin{aligned} d_L(z) = H_0^{-1}\Bigg[& z + \frac{1}{2}(1-q_0)z^2 - \frac{1}{6}\left(1 - q_0 - 3q_0^2 + j_0 + \frac{K}{H_0^2 a_0^2}\right) z^3 \\ & + \frac{1}{24}\Bigg(2 - 2q_0 - 15q_0^2 - 15q_0^3 + 5j_0 + 10q_0 j_0 \\ & + s_0 + \frac{2K(1+3q_0)}{H_0^2 a_0^2}\Bigg) z^4 + \cdots \Bigg]\,, \end{aligned}$$

where j_0 and s_0 are parameters known as the *jerk* and *snap*:

$$j_0 \equiv \frac{1}{H_0^3\, a(t_0)}\, \frac{d^3 a(t)}{dt^3}\bigg|_{t=t_0}\,, \qquad s_0 \equiv \frac{1}{H_0^4\, a(t_0)}\, \frac{d^4 a(t)}{dt^4}\bigg|_{t=t_0}\,.$$

Years ago cosmology was called "a search for two numbers," H_0 and q_0. The determination of H_0 is still a major goal of astronomy, as discussed in the previous section. On the other hand, there is less interest now in q_0. Instead of high-precision distance determinations at moderate redshifts, of order 0.1 to 0.2, which would give an accurate value of q_0, we now have distance determinations of only moderate precision at high redshifts, of order unity, which depend on the whole form of the function $a(t)$ over the past few billion years. For redshifts of order unity, it is not very useful to expand in powers of redshift. In order to interpret these measurements, we will need a dynamical theory of the expansion, to be developed in the next section. As we will see there, modern observations suggest strongly that there are not two but at least three parameters that need to be measured to calculate $a(t)$.

Before turning to this dynamical theory, let's pause a moment to clarify the distinction between different measures of distance. So far, we have encountered the proper distance (1.1.15) and the luminosity distance (1.4.3). There is another sort of distance, which is what we measure when we compare angular sizes with physical dimensions. Inspection of the metric

[1] M. Visser, *Class. Quant. Grav.* **21**, 2603 (2004) [gr-qc/0309109]. The term of third order in z was previously calculated by T. Chiba and T. Nakamura, *Prog. Theor. Phys.* **100**, 1077 (1998).

(1.1.12) shows that a source at co-moving radial coordinate r_1 that emits light at time t_1 and is observed at present to subtend a small angle θ will extend over a proper distance s (normal to the line of sight) equal to $a(t_1)r_1\theta$. The *angular diameter distance* d_A is defined so that θ is given by the usual relation of Euclidean geometry

$$\theta = s/d_A \tag{1.4.10}$$

and we see that

$$d_A = a(t_1)r_1 \ . \tag{1.4.11}$$

Comparison of this result with Eq. (1.4.3) shows that the ratio of the luminosity and angular-diameter distances is simply a function of redshift:

$$d_A/d_L = (1+z)^{-2} \ . \tag{1.4.12}$$

Therefore if we have measured the luminosity distance at a given redshift (and if we are convinced of the correctness of the Robertson–Walker metric), then we learn nothing additional about $a(t)$ if we also measure the angular diameter distance at that redshift. Neither galaxies nor supernovas have well-defined edges, so angular diameter distances are much less useful in studying the cosmological expansion than are luminosity distances. However, as we shall see, they play an important role in the theoretical analysis of both gravitational lenses in Chapter 9 and of the fluctuations in the cosmic microwave radiation background in Chapters 2 and 7. We will see in Section 8.1 that the observation of acoustic oscillations in the matter density may allow a measurement of yet another distance, a *structure distance*, equal to $a(t_0)r_1 = (1+z)d_A$.

1.5 Dynamics of expansion

All our results up to now have been very general, not depending on assumptions about the dynamics of the cosmological expansion. To go further we will need now to apply the gravitational field equations of Einstein, with various tentative assumptions about the cosmic energy density and pressure.

The expansion of the universe is governed by the Einstein field equations (B.71), which can be put in the convenient form

$$R_{\mu\nu} = -8\pi G S_{\mu\nu} \ , \tag{1.5.1}$$

where $R_{\mu\nu}$ is the *Ricci tensor*:

$$R_{\mu\nu} = \frac{\partial \Gamma^\lambda_{\lambda\mu}}{\partial x^\nu} - \frac{\partial \Gamma^\lambda_{\mu\nu}}{\partial x^\lambda} + \Gamma^\lambda_{\mu\sigma}\Gamma^\sigma_{\nu\lambda} - \Gamma^\lambda_{\mu\nu}\Gamma^\sigma_{\lambda\sigma} \ , \tag{1.5.2}$$

and $S_{\mu\nu}$ is given in terms of the energy-momentum tensor $T_{\mu\nu}$ by

$$S_{\mu\nu} \equiv T_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} T^{\lambda}{}_{\lambda} \ . \tag{1.5.3}$$

As we saw in Section 1.1, for the Robertson–Walker metric the components of the affine connection with two or three time indices all vanish, so

$$\begin{aligned} R_{ij} = & \frac{\partial \Gamma^k_{ki}}{\partial x^j} - \left[\frac{\partial \Gamma^k_{ij}}{\partial x^k} + \frac{\partial \Gamma^0_{ij}}{\partial t} \right] \\ & + \left[\Gamma^0_{ik} \Gamma^k_{j0} + \Gamma^k_{i0} \Gamma^0_{jk} + \Gamma^l_{ik} \Gamma^k_{jl} \right] \\ & - \left[\Gamma^k_{ij} \Gamma^l_{kl} + \Gamma^0_{ij} \Gamma^l_{0l} \right] \end{aligned} \tag{1.5.4}$$

$$R_{00} = \frac{\partial \Gamma^i_{i0}}{\partial t} + \Gamma^i_{0j} \Gamma^j_{0i} \tag{1.5.5}$$

We don't need to calculate $R_{i0} = R_{0i}$, because it is a three-vector, and therefore must vanish due to the isotropy of the Robertson–Walker metric. Using the formulas (1.1.17)–(1.1.19) for the non-vanishing components of the affine connection gives

$$\begin{aligned} & \frac{\partial \Gamma^0_{ij}}{\partial t} = \tilde{g}_{ij} \frac{d}{dt} (a\dot{a}) \ , \quad \Gamma^0_{ik} \Gamma^k_{j0} = \tilde{g}_{ij} \dot{a}^2 \ , \quad \Gamma^0_{ij} \Gamma^l_{0l} = 3 \tilde{g}_{ij} \dot{a}^2 \ , \\ & \frac{\partial \Gamma^i_{i0}}{\partial t} = 3 \frac{d}{dt} \left(\frac{\dot{a}}{a} \right) \ , \qquad \Gamma^i_{0j} \Gamma^j_{i0} = 3 \left(\frac{\dot{a}}{a} \right)^2 \ , \end{aligned} \tag{1.5.6}$$

where dots denote time derivatives. Using this in Eqs. (1.5.4) and (1.5.5), we find that the non-vanishing components of the Ricci tensor are

$$R_{ij} = \tilde{R}_{ij} - 2\dot{a}^2 \tilde{g}_{ij} - a\ddot{a} \tilde{g}_{ij} \ , \tag{1.5.7}$$

$$R_{00} = 3 \frac{d}{dt} \left(\frac{\dot{a}}{a} \right) + 3 \left(\frac{\dot{a}}{a} \right)^2 = 3 \frac{\ddot{a}}{a} \ , \tag{1.5.8}$$

where $\tilde{R}_{ij}$ is the purely spatial Ricci tensor

$$\tilde{R}_{ij} = \frac{\partial \Gamma^k_{ki}}{\partial x^j} - \frac{\partial \Gamma^k_{ij}}{\partial x^k} + \Gamma^l_{ik} \Gamma^k_{jl} - \Gamma^l_{ij} \Gamma^k_{kl} \ . \tag{1.5.9}$$

According to Eq. (1.1.19), the spatial components Γ^i_{jk} of the four-dimensional affine connection are here the same as those of the affine connection that would be calculated in three dimensions from the three-metric $\tilde{g}_{ij}$:

$$\Gamma^k_{ij} = K x^k \tilde{g}_{ij} \ . \tag{1.5.10}$$

To calculate $\tilde{R}_{ij}$, we use a trick used earlier in calculating particle trajectories: we calculate $\tilde{R}_{ij}$ where the calculation is simplest, at $\mathbf{x} = 0$, and express the result as a relation that is invariant under all transformations of the spatial coordinates, so that the homogeneity of the three-dimensional metric insures that this relation is valid everywhere. The spatial Ricci tensor at $\mathbf{x} = 0$ is

$$\tilde{R}_{ij} = \frac{\partial \Gamma^l_{li}}{\partial x^j} - \frac{\partial \Gamma^l_{ji}}{\partial x^l} = K\delta_{ij} - 3K\delta_{ij} = -2K\delta_{ij} \ . \tag{1.5.11}$$

At $\mathbf{x} = 0$ the spatial metric $\tilde{g}_{ij}$ is just δ_{ij}, so this can be rewritten as

$$\tilde{R}_{ij} = -2K\tilde{g}_{ij} \ , \tag{1.5.12}$$

which, since it is an equality between two three-tensors, is then true in all spatial coordinate systems, including systems in which the point $\mathbf{x} = 0$ is transformed into any other point. Hence Eq. (1.5.12) is true everywhere, and together with Eq. (1.5.7) gives

$$R_{ij} = -\left[2K + 2\dot{a}^2 + a\ddot{a}\right]\tilde{g}_{ij} \ . \tag{1.5.13}$$

We also need the values of S_{ij} and S_{00}. For this, we use Eq. (1.1.31) in the form

$$T_{00} = \rho \ , \qquad T_{i0} = 0 \ , \qquad T_{ij} = a^2 p\, \tilde{g}_{ij} \ , \tag{1.5.14}$$

where $\rho(t)$ and $p(t)$ are the proper energy density and pressure. Eq. (1.5.3) gives then

$$S_{ij} = T_{ij} - \frac{1}{2}\tilde{g}_{ij}a^2\left(T^k{}_k + T^0{}_0\right) = a^2 p\tilde{g}_{ij} - \frac{1}{2}a^2\tilde{g}_{ij}(3p-\rho) = \frac{1}{2}(\rho - p)\, a^2\, \tilde{g}_{ij} \ , \tag{1.5.15}$$

$$S_{00} = T_{00} + \frac{1}{2}\left(T^k{}_k + T^0{}_0\right) = \rho + \frac{1}{2}(3p - \rho) = \frac{1}{2}(\rho + 3p) \ , \tag{1.5.16}$$

and $S_{i0} = 0$. The Einstein equations are therefore

$$-\frac{2K}{a^2} - \frac{2\dot{a}^2}{a^2} - \frac{\ddot{a}}{a} = -4\pi G(\rho - p) \ , \tag{1.5.17}$$

$$\frac{3\ddot{a}}{a} = -4\pi G(3p + \rho) \ . \tag{1.5.18}$$

We can eliminate the second derivative terms by adding three times the first equation to the second, and find

$$\dot{a}^2 + K = \frac{8\pi G \rho a^2}{3} \ . \tag{1.5.19}$$

This is the fundamental Friedmann equation[1] governing the expansion of the universe.

The remaining information in Eqs. (1.5.17) and (1.5.18) just reproduces the conservation law (1.1.32):

$$\dot{\rho} = -\frac{3\dot{a}}{a}(\rho + p) \ . \tag{1.5.20}$$

(This should come as no surprise. Under all circumstances, the energy-momentum conservation law may be derived as a consequence of the Einstein field equations.) Given p as a function of ρ, we can solve Eq. (1.5.20) to find ρ as a function of a, and then use this in Eq. (1.5.19) to find a as a function of t.

There is another way of deriving Eq. (1.5.19), at least for the case of non-relativistic matter. Imagine a co-moving ball cut out from the expanding universe, with some typical galaxy at its center, and suppose it then emptied of the matter it contains. According to Birkhoff's theorem,[2] in any system that is spherically symmetric around some point, the metric in an empty ball centered on this point must be that of flat space. This holds whatever is happening outside the empty ball, as long as it is spherically symmetric. Now imagine putting the matter back in the ball, with a velocity proportional to distance from the center of symmetry, taken as $\mathbf{X} = 0$:

$$\dot{\mathbf{X}} = H(t)\mathbf{X} \ . \tag{1.5.21}$$

(Here the components X^i of $\mathbf{X}$ are ordinary Cartesian coordinates, not the co-moving coordinates x^i used in the Robertson–Walker metric. Note that this is the one pattern of velocities consistent with the principle of homogeneity: The velocity of a co-moving particle at $\mathbf{X}_1$ relative to a co-moving particle at $\mathbf{X}_2$ is $\dot{\mathbf{X}}_1 - \dot{\mathbf{X}}_2 = H(t)(\mathbf{X}_1 - \mathbf{X}_2)$.) The solution of Eq. (1.5.21) is

$$\mathbf{X}(t) = \left(\frac{a(t)}{a(t_0)}\right)\mathbf{X}(t_0) \ , \tag{1.5.22}$$

where $a(t)$ is the solution of the equation

$$\dot{a}(t)/a(t) = H(t) \ . \tag{1.5.23}$$

[1] A. Friedmann, *Z. Phys.* **16**, 377 (1922); *ibid* **21**, 326 (1924).

[2] G&C, Section 11.7.

As long as the radius of the ball is chosen to be not too large, the expansion velocity (1.5.21) of the matter we put into it will be non-relativistic, and the gravitational field will be weak, so that we can follow its motion using Newtonian mechanics. The kinetic energy of a co-moving particle of mass m at $\mathbf{X}$ is

$$K.E. = \frac{1}{2} m \dot{\mathbf{X}}^2 = \frac{m \dot{a}^2 \mathbf{X}^2}{2\, a^2} \ . \tag{1.5.24}$$

The mass interior to the position of the particle is $M(\mathbf{X}) = 4\pi\rho|\mathbf{X}|^3/3$, so the potential energy of the particle is

$$P.E. = -\frac{G\, m\, M(\mathbf{X})}{|\mathbf{X}|} = -\frac{4\pi\, G\, m\, \rho\, |\mathbf{X}|^2}{3} \ . \tag{1.5.25}$$

The condition of energy conservation thus tells us that

$$E = K.E. + P.E. = \frac{m\, |\mathbf{X}(t_0)|^2}{a^2(t_0)} \left[\frac{\dot{a}^2}{2} - \frac{4\pi\, G\, \rho\, a^2}{3} \right] = \text{constant} \ . \tag{1.5.26}$$

This is the same as Eq. (1.5.19), providing we identify the particle energy as

$$E = -\frac{K\, m\, |\mathbf{X}(t_0)|^2}{2\, a^2(t_0)} \ . \tag{1.5.27}$$

Particles will be able to escape to infinity if and only if $E \geq 0$, which requires $K = 0$ or $K = -1$. For $K = +1$ they have less than escape velocity, so the expansion eventually stops, and particles fall back toward each other.

Returning now to the relativistic formalism and an arbitrary dependence of ρ on a, even without knowing this dependence we can use Eq. (1.5.19) to draw important consequences about the general features of the expansion. First, as long as ρ remains positive, it is only possible for the expansion of the universe to stop if $K = +1$, the case of spherical geometry. Also, for any value of the Hubble constant $H_0 \equiv \dot{a}(t_0)/a(t_0)$, we may define a critical present density

$$\rho_{0,\text{crit}} \equiv \frac{3H_0^2}{8\pi G} = 1.878 \times 10^{-29}\, h^2 \ \text{g/cm}^3 \ , \tag{1.5.28}$$

where h is the Hubble constant in units of 100 km s^{-1} Mpc^{-1}. According to Eq. (1.5.19), whatever we assume about the constituents of the universe, the curvature constant K will be $+1$ or 0 or -1 according to whether the present density ρ_0 is greater than, equal to, or less than $\rho_{0,\text{crit}}$. *If* the quantity $3p + \rho$ is positive (as it is for any mixture of matter and radiation, in the absence of a vacuum energy density) then Eq. (1.5.18) shows that $\ddot{a}/a \leq 0$, so the

expansion must have started with $a = 0$ at some moment in the past; the present age of the universe t_0 is *less* than the Hubble time

$$t_0 < H_0^{-1} \ . \tag{1.5.29}$$

Also, if $K = +1$ and the expansion stops, then with $\ddot{a}/a \leq 0$ the universe will again contract to a singularity at which $a = 0$.

We can use Eq. (1.5.18) to give a general formula for the deceleration parameter $q_0 \equiv -\ddot{a}(t_0)a(t_0)/\dot{a}^2(t_0)$:

$$q_0 = \frac{4\pi G(\rho_0 + 3p_0)}{3H_0^2} = \frac{\rho_0 + 3p_0}{2\rho_{0,\text{crit}}} \ , \tag{1.5.30}$$

with a subscript 0 denoting a present value. If the present density of the universe were dominated by non-relativistic matter then $p_0 \ll \rho_0$, and the curvature constant K would be $+1$ or 0 or -1 according to whether $q_0 > \frac{1}{2}$ or $q_0 = \frac{1}{2}$ or $q_0 < \frac{1}{2}$. On the other hand, if the present density of the universe were dominated by relativistic matter then $p_0 = \rho_0/3$, and the critical value of the deceleration parameter at which $K = 0$ would be $q_0 = 1$. Finally, if the present density of the universe were dominated by vacuum energy then $p_0 = -\rho_0$, and the value of the deceleration parameter at which $K = 0$ would be $q_0 = -1$.

There is a peculiar aspect to these results. The contribution of non-relativistic and relativistic matter to the quantity ρa^2 in Eq. (1.5.19) grows as a^{-1} and a^{-2}, respectively, as $a \to 0$, so at sufficiently early times in the expansion we may certainly neglect the constant K, and Eq. (1.5.19) gives

$$\frac{\dot{a}^2}{a^2} \to \frac{8\pi G\rho}{3} \ . \tag{1.5.31}$$

That is, at these early times the density becomes essentially equal to the critical density $3H^2/8\pi G$, where $H \equiv \dot{a}/a$ is the value of the Hubble "constant" at those times. On the other hand, we will see later that the total energy density of the present universe is still a fair fraction of the critical density. How is it that after billions of years, ρ is still not very different from ρ_{crit}? This is sometimes called the *flatness problem*.

The simplest solution to the flatness problem is just that we are in a spatially flat universe, in which $K = 0$ and ρ is always precisely equal to ρ_{crit}. A more popular solution is provided by the inflationary theories discussed in Chapter 4. In these theories K may not vanish, and ρ may not start out close to ρ_{crit}, but there is an early period of rapid growth in which ρ/ρ_{crit} rapidly approaches unity. In inflationary theories it is expected though not required that ρ should now be very close to ρ_{crit}, in which case it is a good approximation to take $K = 0$.

For $K = 0$ we get very simple solutions to Eq. (1.5.19) in the three special cases listed in Section 1.1:

Non-relativistic matter: Here $\rho = \rho_0(a/a_0)^{-3}$, and the solution of Eq. (1.5.19) with $K = 0$ is

$$a(t) \propto t^{2/3} \ . \tag{1.5.32}$$

This gives $q_0 \equiv -a\ddot{a}/\dot{a}^2 = 1/2$, and a simple relation between the age of the universe and the Hubble constant

$$t_0 = \frac{2}{3H_0} = 6.52 \times 10^9 \, h^{-1} \, \text{yr} \ . \tag{1.5.33}$$

Eqs. (1.5.32) and (1.5.18) show that for $K = 0$, the energy density at time t is $\rho = 1/6\pi G t^2$. This is known as the *Einstein–de Sitter model*. It was for many years the most popular cosmological model, though as we shall see, the age (1.5.33) is uncomfortably short compared with the ages of certain stars.

Relativistic matter: Here $\rho = \rho_0(a/a_0)^{-4}$, and the solution of Eq. (1.5.19) with $K = 0$ is

$$a(t) \propto \sqrt{t} \ . \tag{1.5.34}$$

This gives $q_0 = +1$, while the age of the universe and the Hubble constant are related by

$$t_0 = \frac{1}{2H_0} \ . \tag{1.5.35}$$

The energy density at time t is $\rho = 3/32\pi G t^2$.

Vacuum energy: Lorentz invariance requires that in locally inertial coordinate systems the energy-momentum tensor $T_V^{\mu\nu}$ of the vacuum must be proportional to the Minkowski metric $\eta^{\mu\nu}$ (for which $\eta^{ij} = \eta_{ij} = \delta_{ij}$, $\eta^{i0} = \eta_{i0} = \eta^{0i} = \eta_{0i} = 0$, $\eta^{00} = \eta_{00} = -1$), and so in general coordinate systems $T_V^{\mu\nu}$ must be proportional to $g^{\mu\nu}$. Comparing this with Eq. (B.43) shows that the vacuum has $p_V = -\rho_V$, so that $T_V^{\mu\nu} = -\rho_V \, g^{\mu\nu}$. In the absence of any other form of energy this would satisfy the conservation law $0 = T_V^{\mu\nu}{}_{;\mu} = g^{\mu\nu}\partial\rho_V/\partial x^\mu$, so that ρ_V would be a constant, independent of spacetime position. Eq. (1.5.19) for $K = 0$ requires that $\rho_V > 0$, and has the solutions

$$a(t) \propto \exp(Ht) \tag{1.5.36}$$

where H is the Hubble constant, now really a constant, given by

$$H = \sqrt{\frac{8\pi G \rho_V}{3}} \ . \tag{1.5.37}$$

Here $q_0 = -1$, and the age of the universe in this case is infinite. This is known as the *de Sitter model*.[3] Of course, there is *some* matter in the universe, so even if the energy density of the universe is now dominated by a constant vacuum energy, there was a time in the past when matter and/or radiation were more important, and so the expansion has a finite age, although greater than it would be without a vacuum energy.

More generally, for arbitrary K and a mixture of vacuum energy and relativistic and non-relativistic matter, making up fractions Ω_Λ, Ω_M, and Ω_R of the critical energy density,[4] we have

$$\rho = \frac{3\,H_0^2}{8\pi G}\left[\Omega_\Lambda + \Omega_M\left(\frac{a_0}{a}\right)^3 + \Omega_R\left(\frac{a_0}{a}\right)^4\right] , \tag{1.5.38}$$

where the present energy densities in the vacuum, non-relativistic matter, and and relativistic matter (i.e., radiation) are, respectively,

$$\rho_{V0} \equiv \frac{3H_0^2\Omega_\Lambda}{8\pi G} , \quad \rho_{M0} \equiv \frac{3H_0^2\Omega_M}{8\pi G} , \quad \rho_{R0} \equiv \frac{3H_0^2\Omega_R}{8\pi G} , \tag{1.5.39}$$

and, according to Eq. (1.5.19),

$$\Omega_\Lambda + \Omega_M + \Omega_R + \Omega_K = 1 , \quad \Omega_K \equiv -\frac{K}{a_0^2H_0^2} . \tag{1.5.40}$$

Using this in Eq. (1.5.19) gives

$$\begin{aligned} dt &= \frac{dx}{H_0 x\sqrt{\Omega_\Lambda + \Omega_K x^{-2} + \Omega_M x^{-3} + \Omega_R x^{-4}}} \\ &= \frac{-dz}{H_0(1+z)\sqrt{\Omega_\Lambda + \Omega_K(1+z)^2 + \Omega_M(1+z)^3 + \Omega_R(1+z)^4}} , \end{aligned} \tag{1.5.41}$$

where $x \equiv a/a_0 = 1/(1+z)$. Therefore, if we define the zero of time as corresponding to an infinite redshift, then the time at which light was emitted that reaches us with redshift z is given by

$$t(z) = \frac{1}{H_0}\int_0^{1/(1+z)} \frac{dx}{x\sqrt{\Omega_\Lambda + \Omega_K x^{-2} + \Omega_M x^{-3} + \Omega_R x^{-4}}} . \tag{1.5.42}$$

[3] W. de Sitter, *Proc. Roy. Acad. Sci.* (Amsterdam), **19**, 1217 (1917); *ibid.* **20**, 229 (1917); *ibid.* **20**, 1309 (1917); *Mon. Not. Roy. Astron. Soc.*, **78**, 2 (1917).

[4] The use of the symbol Ω_Λ instead of Ω_V for the ratio of the vacuum energy density to the critical energy density has become standard, because of a connection with the cosmological constant discussed in a historical note below.

In particular, by setting $z = 0$, we find the present age of the universe:

$$t_0 = \frac{1}{H_0}\int_0^1 \frac{dx}{x\sqrt{\Omega_\Lambda + \Omega_K x^{-2} + \Omega_M x^{-3} + \Omega_R x^{-4}}}\ . \qquad (1.5.43)$$

In order to calculate luminosity or angular diameter distances, we also need to know the radial coordinate $r(z)$ of a source that is observed now with redshift z. According to Eqs. (1.2.2) and (1.5.41), this is

$$\begin{aligned} r(z) &= S\left[\int_{t(z)}^{t_0} \frac{dt}{a(t)}\right] \\ &= S\left[\frac{1}{a_0 H_0}\int_{1/(1+z)}^1 \frac{dx}{x^2\sqrt{\Omega_\Lambda + \Omega_K x^{-2} + \Omega_M x^{-3} + \Omega_R x^{-4}}}\right], \end{aligned}$$

where

$$S[y] \equiv \begin{cases} \sin y & K = +1 \\ y & K = 0 \\ \sinh y & K = -1\ . \end{cases}$$

This can be written more conveniently by using Eq. (1.5.40) to express $a_0 H_0$ in terms of Ω_K. We then have a single formula

$$\begin{aligned} a_0 r(z) &= \frac{1}{H_0 \Omega_K^{1/2}} \\ &\quad \times \sinh\left[\Omega_K^{1/2}\int_{1/(1+z)}^1 \frac{dx}{x^2\sqrt{\Omega_\Lambda + \Omega_K x^{-2} + \Omega_M x^{-3} + \Omega_R x^{-4}}}\right], \end{aligned} \qquad (1.5.44)$$

which can be used for any curvature. (Eq. (1.5.43) has a smooth limit for $\Omega_K \to 0$, which gives the result for zero curvature. Also, for $\Omega_K < 0$, the argument of the hyperbolic sine is imaginary, and we can use $\sinh ix = i \sin x$.) Using Eq. (1.5.44) in Eq. (1.4.3) gives the luminosity distance of a source observed with redshift z as

$$\begin{aligned} d_L(z) = a_0 r(z)(1+z) &= \frac{1+z}{H_0 \Omega_K^{1/2}} \\ &\quad \times \sinh\left[\Omega_K^{1/2}\int_{1/(1+z)}^1 \frac{dx}{x^2\sqrt{\Omega_\Lambda + \Omega_K x^{-2} + \Omega_M x^{-3} + \Omega_R x^{-4}}}\right]. \end{aligned} \qquad (1.5.45)$$

For $K = 0$ we have $\Omega_K = 0$ and Eq. (1.5.45) becomes

$$d_L(z) = a_0 r_1 (1+z) = \frac{1+z}{H_0} \int_{1/(1+z)}^{1} \frac{dx}{x^2 \sqrt{\Omega_\Lambda + \Omega_M x^{-3} + \Omega_R x^{-4}}} \,. \tag{1.5.46}$$

As we will see in Section 2.1, Ω_R is much less than Ω_M, and the integral (1.5.46) converges at its lower bound for $z \to \infty$ whether or not Ω_R vanishes, so it is a good approximation to take $\Omega_R = 0$ here.

It is of some interest to express the deceleration parameter q_0 in terms of the Ωs. The p/ρ ratio w for vacuum, matter, and radiation is -1, 0, and $1/3$, respectively, so Eq. (1.5.39) gives the present pressure as

$$p_0 = \frac{3H_0^2}{8\pi G} \left(-\Omega_\Lambda + \frac{1}{3}\Omega_R \right) \,. \tag{1.5.47}$$

Eq. (1.5.30) then gives

$$q_0 = \frac{4\pi G(3p_0 + \rho_0)}{3H_0^2} = \frac{1}{2} \left(\Omega_M - 2\Omega_\Lambda + 2\Omega_R \right) \,. \tag{1.5.48}$$

One of the reasons for our interest in the values of Ω_K, Ω_M, etc. is that they tell us whether the present expansion of the universe will ever stop. According to Eq. (1.5.38), the expansion can only stop if there is a real root of the cubic equation

$$\Omega_\Lambda u^3 + \Omega_K u + \Omega_M = 0 \,, \tag{1.5.49}$$

where $u \equiv a(t)/a(t_0)$ is greater than one. (We are ignoring radiation here, since it will become even less important as the universe expands.) This expression has the value $+1$ for $u = 1$. If $\Omega_\Lambda < 0$ then the left-hand side of Eq. (1.5.49) becomes negative for sufficiently large u, so it must take the value zero at some intermediate value of u, and the expansion will stop when this value of u is reached. Even for $\Omega_\Lambda \geq 0$ it is still possible for the expansion to stop, provided $\Omega_K = 1 - \Omega_\Lambda - \Omega_M$ is sufficiently negative (which, among other things, requires that $K = +1$).

Historical Note 1: If we express the total energy momentum tensor $T_{\mu\nu}$ as the sum of a possible vacuum term $-\rho_V g_{\mu\nu}$ and a term $T^M_{\mu\nu}$ arising from matter (including radiation), then the Einstein equations take the form

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R^\lambda{}_\lambda = -8\pi G T^M_{\mu\nu} + 8\pi G \rho_V g_{\mu\nu} \,. \tag{1.5.50}$$

Thus the effect of a vacuum energy is equivalent to modifying the Einstein field equations to read

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R - \Lambda g_{\mu\nu} = -8\pi G T^M_{\mu\nu} \,, \tag{1.5.51}$$

where

$$\Lambda = 8\,\pi\; G\,\rho_V \;. \tag{1.5.52}$$

The quantity Λ is known as the *cosmological constant*. It was introduced into the field equation by Einstein in 1917 in order to satisfy a condition that at the time was generally regarded as essential, that the universe should be static.[5] According to Eqs. (1.5.18) and (1.5.19), a static universe is only possible if $3p + \rho = 0$ and $K = 8\pi G\rho a^2/3$. If the contents of the universe are limited to vacuum energy and non-relativistic matter, then $\rho = \rho_M + \rho_V$, $p = -\rho_V$, and $\rho_M \geq 0$. It follows that $\rho_M = 2\rho_V \geq 0$, so $K > 0$, which by convention means $K = +1$, so that a takes the value $a_E = 1/\sqrt{8\pi G\rho_V} = 1/\sqrt{\Lambda}$. This is known as the *Einstein model.*

Einstein did not realize it, but his cosmology was unstable: If a is a little less than a_E then ρ_M is a little larger than $2\rho_V$, so Eq. (1.5.18) shows that $\ddot{a}/a < 0$, and a thus begins to decrease. Likewise, if a is a little greater than a_E then it begins to increase. The models with $K = +1$ and $\Lambda > 0$ in which a starts at the Einstein radius $a = a_E$ with $\rho_M = 2\rho_V$ and then expands to infinity (or starts at $a = 0$ and approaches a_E as $t \to \infty$ with just enough matter so that $\rho_M = 2\rho_V$ at the Einstein radius), are known as *Eddington–Lemaître models*.[6] There are also models with $K = +1$ and a little more matter, that start at $a = 0$, spend a long time near the Einstein radius, and then expand again to infinity, approaching a de Sitter model. These are known as *Lemaître models*.[7]

Oddly, de Sitter also invented his cosmological model (with $a \propto \exp(Ht)$) in order to satisfy a supposed need for a static universe. He originally proposed a time-independent metric, given by

$$d\tau^2 = (1 - r^2/R^2)dt^2 - \frac{dr^2}{1 - r^2/R^2} - r^2\,d\theta^2 - r^2\,\sin^2\theta\,d\phi^2 \,, \tag{1.5.53}$$

[5] A. Einstein, *Sitz. Preuss. Akad. Wiss.* 142 (1917). For an English translation, see *The Principle of Relativity* (Methuen, 1923; reprinted by Dover Publications, New York, 1952), p. 35.

[6] A. S. Eddington, *Mon. Not. Roy. Astron. Soc.* **90**, 668 (1930); G. Lemaître, *Ann. Soc. Sci. Brux.* **A47**, 49 (1927); *Mon. Not. Roy. Astron. Soc.* **91**, 483 (1931). The interpretation of the cosmological constant in terms of vacuum energy was stated by Lemaître in *Proc. Nat. Acad. Sci.* **20**, 12L (1934).

[7] G. Lemaître, *op. cit.*

or equivalently, setting $r = R \sin\chi$,

$$d\tau^2 = \cos^2\chi\, dt^2 - R^2\Big(d\chi^2 + \sin^2\chi\, d\theta^2 + \sin^2\chi\ \sin^2\theta\, d\phi^2\Big)\,, \quad (1.5.54)$$

with $R = \sqrt{3/\Lambda}$ constant. De Sitter did not realize at first that this metric has $\Gamma^i_{00} \neq 0$, so that his coordinate system was not co-moving.[8] Only later was it noticed that, using co-moving spatial coordinates and cosmological standard time, de Sitter's model is equivalent to a Robertson–Walker metric with $K = 0$ and $a \propto \exp(t/R)$.

After the discovery of the expansion of the universe, cosmologists lost interest in a static universe, and Einstein came to regret his introduction of a cosmological constant, calling it his greatest mistake. But as we shall see in the next section, there are theoretical reasons to expect a non-vanishing vacuum energy, and there is observational evidence that in fact it does not vanish. Einstein's mistake was not that he introduced the cosmological constant — it was that he thought it was a mistake.

Historical Note 2: There is a cosmological model due to Bondi and Gold[9] and in a somewhat different version to Hoyle,[10] known as the *steady state theory*. In this model nothing physical changes with time, so the Hubble constant really is constant, and hence $a(t) \propto \exp(Ht)$, just as in the de Sitter model. To keep the curvature constant, it is necessary to take $K = 0$. In this model new matter must be continually created to keep ρ constant as the universe expands. Since the discovery of the cosmic microwave background (discussed in Chapter 2) the steady state theory in its original form has been pretty well abandoned.

1.6 Distances at large redshift: Accelerated expansion

We now return to our account of the measurement of distances as a function of redshift, considering now redshifts $z > 0.1$, which are large enough so that we can ignore the peculiar motions of the light sources, and also large enough so that we need to take into account the effects of cosmological expansion on distance determination.

For many years, the chief "standard candles" used at large redshift were the brightest galaxies in rich clusters. It is now well established that the

[8] A. S. Eddington, *The Mathematical Theory of Relativity*, 2nd ed. (Cambridge University Press, Cambridge, 1924), Section 70. It is interesting that Eddington interpreted Slipher's observation that most spiral nebulae exhibit redshifts rather than blueshifts in terms of the de Sitter model, rather than Friedmann's models.

[9] H. Bondi and T. Gold, *Mon. Not. Roy. Astron. Soc.* **108**, 252 (1948).

[10] F. Hoyle, *Mon. Not. Roy. Astron. Soc.* **108**, 372 (1948), *ibid.* **109**, 365 (1949).

absolute luminosity of these brightest galaxies evolves significantly over cosmological time scales. There are also severe selection effects: there is a tendency to pick out larger clusters with brightest galaxies of higher absolute luminosity at large distances. The evolution of brightest galaxies is of interest in itself, and continues to be the object of astronomical study,[1] but the use of these galaxies as distance indicators has been pretty well abandoned. Similarly, although the Tully–Fisher relation discussed in Section 1.3 has been applied to galaxies with redshifts of order unity, at these redshifts it is used to study galactic evolution, rather than to measure cosmological parameters.[2]

Fortunately, the Type Ia supernovae discussed in Section 1.3 provide an excellent replacement as standard candles.[3] They are very bright; the peak blue absolute magnitude averages about -19.2, which compares well with the absolute magnitude -20.3 estimated for our own galaxy. Also, as described in Section 1.3, a Type Ia supernova typically occurs when a white dwarf member of a binary pair has accreted just enough mass to push it over the Chandrasekhar limit, so that the nature of the explosion does not depend much on when in the history of the universe this happens, or on the mass with which the white dwarf started or the nature of the companion star. But it might depend somewhat on the metallicity (the proportion of elements heavier than helium) of the white dwarf, which can depend on the epoch of the explosion. The absolute luminosity of Type Ia supernovae is observed to vary with environmental conditions, but fortunately in the use of supernovae as distance indicators the bulk of this variation is correctable empirically.

Observations of Type Ia supernovae have been compared with theoretical predictions (equivalent to Eq. (1.5.45)) for luminosity distance as a function of redshift at about the same time by two groups: The Supernova Cosmology Project[4] and the High-z Supernova Search Team.[5]

---

[1] See, e.g., D. Zaritsky *et al.*, in *Proceedings of the Sesto 2001 Conference on Tracing Cosmic Evolution with Galaxy Clusters* [astro-ph/0108152]; S. Brough *et al.*, in *Proceedings of the Sesto 2001 Conference on Tracing Cosmic Evolution with Galaxy Clusters* [astro-ph/0108186].

[2] N. P. Vogt *et al.*, *Astrophys. J.* **465**, l15 (1996). For a review and more recent references, see A. Aragón-Salmanca, in *Galaxy Evolution Across the Hubble Time – Proceedings of I. A. U. Symposium 235*, eds. F. Combes and J. Palous [astro-ph/0610587].

[3] For reviews, see S. Perlmutter and B. P. Schmidt, in *Supernovae & Gamma Ray Bursts*, ed. K. Weiler (Springer, 2003) [astro-ph/0303428]; P. Ruiz-Lapuente, *Astrophys. Space Sci.* **290**, 43 (2004) [astro-ph/0304108]; A. V. Filippenko, in *Measuring and Modeling of the Universe* (Carnegie Observatories Astrophysics Series, Vol 2., Cambridge University Press) [astro-ph/0307139]; Lect. Notes Phys. **645**, 191 (2004) [astro-ph/0309739]; N. Panagia, *Nuovo Cimento* **B 210**, 667 (2005) [astro-ph/0502247].

[4] S. Perlmutter *et al.*, *Astrophys. J.* **517**, 565 (1999) [astro-ph/9812133]. Also see S. Perlmutter *et al.*, *Nature* **391**, 51 (1998) [astro-ph/9712212].

[5] A. G. Riess *et al.*, *Astron. J.* **116**, 1009 (1998) [astro-ph/9805201]. Also see B. Schmidt *et al.*, *Astrophys. J.* **507**, 46 (1998) [astro-ph/9805200].

The Supernova Cosmology Project analyzed the relation between apparent luminosity and redshift for 42 Type Ia supernovae, with redshifts z ranging from 0.18 to 0.83, together with a set of closer supernovae from another supernova survey, at redshifts below 0.1. Their original results are shown in Figure 1.1.

With a confidence level of 99%, the data rule out the case $\Omega_\Lambda = 0$ (or $\Omega_\Lambda < 0$). For a flat cosmology with $\Omega_K = \Omega_R = 0$, so that $\Omega_\Lambda + \Omega_M = 1$, the data indicate a value

$$\Omega_M = 0.28^{+0.09}_{-0.08}\ (1\sigma\ \text{statistical})^{+0.05}_{-0.04}\ \text{(identified systematics)}$$

(These results are independent of the Hubble constant or the absolute calibration of the relation between supernova absolute luminosity and time scale, though they do depend on the shape of this relation.) This gives the age (1.5.43) as

$$t_0 = 13.4^{+1.3}_{-1.0} \times 10^9 \left(\frac{70\ \text{km s}^{-1}\ \text{Mpc}^{-1}}{H_0}\right)\ \text{yr}\ .$$

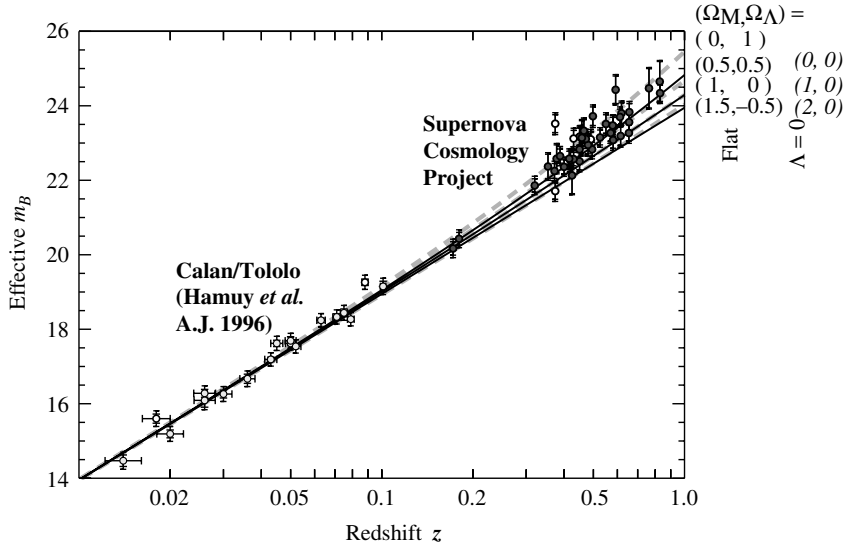

Figure 1.1: Evidence for dark energy, found in 1998 by the Supernova Cosmology Project, from S. Perlmutter *et al.*, *Astrophys. J.* **517**, 565 (1999) [astro-ph/9812133]. Here the effective blue apparent magnitude (corrected for variations in absolute magnitude, as indicated by supernova light curves) are plotted versus redshift for 42 high redshift Type Ia supernovae observed by the Supernova Cosmology Project, along with 18 lower redshift Type Ia supernovae from the Calán–Tololo Supernovae Survey. Horizontal bars indicate the uncertainty in cosmological redshift due to an assumed peculiar velocity uncertainty of 300 km sec^{-1}. Dashed and solid curves give the theoretical effective apparent luminosities for cosmological models with $\Omega_K = 0$ or $\Omega_\Lambda = 0$, respectively, and various possible values of Ω_M.

For $\Omega_M = 0.28$ and $\Omega_\Lambda = 1 - \Omega_M$, Eq. (1.5.48) gives a negative deceleration parameter, $q_0 = -0.58$, indicating that *the expansion of the universe is accelerating*.

The High-z Supernova Search Team originally studied 16 Type Ia supernovae of high redshift (with redshifts ranging from 0.16 to 0.97), including 2 from the Supernova Cosmology Project, together with 34 nearby supernovae, and conclude that $\Omega_\Lambda > 0$ at the 99.7% confidence level, with no assumptions about spatial curvature. Their original results are shown in Figure 1.2.

Their best fit for a flat cosmology is $\Omega_M = 0.28 \pm 0.10$ and $\Omega_\Lambda = 1 - \Omega_M$, giving an age of about $(14.2 \pm 1.5) \times 10^9$ years, including uncertainties in the Cepheid distance scale. Assuming only $\Omega_M \geq 0$, and with a conservative

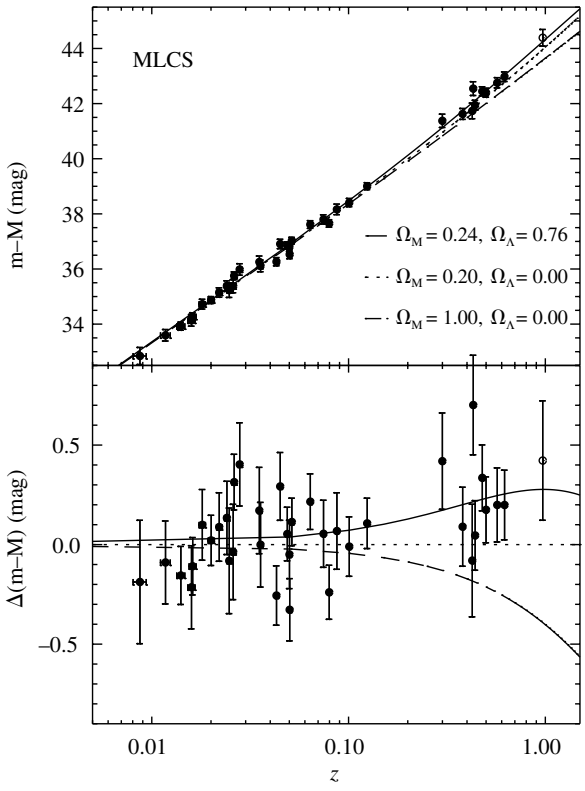

Figure 1.2: Evidence for dark energy, found in 1998 by the High-z Supernova Search Team, from A. G. Riess *et al.*, *Astron. J.* **116**, 1009 (1998) [astro-ph/9805201]. In the upper panel distance modulus is plotted against redshift for a sample of Type Ia supernovae. The curves give the theoretical results for two cosmologies with $\Omega_\Lambda = 0$ and a good-fit flat cosmology with $\Omega_M = 0.24$ and $\Omega_\Lambda = 0.76$. The bottom panel shows the difference between data and a formerly popular Einstein–de Sitter model with $\Omega_M = 0.2$ and $\Omega_\Lambda = 0$, represented by the horizontal dotted line.

fitting method, with 99.5% confidence they conclude that $q_0 < 0$, again strongly indicating an accelerated expansion. Including 8 new supernovae in a sample of 230 supernovae of Type Ia gave[6] $1.4\,\Omega_M - \Omega_\Lambda = -0.35 \pm 0.14$, providing further evidence that $\Omega_\Lambda > 0$. The case for vacuum energy was then strenghtened when the Supernova Cosmology Project,[7] including a new set of supernova, found for a flat universe that $\Omega_\Lambda = 0.75^{+0.06}_{-0.07}(\text{stat.}) \pm 0.032(\text{syst.})$.

Both groups agree that their results are chiefly sensitive to a linear combination of Ω_Λ and Ω_M, given as $0.8\Omega_M - 0.6\Omega_\Lambda$ by the Supernova Cosmology Project and $\Omega_M - \Omega_\Lambda$ or $1.4\Omega_M - \Omega_\Lambda$ by the High-z Supernova Search Team. The minus sign in these linear combinations, as in Eq. (1.5.48), reflects the fact that matter and vacuum energy have opposite effects on the cosmological acceleration: Matter causes it to slow down, while a positive vacuum energy causes it to accelerate. The negative values found for these linear combinations shows the presence of a component of energy something like vacuum energy, with $p \simeq -\rho$. This is often called *dark energy*.

Incidentally, these linear combinations of Ω_Λ and Ω_M are quite different from the expression $\Omega_M/2 - \Omega_\Lambda$, which according to Eq. (1.5.48) gives the deceleration parameter q_0 that was the target of much cosmological work of the past. Thus the observations of Type Ia supernovae at cosmological distances should not be regarded as simply measurements of q_0.

The High-z Supernova Search Team subsequently began to use the same survey observations to follow the time development of supernovae that were used to find them.[8] They discovered 23 new high redshift supernovae of Type Ia, including 15 with $z > 0.7$. Using these new supernovae along with the 230 used earlier by Tonry *et al.*, and with the assumption that $\Omega_M + \Omega_\Lambda = 1$, they found the best-fit values $\Omega_M = 0.33$ and $\Omega_\Lambda = 0.67$.

The crucial feature of the supernova data that indicates that $\Omega_\Lambda > \Omega_M$ is that the apparent luminosity of Type Ia supernovae falls off more rapidly with redshift than would be expected in an Einstein–de Sitter cosmology with $\Omega_M = 1$ and $\Omega_\Lambda = 0$. We can see the effect of vacuum energy on apparent luminosity by comparing the luminosity distance calculated in two extreme cases, both with no matter or radiation. For a vacuum-dominated flat model with $\Omega_\Lambda = 1$ and $\Omega_K = \Omega_M = \Omega_R = 0$, Eq. (1.5.46) gives

$$d_L(z) = \frac{z + z^2}{H_0} \quad \text{(vacuum dominated)}\,, \tag{1.6.1}$$

[6] J. L. Tonry *et al.*, *Astrophys. J.* **594**, 1 (2003) [astro-ph/0305008].
[7] R. Knop *et al.*, *Astrophys. J.* **598**, 102 (2003) [astro-ph/0309368].
[8] B. J. Barris *et al.*, *Astrophys. J.* **502**, 571 (2004) [astro-ph/0310843].

while for an empty model with $\Omega_K = 1$ and $\Omega_\Lambda = \Omega_M = \Omega_R = 0$, Eq, (1.5.46) gives

$$d_L(z) = \frac{z + z^2/2}{H_0} \quad \text{(empty)} , \tag{1.6.2}$$

Evidently for all z, vacuum energy increases the luminosity distance. The same increase is seen if we compare the more realistic case $\Omega_\Lambda = 0.7$, $\Omega_M = 0.3$, $\Omega_K = \Omega_R = 0$ with the corresponding case without vacuum energy and $\Omega_K = 0.7$, $\Omega_M = 0.3$, $\Omega_\Lambda = \Omega_R = 0$, as can be seen in Figure 1.3.

Both the Supernova Cosmology Project and the High-z Supernova Search Team found that curve of measured luminosity distances *vs.* redshift of Type Ia supernovae was closer to the upper than the lower curve in Figure 1.3. Indeed, according to Eq. (1.4.9), the negative value of q_0 found by all groups corresponds to the fact that the apparent luminosity of the type Ia supernovae seen at moderate redshifts is *less* than in the empty model, for which $q_0 = 0$, in contrast with what had been expected, that the expansion is dominated by matter, in which case we would have had $q_0 > 0$, and the apparent luminosities at moderate redshifts would have been larger than for $q_0 = 0$.

The connection between an accelerating expansion and a reduced apparent luminosity can be understood on the basis of the naive Newtonian cosmological model discussed in Section 1.5. In this model, the redshift we observe from a distant galaxy depends on the speed the galaxy had when the light we observe was emitted, but the apparent luminosity is inversely proportional to the square of the distance of this galaxy *now*, because the galaxy's light is now spread over an area equal to 4π times this squared distance. If the galaxies we observe have been traveling at constant speed since the beginning, as in the empty model, then the distance of any galaxy from

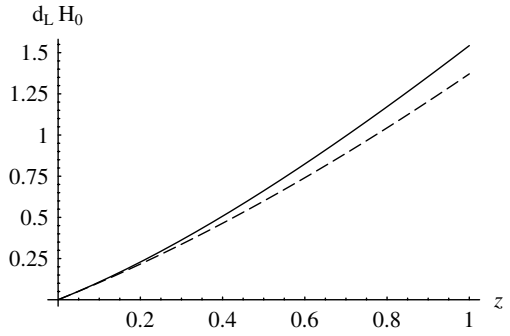

Figure 1.3: Luminosity distance versus redshift for two cosmological models. The upper solid curve is for the case $\Omega_\Lambda = 0.7$, $\Omega_M = 0.3$, $\Omega_K = \Omega_R = 0$; the lower dashed curve is for an empty model, with $\Omega_K = 1$, $\Omega_\Lambda = \Omega_M = \Omega_R = 0$. The vertical axis gives the luminosity distance times the Hubble constant.

us now would be proportional to its speed when the light was emitted. In the absence of a vacuum energy, we would expect the galaxies to be slowing down under the influence of their mutual gravitational attraction, so that the speed we observe would be greater than the speed they have had since the light was emitted, and their distances now would therefore be smaller than they would be if the speeds were constant. Thus in the absence of vacuum energy we would expect an enhanced apparent luminosity of the supernovae in these galaxies. In fact, it seems that the luminosity distances of supernovae are *larger* than they would be if the speeds of their host galaxies were constant, indicating that these galaxies have not been slowing down, but speeding up. This is just the effect that would be expected from a positive vacuum energy.

Of course, it is also possible that the reduction in apparent luminosity is due to absorption or scattering of light by intervening material rather than an accelerated expansion. It is possible to distinguish such effects from a true increase in luminosity distance by the change in the apparent color produced by such absorption or scattering, but this is a complicated business.[9] This concern has been allayed by careful color measurements.[10] But it is still possible to invent intergalactic media (so-called gray dust) that would reduce the apparent luminosity while leaving the color unchanged.

This concern has been largely put to rest, first by the study[11] of the supernovae SN1997ff found in the Hubble Deep Field[12] in a galaxy with a redshift $z = 1.7 \pm 0.1$, the greatest yet found for any supernova, and then by the discovery and analysis by a new team, the Higher-z Supernova Team,[13] of 16 new Type Ia supernovae, of which six have $z > 1.25$. These redshifts are so large that during a good part of the time that the light from these supernovae has been on its way to us, the energy density of the universe would have been dominated by matter rather than by a cosmological constant, and so the expansion of the universe would have been decelerating rather than accelerating as at present. Thus if the interpretation of the results of the two groups at smaller redshifts in terms of Ω_M and Ω_Λ is correct, then the apparent luminosity of these supernovae should be *larger* than would be given by a linear relation between luminosity distance and redshift, a result that could not be produced by absorption or scattering of light. We see this in Figure 1.4, which shows the difference between the luminosity distance (in units H_0^{-1}) for the realistic case with $\Omega_\Lambda = 0.7$, $\Omega_M = 0.3$, $\Omega_K = \Omega_R = 0$ and for

[9] See e.g., A. Aguirre, *Astrophys. J.* **525**, 583 (1999) [astro-ph/9904319].

[10] R. Knop *et al.*, ref. 7; also see M. Sullivan *et al.*, *Mon. Not. Roy. Astron. Soc.* **340**, 1057 (2003) [astro-ph/0211444].

[11] A. G. Riess *et al.*, *Astrophys. J.* **560**, 49 (2001) [astro-ph/0104455].

[12] R. L. Gilliland, P. E. Nugent, and M. M. Phillips, *Astrophys. J.* **521**, 30 (1999).

[13] A. G. Riess *et al.*, *Astrophys. J.* **607**, 665 (2004) [astro-ph/0402512].

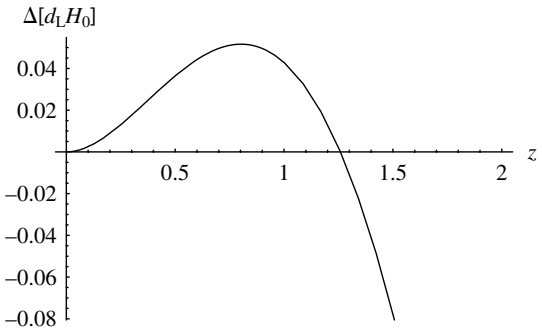

Figure 1.4: The luminosity distance times H_0 for the realistic case $\Omega_\Lambda = 0.7$, $\Omega_M = 0.3$, $\Omega_K = \Omega_R = 0$, minus its value for the empty case $\Omega_K = 1$, $\Omega_\Lambda = \Omega_M = \Omega_R = 0$, plotted against redshift.

the empty model with $\Omega_K = 1$, $\Omega_M = \Omega_K = \Omega_R = 0$. We see that luminosity distances for the realistic model are greater than for the empty model for moderate redshift, but become less than for the empty model for $z > 1.25$. This is just what is seen. The apparent luminosity of all supernovae is consistent with the parameters $\Omega_M \approx 0.3$, $\Omega_\Lambda \approx 0.7$ found in the 1998 studies, but not consistent with what would be expected for gray dust and $\Omega_\Lambda = 0$. These conclusions have subsequently been strengthened by the measurement of luminosity distances of additional Type Ia supernovae with redshifts near 0.5.[14] In 2006 Riess *et al.*[15] announced the discovery with the Hubble Space Telescope of 21 new Type Ia supernovae, which included 13 supernovae with redshifts $z \geq 1$ measured spectroscopically (not just photometrically). Their measured luminosity distances and redshifts, together with data on previously discovered Type Ia supernovae, gave further evidence of a transition from a matter-dominated to a vacuum energy-dominated expansion, and showed that the pressure/density ratio of the vacuum energy for $z > 1$ is consistent with $w = -1$, and not rapidly evolving.

Another serious concern arises from the possibility that the absolute luminosity of Type Ia supernovae may depend on when the supernovae occur. Because Type Ia supernovae occur at a characteristic moment in the history of a star, evolution effects on the luminosities of these supernovae are not expected to be as important as for whole galaxies, which at great distances are seen at an earlier stage in their history.[16] Even so, the absolute luminosity of a Type Ia supernova is affected by the chemical composition

[14] A. Clocchiatti *et al.*, *Astrophys. J.* **642**, 1 (2006) [astro-ph/0510155].

[15] A. Riess *et al.*, *Astrophys. J.* **659**, 98 (2007) [astro-ph/0611572].

[16] D. Branch, S. Perlmutter, E. Baron, and P. Nugent, contribution to the *Supernova Acceleration Probe Yellow Book* (Snowmass, 2001) [astro-ph/0109070].

of the two progenitor stars of the supernova, which is in turn affected by the evolution of the host galaxy.[17] Such effects are mitigated by taking account of the correlation of supernova absolute luminosity with decay time and with intrinsic color, both of which presumably depend on the progenitor's chemical composition. Also, evidence for dark energy has been found in studies of subsets of Type Ia supernovae found in very different environments with very different histories.[18] The study of the seven supernovae with $z > 1.25$ mentioned above rules out models with $\Omega_\Lambda = 0$ and any sort of dramatic monotonic evolution of supernovae absolute luminosities that would mimic the effects of dark energy.

There are other effects that might possibly impact the observed relation between supernova apparent luminosities and redshifts:

1. The effect of weak gravitational lensing on the implications of the supernova observations is expected to be small,[19] except perhaps for small area surveys.[20] (Gravitational lensing is discussed in Chapter 9.) It had been thought that the apparent luminosity of the most distant observed supernova, SN1997ff, may be enhanced by gravitational lensing,[21] conceivably reopening the possibility that the reduction of the apparent luminosity of the nearer supernovae *is* due to gray dust. However, a subsequent analysis by the same group[22] reported that the magnification of this supernova due to gravitational lensing is less than had been thought, and that the effect of the corrections due to gravitational lensing on current cosmological studies is small. Members of the High-z Supernova project have reported that instead this effect is likely to improve agreement with the estimate that $\Omega_M = 0.35$ and $\Omega_\Lambda = 0.65$.[23]

2. It has been argued that inhomogeneities in the cosmic distribution of matter could produce an accelerating expansion, without the need for any sort of exotic vacuum energy.[24] Given the high degree of

[17] P. Podsiadlowski *et al.*, astro-ph/0608324. Evolution may also affect the extinction of light by dust in the host galaxy; see T. Totani and C. Kobayashi, *Astrophys. J.* **526**, 65 (1999).

[18] M. Sullivan *et al.*, ref. 10.

[19] A. J. Barber, *Astron. Soc. Pacific Conf. Ser.* **237**, 363 (2001) [astro-ph/0109043].

[20] A. Cooray, D. Huterer, and D. E. Holz, *Phys. Rev. Lett.* **96**, 021301 (2006).

[21] E. Mörtstell, C. unnarsson, and A. Goobar, *Astrophys. J.* **561**, 106 (2001); C. Gunnarsson, in *Proceedings of a Conference on New Trends in Theoretical and Observational Cosmology – Tokyo, 2001* [astro-ph/0112340].

[22] J. Jönsson *et al.*, *Astrophys. J.* **639**, 991 (2006) [astro-ph/0506765].

[23] N. Benítez *et al.*, *Astrophys. J.* **577**, L1 (2002) [astro-ph/0207097].

[24] E. W. Kolb, S. Matarrese, A. Notari, and A. Riotto, *Astrophys. J.* **626**, 195 (2005) [hep-th/0503117]; E. W. Kolb, S. Matarrese, and A. Riotto, *New J. Phys.* **8**, 322 (2006) [astro-ph/0506534]; E. Barausse, S. Matarrese, and A. Riotto, *Phys. Rev. D* **71**, 063537 (2005).

homogeneity of the universe when averaged over sufficiently large scales, this seems unlikely.[25]

3. There is some evidence for two classes of type Ia supernovae,[26] with the minority associated perhaps with merging white dwarfs, or with a variation in explosion physics. The effect on cosmological studies remains to be evaluated.

4. Other uncertainties that can degrade the accuracy of measurements of dark energy (without casting doubt on its existence) arise from the circumstance that the shape of the curve of luminosity distance versus redshift is found by numerous observatories, both ground-based and space-based, and there are various flux calibration errors that can arise between these different observatories.

5. The measurement of luminosity distance of any source of light at large redshift has historically been plagued by the fact that measurements are not "bolometric," that is, equally sensitive to all wavelengths, but are rather chiefly sensitive to wavelengths in a limited range. The cosmological redshift changes the apparent colors of sources, and thereby changes the sensitivity with which apparent luminosity is measured. To take this into account, the observed apparent magnitude is corrected with a so-called *K-correction*.[27] The K-correction for supernovae were worked out before the discovery of dark energy,[28] and has been refined subsequently.[29] As the precision of supernovae observations improves, further improvements may also be needed in the K-correction.

These observations of an accelerated expansion are consistent with the existence of a constant vacuum energy, but do not prove that this energy density really is constant. According to Eq. (1.5.18), the existence of an accelerated expansion does however require that a large part of the energy density of the universe is in a form that has $\rho + 3p < 0$, unlike ordinary matter or radiation. This has come to be called *dark energy*.[30]

---

[25] É. É. Flanagan, *Phys. Rev. D* **71**, 103521 (2005) [hep-th/0503202]; G. G. Geshnizjani, D. J. H. Chung, and N. Afshordi, *Phys. Rev. D* **72**, 023517 (2005) [astro-ph/0503553]; C. M. Hirata and U. Seljak, *Phys. Rev. D* **72**, 083501 (2005) [astro-ph/0503582]; A. Ishibashi and R. M. Wald, Class. Quant. Grav. **23**, 235 (2006) [gr-qc/0509108].

[26] D. Howell *et al.*, *Nature* **443**, 308 (2006); S. Jha, A. Riess, & R. P. Kirshner, Astrophys. J. **654**, 122 (2007); R. Quimby, P. Höflich, and J. C. Wheeler, 0705.4467.

[27] For a discussion of the K-correction applied to observations of whole galaxies, and original references, see G&C, p. 443.

[28] A. Kim, A. Goobar, and S. Perlmutter, *Proc. Astron. Soc. Pacific* **108**, 190 (1995) [astro-ph/9505024].

[29] P. Nugent, A. Kim, and S. Perlmutter, *Proc. Astron. Soc. Pacific* **114**, 803 (2002) [astro-ph/0205351].

[30] For a general review, see P. J. E. Peebles and B. Ratra, *Rev. Mod. Phys.* **75**, 559 (2003).

To take into account the possibility that the dark energy density is not constant, it has become conventional to analyze observations in terms of its pressure/density ratio $p_{\rm D.E.}/\rho_{\rm D.E.} \equiv w$. Except in the case of a constant vacuum energy density, for which $w = -1$, there is no special reason why w should be time-independent. (A different, more physical, possibility is explored at the end of Section 1.12.) Still, it is popular to explore cosmological models with w constant but not necessarily equal to -1. As long as the dark energy density and Ω_K are non-negative, the expansion of the universe will continue, with $\dot{a}$ always positive. As shown in Eq. (1.1.34), the dark energy density in this case goes as a^{-3-3w}, so if w is negative (as indicated by the supernova observations) the energy density of radiation and matter must eventually become negligible compared with the dark energy density. For $w < -1/3$, the effect of a possible curvature in the Friedmann equation (1.5.19) also eventually becomes negligible. The solution of this equation for $w > -1$ with $\dot{a} > 0$ then becomes $t - t_1 \to Ca^{(3+3w)/2}$, with $C > 0$, and t_1 an integration constant. This is a continued expansion, with a decreasing expansion rate. But for $w < -1$, sometimes known as the case of *phantom energy*, the solution with $\dot{a} > 0$ is instead $t_1 - t \to Ca^{(3+3w)/2}$, again with $C > 0$. This solution has the remarkable feature that $a(t)$ becomes infinite at the time t_1. In contrast with the case $w \geq -1$, for $w < -1$ all structures — galaxy clusters, galaxy clusters, stars, atoms, atomic nuclei, protons and neutrons — eventually would be ripped apart by the repulsive forces associated with dark energy.[31]

To further study the time dependence of the dark energy, a five year supernova survey, the Supernova Legacy Survey,[32] was begun in 2003 at the Canada–France–Hawaii telescope on Mauna Kea. At the end of the first year, 71 high redshift Type Ia supernovae had been discovered and studied, with the result that $\Omega_M = 0.263 \pm 0.042(\text{stat}) \pm 0.032(\text{sys})$. Combining this supernova data with data from the Sloan Digital Sky Survey (discussed in Chapter 8), and assuming that the dark energy has $w \equiv p/\rho$ time-independent, it is found that if w is constant then $w = -1.023 \pm 0.09(\text{stat}) \pm 0.054(\text{sys})$, consistent with the value $w = -1$ for a constant vacuum energy. At the time of writing, results have just become available for 60 Type Ia supernovae from another supernova survey, ESSENCE.[33] (The acronym is for Equation of State: Supernovae trace Cosmic Expansion). Combining these with the results of the Supernova Legacy Survey, the ESSENCE group found that if w is constant then $w = -1.07 \pm 0.09(\text{stat}, 1\sigma) \pm 0.13(\text{syst})$, and $\Omega_M = 0.267^{-0.028}_{-0.018}(\text{stat}, 1\sigma)$.

[31] R. R. Caldwell, M. Kamionkowski, and N. N. Weinberg, *Phys. Rev. Lett.* **91**, 071301 (2003) [astro-ph/0302506].

[32] P. Astier, *et al.*, *Astron. Astrophys.* **447**, 31 (2006) [astro-ph/0510447].

[33] M. Wood-Vesey *et al.*, astro-ph/0701041

The conclusion that dark energy makes up a large fraction of the energy of the universe has been confirmed by observations of the cosmic microwave background, as discussed in Section 7.2. This conclusion has also received support from the use of a different sort of secondary distance indicator, the emission of X-rays from hot gas in galaxy clusters. In Section 1.9 we will see that the measurement of redshift, temperature, apparent X-ray luminosity and angular diameter of a cluster allows a determination of the ratio of hot gas ("baryons") to all matter in the cluster, with this ratio proportional to $d_A^{-3/2}$, where d_A is the assumed angular-diameter distance of the cluster. This can be turned around: under the assumption that the ratio of hot gas to all matter is the same for all clusters in a sample, X-ray observations can be used to find the dependence of the cluster angular diameter distances on redshift.[34] In this way, observations by the Chandra satellite of X-rays from 26 galaxy clusters with redshifts in the range $0.07 < z < 0.9$ have been used to determine that in a cosmology with a constant vacuum energy and cold dark matter, $\Omega_\Lambda = 0.94^{+0.21}_{-0.25}$, within 68% confidence limits.[35] Relaxing the assumption that the cosmological dark energy density is constant, but assuming $\Omega_K = 0$ and a constant w, and taking the baryon density to have the value indicated by cosmological nucleosynthesis (discussed in Section 3.2), this analysis of the Chandra data yields $1 - \Omega_M = 0.76 \pm 0.04$ and a dark energy pressure/density ratio $w = -1.20^{+0.24}_{-0.28}$.

It is possible that measurements of luminosity distance can be pushed to much larger redshifts by the use of long gamma ray bursts as secondary distance indicators. These bursts definitely do not have uniform absolute luminosity, but there are indications that the absolute gamma ray luminosity is correlated with the peak gamma ray energy and a characteristic time scale.[36]

The discovery of dark energy is of great importance, both in interpreting other observations and as a challenge to fundamental theory. It is profoundly puzzling why the dark energy density is so small. The contribution of quantum fluctuations in known fields up to 300 GeV, roughly the highest energy at which current theories have been verified, gives a vacuum energy

[34] S. Sasaki, Publ. Astron. Soc. Japan **48**, 119 (1996) [astro-ph/9611033]; U.-L. Pen, New Astron. **2**, 309 (1997) [astro-ph/9610147].

[35] S. W. Allen, R. W. Schmidt, H. Ebeling, A. C. Fabian, and L. van Speybroeck, *Mon. Not. Roy. Astron. Soc.* **353**, 457 (2004) [astro-ph/0405340]. For earlier applications of this technique, see K. Rines *et al.*, *Astrophys. J.* **517**, 70 (1999); S. Ettori and A. Fabian, *Astron. Soc. Pac. Conf. Ser.* **200**, 369 (2000); S. W. Allen, R. W. Schmidt, and A. C. Fabian, *Mon. Not. Roy. Astron. Soc.* **334**, L11 (2002); S. Ettori, P, Tozzi, and P. Rosati, Astron. & Astrophys. **398**, 879 (2003). The possibility of a variable ratio of hot gas to all matter is explored by R. Sadat *et al.*, *Astron. & Astrophys.* **437**, 310 (2005); L. D. Ferramacho and A. Blanchard, *Astron. & Astrophys.* **463**, 423 (2007) [astro-ph/0609822].

[36] C. Firmani, V. Avila-Reese, G. Ghisellini, and G. Ghirlanda, *Mon. Not. Roy. Astron. Soc.* **372**, 28 (2006) [astro-ph/0605430]; G. Ghirlanda, G. Ghisellini, and C. Firmani, *New J. Phys.* **8**, 123 (2006) [astro-ph/0610248].

density of order (300 GeV),[4] or about 10^{27} g/cm^3. This of course is vastly larger than the observed dark energy density, $\Omega_V \rho_{0,\text{crit}} \simeq 10^{-29}$ g/cm^3, by a factor of order 10^{56}. There are other unknown contributions to the vacuum energy that might cancel this contribution, coming from fluctuations in fields at higher energies or from the field equations themselves, but this cancelation would have to be precise to about 56 decimal places. There is no known reason for this remarkable cancelation.[37] The discovery of dark energy now adds a second problem: why is the dark energy density comparable to the matter energy density at this particular moment in the history of the universe?

In thinking about these problems, it is crucial to know whether the vacuum energy is really time-independent, or varies with time, a question that may be answered by future studies of distant Type Ia supernovae or other measurements at large redshift. The possibility of a varying dark energy (known as *quintessence*) will be considered further in Section 1.12.

1.7 Cosmic expansion or tired light?

In comparing observations of redshifts and luminosity distances with theory, we rely on the general understanding of redshifts and luminosities outlined in Sections 1.2 and 1.4. One thing that might invalidate this understanding is absorption or scattering, which reduces the number of photons reaching us from distant sources. This possibility is usually taken into account by measuring the color of the source, which would be affected by absorption or scattering, though as mentioned in the previous section there is a possibility of gray dust, which could not be detected in this way. Another possible way that apparent luminosities could be reduced is through the conversion of photons into particles called axions by intergalactic magnetic fields. There is also a more radical possibility. Ever since the discovery of the cosmological redshift, there has been a nagging doubt about its interpretation as evidence of an expanding universe. Is it possible that the universe is really static, and that photons simply suffer a loss of energy and hence a decrease in frequency as they travel to us, the loss of energy and hence the redshift naturally increasing with the distance that the photons have to travel?

It is possible to rule out all these possibilities by comparing the luminosity distance $d_L(z)$ with the angular diameter distance $d_A(z)$ of the same distant source. None of the possibilities mentioned above can affect the angular diameter distance, while the conventional interpretation of

[37] For a survey of efforts to answer this question, see S. Weinberg, *Rev. Mod. Phys.* **61**, 1 (1989).

redshifts and luminosities provides the model-independent result (1.4.12), that $d_L(z)/d_A(z) = (1+z)^2$, so a verification of this ratio can confirm the conventional understanding of $d_L(z)$.

We can check this formula for $d_L(z)/d_A(z)$ by a "surface brightness test" suggested long ago by Tolman.[1] If a light source has an absolute luminosity per proper area $\mathcal{L}$, then the apparent luminosity of a patch of area A will be $\ell = \mathcal{L}A/4\pi d_L^2$. This patch will subtend a solid angle $\Omega = A/d_A^2$. The *surface brightness* B is defined as the apparent luminosity per solid angle, so

$$B \equiv \frac{\ell}{\Omega} = \frac{\mathcal{L}\, d_A^2}{4\pi\, d_L^2} \ . \tag{1.7.1}$$

In the conventional big bang cosmology the ratio d_A/d_L is given by Eq. (1.4.12), so

$$B = (1+z)^{-4} \left(\frac{\mathcal{L}}{4\pi}\right) \ . \tag{1.7.2}$$

If we can find a class of light sources with a common value for the absolute luminosity per proper area $\mathcal{L}$, then their surface brightness should be found to decrease with redshift precisely as $(1+z)^{-4}$.

For instance, one important difference between "tired light" theories and the conventional big bang theory is that in the conventional theory all rates at the source are decreased by a factor $(1+z)^{-1}$, while in tired light theories there is no such slowing down. One rate that is slowed down at large redshifts in the conventional theory is the rate at which photons are emitted by the source. This is responsible for one of the two factors of $(1+z)^{-1}$ in formula (1.4.1) for apparent luminosity, the other factor being due to the reduction of energy of individual photons. On the other hand, if the rate of photon emission is not affected by the redshift, then in a static Euclidean universe in which photons lose energy as they travel to us, the apparent luminosity of a distant source L at a distance d will be given by $L/4\pi(1+z)d^2$, with only a single factor $1+z$ in the denominator to take account of the photon energy loss. That is, the luminosity distance will be $(1+z)^{1/2}d$, while the angular diameter distance in a Euclidean universe is just d, so here $d_L/d_A = (1+z)^{1/2}$, and the surface brightness of distant galaxies should decrease as $(1+z)^{-1}$.

Lubin and Sandage[2] have used the Hubble Space Telescope to compare the surface brightness of galaxies in three distant clusters with redshifts

[1] R. C. Tolman, *Proc. Nat. Acad. Sci* **16**, 5111 (1930); *Relativity, Thermodynamics, and Cosmology* (Oxford Press, Oxford, 1934): 467.

[2] L. M. Lubin and A. Sandage, Astron. J. **122**, 1084 (2001) [astro-ph/0106566]. Their earlier work is described in A. Sandage and L. M. Lubin, Astron. J. **121**, 2271 (2001); L. M. Lubin and A. Sandage, *ibid*, 2289 (2001) and Astron. J. **122**, 1071 (2001) [astro-ph/0106563.]

0.76, 0.90, and 0.92 with the surface brightness measured in closer galaxies. They detect a decrease of B with increasing z that is consistent with Eq. (1.7.2) with reasonable corrections for the effects of galaxy evolution, and is quite inconsistent with the behavior $B \propto (1+z)^{-1}$ expected in a static universe with "tired light."

In the standard big bang cosmology all rates observed from a distant source are slowed by a factor $1/(1+z)$, not just the rate at which photons are emitted. This slowing has been confirmed[3] for the rate of decline of light from some of the Type Ia supernovae used by the Supernova Cosmology Project discussed in the previous section. The hypothesis that the absolute luminosity is simply correlated with the intrinsic decline time is found to work much better if the observed decline time is taken to be the intrinsic decline time stretched out by a factor $1+z$. Nothing like this would be expected in a static Euclidean universe with redshifts attributed to tired light.

1.8 Ages

As we have seen, a knowledge of the Hubble constant and of the matter and vacuum density parameters Ω_M and Ω_Λ allows us to estimate the age of the universe. In this section we will discuss independent estimates of the age of the universe, based on calculations of the ages of some of the oldest things it contains.

Since metals (elements heavier than helium) found in the outer parts of stars arise chiefly from earlier generations of stars, the oldest stars are generally those whose spectra show small abundances of metals. These are the so-called Population II stars, found in the halo of our galaxy, and in particular in globular clusters. There are two main ways of estimating ages of old stars:

A. Heavy element abundances

If a nucleus decays with decay rate λ, and has an initial abundance $A_{\rm init}$, then the abundance A after a time T is $A = A_{\rm init}\exp(-\lambda T)$. Hence if we knew $A_{\rm init}$ and could measure A, we could determine T from $T = \lambda^{-1}\ln(A_{\rm init}/A)$. Unfortunately neither condition is likely to be satisfied. On the other hand, it is often possible to calculate the *ratio* of the initial abundances $A_{1\,\rm init}$ and $A_{2\,\rm init}$ of two nuclei, and to measure their *relative* present abundance A_1/A_2.

[3] B. Leibundgut *et al.*, Astrophys. J. *466*, L21 (1996); G. Goldhaber *et al.* (Supernova Cosmology Project), *Astrophys. J.* **558**, 359 (2001) [astro-ph/0104382].

This relative abundance is given by

$$\frac{A_1}{A_2} = \left(\frac{A_{1\,\mathrm{init}}}{A_{2\,\mathrm{init}}}\right) \exp\Big((\lambda_2 - \lambda_1)T\Big)\,,$$

so

$$T = \frac{1}{\lambda_2 - \lambda_1}\left[\ln\left(\frac{A_1}{A_2}\right) - \ln\left(\frac{A_{1\,\mathrm{init}}}{A_{2\,\mathrm{init}}}\right)\right] \tag{1.8.1}$$

If the initial abundances are similar and the observed abundances are very different, then the estimated value of T will be insensitive to the precise value of the initial abundance ratio.

The initial relative abundances of heavy, radioactive elements are estimated on the well-founded assumption that these elements are made in the so-called r-process, the rapid addition of neutrons to lighter elements such as iron in core-collapse supernova explosions, after which the neutron-rich isotopes formed in this way undergo multiple beta decays, transforming them to the most deeply bound nuclei with the same number of nucleons. This method has been used to put a lower bound on the age of our galaxy from the terrestrial abundance of ^{235}U, which has a decay rate of 0.971×10^{-9}/yr. To avoid uncertainties in the distribution of ^{235}U in earth, its abundance is measured relative to the isotope ^{238}U, which has a slower decay rate of 0.154×10^{-9}/yr, but behaves the same chemically and is presumably distributed in the same way. The initial abundance ratio of ^{235}U to ^{238}U is estimated to be 1.65 ± 0.15; it is larger than one because three additional neutrons must be added to the progenitor of ^{235}U to form the progenitor of ^{238}U. On the other hand, the larger decay rate of ^{235}U makes it (fortunately) less abundant than ^{238}U now. The present abundance ratio of uranium isotopes on earth is 0.00723, so this uranium has been decaying for a time

$$\frac{\ln(1.65) - \ln(.00723)}{0.971 \times 10^{-9}/\mathrm{yr} - 0.154 \times 10^{-9}/\mathrm{yr}} = 6.6\ \mathrm{Gyr} \qquad [1\ \mathrm{Gyr} = 10^9\ \mathrm{yr}].$$

But the sun is a second (or perhaps third) generation (called "Population I") star, and presumably its uranium was being produced over a long time interval before the formation of the solar system. The uranium abundance ratio has been supplemented with measurements of other abundance ratios on the earth and meteorites, such as ^{232}Th/^{238}U and ^{187}Re/^{187}Os ratios, and analyzed with the length of the period of heavy element formation left as a free parameter. This gives a more stringent (but less certain) lower bound of 9.6 Gyr[1] on the age of the heavy elements in the neighborhood of the solar system.

[1] B. S. Meyer and D. N. Schramm, *Astrophys. J.* **311**, 406 (1986).

A much more stringent lower bound on the age of the galaxy is given by applying these methods to heavy elements in metal-poor stars beyond the solar system. First thorium was observed spectroscopically in a very metal-poor star (and hence presumably old) K giant star, CS 22892-052.[2] The relative abundances in this star of the more stable elements produced in the r-process, as measured from the intensity of absorption lines in the star's spectrum, matches those of the same elements in the solar system, except for a much lower abundance of the heaviest detected element thorium, which (for ^{232}Th) has a half life 14 Gyr. Attributing the decrease in thorium to its radioactive decay, the age of the thorium in this star is estimated as 14.1 ± 3 Gyr. Other estimates of the ages of CS 22892-052 and other metal-poor stars have been made using the measured abundance ratios of thorium to europium and lanthanum.[3]

Uranium-238 decays more rapidly than ^{232}Th, so we can get a more sensitive estimate of the age of a star by using both its uranium and its thorium abundances, providing of course that uranium as well as thorium lines can be observed in the star's spectrum. No uranium absorption lines were observed in the spectrum of CS 22892-052, but absorption lines from singly ionized uranium were subsequently observed in two other metal-poor star with an abundance of r-process elements, CS31082-001 and BD+17°3248. The observed abundance ratio of uranium to thorium in CS31082-001 is $10^{-0.74\pm0.15}$, while the initial abundance ratio has been variously estimated as $10^{-0.255}$ or $10^{-0.10}$. Using these numbers in Eq. (1.8.1) gives this star an age of 12.5 ± 3 Gyr.[4] Subsequent observations indicated ages of 14 ± 2 Gyr,[5] 15.5 ± 3.2 Gyr,[6] and 14.1 ± 2.5 Gyr.[7] In a similar way, the age of BD+17°3248 has been calculated as 13.8 ± 4 Gyr.[8] (See Figure 1.5.) More recently, both uranium and thorium lines have been found in the spectrum of the newly discovered metal-poor star HE 1523-0903; the ratio of thorium and uranium abundance to the abundances of other r-process elements, and to each other, was used to give an age of the star as 13.2 Gyr.[9]

---

[2] C. Sneden *et al.*, *Astrophys. J.* **467**, 819 (1996); *Astrophys. J.* **591**, 936 (2003) [astro-ph/0303542]. A review with references to earlier work on thorium abundances was given by C. Sneden and J. J. Cowan, *Astronomia y Astrofisica (Serie de Conferencia)* **10**, 221 (2001) [astro-ph/0008185].

[3] I. I Ivans *et al.*, *Astrophys. J.* **645**, 613 (2006) [astro-ph/0604180], and earlier references cited therein.

[4] R. Cayrel *et al., Nature* **409**, 691 (2001).

[5] V. Hill *et al.*, *Astron. Astrophys.* **387**, 580 (2002).

[6] H. Schatz *et al.*, *Astrophys. J.* **579**, 626 (2002).

[7] S. Wanajo, *Astrophys. J.* **577**, 853 (2002).

[8] J. J. Cowan, *et al.*, *Astrophys. J.* **572**, 861 (2002) [astro-ph/0202429].

[9] A. Frebel *et al.*, *Astrophys.* **660**, L117 (2007). [astro-ph/0703414].

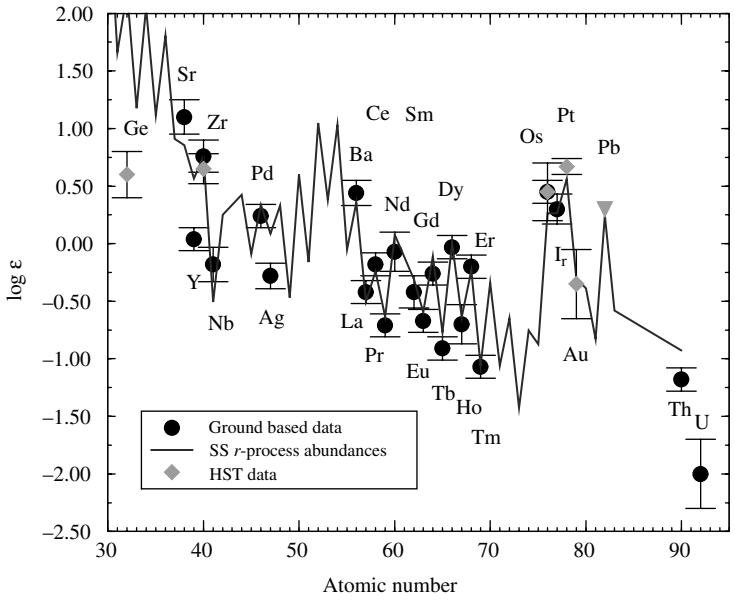

Figure 1.5: Abundances of elements produced by the *r*-process in the star BD+17°3248, obtained by ground-based and Hubble Space Telescope spectroscopic observations. For comparison, the solid curve gives theoretical initial abundances, based on solar system data. Note the low observed abundances of thorium and uranium, compared with the theoretical initial abundances, which indicate an age for BD+17°3248 of 13.8 ± 4 Gyr. From J. J. Cowan *et al.*, *Astrophys. J.* **572**, 861 (2002) [astro-ph/0202429].

B. Main sequence turn-off

The stars that satisfy the main sequence relation between absolute luminosity and surface temperature are still burning hydrogen at their core. When the hydrogen is exhausted at the core, hydrogen-burning continues in a shell around a (temporarily) inert helium core. The star then moves off the main sequence, toward higher luminosity and lower surface temperature. The heavier a star is, the more luminous it is on the main sequence, and the faster it evolves. Thus as time passes, the main sequence of a cluster of stars of different masses but the same age shows a turn-off that moves to lower and lower luminosities. (See Figure 1.6). Roughly, the absolute luminosity of stars at the turn-off point is inversely proportional to the age of the cluster. In particular, observations of the main sequences of a number of globular clusters gave ages variously calculated[10] as 11.5 ± 1.3 Gyr, 12 ± 1 Gyr, 11.8 ± 1.2 Gyr, 14.0 ± 1.2 Gyr, 12 ± 1 Gyr, and 12.2 ± 1.8 Gyr. A summary by Schramm[11] found that most of the discrepancies disappear when

[10] For references, see B. Chaboyer, *Phys. Rep.* **307**, 23 (1998) [astro-ph/9808200].

[11] D. Schramm, in *Critical Dialogues in Cosmology*, N. Turok, ed. (World Scientific, Singapore, 1997): 81

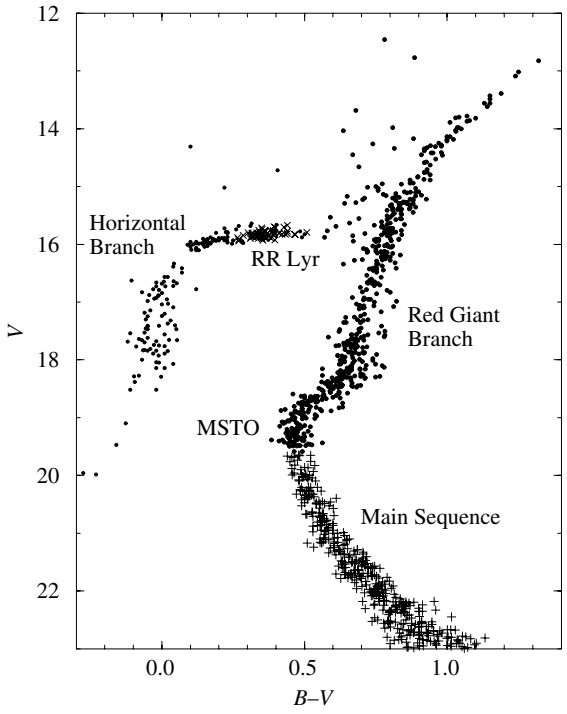

Figure 1.6: Color–magnitude diagram for the globular cluster M15. Visual apparent magnitudes of M15 stars are plotted on the vertical axis. Since all stars in M15 are at about the same distance from earth, the apparent visual magnitude differs from the absolute visual magnitude by a constant term, with absolute luminosities increasing upwards. The difference of apparent blue and visual magnitudes is plotted on the horizontal axis. This is a measure of surface temperature, with temperature decreasing to the right. If M15 were young, the main sequence would continue upwards and to the left; the position of the main sequence turn-off (MSTO) and other features of the diagram indicate that the age of the cluster is 15 ± 3 Gyr. Diagram from B. Chaboyer, *Phys. Rep.* **307**, 23 (1998), based on data of P. R. Durrell and W. E. Harris, *Astron. J.* **105**, 1420 (1993) [astro-ph/9808200].

the various calculations are done with the same input values for parameters like the initial abundance of helium, oxygen, and iron, and gave a consensus value as 14 ± 2(statistical) $\pm$ 2(systematic) Gyr. Note that all these ages are sensitive to the distance scale; a fractional change $\delta d/d$ in estimates of distances would produce a fractional change $\delta L/L = -2\delta d/d$ in estimates of absolute luminosities, and hence a fractional change $\delta t/t \approx +2\delta d/d$ in estimates of ages. Using measurements of distances to nine globular clusters with the Hipparcos satellite yields an estimated galactic age[12] of 13.2 ± 2.0 Gyr.

[12]E. Carretta, R. G. Gratton, G. Clementini, and F. F. Pecci, *Astrophys. J.* **533**, 215 (2000) [astro-ph/9902086].

These ages would pose a problem for what used to be the most popular model, with $\Omega_M = 1$ and no vacuum energy. In this case, the age of the universe is

$$t_0 = \frac{2}{3H_0} = 9.3 \left(\frac{70 \text{ km/sec/Mpc}}{H_0}\right) \text{ Gyr} ,$$

which is somewhat younger than the oldest objects in the galaxy, though not by many standard deviations. Inclusion of a constant vacuum energy helps to avoid this problem; as remarked in Section 1.5, with nothing else in the universe we would have $a(t) \propto \exp(Ht)$, and the age of the universe would be infinite. As we saw in Section 1.6, the supernovae distance–redshift relation indicates that the vacuum energy is now roughly twice the matter energy, giving an age much longer than $2/3H_0$:

$$t_0 = 13.4^{+1.3}_{-1.0} \left(\frac{70 \text{ km/sec/Mpc}}{H_0}\right) \text{ Gyr} ,$$

This removes the danger of a conflict, *provided* that the globular clusters in our galaxy are not much younger than the universe itself. In fact, there is now a truly impressive agreement between the age of the oldest stars and star clusters on one hand and the cosmic age calculated using values of H_0, Ω_M, and Ω_Λ found from the redshift–distance relation. As we will see in Section 7.2, there is also an excellent agreement between these ages and the age calculated using parameters measured in observations of anisotropies in the cosmic microwave background.

So far in this section, we have considered only the present age of our own galaxy. It is also possible to estimate the ages of other galaxies at high redshift, at the time far in the past when the light we now observe left these galaxies. Of course, it is not possible to distinguish individual stars or globular clusters in these galaxies, but the spectrum of the galaxy gives a good idea of the age. We need the whole spectrum to separate the effects of metallicity, scattering, etc., but roughly speaking, the redder the galaxy, the more of its bright bluer stars have left the main sequence, and hence the older it is. In this way, it has been found[13] that the radio galaxies 53W091($z = 1.55$) and 53W069($z = 1.43$) have ages $\simeq$ 3.5 Gyr and 3 to 4 Gyr, respectively. This sets useful lower bounds on the vacuum energy. In a model with non-relativistic matter and a constant vacuum energy, the age of the universe at the time of emission of light that is seen at present with

[13] J. S. Dunlop *et al.*, *Nature* **381**, 581 (1996); J. S. Dunlop, in *The Most Distant Radio Galaxies - KNAW Colloquium, Amsterdam, October 1997*, eds. Best *et al.* [astro-ph/9801114].

redshift z is given by Eq. (1.5.42) as

$$t(z) = \frac{1}{H_0} \int_0^{\frac{1}{1+z}} \frac{dx}{x\sqrt{\Omega_\Lambda + \Omega_K x^{-2} + \Omega_M x^{-3} + \Omega_R x^{-4}}} \,. \quad (1.8.2)$$

Any galaxy observed with redshift z must have been younger than this at the time that its light was emitted. For instance, for a flat universe with $\Omega_K = \Omega_R = 0$, so that $\Omega_M = 1 - \Omega_\Lambda$, the existence of a galaxy at $z = 1.55$ with age $\simeq 3.5$ Gyr sets a lower bound[14] on Ω_Λ of about 0.6 for $H_0 = 70$ km s^{-1} Mpc^{-1}.

Eventually the accuracy of these age determinations may become good enough to allow us to measure at least the dependence of redshift on the cosmic age. Of course, galaxies form at various times in the history of the universe, so the age of any one galaxy does not allow us to infer the age of the universe at the time light we now see left that galaxy. However, the homogeneity of the universe implies that the *distribution* of cosmic times of formation for any one variety of galaxy is the same anywhere in the universe. From differences in the distributions of ages of a suitable species of galaxy at different redshifts, we can then infer the difference of cosmic age t at these redshifts. The Robertson–Walker scale factor $a(t)$ is related to the redshift $z(t)$ observed now of objects that emitted light when the cosmic age was t by $1 + z(t) = a(t_0)/a(t)$, so $\dot{z} = -H(t)(1+z)$. To calculate $\ddot{z}$, we note that for $K = 0$, $H^2(t) = 8\pi G\rho(t)/3$, and $\dot{\rho} = -3H(\rho + p)$,

$$\dot{H}(t) = -4\pi G\Big(\rho(t) + p(t)\Big) \,. \quad (1.8.3)$$

Then for $K = 0$

$$\ddot{z} = \frac{\dot{z}^2}{1+z}\left(\frac{5}{2} + \frac{3p}{2\rho}\right) \,. \quad (1.8.4)$$

Thus measurements of differences in t for various differences in redshift may allow a measurement of the ratio p/ρ at various times in the recent history of the universe.[15]

1.9 Masses

We saw in Section 1.6 that the observed dependence of luminosity distance on redshift suggests that the fraction Ω_M of the critical density provided by

[14] L. M. Krauss, *Astrophys. J.* **489**, 486 (1997); J. S. Alcaniz and J. A. S. Lima, *Astrophys. J.* **521**, L87 (1999) [astro-ph/9902298].

[15] R. Jiminez and A. Loeb, *Astrophys. J.* **573**, 37 (2002) [astro-ph/0106145].

non-relativistic matter is roughly 30%. In this section we will consider other independent ways that Ω_M is measured.

A. Virialized clusters of galaxies

The classic approach[1] to the measurement of Ω_M is to use the virial theorem to estimate the masses of various clusters of galaxies, calculate a mean ratio of mass to absolute luminosity, and then use observations of the total luminosity of the sky to estimate the total mass density, under the assumption that the mass-to-light ratio of clusters of galaxies is typical of the universe as a whole.

To derive the virial theorem, consider a non-relativistic gravitationally bound system of point masses m_n (either galaxies, or stars, or single particles) with positions relative to the center of mass (in an ordinary Cartesian coordinate system) $\mathbf{X}_n$. The equations of motion are

$$m_n \ddot{X}_n^i = -\frac{\partial V}{\partial X_n^i} \,, \tag{1.9.1}$$

where the potential energy V is

$$V = -\frac{1}{2} \sum_{n \neq \ell} \frac{G m_n m_\ell}{|\mathbf{X}_n - \mathbf{X}_\ell|} \,. \tag{1.9.2}$$

Multiplying Eq. (1.9.1) with X_n^i and summing over n and i gives

$$-\sum_n X_n^i \frac{\partial V}{\partial X_n^i} = \sum_n m_n \mathbf{X}_n \cdot \ddot{\mathbf{X}}_n = \frac{1}{2} \frac{d^2}{dt^2} \sum_n m_n \mathbf{X}_n^2 - 2T \,, \tag{1.9.3}$$

where T is the internal kinetic energy (not counting any motion of the center of mass)

$$T = \frac{1}{2} \sum_n m_n \dot{\mathbf{X}}_n^2 \,. \tag{1.9.4}$$

Let us assume that the system has reached a state of equilibrium ("become virialized"), so that although the individual masses are moving there is no further statistical evolution, and in particular that

$$0 = \frac{d^2}{dt^2} \sum_n m_n \mathbf{X}_n^2 \tag{1.9.5}$$

[1] F. Zwicky, *Astrophys. J.* **86**, 217 (1937); J. H. Oort, in *La Structure et l'Evolution de l'Universe* (Institut International de Physique Solvay, R. Stoops, Brussels, 1958): 163.

(This is why it was important to specify that $\mathbf{X}_n$ is measured relative to the center of mass; otherwise a motion of the whole cluster would give the sum a term proportional to t^2, invalidating Eq. (1.9.5).) But V is of order -1 in the coordinates, so the left-hand side of Eq. (1.9.3) is just V, giving the *virial theorem*:

$$2T + V = 0 \ . \tag{1.9.6}$$

We may express T and V as

$$T = \frac{1}{2} M \langle v^2 \rangle \ , \qquad V = -\frac{1}{2} G M^2 \langle \frac{1}{r} \rangle \ , \tag{1.9.7}$$

where $\langle v^2 \rangle$ is the mean (mass weighted) square velocity relative to the center of mass, $\langle 1/r \rangle$ is the mean inverse separation, and $M = \sum_n m_n$ is the total mass. Eq. (1.9.6) thus gives the virial formula for M:

$$M = \frac{2 \langle v^2 \rangle}{G \langle 1/r \rangle} \ . \tag{1.9.8}$$

This derivation does not apply to irregular clusters of galaxies, like the nearby one in Virgo. Clusters like this do not seem to have settled into a configuration in which the condition (1.9.5) is satisfied, and therefore probably do not satisfy the virial theorem. On the other hand, the virial theorem probably does apply at least approximately to other clusters of galaxies, like the one in Coma, which appear more or less spherical. According to general ideas of statistical equilibrium, we may expect the rms velocity dispersion $\sqrt{\langle v^2 \rangle}$ of the dominant masses in such clusters to equal the velocity dispersion of the visible galaxies in the cluster, which can be measured from the spread of their Doppler shifts, and also to equal the velocity dispersion of the ionized intergalactic gas in the cluster, which since the advent of X-ray astronomy can be measured from the X-ray spectrum of the gas. The values obtained in these ways for $\langle v^2 \rangle$ are independent of the distance scale. On the other hand, values for $\langle 1/r \rangle$ are obtained from angular separations: the true transverse proper distance d is given in terms of the angular separation θ by $d = \theta d_A$, where d_A is the angular diameter distance (1.4.11). For clusters with $z \ll 1$, Eqs. (1.4.9) and (1.4.11) give $d_A \simeq z/H_0$, so $d \simeq \theta z/H_0$. Thus the estimated values of $\langle 1/r \rangle$ for galaxy clusters with $z \ll 1$ scale as H_0, and the values of M inferred from Eq. (1.9.8) scale as $1/H_0$. The absolute luminosity L of a cluster of galaxies with redshift z and apparent luminosity ℓ is given for $z \ll 1$ by Eqs. (1.4.2) and (1.4.9) as $L = 4\pi z^2 \ell / H_0^2$, so the values of L scale as H_0^{-2}, and the mass-to-light ratios obtained in this way therefore scale as $H_0^{-1}/H_0^{-2} = H_0$.

Estimates of M/L for rich clusters have generally given results of order 200 to 300 $h\ M_\odot/L_\odot$, where h is the Hubble constant in units of 100 km s^{-1} Mpc^{-1}, and $M_\odot$ and $L_\odot$ are the mass and absolute luminosity of the sun. For instance, a 1996 study[2] of 16 clusters of galaxies with redshifts between 0.17 and 0.55 gave $M/L = (295 \pm 53)\, h\, M_\odot/L_\odot$. Some of the same group[3] have corrected this result for various biases, and now find $M/L = (213 \pm 59)\, h\, M_\odot/L_\odot$. A more recent application[4] of the virial theorem to 459 clusters has found a value $M/L \simeq 348\, h\, M_\odot/L_\odot$.

All these values of M/L for clusters of galaxies are very much larger than the mass-to-light ratios of the visible regions of individual galaxies.[5] The mass-to-light ratios of individual elliptical galaxies can be measured using the virial theorem, with $\sqrt{< v^2 >}$ taken as the velocity dispersion of stars contained in the galaxy; this gives mass-to-light ratios generally in the range of 10 to 20 $h\, M_\odot/L_\odot$.[6] All of the visible light from clusters comes from their galaxies, so we must conclude that most of the mass in clusters of galaxies is in some non-luminous form, either in the outer non-luminous parts of galaxies or in intergalactic space. It has been argued that this mass is in large dark halos surrounding galaxies, extending to 200 kpc for bright galaxies.[7] The nature of this dark matter is an outstanding problem of cosmology, to which we will frequently return.

Incidentally, the large value of M/L given by the virial theorem for elliptical galaxies shows that most of the mass of these galaxies is not in the form of stars as bright as the sun. It is harder to estimate M/L for spiral galaxies, but since the work of Vera Rubin[8] it has been known that most of their mass is also not in luminous stars.[9] If most of the mass of a spiral galaxy were in the luminous central regions of the galaxy, then the rotational speeds of stars outside this region would follow the Kepler law, $v \propto r^{-1/2}$. Instead, it is observed that v outside the central region is roughly constant, even beyond the visible disk of the galaxy, which is what would be expected for a spherical halo with a mass density that decreases only as $1/r^2$, in which case most of the mass of the galaxy would be in the dark outer

[2] R. G. Carlberg *et al.*, *Astrophys. J.* **462**, 32 (1996).

[3] R. G. Carlberg, H. K. C. Yee, and E. Ellingson, *Astrophys. J.* **478**, 462 (1997).

[4] H. Andernach, M. Plionis, O. López-Cruz, E. Tago, and S. Basilakos, *Astron. Soc. Pacific Conf. Ser.* **329**, 289 (2005) [astro-ph/0407098].

[5] This conclusion was first reached in a study of the Coma cluster by F. Zwicky, *Helv. Phys. Acta* **6**, 110 (1933).

[6] T. R. Lauer, *Astrophys. J.* **292**, 104 (1985); J. Binney and S. Tremaine, *Galactic Dynamics* (Princeton University Press, Princeton, 1987).

[7] N. A. Bahcall, L. M. Lubin, and V. Dorman, *Astrophys. J.* **447**, L81 (1995).

[8] V. C. Rubin, W. K. Ford, and N. Thonnard, *Astrophys. J.* **225**, L107; **238**, 471 (1980).

[9] M. Persic and P. Salucci, *Astrophys. J.* Supp. **99**, 501 (1995); M. Persic, P. Salucci, and F. Stel, *Mon. Not. Roy. Astron. Soc.* **281**, 27P (1996).

parts of the halo. There is some evidence from the absence of gravitational microlensing by the halo (discussed in Section 9.2) that this mass is not in the form of dark stars either, but it is still possible that most of the matter in galaxies is baryonic. We will not go into this in detail, because the formation of galaxies involves cooling processes that requires baryonic matter of the same sort as in stars, so that we would only expect the value of M/L for galaxies to be similar to the value for the universe as a whole if the matter of the universe were mostly baryonic.

In using the value of M/L derived from the virial theorem for clusters of galaxies to find the mass density of the universe, we cannot just add up the luminosity per volume of clusters, because most of the light of the universe comes from "field" galaxies that are not in clusters. Instead, if we assume that the field galaxies are accompanied by the same amount of dark matter as the galaxies in clusters, as argued in ref. 7, then we can find Ω_M by using the value of M/L for clusters together with an estimate of the total luminosity density $\mathcal{L}$ to estimate the total mass density as

$$\rho_M = (M/L)\mathcal{L} \; . \tag{1.9.9}$$

Since values of absolute luminosities inferred from apparent luminosities and redshifts scale as H_0^{-2}, and distances inferred from redshifts scale as H_0^{-1}, the total luminosity density of the universe calculated by adding up the absolute luminosities of galaxies per volume scales as $H_0^{-2}/(H_0^{-1})^3 = H_0$. For example, a 1999 estimate[10] gave $\mathcal{L} = 2 \pm 0.2 \times 10^8 \, h \, L_\odot \, \mathrm{Mpc}^{-3}$. For the purpose of calculating Ω_M it is more convenient to write this as a ratio of the critical mass density to the luminosity density:

$$\rho_{0,\mathrm{crit}}/\mathcal{L} = (1390 \pm 140) \, h \, M_\odot/L_\odot \; .$$

(Here we use $M_\odot = 1.989 \times 10^{33}$ g, 1 Mpc $= 3.0857 \times 10^{24}$ cm, and $\rho_{0,\mathrm{crit}} = 1.878 \times 10^{-29} \, h^2$ g/cm^3.) Taking $M/L = (213 \pm 53) h M_\odot/L_\odot$ gives then

$$\Omega_M = \frac{M/L}{\rho_{0,\mathrm{crit}}/\mathcal{L}} = 0.15 \pm 0.02 \pm .04 \; ,$$

with the first uncertainty arising from $\mathcal{L}$ and the second from M/L. It is important to note that this is independent of the Hubble constant, as both $\mathcal{L}M/L$ and $\rho_{0,\mathrm{crit}}$ scale as H_0^2.

This estimate of Ω_M is somewhat lower than those derived from the redshift–luminosity relation of supernovae and from the anisotropies in the

[10] S. Folkes *et al.*, *Mon. Not. Roy. Astron. Soc.* **308**, 459 (1999); M. L. Blanton *et al.*, *Astron. J.* **121**, 2358 (2001).

cosmic microwave background, to be discussed in Section 2.6 and Chapter 7. But all these estimates agree that Ω_M is distinctly less than unity.

B. X-ray luminosity of clusters of galaxies

Does the dark matter in clusters of galaxies consist of ordinary nuclei and electrons? We can find the ratio of the fraction Ω_B of the critical density provided by baryonic matter (nuclei and electrons) to the fraction Ω_M provided by all forms of non-relativistic matter by studying the X-rays from clusters of galaxies, for it is only the collisions of ordinary baryonic particles that produces these X-rays. Because these collision processes involve pairs of particles of baryonic matter, the absolute X-ray luminosity per unit proper volume takes the form

$$\mathcal{L}_X = \Lambda\Big(T_B\Big)\rho_B^2 \ , \tag{1.9.10}$$

where T_B and ρ_B are the temperature and density of the baryonic matter, and $\Lambda(T)$ is a known function of temperature and fundamental constants. The baryonic density satisfies the equation of hydrostatic equilibrium, which (assuming spherical symmetry) follows from the balance of pressure and gravitational forces acting on the baryons in a small area A and between radii r and $r + \delta r$:

$$A\Big(p_B(r+\delta r) - p_B(r)\Big) = -\frac{A\delta r\,\rho_B(r)\,G}{r^2}\int_0^r 4\pi r^2\,\rho_M(r)\,dr \ ,$$

or, canceling factors of A and δr and using the ideal gas law $p_B = k_{\mathcal{B}}\,T_B\rho_B/m_B$,

$$\frac{d}{dr}\left(\frac{k_{\mathcal{B}}\,T_B(r)\rho_B(r)}{m_B}\right) = -\frac{G\rho_B(r)}{r^2}\int_0^r 4\pi r^2\rho_M(r)\,dr \ ,$$

where $\rho_M(r)$ is the total mass density, $k_{\mathcal{B}}$ is Boltzmann's constant, m_B is a characteristic mass of the baryonic gas particles, and r is here the proper distance to the center of the cluster. Multiplying by $r^2/\rho_B(r)$ and differentiating with respect to r yields

$$\frac{d}{dr}\left[\frac{r^2}{\rho_B(r)}\frac{d}{dr}\left(\frac{k_{\mathcal{B}}\,T_B(r)\rho_B(r)}{m_B}\right)\right] = -4\pi\,Gr^2\rho_M(r) \ . \tag{1.9.11}$$

If we make the assumption that cold dark matter particles, or whatever particles dominate the dark intergalactic matter, have an isotropic velocity

distribution (which is not very well motivated), then the same derivation applies to these particles, and their density $\rho_D = \rho_M - \rho_B$ satisfies the non-linear differential equation

$$\frac{d}{dr}\left[\frac{r^2}{\rho_D(r)}\frac{d}{dr}\left(\frac{k_{\mathcal{B}}\,T_D(r)\rho_D(r)}{m_D}\right)\right] = -4\pi G r^2 \rho_M(r)\ . \qquad (1.9.12)$$

where $T_D(r)$ and m_D are the temperature and mass of the dark matter particles. With perfect X-ray data and a knowledge of the distance of the source, one could measure the X-ray luminosity density $\mathcal{L}_X(r)$ and (using the X-ray spectrum) the baryon temperature $T_B(r)$ at each point in the cluster, then use Eq. (1.9.10) to find the baryon density $\rho_B(r)$ at each point, and then use Eq. (1.9.11) to find the total mass density at each point. We could then calculate the fractional baryon density ρ_B/ρ_M, and if we were interested we could also use Eq. (1.9.12) to find the velocity dispersion $k_{\mathcal{B}}T_D(r)/m_D$ of the dark matter.

In practice, it is usually necessary to use some sort of cluster model. In the simplest sort of model, one assumes an *isothermal sphere*: the temperatures T_B and T_D are taken to be independent of position, at least near the center of the cluster where most of the X-rays come from. It is also often assumed that the same gravitational effects that causes the concentration of the hot intergalactic gas in the cluster is also responsible for the concentration of the dark matter, so that the densities $\rho_B(r)$ and $\rho_M(r)$ are the same, up to a constant factor, which represents the cosmic ratio Ω_B/Ω_M of baryons to all non-relativistic matter. (These gravitational effects are believed to be a so-called "violent relaxation,"[11] caused by close encounters of clumps of matter whose gravitational attraction cannot be represented as an interaction with a smoothed average gravitational field. The condensation of galaxies out of this mixture requires quite different cooling processes that can affect only the baryonic gas, which is why galaxies have a lower proportion of dark matter and a lower mass-to-light ratio.) Comparison of Eqs. (1.9.11) and (1.9.12) shows that $\rho_B(r)$ and $\rho_D(r)$ will be proportional to each other, and hence also to $\rho_M(r)$ if the velocity dispersions of the dark matter and hot baryonic gas are the same:

$$k_{\mathcal{B}}T_M/m_M = k_{\mathcal{B}}T_D/m_D \equiv \sigma^2\ . \qquad (1.9.13)$$

Equations (1.9.11) and (1.9.12) both then tell us that

$$\rho_M(r) = \rho_M(0)\,F(r/r_0) \qquad (1.9.14)$$

[11] D. Lynden-Bell, *Mon. Not. Roy. Astron. Soc.* **136**, 101 (1967).

where $F(0) \equiv 1$; r_0 is a *core radius*, defined conventionally by

$$r_0 \equiv \sqrt{\frac{9\sigma^2}{4\pi G \rho_M(0)}} \; ; \tag{1.9.15}$$

and $F(u)$ is a function satisfying the differential equation

$$\frac{d}{du}\left(\frac{u^2}{F(u)}\frac{d\,F(u)}{du}\right) = -9u^2 F(u) \; . \tag{1.9.16}$$

We must also impose the boundary condition that ρ_M is analytic in the coordinate $\mathbf{X}$ at $\mathbf{X} = 0$, which for a function only of r means that it is given near $r = 0$ by a power series in r^2, so that $F(u)$ is given near $u = 0$ by a power series in u^2, $F(u) = 1 + O(u^2)$. Together with this boundary condition, Eq. (1.9.16) defines a unique function[12] that for small u has the approximate behavior[13]

$$F(u) \simeq (1 + u^2)^{-3/2} \; . \tag{1.9.17}$$

The solution to Eq. (1.9.16) is shown together with the approximation (1.9.17) in Figure 1.7.

For large u, $F(u)$ approaches the exact solution $2/9u^2$. Taken literally, this would make the integral for the total mass diverge at large r, which shows that the assumption of constant σ^2 must break down at some large radius. Often the function $F(u)$ is taken simply as[14]

$$F(u) = (1 + u^2)^{-3\beta/2} \; ,$$

where β is an exponent of order unity.

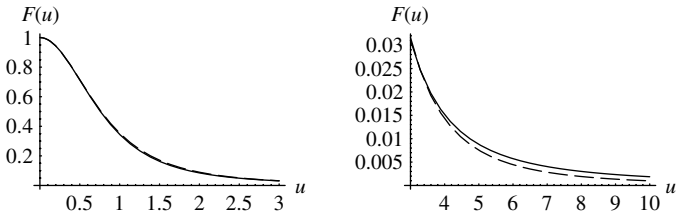

Figure 1.7: The solution to Eq. (1.9.16) (solid line) and the approximation (1.9.17) (dashed line). For the lower values of u in the figure at the left, the two curves are indistinguishable.

[12] For a tabulation of values of $F(u)$, see e.g. J. Binney and S. Tremaine, *Galactic Dynamics* (Princeton University, Princeton, 1987): Table 4.1.

[13] I. R. King, *Astron. J.* **67**, 471 (1962).

[14] A. Cavaliere and R. Fusco-Fermiano, *Astron. Astrophys.* **49**, 137 (1976).

Also, Eq. (1.9.11) has the solution

$$\rho_B(r) = \rho_B(0) F(r/r_0) \ , \tag{1.9.18}$$

with the same function $F(u)$ and the same core radius r_0. We can measure the core radius from the X-ray image of the cluster, and measure σ^2 from the X-ray spectrum, so that Eq. (1.9.15) can be used to find the central density $\rho_M(0)$ of all non-relativistic matter. The central density $\rho_B(0)$ of the baryonic matter can then be found from the total X-ray luminosity, which with these approximations (and using Eq. (1.9.10)) is

$$L_X \equiv \int d^3x \, \mathcal{L}_X = 4\pi \, \Lambda(T_B) \, r_0^3 \, \rho_B(0)^2 \mathcal{I} \ , \tag{1.9.19}$$

where

$$\mathcal{I} \equiv \int_0^\infty u^2 F^2(u) \, du \ . \tag{1.9.20}$$

Even though the solution of Eq. (1.9.16) gives an infinite mass, it gives a finite total X-ray luminosity, with $\mathcal{I} = 0.1961$. (The approximation (1.9.17) would give $\mathcal{I} = \pi/16 = 0.1963$.)

For a cluster at redshift z, the core radius r_0 inferred from observation of the angular size of the cluster will be proportional to the angular diameter distance $d_A(z)$, while the temperature and velocity dispersion found from the X-ray spectrum will not depend on the assumed distance. Thus the value of the central total matter density $\rho_M(0)$ given by Eq. (1.9.15) will be proportional to $1/d_A^2(z)$. On the other hand, the absolute X-ray luminosity L_X inferred from the apparent X-ray luminosity will (like all absolute luminosities) be proportional to the value assumed for $d_L^2(z)$, the square of the luminosity distance, so with $r_0 \propto d_A$, the central baryon density $\rho_B(0)$ given by Eq. (1.9.19) will be proportional to $[d_L^2(z)/d_A^3(z)]^{1/2}$. The value of the ratio of central densities inferred from observations of a given cluster at redshift z will therefore have a dependence on the distance assumed for the cluster given by

$$\frac{\rho_B(0)}{\rho_M(0)} \propto d_L(z) d_A^{1/2}(z) = (1+z)^2 d_A^{3/2}(z) \ , \tag{1.9.21}$$

in which we have used the relation (1.4.12) between luminosity and angular diameter distances.

For $z \ll 1$, we have $d_A(z) \simeq d_L(z) \simeq z/H_0$, and so according to Eq. (1.9.21) the value of $\rho_B(0)/\rho_M(0)$ obtained from observations of clusters of small redshift will be proportional to the assumed value of $H_0^{-3/2}$. It is believed that most of the baryonic mass in a cluster of galaxies is in the

hot gas outside the galaxies, and if we suppose that this mass is the same fraction of the total mass as in the universe as a whole,[15] then we should get the same value of $\rho_B(0)/\rho_M(0)$, equal to Ω_B/Ω_M, for all clusters, whatever value we assume for H_0, but this value of Ω_B/Ω_M will be proportional to the assumed value of $H_0^{-3/2}$. For example, Schindler[16] quotes various studies that give $\rho_B(0)/\rho_M(0)$ as 0.14, 0.11, 0.12, and 0.12 for $H_0 = 65$ km s^{-1} Mpc^{-1}, so if we take the average 0.12 of these values as the cosmic value of Ω_B/Ω_M for $H_0 = 65$ km s^{-1} Mpc^{-1}, then for a general Hubble constant we find

$$\Omega_B/\Omega_M \simeq 0.06\, h^{-3/2}\ , \tag{1.9.22}$$

where h as usual is Hubble's constant in units of 100 km s^{-1} Mpc^{-1}. We can thus conclude pretty definitely that *only a small fraction of the mass in clusters of galaxies is in a baryonic form that can emit X-rays.*

On the other hand, when we study clusters with a range of redshifts that are not all small, we will not get a uniform value of $\rho_B(0)/\rho_M(0)$ unless we use values of $d_A(z)$ with the correct dependence on z. As remarked in Section 1.6, observations of clusters have been used in this way to learn about the z-dependence of $d_A(z)$.

It should be mentioned that computer simulations that treat galaxy clusters as assemblages of collisionless particles do not show evidence for a central core,[17] but instead indicate that the dark matter density at small distances r from the center should diverge as r^{-1} to $r^{-3/2}$. On the other hand, it has been shown[18] that the density of a baryonic gas in hydrostatic equilibrium in the gravitational field of such a distribution of dark matter does exhibit the core expected from Eq. (1.9.18). In any case, the dark matter and baryonic gas densities do have the same distributions at distances from the center that are larger than r_0.

As we will see in Section 3.2, it is possible to infer a value for $\Omega_B h^2$ from the abundances of deuterium and other light isotopes, which together with Eq. (1.9.22) can be used to derive a value for $\Omega_M h^{1/2}$. There are several other methods for estimating Ω_M or $\Omega_M h^2$ that will be discussed elsewhere

[15] This is argued by S. D. M. White, J. F. Navarro, A. E. Evrard, and C. S. Frenk, *Nature* **366**, 429 (1993). Calculations supporting this assumption are described in Section 8.3.

[16] S. Schindler, in *Space Science Reviews* **100**, 299 (2002), ed. P. Jetzer, K. Pretzl, and R. von Steiger (Kluwer) [astro-ph/0107028].

[17] J. F. Navarro, C. S. Frenk, and S. D. M. White, *Astrophys. J.* **462**, 563 (1996) [astro-ph/9508025]; **490**, 493 (1997) [astro-ph/9610188]; T. Fukushige and J. Makino, *Astrophys. J.* **477**, L9 (1997) [astro-ph/9610005]; B. Moore *et al., Mon. Not. Roy. Astron. Soc.* **499**, L5 (1998).

[18] N. Makino, S. Sasaki, and Y. Suto, *Astrophys. J.* **497**, 555 (1998). Also see Y. Suto, S. Sasaki, and M. Makino, *Astrophys. J.* **509**, 544 (1998); E. Komatsu and U. Seljak, *Mon. Not. Roy. Astron. Soc.* **327**, 1353 (2001).

in this book, using gravitational lenses (Section 9.3), the Sunyaev–Zel'dovich effect (Section 2.5), and anisotropies in the cosmic microwave background (Sections 2.6 and 7.2), the last of which also gives a value for $\Omega_B h^2$. In addition to these, there are methods[19] based on the evolution of clusters of galaxies, cosmic flows, cluster correlations, etc., that depend on detailed dynamical theories of structure formation.

1.10 Intergalactic absorption

Some of the cosmic gas of nuclei and electrons from which the first galaxies and clusters of galaxies condensed must be still out there in intergalactic space. Atoms or molecules in this gas could be observed through the resonant absorption of light or radio waves from more distant galaxies or quasars, but it is believed that most of the gas was ionized by light from a first generation of hot massive stars, now long gone, that are sometimes called stars of Population III. It now appears that some quasars formed before this ionization was complete, giving us the opportunity to observe the intergalactic gas through resonant absorption of the light from these very distant quasars.

Let us suppose that an atomic transition in a distant source produces a ray of light of frequency ν_1 that leaves the source at time t_1 and arrives at the Earth with frequency ν_0 at time t_0. At time t along its journey the light will have frequency redshifted to $\nu_1 a(t_1)/a(t)$, so if the intergalactic medium absorbs light of frequency ν at a rate (per proper time) $\Lambda(\nu, t)$, and does not emit light, then the intensity $I(t)$ of the light ray will decrease according to the equation

$$\dot{I}(t) = -\Lambda\Big(\nu_1 a(t_1)/a(t), t\Big)\, I(t) \ .$$

But if the intergalactic gas is at a non-zero temperature $T(t)$, then photons will also be added to the light ray through the process of stimulated emission, as a rate per photon given by the Einstein formula[1] $\exp(-h\nu/k_B T)\,\Lambda(\nu, t)$, so the intensity of the light ray will satisfy

$$\dot{I}(t) = -\left[1 - \exp\left(-\frac{h\nu_1 a(t_1)}{k_B T(t) a(t)}\right)\right] \Lambda\Big(\nu_1 a(t_1)/a(t), t\Big)\, I(t) \quad (1.10.1)$$

The intensity observed at the earth will then be

$$I(t_0) = \exp(-\tau) I(t_1) \ , \quad (1.10.2)$$

[19] For surveys, see N. A. Bahcall, *Astrophys. J.* **535**, 593 (2000) [astro-ph/0001076]; M. Turner, *Astrophys. J.* **576**, L101 (2002) [astro-ph/0106035]; S. Schindler, *op. cit.*; K. A. Olive, lectures given at Theoretical Advanced Study Institute on Elementary Particle Physics, Boulder, June 2002 [astro-ph/0301505].

[1] A. Einstein, *Phys. Z.* **18**, 121 (1917).

where τ is the *optical depth*:

$$\tau = \int_{t_1}^{t_0} \left[1 - \exp\left(-\frac{h\nu_1 a(t_1)}{k_{\mathcal{B}} T(t) a(t)}\right)\right] \Lambda(\nu_1 a(t_1)/a(t), t)\, dt \ . \quad (1.10.3)$$

The absorption rate is given by

$$\Lambda(\nu, t) = n(t)\, \sigma(\nu) \ , \quad (1.10.4)$$

where $\sigma(\nu)$ is the absorption cross section at frequency ν, and $n(t)$ is the number density (per proper volume) of absorbing atoms. Often the absorption cross section is sharply peaked at some frequency ν_R, so the absorption takes place only close to a time t_R, given by

$$a(t_R) = \nu_1 a(t_1)/\nu_R \ . \quad (1.10.5)$$

Therefore the optical depth can be approximated as

$$\tau \simeq n(t_R) \left[1 - \exp\Big(-h\nu_R/k_{\mathcal{B}} T(t_R)\Big)\right] \int \sigma\Big(\nu_1 a(t_1)/a(t)\Big)\, dt \ .$$

By changing the variable of integration from time to frequency, we can write this as

$$\tau \simeq n(t_R) \left[1 - \exp\Big(-h\nu_R/k_{\mathcal{B}} T(t_R)\Big)\right] [a(t_R)/\dot{a}(t_R)] \mathcal{I}_R \ , \quad (1.10.6)$$

where

$$\mathcal{I}_R \equiv \frac{1}{\nu_R} \int \sigma(\nu)\, d\nu \ , \quad (1.10.7)$$

the integral being taken over a small range of frequencies containing the absorption line. The only thing in the formula for τ that depends on a cosmological model is the Hubble expansion rate $\dot{a}(t_R)/a(t_R)$ at the time of absorption, given by Eq. (1.5.19) and (1.5.38) as

$$\frac{\dot{a}(t_R)}{a(t_R)} = H_0 \sqrt{\Omega_\Lambda + \Omega_K (1+z_R)^2 + \Omega_M (1+z_R)^3 + \Omega_R (1+z_R)^4} \ , \quad (1.10.8)$$

where $z_R = a(t_0)/a(t_R) - 1 = \nu_R/\nu_0 - 1$ is the redshift of the location of the resonant absorption. For a source at redshift z, the absorption takes place over a range of *observed* frequencies $\nu_0 = \nu_1/(1+z)$ given by the condition that the time t_R defined by Eq. (1.10.5) should be between t_1 and t_0:

$$\nu_R/(1+z) \leq \nu_0 \leq \nu_R \ . \quad (1.10.9)$$

For example, in 1959 Field[2] suggested looking for the effects of absorption of radio frequencies in the 21 cm transition in hydrogen atoms, caused in transitions from the spin zero to spin one hyperfine states in the $1s$ state of intergalactic hydrogen. Here $\nu_R = 1420$ MHz, so the radio spectrum of the galaxy Cygnus A at a redshift $z = 0.056$ should show an absorption trough (1.10.9) from 1342 MHz to 1420 MHz. Unfortunately, the temperature of neutral hydrogen in intergalactic space is much larger than $h\nu_R/k_{\mathcal{B}} = 0.068$K, so the optical depth (1.10.6) is suppressed by a factor $\simeq 0.068\ \mathrm{K}/T(t_R)$. No sign of this absorption trough has been discovered. It is hoped that in the future a new generation of low frequency radio telescopes with good angular resolution may be able to use the emission and absorption of 21 cm radiation at large redshifts to study both the growth of structure and primordial density perturbations from which they grew.[3] For instance, by 2010 the Low Frequency Array (LOFAR) should be able to study 21 cm radiation from sources at redshift between 5 and 15 with good sensitivity and high angular resolution.[4]

For the present, a much better probe of intergalactic hydrogen atoms is provided by absorption of photons in the Lyman α transition from the $1s$ ground state to the $2p$ excited state, known as the *Gunn–Peterson effect.*[5] This has a resonant frequency in the ultraviolet, $\nu_R = 2.47 \times 10^{15}$ Hz, corresponding to a wavelength 1,215 Å, but for a source of redshift $z > 1.5$ the lower part or the absorption trough (1.10.9) will be observable on the Earth's surface at wavelengths greater than 3,000 Å, in the visible or infrared part of the spectrum. Here $h\nu_R/k_{\mathcal{B}} = 118{,}000$ K, which is likely to be larger than the temperature of the intergalactic medium, in which case the factor $1 - \exp\Big(-h\nu_R/k_{\mathcal{B}}T(t_R)\Big)$ in Eq. (1.10.6) can be set equal to unity. The integral (1.10.7) here has the value 4.5×10^{-18} cm^2, so Eq. (1.10.6) gives the optical depth just above the lower end of the absorption trough (1.10.9) as

$$\tau_{\nu_0=\nu_R/(1+z)+} = \left(\frac{n(t_R)}{2.4\,h\,\times 10^{-11}\ \mathrm{cm}^{-3}}\right)\Big(\Omega_\Lambda + \Omega_K(1+z)^2 + \Omega_M(1+z)^3 + \Omega_R(1+z)^4\Big)^{-1/2}, \tag{1.10.10}$$

where again h is Hubble's constant in units of 100 km s^{-1} Mpc^{-1}. For instance, if a fraction f of the baryons of the universe at a time corresponding

[2] G. Field, *Astrophys. J.* **129**, 525 (1959).

[3] A. Loeb and M. Zaldarriaga, *Phys. Rev. Lett.* **92**, 211301 (2004) [astro-ph/0312134]; S. Furlanetto, S. P. Oh, and F. Briggs, *Phys. Rep.* **433**, 181 (2006) [astro-ph/0608032].

[4] H. J. A. Röttgering *et al.*, in *Cosmology, Galaxy Formation, and Astroparticle Physics on the Pathway to the SKA*, eds. H.-R. Klöckner *et al.* [astro-ph/0610596].

[5] J. E. Gunn and B. A. Peterson, *Astrophys. J.* **142**, 1633 (1965). Also see I. S. Shklovsky, *Astron. Zh.* **41**, 408 (1964); P. A. G. Scheuer, *Nature* **207**, 963 (1965).

to $z = 5$ were in the form of neutral intergalactic hydrogen atoms, and $\Omega_B h^2 = 0.02$, then the number density of hydrogen atoms at $z = 5$ would be $4.8f \times 10^{-5}$ cm^{-3}. Taking $h = 0.65$, $\Omega_\Lambda = 0.7$, $\Omega_M = 0.3$, and $\Omega_K = \Omega_R = 0$, the optical depth (1.10.10) would be $3.8f \times 10^5$. Thus with these parameters intergalactic neutral hydrogen that makes up a fraction of baryonic matter $f \gg 2.6 \times 10^{-6}$ would completely block any light with a frequency above the redshifted Lyman α line from sources beyond $z = 5$. Evidently the Gunn–Peterson effect provides a very sensitive probe of even a small proportion of neutral hydrogen atoms.

For many years the search for the Lyman α absorption trough was unsuccessful. Quasar spectra show numerous Lyman α absorption lines, forming what are sometimes called "Lyman α forests," which are believed to arise from clouds of neutral hydrogen atoms along the line of sight, but for quasars out to $z \approx 5$ there was no sign of a general suppression of frequencies above the redshifted Lyman α frequency,[6] that would be produced by even a small fraction f of the baryons in the universe in the form of neutral intergalactic hydrogen atoms. Then in 2001 the spectrum of the quasar SDSSp J103027.10+052455.0 with redshift $z = 6.28$ discovered by the Sloan Digital Sky Survey was found to show clear signs of a complete suppression of light in the wavelength range from just below the redshifted Lyman α wavelength at 8,845 Å down to 8,450 Å, indicating a significant fraction f of baryons in the form of neutral intergalactic hydrogen atoms at redshifts greater than $8,450/1,215 - 1 = 5.95$.[7] (See Figure 1.8.) Thus a redshift of order 6 may mark the end of a "dark age," in which the absorption of light by neutral hydrogen atoms made the universe opaque to light with frequencies above the redshifted Lyman α frequency. Further evidence for this conclusion is supplied by the spectrum of intense gamma ray sources, known as *gamma ray bursters*, at large redshifts.[8]

This does not mean that all or even most of the hydrogen in the universe was in the form of neutral atoms at $z > 6$. As we have seen, even small concentrations of neutral hydrogen could have produced an absorption trough in the spectrum of distant quasars. In fact, we shall see in Chapter 7 that there is now some evidence from the study of the cosmic microwave background that hydrogen became mostly ionized at redshifts considerably larger than $z \approx 6$, perhaps around $z \approx 10$.

[6] A. Songalia, E. Hu, L. Cowie, and R. McMahon, *Astrophys. J.* **525**, L5 (1999).

[7] R. H. Becker *et al.*, *Astron. J.* **122**, 2850 (2001) [astro-ph/0108097]. See S. G. Djorgovski *et al.*, *Astrophys. J.* **560**, L5 (2001) [astro-ph/0108069], for a hint of absorption by neutral hydrogen at slightly smaller redshifts. Also see X. Fan *et al.*, *Astrophys. J.* **123**, 1247 (2002) [astro-ph/0111184].

[8] T. Totani *et al.*, *Publ. Astron. Soc. Pacific* **58**, 485 (2006) [astro-ph/0512154].

The clouds of neutral hydrogen at redshifts $z < 6$ which produce the Lyman α forest can provide an independent means of measuring Ω_M and Ω_Λ. The idea goes back to a 1979 paper of Alcock and Paczyński.[9] Suppose we observe a luminous object at a redshift z that extends a proper distance $D_\perp$ perpendicular to the line of sight and a proper distance $D_\parallel$ along the line of sight. According to the definition of the angular diameter distance, the object will subtend an angle

$$\Delta\theta = D_\perp / d_A(z) \ . \tag{1.10.11}$$

Also, when we observe light from the whole object at the same time t_0, the difference in the time t_1 that the light was emitted from the far and near points of the object will be $\Delta t_1 = D_\parallel$. The redshift is $a(t_0)/a(t_1) - 1$, so the absolute value of the difference of redshift from the far and near points of the object will be

$$\Delta z = \frac{a(t_0)}{a^2(t_1)} \dot{a}(t_1) \Delta t_1 = (1+z) H(z) D_\parallel \ , \tag{1.10.12}$$

where $H(z) \equiv \dot{a}(t_1)/a(t_1)$ is the Hubble constant at the time of emission. Taking the ratio, we have

$$\frac{\Delta z}{\Delta\theta} = (1+z)\, H(z)\, d_A(z) \Big(D_\parallel / D_\perp \Big) \tag{1.10.13}$$

It is then only necessary to use Eq. (1.5.19) to write $H(z)$ as

$$\begin{aligned} H(z) &= \sqrt{\left(\frac{8\pi G}{3}\right)\Big(\rho_{M0}(1+z)^3 + \rho_V + \rho_{R0}(1+z)^4\Big) - \frac{K}{a_0^2}(1+z)^2} \\ &= H_0 \sqrt{\Omega_M (1+z)^3 + \Omega_\Lambda + \Omega_R (1+z)^4 + \Omega_K (1+z)^2} \ , \end{aligned} \tag{1.10.14}$$

and use Eqs. (1.4.12) and (1.5.45) to write $d_A(z)$ as

$$\begin{aligned} d_A(z) = {} & \frac{1}{(1+z) H_0 \Omega_K^{1/2}} \\ & \times \sinh \left[\Omega_K^{1/2} \int_{1/(1+z)}^1 \frac{dx}{x^2 \sqrt{\Omega_\Lambda + \Omega_K x^{-2} + \Omega_M x^{-3} + \Omega_R x^{-4}}} \right] . \end{aligned} \tag{1.10.15}$$

[9] C. Alcock and B. Paczyński, *Nature* **281**, 358 (1979).

The Hubble constant H_0 cancels in the product, and we find a result that depends only on z, $D_\parallel/D_\perp$, and the Ωs:

$$\frac{\Delta z}{\Delta\theta} = \Big(D_\parallel/D_\perp\Big)\Omega_K^{-1/2}\sqrt{\Omega_M(1+z)^3+\Omega_\Lambda+\Omega_R(1+z)^4+\Omega_K(1+z)^2} \times \sinh\left[\Omega_K^{1/2}\int_{1/(1+z)}^{1}\frac{dx}{x^2\sqrt{\Omega_\Lambda+\Omega_K x^{-2}+\Omega_M x^{-3}+\Omega_R x^{-4}}}\right]. \tag{1.10.16}$$

For instance, if the object is known to be a sphere, such as a spherical cluster of galaxies, then $D_\parallel/D_\perp = 1$, and we can use a measurement of Δz and $\Delta\theta$ to set a model-independent constraint on the Ωs, with no need to worry about effects of evolution or intergalactic absorption.

Unfortunately, it is not so easy to find spherical objects at large redshift. But there are various objects whose distribution functions *are* spherically symmetric. For instance, the distribution of field galaxies is presumably spherically symmetric about any point in space, and it has been proposed that the application of the Alcock–Paczyński method to galaxies might allow a determination of the cosmological constant.[10] This method has been applied[11] instead to the distribution of quasars measured in the 2dF QSO Redshift Survey.[12] Assuming $K = 0$, this analysis gives $\Omega_\Lambda = 0.71^{0.09}_{-0.17}$.

Recently the Alcock–Paczyński idea has been applied to the distribution function of Lyman α clouds.[13] As already mentioned, these are intergalactic clouds containing neutral hydrogen atoms, which absorb light from more distant quasars along the line of sight in $1s \to 2p$ transitions, showing up as dark lines in the spectrum of the quasar at wavelengths $1215\,(1+z)$ Å for clouds at redshift z. Suppose we measure the number density $N(z,\hat{n})$ of Lyman α clouds at various redshifts z in various directions $\hat{n}$. Assuming a spherically symmetric distribution of Lyman α clouds, the mean value of the product of the number densities of these clouds at two nearby points with redshifts z and $z+\Delta z$ (with $\Delta z \ll 1$) and directions $\hat{n}$ and $\hat{n}+\Delta\hat{n}$ separated by a small angle $\Delta\theta$ will be a function only of z and the proper distance between the points, and will be analytic in the components of the vector

[10] W. E. Ballinger, J. A. Peacock, and A. F. Heavens, *Mon. Not. Roy. Astron. Soc.* **281**, 877 (1996).

[11] P. J. Outram *et al.*, *Mon. Not. Roy. Astron. Soc.* **348**, 745 (2004) [astro-ph/0310873].

[12] S. M. Croom *et al.*, *Mon. Not. Roy. Astron. Soc.* **349**, 1397 (2004); available at `www.2dfquasar.org`.

[13] L. Hui, A. Stebbins, and S. Burles, *Astrophys. J.* **511**, L5 (1999); P. McDonald and J. Miralda-Escudeé, *Astrophys. J.* **518**, 24 (1999); W-C. Lin and M. L. Norman, talk at the Theoretical Astrophysics in Southern California meeting, Santa Barbara, October 2002 [astro-ph/0211177]; P. McDonald, *Astrophys. J.* **585**, 34 (2003).

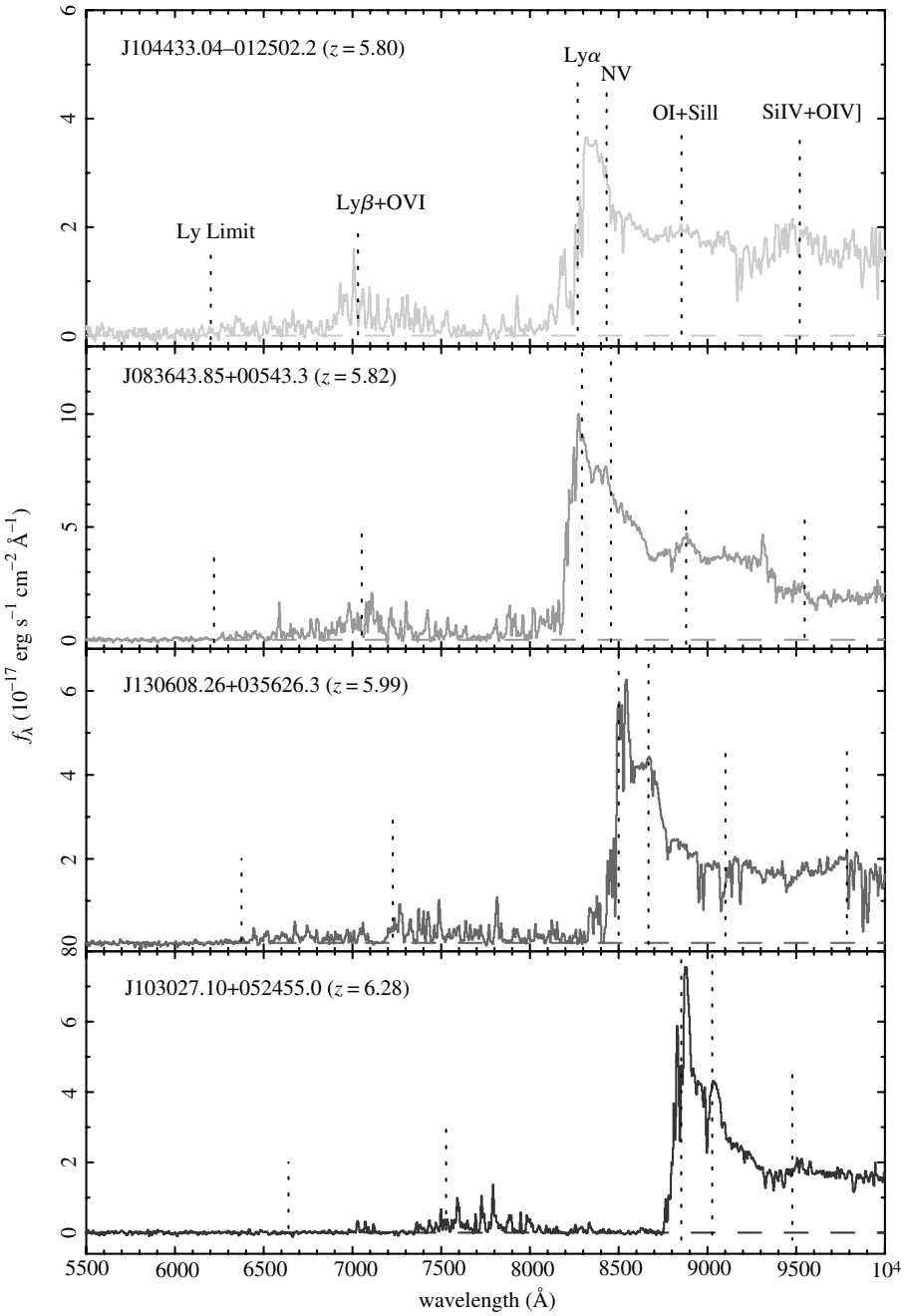

Figure 1.8: Observed intensity versus wavelength for four high-redshift quasars, from R. H. Becker *et al.*, *Astron. J.* **122**, 2850 (2001) [astro-ph/0108097]. Vertical dashed lines indicate the redshifted wavelengths for various spectral lines. In the direction of the quasar with $z = 6.28$ the intensity drops to zero within experimental accuracy just to the left of the Lyman α line at 8845 Å, a feature not seen for the quasar with $z = 5.99$, indicating the onset of patches of nearly complete ionization at a redshift between 5.99 and 6.28.

separating these components, so for small separations it can be written

$$\langle N(z,\hat{n})\, N(z+\Delta z,\hat{n}+\Delta\hat{n})\rangle \simeq \langle N^2(z,\hat{n})\rangle \left[1-\frac{D_\perp^2+D_\parallel^2}{L^2(z)}\right], \quad (1.10.17)$$

where $D_\perp$ and $D_\parallel$ are given by Eqs. (1.10.11) and (1.10.12), and L is some correlation length. This can be written in terms of the observed Δz and $\Delta\theta$, as

$$\langle N(z,\hat{n})\, N(z+\Delta z,\hat{n}+\Delta\hat{n})\rangle \simeq \langle N^2(z,\hat{n})\rangle \left[1-\frac{\Delta z^2}{L_z^2(z)}-\frac{\Delta\theta^2}{L_\theta^2(z)}\right], \quad (1.10.18)$$

where L_z and L_θ are correlation lengths for redshift and angle

$$L_\theta(z)=\frac{L(z)}{d_A(z)}, \qquad L_z(z)=L(z)(1+z)H(z)\,. \quad (1.10.19)$$

By measuring this product for various redshifts and directions, we can infer a value for the ratio of correlation lengths, which is independent of L:

$$\frac{L_z(z)}{L_\theta(z)}=\Omega_K^{-1/2}\sqrt{\Omega_M(1+z)^3+\Omega_\Lambda+\Omega_R(1+z)^4+\Omega_K(1+z)^2}$$
$$\times\sinh\left[\Omega_K^{1/2}\int_{1/(1+z)}^{1}\frac{dx}{x^2\sqrt{\Omega_\Lambda+\Omega_K x^{-2}+\Omega_M x^{-3}+\Omega_R x^{-4}}}\right]. \quad (1.10.20)$$

This method has been applied[14] to five pairs of close quasars, with redshifts in the range from 2.5 to 3.5 and separations ranging from 33 to 180 arcseconds. Use of this limited sample sets only weak constraints on the Ωs, but it rules out $\Omega_\Lambda = 0$ at the level of 2 standard deviations.

1.11 Number counts

A uniform distribution of sources with a smooth distribution of absolute luminosity leads in ordinary Euclidean space to a unique distribution in apparent luminosity. If there are $N(L)dL$ sources per unit volume with absolute luminosity between L and $L+dL$, then the number $n(>\ell)$ of

[14] A. Lidz, L. Hui, A. P. S. Crotts, and M. Zaldarriaga, astro-ph/0309204 (unpublished).

sources observed with apparent luminosity greater than ℓ is given by

$$n(>\ell) = \int_0^\infty N(L)\,dL \int_0^{\sqrt{L/4\pi\ell}} 4\pi r^2\,dr$$
$$= \frac{1}{3\sqrt{4\pi}\,\ell^{3/2}} \int_0^\infty L^{3/2}\,N(L)\,dL \qquad (1.11.1)$$

Thus whatever the distribution in absolute luminosity, we expect that $n(>\ell) \propto \ell^{-3/2}$.

This analysis needs several changes in a cosmological setting:

1. Instead of the volume element $r^2 \sin\theta\,dr\,d\theta\,d\phi$, the proper volume element here is $(\mathrm{Det}\,g^{(3)})^{1/2} dr\,d\theta\,d\phi$, where $g^{(3)}_{ij} \equiv a^2\tilde{g}_{ij}$ is the three-dimensional metric, with non-vanishing components $g^{(3)}_{rr} = a^2/(1-Kr^2)$, $g^{(3)}_{\theta\theta} = a^2r^2$, $g^{(3)}_{\phi\phi} = a^2r^2\sin^2\theta$, so

$$dV = \frac{a^3(t)\,r^2\,\sin\theta\,dr\,d\theta\,d\phi}{\sqrt{1-Kr^2}}\ . \qquad (1.11.2)$$

2. The apparent luminosity is related to the absolute luminosity by

$$\ell = \frac{L}{4\pi\,d_L^2(z)}\ , \qquad (1.11.3)$$

where $d_L(z)$ is the luminosity distance (1.4.3).

3. Except in the steady state cosmology, the number density of sources changes with time, even if only through the cosmic expansion.

4. We can often measure the redshift z as well as the apparent luminosity.

Eq. (1.11.2) gives the number of sources with redshift between z and $z + dz$ and apparent luminosity between ℓ and $\ell + d\ell$ as

$$n(z,\ell)\,dz\,d\ell = 4\pi\mathcal{N}(t,L)dL\,\frac{a^3(t)\,r^2\,dr}{\sqrt{1-Kr^2}}\ , \qquad (1.11.4)$$

where $\mathcal{N}(t,L)\,dL$ is the number of sources per proper volume at time t with absolute luminosity between L and $L + dL$; t and z are related by $1 + z = a(t_0)/a(t)$, and t and r are related by Eq. (1.2.2):

$$\int_t^{t_0} \frac{dt'}{a(t')} = \int_0^r \frac{dr'}{\sqrt{1-Kr'^2}}\ . \qquad (1.11.5)$$

We use (1.11.5) to express the differential dr in terms of dt, and then express dt in terms of dz:

$$\frac{dr}{\sqrt{1-Kr^2}} = -\frac{dt}{a(t)} = \frac{dz}{H(z)\,a_0} \ ,$$

where $H(z) \equiv \dot{a}(t)/a(t)$ and $a_0 \equiv a(t_0)$. As a reminder, for a universe containing radiation, matter, and a constant vacuum energy, Eq. (1.5.41) gives

$$H(z) = H_0\sqrt{\Omega_\Lambda + \Omega_K(1+z)^2 + \Omega_M(1+z)^3 + \Omega_R(1+z)^4} \ .$$

Canceling dz in Eq. (1.11.4), we then have

$$n(z,\ell)\,d\ell = \frac{4\,\pi\,\mathcal{N}\Big(t(z),L\Big)\,r^2(z)a_0^2\,dL}{(1+z)^3\,H(z)} \ ,$$

We next use Eq. (1.11.3) to write (with z now held fixed):

$$dL = 4\pi\,d_L^2(z)\,d\ell \ ,$$

so that canceling $d\ell$ gives

$$n(z,\ell) = \frac{16\,\pi^2\,\mathcal{N}\Big(t(z),4\pi\,d_L^2(z)\ell\Big)\,d_L^4(z)}{H(z)\,(1+z)^5} \ , \tag{1.11.6}$$

in which we have used Eq. (1.4.3) to express $a_0 r$ in terms of d_L.

In particular, for a sample of sources that are not evolving at a time $t(z)$, the time dependence of the number density $\mathcal{N}$ is just proportional to $a^{-3} \propto (1+z)^3$:

$$\mathcal{N}\Big(t(z),L\Big) = (1+z)^3\mathcal{N}_0(L) \ . \tag{1.11.7}$$

If all members of this sample are bright enough to be visible at a redshift z, then the total number of sources observed with redshifts between z and $z+dz$ will be $n(z)\,dz$, where

$$n(z) \equiv \int_0^\infty n(z,\ell)\,d\ell = \frac{4\,\pi\,\mathcal{N}_0\,d_L^2(z)}{H(z)\,(1+z)^2} \tag{1.11.8}$$

where $d_L(z)$ is given by Eq. (1.5.45), and

$$\mathcal{N}_0 \equiv \int_0^\infty \mathcal{N}_0(L)\,dL \ . \tag{1.11.9}$$

In principle, even without knowing $\mathcal{N}_0$ or H_0, if $n(z)$ were accurately measured we could compare the observed *shape* of this function with Eq. (1.11.8) to find the Ωs.

There are several obvious dangers in using Eq. (1.11.8) in this way. For one thing, it is necessary to avoid missing sources that have high redshift and hence low apparent luminosity. Also, evolution in the number of sources can introduce an additional dependence on the light emission time t, and hence on z. In 1986 Loh and Spillar[1] carried out a survey of galaxy numbers as a function of redshift. The redshifts were measured photometrically (i. e., from their luminosities at various colors rather than by the shift of specific spectral lines), which generally gives less reliable results. Comparing their results with Eq. (1.11.8) in the case $\Omega_K = \Omega_R = 0$ (so that $\Omega_\Lambda + \Omega_M = 1$), they found that $\Omega_\Lambda/\Omega_M = 0.1^{-0.4}_{+0.2}$. By now it has been realized that the evolution of sources cannot be neglected at redshifts large enough for $n(z)$ to be sensitive to cosmological parameters, and this result for Ω_Λ/Ω_M has been abandoned.[2]

Useful results can be obtained when evolution is taken into account. One group[3] used number counts of very faint galaxies[4] as a function of apparent luminosity to estimate the free parameters in a model of galactic luminosity evolution (assuming the number of galaxies per coordinate volume to be constant), and then used this model together with a redshift survey[5] extending to $z \simeq 0.47$ to conclude that Ω_M is small and that Ω_Λ is in the range of 0.5 to 1. More recently, several surveys[6] of numbers of galaxies at different redshifts that yield important results about galactic evolution, and with the use of dynamical models they can yield information about Ω_M and Ω_Λ.[7] But it appears that number counts of galaxies will be more useful in learning about galactic evolution than in making precise determinations of cosmological parameters. In a dramatic application of this approach,[8] a

[1] E. D. Loh, *Phys. Rev. Lett.* **57**, 2865 (1986); E. D. Loh and E. J. Spillar, *Astrophys. J.* **284**, 439 (1986).

[2] For a discussion of future prospects for measuring Ω_Λ in redshift surveys, see W. E. Ballinger, J. A. Peacock, and A. F. Heavens, *Mon. Not. Roy. Astron. Soc.* **282**, 877 (1996).

[3] M. Fukugita, F. Takahara, K. Yamashita, and Y. Yoshii, *Astrophys. J.* **361**, L1 (1990).

[4] J. A. Tyson, *Astron. J.* **96**, 1 (1988).

[5] T. J. Broadhurst, R. S. Ellis, and T. Shanks, *Mon. Not. Roy. Astron. Soc.* **235**, 827 (1988).

[6] G. Efstathiou, R. S. Ellis, B. A. Peterson, *Mon. Not. Roy. Astron. Soc.* **232**, 431 (1988); J. Loveday, B. A. Peterson, G. Efstathiou, and S. J. Maddox, *Astrophys. J.* **390**, 338 (1992); L. da Costa, in *Proceedings of the Conference on Evolution of Large Scale Structure*, Garching, August 1998 [astro-ph/9812258]; S. Borgani, P. Rosati, P. Tozzi, and C. Norman, *Astrophys. J.* **517**, 40 (1999) [astro-ph/9901017]; S. J. Oliver, in *Highlights of the ISO Mission: Special Scientific Session of the IAU General Assembly*. eds. D. Lemke *et al.* (Kluwer) [astro-ph/9901272]; M. Colless, in Publ. Astron. Soc. Australia [astro-ph/9911326]; S. Rawlings, astro-ph/0008067.

[7] W. J. Percival *et al.*, *Mon. Not. Roy. Astron. Soc.* **327**, 1297 (2001) [astro-ph/0105252]; S. Borgnani *et al.*, *Astrophys. J.* **561**, 13 (2001) [astro-ph/0106428].

[8] R. J. Bouwens and G. D. Illingworth, *Nature* **443**, 189 (2006).

search at the Lick Observatory for galaxies with redshifts in the range $z \approx$ 7 to 8 found at most just one galaxy, while it is estimated that if Eq. (1.11.7) were valid then, on the basis of the number of galaxies observed (with the same conservative selection criteria) at redshifts $z \approx 6$, ten galaxies should have been found with $z \approx 7$ to 8. The implication is that there must have been a spurt in the formation of luminous galaxies at a redshift in the range 6 to 7. This fits in well with the conclusion discussed in Section 1.10, that the ionization of intergalactic hydrogen became essentially complete at a redshift of order 6, presumably due to ultraviolet radiation from massive stars formed around that time.

* * *

Historically the first important application of number counts was in radio source surveys, where redshifts are not generally available. These surveys take place at a fixed receiving frequency ν, corresponding to a variable emitted frequency $\nu(1+z)$, so the source counts are affected by the frequency dependence of the distribution of intrinsic source powers.

If a source with a redshift z emits a power[9] $P(\nu)d\nu$ between frequencies ν and $\nu + d\nu$, then the power received at the origin per unit antenna area between frequencies ν and $\nu + d\nu$ is

$$S(\nu)d\nu = \frac{P\Big(\nu(1+z)\Big)\, d\nu\,(1+z)}{4\pi\, d_L^2(z)}\ . \tag{1.11.10}$$

Many radio sources have a "straight" spectrum, i.e.

$$P(\nu) \propto \nu^{-\alpha} \tag{1.11.11}$$

with the spectral index α typically about 0.7 to 0.8. This allows a great simplification in Eq. (1.11.10):

$$S(\nu)d\nu = \frac{P(\nu)\, d\nu}{4\pi\ d_L^2(z)(1+z)^{\alpha-1}}\ . \tag{1.11.12}$$

From now on we will take the observed frequency ν as fixed, and write $S(\nu) = S$ and $P(\nu) = P$. Canceling $d\nu$, Eq. (1.11.12) then reads

$$S = \frac{P}{4\pi\ d_L^2(z)\,(1+z)^{\alpha-1}}\ . \tag{1.11.13}$$

[9] In G&C, P was defined as the power emitted per solid angle, while here it is the power emitted in all directions.

If at time t there are $N(P,t)\,dP$ sources per proper volume with power between P and $P+dP$, then the number of sources observed with power per antenna area greater than S is

$$n(>S) = \int_0^\infty dP \int N(P,t)\,\frac{4\pi r^2 a^3(t)\,dr}{\sqrt{1-Kr^2}}\ , \tag{1.11.14}$$

with the upper limit on the integral over r set by the condition that

$$a_0^2\, r^2\,(1+z)^{1+\alpha} < \frac{P}{4\pi\, S}\ . \tag{1.11.15}$$

Of course, r, z, and t are related by the familiar formulas

$$\int_0^r \frac{dr'}{\sqrt{1-Kr'^2}} = \int_t^{t_0} \frac{dt'}{a(t')}\ , \qquad 1+z = a(t_0)/a(t)\ . \tag{1.11.16}$$

This becomes much simpler if we assume that the time-dependence of the source number density can be parameterized as

$$N(P,t) = N(P)\left(\frac{a(t)}{a_0}\right)^\beta\ . \tag{1.11.17}$$

For instance, if sources do not evolve and are neither created nor destroyed, then $\beta = -3$, while in the steady-state model $\beta = 0$. Eq. (1.11.14) now reads

$$n(>S) = a_0^3 \int_0^\infty N(P)\,dP \int \frac{4\pi r^2 (1+z)^{-\beta-3}\,dr}{\sqrt{1-Kr^2}}\ , \tag{1.11.18}$$

with the same P/S-dependent upper limit (1.11.15) on r.

The coordinate r is given in terms of z by the power series (1.4.8)

$$a_0 H_0\, r = z - \tfrac{1}{2}(1+q_0)z^2 + \dots\ . \tag{1.11.19}$$

We can then convert the integral over r to one over z, with

$$a_0 H_0\, dr = dz\,[1-(1+q_0)z + \dots]\ , \tag{1.11.20}$$

and the upper limit on z is given by

$$z^2\,[1+z(\alpha-q_0)+\dots] < \frac{PH_0^2}{4\pi\, S}\ ,$$

or, in other words,

$$z < \sqrt{\frac{PH_0^2}{4\pi\, S}}\left[1-\frac{1}{2}(\alpha-q_0)\sqrt{\frac{PH_0^2}{4\pi\, S}}+\dots\right]\ . \tag{1.11.21}$$

Then Eq. (1.11.18) becomes

$$n(>S) = \frac{4\pi}{H_0^3}\int_0^\infty N(P)\,dP \times\left[\frac{1}{3}\left(\frac{PH_0^2}{4\pi S}\right)^{3/2}\left(1-\frac{3}{2}(\alpha-q_0)\left(\frac{PH_0^2}{4\pi S}\right)^{1/2}+\dots\right) -\frac{1}{4}\left(\frac{PH_0^2}{4\pi S}\right)^{2}(\beta+5+2q_0)+\dots\right],$$

or, collecting terms,

$$n(>S) = \frac{1}{3\sqrt{4\pi}\,S^{3/2}}\int_0^\infty P^{3/2}N(P)\,dP \times\left[1-\frac{3}{4}\Big(5+\beta+2\alpha\Big)\left(\frac{PH_0^2}{4\pi S}\right)^{1/2}+\dots\right]. \quad (1.11.22)$$

We see that $n(>S)$ has a term with the familiar $S^{-3/2}$ dependence, plus a correction proportional to S^{-2} with a coefficient proportional to $5+\beta+2\alpha$. It is noteworthy that this coefficient is independent of q_0 or K. For the standard cosmology with no evolution of sources $\beta=-3$, and we have mentioned that $\alpha\approx 0.75$, so $5+\beta+2\alpha=3.5$. Although the precise value is uncertain, this coefficient is definitely positive, which means that for faint sources $n(>S)$ should fall off more *slowly* than $S^{-3/2}$. This is definitely not what is observed.[10] It has been known for many years that for $S>5\times10^{-26}\mathrm{Wm}^{-2}/\mathrm{Hz}$, the source count function $N(>S)$ falls off more rapidly than $S^{-3/2}$. The conclusion is inevitable that the number of radio sources per co-moving volume is decreasing, with $\beta<-6.5$. Radio source counts are useful in studying this evolution, but not for measuring cosmological parameters.

On the other hand, for the steady state cosmology (discussed in Section 1.5) we have $\beta=0$, so the coefficient $5+\beta+2\alpha\approx 6.5$, and the predicted number count $N(>S)$ decreases even more slowly with S, making the disagreement with experiment even worse than for the standard cosmology with no evolution of sources. Here it is not possible to save the situation by appealing to evolution, because the essence of the steady state model is that on the average there is no evolution. This observation

[10] For a list of major radio source surveys, and references to the original literature, see G&C, Sec. 14.8.

discredited the steady state model even before the discovery of the cosmic microwave radiation background.

1.12 Quintessence

So far, we have taken into account only non-relativistic matter, radiation, and a constant vacuum energy in calculating the rate of expansion of the universe. It appears that the vacuum energy is not only much smaller than would be expected from order-of-magnitude estimates based on the quantum theory of fields, but is only a few times greater than the present matter density. This has led to a widespread speculation that the vacuum energy is not in fact constant; it may now be small because the universe is old. A time-varying vacuum energy is sometimes called *quintessence*.[1]

The natural way to introduce a varying vacuum energy is to assume the existence of one or more scalar fields, on which the vacuum energy depends, and whose cosmic expectation values change with time. Scalar fields of this sort play a crucial part in the modern theory of weak and electromagnetic interactions, and are also introduced in theories of inflation, as discussed in Chapters 4 and 10.

For simplicity, let us consider a single real scalar field $\varphi(\mathbf{x}, t)$. We will be concerned here with fields that are vary little on elementary particle spacetime scales, so the action of these field is taken to have a minimum number of spacetime derivatives:

$$I_\varphi = -\int d^4x \, \sqrt{-\mathrm{Det}g} \left[\frac{1}{2} g^{\lambda\kappa} \frac{\partial\varphi}{\partial x^\lambda} \frac{\partial\varphi}{\partial x^\kappa} + V(\varphi) \right] , \qquad (1.12.1)$$

with an unspecified potential function $V(\varphi)$. We are interested here in the case of a Robertson–Walker metric, and a scalar field that depends only on time, not position. In this case the formulas (B.66) and (B.67) for the scalar field energy density and pressure become

$$\rho_\varphi = \frac{1}{2}\dot{\varphi}^2 + V(\varphi) \qquad (1.12.2)$$

$$p_\varphi = \frac{1}{2}\dot{\varphi}^2 - V(\varphi) \, . \qquad (1.12.3)$$

It follows immediately that $(1 + w)\rho_\varphi \geq 0$, where $w \equiv p_\varphi/\rho_\varphi$, so as long as $\rho_\varphi \geq 0$ this model has $w \geq -1$, and the phantom energy disaster discussed in Section 1.6 does not occur.

[1] For reviews with references to the original literature, see B. Ratra and P. J. E. Peebles, *Rev. Mod. Phys.* **75**, 559 (2003); E. V. Linder, 0704.2064.

The equation (1.1.32) of energy conservation here reads

$$\ddot{\varphi} + 3H\dot{\varphi} + V'(\varphi) = 0 \,, \tag{1.12.4}$$

(where as usual $H(t) \equiv \dot{a}(t)/a(t)$), which is the same as the field equation derived from the action (1.12.1). This is the equation of a particle of unit mass with one-dimensional coordinate φ, moving in a potential $V(\varphi)$ with a frictional force $-3H\dot{\varphi}$. The field will run toward lower values of $V(\varphi)$, finally coming to rest if it can reach any field value where $V(\varphi)$ is at least a local minimum. Unfortunately, we do not know any reason why the value of $V(\varphi)$ where it is stationary should be small.

Nevertheless, there are potentials that have some attractive properties once we adjust an additive constant in the potential to make them vanish at their stationary point. The original and simplest example is provided by a potential[2]

$$V(\varphi) = M^{4+\alpha}\varphi^{-\alpha} \,, \tag{1.12.5}$$

where α is positive but otherwise arbitrary, and M is a constant with the units of mass (taking $\hbar = c = 1$), which gives $V(\varphi)$ the dimensions of an energy density. There is no special reason to believe that the potential has this form, and in particular there is no known reason for excluding an additive constant (including effects of quantum fluctuations in all other fields), which would give the potential a non-zero value at its stationary point, at $\varphi = \infty$. Nevertheless, it may be illuminating to work out the consequences of this one specific model of quintessence.

For any potential it is necessary to assume that at sufficiently early times ρ_φ was much less than the energy density ρ_R of radiation because, as we will see in Section 3.2, any appreciable increase in the energy density at the time of cosmological nucleosynthesis would lead to a helium abundance exceeding what is observed. At these early times the energy density of radiation (including particles like neutrinos with masses less than $k_{\mathcal{B}}T$) is also greater than that of non-relativistic matter, so Eq. (1.5.34) gives $a(t) \propto t^{1/2}$, and therefore $H = 1/2t$. The field equation (1.12.4) with potential (1.12.5) then reads

$$\ddot{\varphi} + \frac{3}{2t}\dot{\varphi} - \alpha M^{4+\alpha}\varphi^{-\alpha-1} = 0 \,. \tag{1.12.6}$$

[2] P. J. E. Peebles and B. Ratra, *Astrophys. J.* **325**, L17 (1988); B. Ratra and P. J. E. Peebles, *Phys. Rev.* **D 37**, 3406 (1988); C. Wetterich, *Nucl. Phys.* **B302**, 668 (1988). Quintessence models with this potential were intensively studied by I. Zlatev, L. Wang, and P. J. Steinhardt, *Phys. Rev. Lett.* **82**, 896 (1999); P. J. Steinhardt, L. Wang, and I. Zlatev, *Phys. Rev.* **D 59**, 123504 (1999).

This has a solution

$$\varphi = \left(\frac{\alpha(2+\alpha)^2 M^{4+\alpha} t^2}{6+\alpha} \right)^{\frac{1}{2+\alpha}} \quad . \tag{1.12.7}$$

Both $\dot{\varphi}^2$ and $V(\varphi)$ then go as $t^{-2\alpha/(2+\alpha)}$, and therefore at very early times ρ_φ must have been less than ρ_R, which goes as t^{-2}. This solution is not unique, but it is an attractor, in the sense that any other solution that comes close to it will approach it as t increases. (To see this, note that a small perturbation $\delta\varphi$ of the solution (1.12.7) will satisfy

$$0 = \delta\ddot{\varphi} + \frac{3}{2t}\delta\dot{\varphi} + \alpha(1+\alpha)M^{4+\alpha}\varphi^{-\alpha-2}\delta\varphi = \delta\ddot{\varphi} + \frac{3}{2t}\delta\dot{\varphi} + \frac{(6+\alpha)(1+\alpha)}{(2+\alpha)^2 t^2}\delta\varphi \ .$$

This has two independent solutions of the form

$$\delta\varphi \propto t^\gamma \ , \quad \gamma = -\frac{1}{4} \pm \sqrt{\frac{1}{16} - \frac{(6+\alpha)(1+\alpha)}{(2+\alpha)^2}} \ .$$

The square root is imaginary for $\alpha > 0$, so both solutions for $\delta\varphi$ decay as $t^{-1/4}$ for increasing t, while φ itself is increasing.) For this reason, the particular solution of Eq. (1.12.6) that goes as Eq. (1.12.7) for $t \to 0$ is known as the *tracker solution*. There is no particular physical reason to require that the initial conditions for the scalar field are such that the scalar field has approached the tracker solution by the present moment (the set of such initial conditions is called the "basin of attraction"), but since this requirement would make the present evolution of the scalar field insensitive to the initial conditions, it has the practical advantage of providing a model of quintessence with just two free parameters: M and α.

Nothing much changes when the radiation energy density drops below the energy density of non-relativistic matter. The tracker solution for the scalar field continues to grow as $t^{2/(2+\alpha)}$ (though with a different constant factor), so $\dot{\varphi}^2$ and $V(\varphi)$ continue to fall as $t^{-2\alpha/(2+\alpha)}$. But ρ_M and ρ_R are decreasing faster, like t^{-2} and $t^{-8/3}$, respectively, so eventually ρ_M and ρ_R will fall below ρ_φ. It is interesting that the value of φ where ρ_φ becomes equal to ρ_M is independent of the unknown constant M. When the expansion is dominated by matter ρ_M is given by Eq. (1.5.31) as $1/6\pi G t^2$, while (1.1.2), (1.12.5) and (1.12.7) give $\rho_\varphi \approx M^{2(4+\alpha)/(2+\alpha)} t^{-2\alpha/(2+\alpha)}$, so the time t_c at which $\rho_\varphi = \rho_M$ is of order

$$t_c \approx M^{-(4+\alpha)/2} G^{-(2+\alpha)/4} \ . \tag{1.12.8}$$

Using this in Eq. (1.12.7) then gives

$$\varphi(t_c) \approx G^{-1/2} \ . \tag{1.12.9}$$

Once ρ_M falls well below ρ_φ, the equation of motion of $\varphi(t)$ becomes

$$\ddot{\varphi} + \sqrt{24\pi G\rho_\varphi}\,\dot{\varphi} - \alpha M^{4+\alpha}\varphi^{-\alpha-1} = 0 \ , \tag{1.12.10}$$

with ρ_φ given by Eq. (1.12.2). The tracker solution in this era has a complicated time dependence, but it becomes simple again at sufficiently late times, times that may be later than the present. We can guess that the damping term proportional to $\dot{\varphi}$ in this equation will eventually slow the growth of φ, so that $\dot{\varphi}^2$ will become less than $V(\varphi)$, and also guess that the inertial term proportional to $\ddot{\varphi}$ will become negligible compared to the damping and potential terms. (Similar "slow roll" conditions will play an important role in the theory of inflation, described in Chapters 4 and 10.) Equation (1.12.10) then becomes

$$\sqrt{24\pi G M^{4+\alpha}\varphi^{-\alpha}}\,\dot{\varphi} = \alpha M^{4+\alpha}\varphi^{-\alpha-1} \ ,$$

and so

$$\dot{\varphi} = \frac{\alpha M^{2+\alpha/2}\varphi^{-\alpha/2-1}}{\sqrt{24\pi G}} \ . \tag{1.12.11}$$

The solution is

$$\varphi = M\left(\frac{\alpha(2+\alpha/2)\,t}{\sqrt{24\pi G}}\right)^{1/(2+\alpha/2)} \ . \tag{1.12.12}$$

(In general this involves a redefinition of the zero of time, to avoid a possible integration constant that might be added to t.) We can now check the approximations used in deriving Eq. (1.12.11), of which this is the solution. From Eq. (1.12.12) we see that $\dot{\varphi}^2 \propto t^{-(2+\alpha)/(2+\alpha/2)}$ while $V(\varphi) \propto t^{-\alpha/(2+\alpha/2)}$, so the kinetic energy term in Eq. (1.12.2) does become small compared with the potential term at late times. Also, $\ddot{\varphi} \propto t^{-(3+\alpha)/(2+\alpha/2)}$ while $V'(\varphi) \propto t^{-(1+\alpha)/(2+\alpha/2)}$, so the inertial term in Eq. (1.12.10) does become small compared with the potential term at late times. Eq. (1.12.12) is therefore a valid asymptotic solution of Eq. (1.12.10) for $t \to \infty$. Numerical calculations show that it is not only a solution for $t \to \infty$; it is the asymptotic form approached for $t \to \infty$ by the tracker solution.

With $\rho_\varphi \propto t^{-\alpha/(2+\alpha/2)}$ dominating the expansion rate at late times, we have $\dot{a}/a \propto t^{-\alpha/2(2+\alpha/2)}$, so

$$\ln a \propto t^{2/(2+\alpha/2)} \ . \tag{1.12.13}$$

This is a similar but less rapid growth of a than would be produced by a cosmological constant, for which $\ln a \propto t$. The difference between the deceleration parameter q_0 and the value -1 for an expansion dominated by a cosmological constant vanishes as $t^{-(2+\alpha)/(2+\alpha/2)}$. Note that the radiation and matter densities decrease as $1/a^4$ and $1/a^3$ respectively, and the curvature decreases as $1/a^2$, all of which have a much faster rate of decrease with time than the power-law decrease of ρ_φ, so the expansion rate is indeed dominated by ρ_φ at late times, justifying the derivation of Eq. (1.12.10).

We have found that, at least for a range of initial conditions, the potential (1.12.5) leads to an expansion that is dominated by radiation and then matter at early times, but becomes dominated by the scalar field energy at late times. But to get agreement with observation it is necessary arbitrarily to exclude a large constant term that might be added to (1.12.5), and also to adjust the value of M to make the critical time (1.12.8) at which the values of ρ_φ and ρ_M cross be close to the present moment $t_0 \approx 1/H_0$. Specifically, Eq. (1.12.8) shows that we need the constant factor in $V(\varphi)$ to take the value

$$M^{4+\alpha} \approx G^{-1-\alpha/2} H_0^2 \,. \tag{1.12.14}$$

There is no known reason why this should be the case.

Several groups of observers are now planning programs to discover whether the vacuum energy density is constant, as in the case of a cosmological constant, or changing with time. In such programs, one would compare the observed luminosity distance (or angular diameter distance) with a formula obtained by replacing the term Ω_Λ in the argument of the square root in Eq. (1.5.45) with a time-varying dark energy term. These observations will not actually measure the value w_0 of w at the present time, much less the present time derivatives $\dot{w}_0$, $\ddot{w}_0$, etc., because for that purpose it would be necessary to have extremely precise measurements of the luminosity distance or angular-diameter distance for small redshifts. Instead, measurements will be made with only moderate precision, but over a fairly large range of redshifts. To compare such measurements with theory, one needs a model of the time-variation of the dark energy. One model is simply to assume that w is constant, or perhaps varying linearly with time or redshift, but there is no physical model that entails such behavior.[3] It seems preferable to compare observation with the model of a scalar field rolling down a potential, which (whatever reservations may have about its naturalness) at least provides a physically possible model of varying dark

[3] Other assumptions about the form of w as a function of redshift that can mimic scalar field models have been considered by J. Weller and A. Albrecht, *Phys. Rev. D* **65**, 103512 (2002); E. V. Linder, *Phys. Rev. Lett.* **90**, 091301 (2003) [astro-ph/0208512].

energy.[4] Because these observations are difficult, it pays to adopt scalar field models with just two parameters, which can if we like be expressed in terms of $\Omega_\Lambda = 1 - \Omega_M$ (assuming flatness and neglecting the radiation energy density) and w_0.

One possibility is to suppose that over the latest e-folding of cosmic expansion the scalar field φ has taken values for which $V(\varphi)$ is only slowly varying. If $V(\varphi)$ were constant, we would have a constant vacuum energy, with $w = -1$, and the only parameter to measure would be Ω_V. For a two-parameter fit, we can take $V(\varphi)$ to vary linearly with φ:

$$V(\varphi) = V_0 + \Big(\varphi - \varphi_0\Big) V_0' \,. \tag{1.12.15}$$

This is valid if the fractional change in $V'(\varphi)$ in a time interval of order $1/H_0$ is small; that is, if $|V_0''\dot{\varphi}_0| \ll H_0|V_0'|$.

The field equation (1.12.4) for $\varphi(t)$ can be put in a convenient dimensionless form by replacing the dependent variable t and independent variable φ with dimensionless variables x and ω, defined by

$$x \equiv H_0\sqrt{\Omega_M}\, t \,, \qquad \omega \equiv \frac{8\pi G V(\varphi)}{3\Omega_M H_0^2} \,. \tag{1.12.16}$$

Because V is linear in φ, we have

$$\dot{\varphi} = \frac{3\Omega_M H_0^2 \dot{\omega}}{8\pi G V_0'} = \frac{3\Omega_M^{3/2} H_0^3}{8\pi G V_0'} \frac{d\omega}{dx} \,.$$

Then Eq. (1.12.4) becomes

$$\frac{d^2\omega}{dx^2} + 3\mathcal{H}\frac{d\omega}{dx} + \lambda = 0 \,, \tag{1.12.17}$$

where λ is the dimensionless parameter

$$\lambda \equiv \frac{8\pi G V_0'^2}{3H_0^4\Omega_M^2} \,, \tag{1.12.18}$$

and $\mathcal{H}$ is a function of ω and $d\omega/dx$:

$$\mathcal{H} \equiv \frac{H}{H_0\sqrt{\Omega_M}} = \sqrt{(1+z)^3 + \omega + \frac{1}{2\lambda}\left(\frac{d\omega}{dx}\right)^2} \,. \tag{1.12.19}$$

[4] This approach is followed by D. Huterer and H. V. Peiris, *Phys. Rev. D* **75**, 083502 (2007) [astro-ph/0610427]; R. Crittenden, E. Majerotto, and F. Piazza, astro-ph/0702003.

We will also need the differential equation for the redshift:

$$\frac{dz}{dx} = -\mathcal{H}(1+z) \ . \tag{1.12.20}$$

In general, even if we wrote all derivatives with respect to x in terms of derivatives with respect to z, to solve these equations we would need initial conditions for ω and $d\omega/dz$ at some initial z, which with λ would give a three-parameter set of solutions. However, assuming that for large redshift the energy density is dominated by matter rather than vacuum energy (which as we shall see is the case), the derivative $d\omega/dx$ sufficiently late in the matter-dominated era becomes quite insensitive to initial conditions.[5] For $z \gg 1$, Eq. (1.12.19) gives

$$\mathcal{H} \to (1+z)^{3/2} \ , \tag{1.12.21}$$

and (1.12.17) and (1.12.20) then have the solution

$$1+z \to \left(\frac{3x}{2}\right)^{-2/3} \ , \qquad \frac{d\omega}{dx} \to -\frac{\lambda x}{3} \ . \tag{1.12.22}$$

(An integration constant in the solution for z has been absorbed into the definition of x, setting the zero of time. An integration constant in the solution for $d\omega/dx$ has been dropped, because it gives a term in $d\omega/dx$ that dies away with increasing time as $x^{-2} \propto t^{-2}$.) The free parameters in our solution are then λ, together with the value of ω at some arbitrary initial value x_1 of x, taken sufficiently small so that at x_1 the energy density is dominated by matter rather than vacuum energy. (Note that the constant V_0 appears nowhere in these equations; it contributes a term to $\omega(x_1)$, but there is no need to isolate this term.) One must adopt various trial values of λ and $\omega(x_1)$; use Eq. (1.12.22) to calculate $1+z$ and $d\omega/dx$ at $x = x_1$; with these initial conditions, integrate the differential equations (1.12.17) and (1.12.20) numerically from x_1 to a value x_0 where $z = 0$; and then if we like calculate the values of $\Omega_V = 1 - \Omega_M$ and the present value w_0 of the ratio p_φ/ρ_φ for this particular solution,[6] using

$$\frac{\Omega_\Lambda}{\Omega_M} = \omega(x_0) + \frac{1}{2\lambda}\left(\frac{d\omega}{dx}\right)^2_{x=x_0} \ , \quad w_0 = \frac{(d\omega/dx)^2_{x=x_0} - 2\lambda\omega(x_0)}{(d\omega/dx)^2_{x=x_0} + 2\lambda\omega(x_0)} \ . \tag{1.12.23}$$

[5] R. Cahn, private communication. Cahn has also shown that the approximation of neglecting the second derivative term in the field equation does not work well in this context.

[6] As already mentioned, with models of this sort one can only have $w_0 > -1$. To compare the case $w_0 < -1$ with observation, it is necessary to adopt a model with the opposite sign for the derivative term in the action (1.12.1). The analysis given here can then be applied, with only obvious sign changes here and there.

The ratio of the dark energy at a given time to its value at the present is

$$\xi \equiv \frac{\rho_V(t)}{\rho_V(t_0)} = \frac{(d\omega(x)/dx)^2 + 2\lambda\,\omega(x)}{(d\omega/dx)^2_{x=x_0} + 2\,\lambda\omega(x_0)} \tag{1.12.24}$$

For instance, if we take $\Omega_\Lambda = 1 - \Omega_M = 0.76$ and $w_0 = -0.777$, the ratio ξ of the dark energy density to its value at present rises to 1.273 at $z = 1$ and to 1.340 at infinite redshift.[7] The leveling off of $\xi(z)$ for large z occurs because the growth of the matter density for increasing redshift makes the expansion rate grow, so that the friction term $3H\dot{\varphi}$ in Eq. (1.12.4) freezes the value of the scalar field at early times.

It should not be thought that the leveling off of the dark energy for large z for the potential (1.12.15) means that in analyzing dark energy observations with this potential one must give up the idea motivating theories of quintessence, that the vacuum energy is now small because the universe is old. In fact, for the potential $V(\varphi) \propto \varphi^{-\alpha}$, for typical initial conditions the quintessence energy drops at first precipitously, and then levels off while the scalar field rolls slowly down the potential until the field approaches the tracker solution, with the tracker solution not reached by the present time if α is small.[8] The condition $|V_0''\dot{\varphi}_0| \ll H_0|V_0'|$ for treating this potential as linear over a time of order $1/H_0$ is satisfied if $\alpha(1+\alpha)\varphi_0^{-2} \ll 8\pi G$, which in light of Eq. (1.12.9) is likely to be satisfied if $\alpha < 1$.

Another possible two-parameter model is provided by the same potential, $V(\varphi) \propto \varphi^{-\alpha}$, but now under the assumption that the tracker solution is reached by some early time (say, for $z \leq 10$) in the matter-dominated era. With this assumption the observable history of dark energy is insensitive to initial conditions, so the model has just two parameters: M and α. The equations of this model can be put in dimensionless form by writing the coupling constant of this potential in terms of a dimensionless parameter β as

$$M^{4+\alpha} \equiv \beta\,\Omega_M\,H_0^2\,(8\pi G)^{-1-\alpha/2} \tag{1.12.25}$$

and replacing the dependent variable t and independent variable φ with dimensionless variables x and f, defined by

$$t \equiv x/H_0\sqrt{\Omega_M}\,, \qquad \varphi(t) \equiv f(x)/\sqrt{8\pi G}\,. \tag{1.12.26}$$

[7] Numerical results for various values of redshift are given in Table 1.1. These results for the linear potential were calculated by R. Cahn, private communication.

[8] Steinhardt, Wang, and Zlatev, ref. 2.

The field equation (1.12.4) (with no slow roll approximation) in the era dominated by matter and vacuum energy then takes the form

$$\frac{d^2f}{dx^2} + 3\mathcal{H}\frac{df}{dx} - \alpha\beta f^{-\alpha-1} = 0 \ , \tag{1.12.27}$$

where

$$\mathcal{H} \equiv H/\sqrt{\Omega_M}H_0 = \sqrt{\frac{1}{6}\left(\frac{df}{dx}\right)^2 + \frac{\beta}{3}f^{-\alpha} + (1+z)^3} \tag{1.12.28}$$

in which we also need

$$\frac{dz}{dx} = -\mathcal{H}(1+z) \ . \tag{1.12.29}$$

Because the large z solution (1.12.7) is an attractor, the initial conditions introduce no new free parameters; in terms of these dimensionless variables, the initial conditions are that, for $x \to 0$,

$$f \to \left[\frac{\alpha\beta(\alpha+2)^2x^2}{2(\alpha+4)}\right]^{1/(\alpha+2)} \ , \quad 1+z \to \left(\frac{2}{3x}\right)^{2/3} \ . \tag{1.12.30}$$

We need to integrate the equations (1.12.27) and (1.12.29) from some small x (say, $x = 0.01$) to a value x_0 at which $z = 0$, with the initial conditions (1.12.30), and then evaluate $\Omega_M = 1 - \Omega_\Lambda$ from the condition that $\mathcal{H}(x_0) = 1/\sqrt{\Omega_M}$. We can also evaluate the present value w_0 of $w \equiv p_\varphi/\rho_\varphi$ from the formula

$$w_0 = \frac{f'^2(x_0)f(x_0)^\alpha/2\beta - 1}{f'^2(x_0)f(x_0)^\alpha/2\beta + 1} \ , \tag{1.12.31}$$

and then replace the parameters α and β with Ω_M and w_0. For instance, if we arbitrarily take $\alpha = 1$, then to get the realistic value $\Omega_M = 0.24$ we must take $\beta = 9.93$, in which case $w_0 = -0.777$. Of course, we can get any other values of w_0 greater than -1 by choosing different values of α and re-adjusting β to give the same value of Ω_M (though for small α, the range of initial conditions that allow the tracker solution to be reached well before the present is relatively small.) For instance, for $\alpha = 1/2$ we must take $\beta = 7.82$ to have $\Omega_M = 0.24$, and in this case we calculate that $w_0 = -0.87$. (For the case $w < -1$, see footnote 6.) The ratios of dark energy to its value at present calculated in this way for $\Omega_M = 0.24$ and $w_0 = -0.777$ are shown in Table 1.1, along with the values calculated with the same choice of Ω_M and w_0 for both the case of constant w and for the linear potential (1.12.15). The tracker and linear models evidently represent opposite extreme assumptions about the time-dependence of dark

Table 1.1: Ratio of dark energy to its present value, for the tracker solution with the potential (1.12.5), and for the linear potential (1.12.15), calculated for $\Omega_M = 1 - \Omega_\Lambda = 0.24$ and $w_0 = -0.777$, compared with the results for a constant $w = -0.777$.

z	tracker	linear	constant w
0	1	1	1
0.1	1.067	1.062	1.066
0.5	1.347	1.200	1.312
1	1.712	1.273	1.590
2	2.469	1.318	2.086
3	3.224	1.331	2.528
$\gg 1$	$\gg 1$	1.340	$\gg 1$

energy, but both are better motivated physically than the assumption of a constant w.

1.13 Horizons

Modern cosmological theories can exhibit horizons of two different types, which limit the distances at which past events can be observed or at which it will ever be possible to observe future events. These are called by Rindler[1] *particle horizons* and *event horizons*, respectively.

According to Eq. (1.2.2), if the big bang started at a time $t = 0$, then the greatest value $r_{\max}(t)$ of the Robertson–Walker radial coordinate from which an observer at time t will be able to receive signals traveling at the speed of light is given by the condition

$$\int_0^t \frac{dt'}{a(t')} = \int_0^{r_{\max}(t)} \frac{dr}{\sqrt{1 - Kr^2}} \tag{1.13.1}$$

Thus there is a particle horizon unless the integral $\int dt/a(t)$ does not converge at $t = 0$. It does converge in conventional cosmological theories; whatever the contribution of matter or vacuum energy at the present, it is likely that the energy density will be dominated by radiation at early times, in which case $a(t) \propto t^{1/2}$, and the integral converges. The proper distance

[1] W. Rindler, *Mon. Not. Roy. Astron. Soc.* **116**, 663 (1956).

of the horizon is given by Eq. (1.1.15) and (1.13.1) as

$$d_{\max}(t) = a(t) \int_0^{r_{\max}(t)} \frac{dr}{\sqrt{1-Kr^2}} = a(t) \int_0^t \frac{dt'}{a(t')} \,. \tag{1.13.2}$$

For instance, during the radiation-dominated era $a(t) \propto t^{1/2}$, so $d_{\max}(t) = 2t = 1/H$. Well into the matter-dominated era most of the integral over time in Eq. (1.13.1) comes from a time when $a \propto t^{2/3}$, so that $d_{\max}(t) \simeq 3t = 2/H$. At present most of the integral over t' comes from a period when the expansion is dominated by matter and the vacuum energy, and perhaps curvature as well. According to Eq. (1.5.41), the particle horizon distance at present is

$$d_{\max}(t_0) = \frac{1}{H_0} \int_0^1 \frac{dx}{x^2\sqrt{\Omega_\Lambda + \Omega_K x^{-2} + \Omega_M x^{-3}}} \,. \tag{1.13.3}$$

We will see in Chapter 4 that there may have been a time before the radiation-dominated era in which there was nothing in the universe but vacuum energy, in which case the particle horizon distance would actually be infinite. But as far as telescopic observations are concerned, Eq. (1.13.3) gives the proper distance beyond which we cannot now see.

Just as there are past events that we cannot now see, there may be events that we never will see. Again returning to Eq. (1.2.2), if the universe re-collapses at a time T, then the greatest value $r_{\rm MAX}$ of r from which an observer will be able to receive signals traveling at the speed of light emitted at any time later than t is given by the condition

$$\int_t^T \frac{dt'}{a(t')} = \int_0^{r_{\rm MAX}(t)} \frac{dr}{\sqrt{1-Kr^2}} \tag{1.13.4}$$

Even if the future is infinite, if the integral $\int dt/a(t)$ converges at $t = \infty$ there will be an event horizon given by

$$\int_t^\infty \frac{dt'}{a(t')} = \int_0^{r_{\rm MAX}(t)} \frac{dr}{\sqrt{1-Kr^2}} \tag{1.13.5}$$

Since co-moving sources are labeled with a fixed value of r, the condition $r < r_{\rm MAX}$ limits the events occurring at time t that we can ever observe. In the case where the universe does not recollapse, the proper distance to the event horizon is given by

$$d_{\rm MAX}(t) = a(t) \int_0^{r_{\rm MAX}(t)} \frac{dr}{\sqrt{1-Kr^2}} = a(t) \int_t^\infty \frac{dt'}{a(t')} \,. \tag{1.13.6}$$

In the absence of a cosmological constant, $a(t)$ grows like $t^{2/3}$, and the integral diverges, so that there is no event horizon. But with a cosmological constant $a(t)$ will eventually grow as $\exp(Ht)$ with $H = H_0\Omega_\Lambda^{1/2}$ constant, and there really is an event horizon, which approaches the value $d_{\rm MAX}(\infty) = 1/H$. As time passes all sources of light outside our gravitationally bound Local Group will move beyond this distance, and become unobservable. The same is true for the quintessence theory described in the previous section. In that case $a(t)$ eventually grows as $\exp(\text{constant} \times t^{2/(2+\alpha/2)})$, so for any $\alpha \geq 0$ the integral (1.13.6) again converges.

If a source is at a radial coordinate r in a Robertson–Walker coordinate system based on us, then we are at a radial coordinate r in a Robertson–Walker coordinate system based on the source. Hence Eq. (1.13.4) or (1.13.5) also gives the greatest radial coordinate to which, starting at time t, we will ever be able to travel.