# 1

# From Newton to Hubble



The history of ideas on the structure and origin of the Universe shows that humankind has always put itself at the center of creation. As astronomical evidence has accumulated, these anthropocentric convictions have had to be abandoned one by one. From the natural idea that the solid Earth is at rest and the celestial objects all rotate around us, we have come to understand that we inhabit an average-sized planet orbiting an average-sized sun, that the Solar System is in the periphery of a rotating galaxy of average size, flying at hundreds of kilometres per second towards an unknown goal in an immense Universe, containing billions of similar galaxies.

Cosmology aims to explain the origin and evolution of the entire contents of the Universe, the underlying physical processes, and thereby to obtain a deeper understanding of the laws of physics assumed to hold throughout the Universe. Unfortunately, we have only one universe to study, the one we live in, and we cannot make experiments with it, only observations. This puts serious limits on what we can learn about the origin. If there are other universes we will never know.

Although the history of cosmology is long and fascinating, we shall not trace it in detail, nor any further back than Newton, accounting (in Section 1.1) only for those ideas which have fertilized modern cosmology directly, or which happened to be right although they failed to earn timely recognition. In the early days of cosmology, when little was known about the Universe, the field was really just a branch of philosophy.

Having a rigid Earth to stand on is a very valuable asset. How can we describe motion except in relation to a fixed point? Important understanding has come from the study of inertial systems, in uniform motion with respect to one another. From the work of Einstein on inertial systems, the theory of special relativity was born. In Section 1.2 we discuss inertial frames, and see how expansion and contraction are natural consequences of the homogeneity and isotropy of the Universe.

*Introduction to Cosmology*, Fourth Edition. Matts Roos

A classic problem is why the night sky is dark and not blazing like the disc of the Sun, as simple theory in the past would have it. In Section 1.3 we shall discuss this so-called Olbers' paradox, and the modern understanding of it.

The beginning of modern cosmology may be fixed at the publication in 1929 of Hubble's law, which was based on observations of the redshift of spectral lines from remote galaxies. This was subsequently interpreted as evidence for the expansion of the Universe, thus ruling out a static Universe and thereby setting the primary requirement on theory. This will be explained in Section 1.4. In Section 1.5 we turn to determinations of cosmic timescales and the implications of Hubble's law for our knowledge of the age of the Universe.

In Section 1.6 we describe Newton's theory of gravitation, which is the earliest explanation of a gravitational force. We shall 'modernize' it by introducing Hubble's law into it. In fact, we shall see that this leads to a cosmology which already contains many features of current Big Bang cosmologies.

## 1.1 Historical Cosmology

At the time of *Isaac Newton* (1642–1727) the heliocentric Universe of *Nicolaus Copernicus* (1473–1543), *Galileo Galilei* (1564–1642) and *Johannes Kepler* (1571–1630) had been accepted, because no sensible description of the motion of the planets could be found if the Earth was at rest at the center of the Solar System. Humankind was thus dethroned to live on an average-sized planet orbiting around an average-sized sun.

The stars were understood to be suns like ours with fixed positions in a static Universe. The Milky Way had been resolved into an accumulation of faint stars with the telescope of Galileo. The *anthropocentric view* still persisted, however, in locating the Solar System at the center of the Universe.

**Newton's Cosmology.** The first theory of gravitation appeared when Newton published his *Philosophiae Naturalis Principia Mathematica* in 1687. With this theory he could explain the empirical laws of Kepler: that the planets moved in elliptical orbits with the Sun at one of the focal points. An early success of this theory came when *Edmund Halley* (1656–1742) successfully predicted that the comet sighted in 1456, 1531, 1607 and 1682 would return in 1758. Actually, the first observation confirming the heliocentric theory came in 1727 when *James Bradley* (1693–1762) discovered the aberration of starlight, and explained it as due to the changes in the velocity of the Earth in its annual orbit. In our time, Newton's theory of gravitation still suffices to describe most of planetary and satellite mechanics, and it constitutes the nonrelativistic limit of Einstein's relativistic theory of gravitation.

Newton considered the stars to be suns evenly distributed throughout infinite space in spite of the obvious concentration of stars in the Milky Way. A distribution is called *homogeneous* if it is uniformly distributed, and it is called *isotropic* if it has the same properties in all spatial directions. Thus in a homogeneous and isotropic space the distribution of matter would look the same to observers located anywhere—no point would be preferential. Each local region of an isotropic universe contains information which remains true also on a global scale. Clearly, matter introduces lumpiness which grossly violates homogeneity on the scale of stars, but on some larger scale isotropy and homogeneity may still be a good approximation. Going one step further, one may postulate what is called the *cosmological principle*, or sometimes the *Copernican principle*.

> *The Universe is homogeneous and isotropic in three-dimensional space, has always been so, and will always remain so.*

It has always been debated whether this principle is true, and on what scale. On the galactic scale visible matter is lumpy, and on larger scales galaxies form gravitationally bound clusters and narrow strings separated by voids. But galaxies also appear to form loose groups of three to five or more galaxies. Several surveys have now reached agreement that the distribution of these galaxy groups appears to be homogeneous and isotropic within a sphere of 170 Mpc radius [1]. This is an order of magnitude larger than the supercluster to which our Galaxy and our local galaxy group or Local Supercluster (LSC) belong, and which is centered in the constellation of Virgo. Based on his theory of gravitation, Newton formulated a cosmology in 1691. Since all massive bodies attract each other, a finite system of stars distributed over a finite region of space should collapse under their mutual attraction. But this was not observed, in fact the stars were known to have had fixed positions since antiquity, and Newton sought a reason for this stability. He concluded, erroneously, that the self-gravitation within a finite system of stars would be compensated for by the attraction of a sufficient number of stars outside the system, distributed evenly throughout infinite space. However, the total number of stars could not be infinite because then their attraction would also be infinite, making the static Universe unstable. It was understood only much later that the addition of external layers of stars would have no influence on the dynamics of the interior. The right conclusion is that the Universe cannot be static, an idea which would have been too revolutionary at the time.

Newton's contemporary and competitor *Gottfried Wilhelm von Leibnitz* (1646–1716) also regarded the Universe to be spanned by an abstract infinite space, but in contrast to Newton he maintained that the stars must be infinite in number and distributed all over space, otherwise the Universe would be bounded and have a center, contrary to contemporary philosophy. Finiteness was considered equivalent to boundedness, and infinity to unboundedness.

**Rotating Galaxies.** The first description of the Milky Way as a rotating galaxy can be traced to *Thomas Wright* (1711–1786), who wrote *An Original Theory or New Hypothesis of the Universe* in 1750, suggesting that the stars are

> *all moving the same way and not much deviating from the same plane, as the planets in their heliocentric motion do round the solar body.*

Wright's galactic picture had a direct impact on *Immanuel Kant* (1724–1804). In 1755 Kant went a step further, suggesting that the diffuse nebulae which Galileo had already observed could be distant galaxies rather than nearby clouds of incandescent gas. This implied that the Universe could be homogeneous on the scale of galactic distances in support of the cosmological principle.

Kant also pondered over the reason for transversal velocities such as the movement of the Moon. If the Milky Way was the outcome of a gaseous nebula contracting under Newton's law of gravitation, why was all movement not directed towards a common center? Perhaps there also existed repulsive forces of gravitation which would scatter bodies onto trajectories other than radial ones, and perhaps such forces at large distances would compensate for the infinite attraction of an infinite number of stars? Note that the idea of a contracting gaseous nebula constituted the first example of a nonstatic system of stars, but at galactic scale with the Universe still static.

Kant thought that he had settled the argument between Newton and Leibnitz about the finiteness or infiniteness of the system of stars. He claimed that either type of system embedded in an infinite space could not be stable and homogeneous, and thus the question of infinity was irrelevant. Similar thoughts can be traced to the scholar *Yang Shen* in China at about the same time, then unknown to Western civilization [2].

The infinity argument was, however, not properly understood until *Bernhard Riemann* (1826–1866) pointed out that the world could be *finite* yet *unbounded*, provided the geometry of the space had a positive curvature, however small. On the basis of Riemann's geometry, *Albert Einstein* (1879–1955) subsequently established the connection between the geometry of space and the distribution of matter.

Kant's repulsive force would have produced trajectories in random directions, but all the planets and satellites in the Solar System exhibit transversal motion in one and the same direction. This was noticed by *Pierre Simon de Laplace* (1749–1827), who refuted Kant's hypothesis by a simple probabilistic argument in 1825: the observed movements were just too improbable if they were due to random scattering by a repulsive force. Laplace also showed that the large transversal velocities and their direction had their origin in the rotation of the primordial gaseous nebula and the law of conservation of angular momentum. Thus no repulsive force is needed to explain the transversal motion of the planets and their moons, no nebula could contract to a point, and the Moon would not be expected to fall down upon us.

This leads to the question of the origin of time: what was the first cause of the rotation of the nebula and when did it all start? This is the question modern cosmology attempts to answer by tracing the evolution of the Universe backwards in time and by reintroducing the idea of a repulsive force in the form of a cosmological constant needed for other purposes.

**Black Holes.** The implications of Newton's gravity were quite well understood by *John Michell* (1724–1793), who pointed out in 1783 that a sufficiently massive and compact star would have such a strong gravitational field that nothing could escape from its surface. Combining the corpuscular theory of light with Newton's theory, he found that a star with the solar density and escape velocity $c$ would have a radius of $486R_\odot$ and a mass of 120 million solar masses. This was the first mention of a type of star much later to be called a *black hole* (to be discussed in Section 3.4). In 1796 Laplace independently presented the same idea.

**Galactic and Extragalactic Astronomy.** Newton should also be credited with the invention of the reflecting telescope—he even built one—but the first one of importance was built one century later by *William Herschel* (1738–1822). With this instrument, observational astronomy took a big leap forward: Herschel and his son John could map the nearby stars well enough in 1785 to conclude correctly that the Milky Way was a disc-shaped star system. They also concluded erroneously that the Solar System was at its center, but many more observations were needed before it was corrected. Herschel made many important discoveries, among them the planet Uranus, and some 700 binary stars whose movements confirmed the validity of Newton's theory of gravitation outside the Solar System. He also observed some 250 diffuse nebulae, which he first believed were distant galaxies, but which he and many other astronomers later considered to be nearby incandescent gaseous clouds belonging to our Galaxy. The main problem was then to explain why they avoided the directions of the galactic disc, since they were evenly distributed in all other directions.

The view of Kant that the nebulae were distant galaxies was also defended by *Johann Heinrich Lambert* (1728–1777). He came to the conclusion that the Solar System along, with the other stars in our Galaxy, orbited around the galactic center, thus departing from the heliocentric view. The correct reason for the absence of nebulae in the galactic plane was only given by *Richard Anthony Proctor* (1837–1888), who proposed the presence of interstellar dust. The arguments for or against the interpretation of nebulae as distant galaxies nevertheless raged throughout the 19th century because it was not understood how stars in galaxies more luminous than the whole galaxy could exist—these were observations of supernovae. Only in 1925 did *Edwin P. Hubble* (1889–1953) resolve the conflict indisputably by discovering Cepheids and ordinary stars in nebulae, and by determining the distance to several galaxies, among them the celebrated M31 galaxy in the *Andromeda*. Although this distance was off by a factor of two, the conclusion was qualitatively correct.

In spite of the work of Kant and Lambert, the heliocentric picture of the Galaxy—or almost heliocentric since the Sun was located quite close to Herschel's galactic center—remained long into our century. A decisive change came with the observations in 1915–1919 by *Harlow Shapley* (1895–1972) of the distribution of *globular clusters* hosting $10^5$–$10^7$ stars. He found that perpendicular to the galactic plane they were uniformly distributed, but along the plane these clusters had a distribution which peaked in the direction of the Sagittarius. This defined the center of the Galaxy to be quite far from the Solar System: we are at a distance of about two-thirds of the galactic radius. Thus the anthropocentric world picture received its second blow—and not the last one—if we count Copernicus's heliocentric picture as the first one. Note that Shapley still believed our Galaxy to be at the center of the astronomical Universe.

**The End of Newtonian Cosmology.** In 1883 *Ernst Mach* (1838–1916) published a historical and critical analysis of mechanics in which he rejected Newton's concept of an absolute space, precisely because it was unobservable. Mach demanded that the laws of physics should be based only on concepts which could be related to observations. Since motion still had to be referred to some frame at rest, he proposed replacing absolute space by an idealized rigid frame of fixed stars. Thus 'uniform motion' was to be understood as motion relative to the whole Universe. Although Mach clearly realized that all motion is relative, it was left to Einstein to take the full step of studying the laws of physics as seen by observers in inertial frames in relative motion with respect to each other.

Einstein published his General Theory of Relativity in 1917, but the only solution he found to the highly nonlinear differential equations was that of a static Universe. This was not so unsatisfactory though, because the then known Universe comprised only the stars in our Galaxy, which indeed was seen as static, and some nebulae of ill-known distance and controversial nature. Einstein firmly believed in a static Universe until he met Hubble in 1929 and was overwhelmed by the evidence for what was to be called Hubble's law.

Immediately after general relativity became known, *Willem de Sitter* (1872–1934) published (in 1917) another solution, for the case of empty space-time in an exponential state of expansion. In 1922 the Russian meteorologist *Alexandr Friedmann* (1888–1925) found a range of intermediate solutions to the Einstein equation which describe the standard cosmology today. Curiously, this work was ignored for a decade although it was published in widely read journals.

In 1924 Hubble had measured the distances to nine spiral galaxies, and he found that they were extremely far away. The nearest one, M31 in the Andromeda, is now known to be at a distance of 20 galactic diameters (Hubble's value was about 8) and the farther ones at hundreds of galactic diameters. These observations established that the spiral nebulae are, as Kant had conjectured, stellar systems comparable in mass and size with the Milky Way, and their spatial distribution confirmed the expectations of the cosmological principle on the scale of galactic distances.

In 1926–1927 *Bertil Lindblad* (1895–1965) and *Jan Hendrik Oort* (1900–1992) verified Laplace's hypothesis that the Galaxy indeed rotated, and they determined the period to be $10^8$ yr and the mass to be about $10^{11} M_\odot$. The conclusive demonstration that the Milky Way is an average-sized galaxy, in no way exceptional or central, was given only in 1952 by Walter Baade. This we may count as the third breakdown of the anthropocentric world picture.

The later history of cosmology up until 1990 has been excellently summarized by Peebles [3].

To give the reader an idea of where in the Universe we are, what is nearby and what is far away, some cosmic distances are listed in Table A.1 in the appendix. On a cosmological scale we are not really interested in objects smaller than a galaxy! We generally measure cosmic distances in *parsec* (pc) units (kpc for $10^3$ pc and Mpc for $10^6$ pc). A parsec is the distance at which one second of arc is subtended by a length equalling the mean distance between the Sun and the Earth. The parsec unit is given in Table A.2 in the appendix, where the values of some useful cosmological and astrophysical constants are listed.

## 1.2 Inertial Frames and the Cosmological Principle

Newton's first law—the law of inertia—states that a system on which no forces act is either at rest or in uniform motion. Such systems are called *inertial frames*.

Accelerated or rotating frames are not inertial frames. Newton considered that 'at rest' and 'in motion' implicitly referred to an *absolute space* which was unobservable but which had a real existence independent of humankind. Mach rejected the notion of an empty, unobservable space, and only Einstein was able to clarify the physics of motion of observers in inertial frames.

It may be interesting to follow a nonrelativistic argument about the static or nonstatic nature of the Universe which is a direct consequence of the cosmological principle.

Consider an observer 'A' in an inertial frame who measures the density of galaxies and their velocities in the space around him. Because the distribution of galaxies is observed to be homogeneous and isotropic on very large scales (strictly speaking, this is actually true for galaxy groups [1]), he would see the same mean density of galaxies (at one time $t$) in two different directions $\boldsymbol{r}$ and $\boldsymbol{r}'$:

$$\rho_{\mathrm{A}}(\boldsymbol{r}, t) = \rho_{\mathrm{A}}(\boldsymbol{r}', t).$$

Another observer 'B' in another inertial frame (see Figure 1.1) looking in the direction $\boldsymbol{r}$ from her location would also see the same mean density of galaxies:

$$\rho_{\mathrm{B}}(\boldsymbol{r}', t) = \rho_{\mathrm{A}}(\boldsymbol{r}, t).$$

The velocity distributions of galaxies would also look the same to both observers, in fact in all directions, for instance in the $\boldsymbol{r}'$ direction:

$$\boldsymbol{v}_{\mathrm{B}}(\boldsymbol{r}', t) = \boldsymbol{v}_{\mathrm{A}}(\boldsymbol{r}', t).$$

Suppose that the B frame has the relative velocity $\boldsymbol{v}_{\mathrm{A}}\ (\boldsymbol{r}'', t)$ as seen from the A frame along the radius vector $\boldsymbol{r}'' = \boldsymbol{r} - \boldsymbol{r}'$. If all velocities are nonrelativistic, i.e. small compared with the speed of light, we can write

$$\boldsymbol{v}_{\mathrm{A}}(\boldsymbol{r}', t) = \boldsymbol{v}_{\mathrm{A}}(\boldsymbol{r} - \boldsymbol{r}'', t) = \boldsymbol{v}_{\mathrm{A}}(\boldsymbol{r}, t) - \boldsymbol{v}_{\mathrm{A}}(\boldsymbol{r}'', t).$$

This equation is true only if $\boldsymbol{v}_{\mathrm{A}}\ (\boldsymbol{r}, t)$ has a specific form: it must be proportional to $\boldsymbol{r}$,

$$\boldsymbol{v}_{\mathrm{A}}(\boldsymbol{r}, t) = f(t)\boldsymbol{r}, \tag{1.1}$$

where $f(t)$ is an arbitrary function. Why is this so?

Let this universe start to expand. From the vantage point of A (or B equally well, since all points of observation are equal), nearby galaxies will appear to recede slowly.

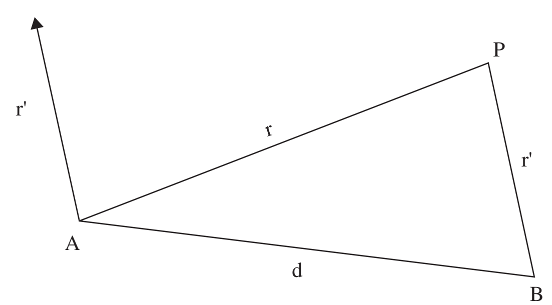


**Figure 1.1** Two observers at A and B making observations in the directions $\boldsymbol{r}$, $\boldsymbol{r}'$.

But in order to preserve uniformity, distant ones must recede faster, in fact their recession velocities must increase linearly with distance. That is the content of Equation (1.1).

If $f(t) > 0$, the Universe would be seen by both observers to expand, each galaxy having a radial velocity proportional to its radial distance $\boldsymbol{r}$. If $f(t) < 0$, the Universe would be seen to contract with velocities in the reversed direction. Thus we have seen that expansion and contraction are natural consequences of the cosmological principle. If $f(t)$ is a positive constant, Equation (1.1) is Hubble's law.

Actually, it is somewhat misleading to say that the galaxies recede when, rather, it is space itself which expands or contracts. This distinction is important when we come to general relativity.

A useful lesson may be learned from studying the limited gravitational system consisting of the Earth and rockets launched into space. This system is not quite like the previous example because it is not homogeneous, and because the motion of a rocket or a satellite in Earth's gravitational field is different from the motion of galaxies in the gravitational field of the Universe. Thus to simplify the case we only consider radial velocities, and we ignore Earth's rotation. Suppose the rockets have initial velocities low enough to make them fall back onto Earth. The rocket–Earth gravitational system is then *closed* and contracting, corresponding to $f(t) < 0$.

When the kinetic energy is large enough to balance gravity, our idealized rocket becomes a satellite, staying above Earth at a fixed height (real satellites circulate in stable Keplerian orbits at various altitudes if their launch velocities are in the range 8–11 km s$^{-1}$). This corresponds to the static solution $f(t) = 0$ for the rocket–Earth gravitational system.

If the launch velocities are increased beyond about 11 km s$^{-1}$, the potential energy of Earth's gravitational field no longer suffices to keep the rockets bound to Earth. Beyond this speed, called the *second cosmic velocity* by rocket engineers, the rockets escape for good. This is an expanding or *open* gravitational system, corresponding to $f(t) > 0$.

The static case is different if we consider the Universe as a whole. According to the cosmological principle, no point is preferred, and therefore there exists no center around which bodies can gravitate in steady-state orbits. Thus the Universe is either expanding or contracting, the static solution being unstable and therefore unlikely.

## 1.3 Olbers' Paradox

Let us turn to an early problem still discussed today, which is associated with the name of *Wilhelm Olbers* (1758–1840), although it seems to have been known already to Kepler in the 17th century, and a treatise on it was published by *Jean-Philippe Loys de Chéseaux* in 1744, as related in the book by E. Harrison [4]. Why is the night sky dark if the Universe is infinite, static and uniformly filled with stars? They should fill up the total field of visibility so that the night sky would be as bright as the Sun, and we would find ourselves in the middle of a heat bath of the temperature of the surface of the Sun. Obviously, at least one of the above assumptions about the Universe must be wrong.

The question of the total number of shining stars was already pondered by Newton and Leibnitz. Let us follow in some detail the argument published by Olbers in 1823. The *absolute luminosity* of a star is defined as the amount of luminous energy radiated per unit time, and the *surface brightness* $B$ as luminosity per unit surface. Let the *apparent luminosity* of a star of absolute luminosity $L$ at distance $r$ from an observer be $l = L/4\pi r^2$.

Suppose that the number of stars with average luminosity $L$ is $N$ and their average density in a volume $V$ is $n = N/V$. If the surface area of an average star is $A$, then its brightness is $B = L/A$. The Sun may be taken to be such an average star, mainly because we know it so well.

The number of stars in a spherical shell of radius $r$ and thickness $\mathrm{d}r$ is then $4\pi r^2 n\ \mathrm{d}r$. Their total radiation as observed at the origin of a static universe of infinite extent is then found by integrating the spherical shells from 0 to $\infty$:

$$\int_0^\infty 4\pi r^2 nl\ \mathrm{d}r = \int_0^\infty nL\ \mathrm{d}r = \infty. \tag{1.2}$$

On the other hand, a finite number of visible stars each taking up an angle $A/r^2$ could cover an infinite number of more distant stars, so it is not correct to integrate $r$ to $\infty$. Let us integrate only up to such a distance $R$ that the whole sky of angle $4\pi$ would be evenly tiled by the star discs. The condition for this is

$$\int_0^R 4\pi r^2 n \frac{A}{r^2}\ \mathrm{d}r = 4\pi.$$

It then follows that the distance is $R = 1/An$. The integrated brightness from these visible stars alone is then

$$\int_0^R nL\ \mathrm{d}r = L/A, \tag{1.3}$$

or equal to the brightness of the Sun. But the night sky is indeed dark, so we are faced with a paradox.

Olbers' own explanation was that invisible interstellar dust absorbed the light. That would make the intensity of starlight decrease exponentially with distance. But one can show that the amount of dust needed would be so great that the Sun would also be obscured. Moreover, the radiation would heat the dust so that it would start to glow soon enough, thereby becoming visible in the infrared.

A large number of different solutions to this paradox have been proposed in the past, some of the wrong ones lingering on into the present day. Let us here follow a valid line of reasoning due to Lord Kelvin (1824–1907), as retold and improved in a popular book by E. Harrison [4].

A star at distance $r$ covers the fraction $A/4\pi r^2$ of the sky. Multiplying this by the number of stars in the shell, $4\pi r^2 n\ \mathrm{d}r$, we obtain the fraction of the whole sky covered by stars viewed by an observer at the center, $An\ \mathrm{d}r$. Since $n$ is the star count per volume element, $An$ has the dimensions of number of stars per linear distance. The inverse of this,

$$\ell = 1/An, \tag{1.4}$$

is the mean radial distance between stars, or the *mean free path* of photons emitted from one star and being absorbed in collisions with another. We can also define a mean collision time:

$$\overline{\tau} = \ell/c. \tag{1.5}$$

The value of $\overline{\tau}$ can be roughly estimated from the properties of the Sun, with radius $R_\odot$ and density $\rho_\odot$. Let the present mean density of luminous matter in the Universe be $\rho_0$ and the distance to the farthest visible star $r_*$. Then the collision time inside this volume of size $\frac{4}{3}\pi r_*^3$ is

$$\overline{\tau} \simeq \overline{\tau}_\odot = \frac{1}{A_\odot nc} = \frac{1}{\pi R_\odot^2}\frac{4\pi r_*^3}{3Nc} = \frac{4\rho_\odot R_\odot}{3\rho_0 c}. \tag{1.6}$$

Taking the solar parameters from Table A.2 in the appendix we obtain approximately $10^{23}$ yr.

The probability that a photon does not collide but arrives safely to be observed by us after a flight distance $r$ can be derived from the assumption that the photon encounters obstacles randomly, that the collisions occur independently and at a constant rate $\ell^{-1}$ per unit distance. The probability $P(r)$ that the distance to the first collision is $r$ is then given by the exponential distribution

$$P(r) = \ell^{-1}\mathrm{e}^{-r/\ell}. \tag{1.7}$$

Thus flight distances much longer than $\ell$ are improbable.

Applying this to photons emitted in a spherical shell of thickness d$r$, and integrating the spherical shell from zero radius to $r_*$, the fraction of all photons emitted in the direction of the center of the sphere and arriving there to be detected is

$$f(r_*) = \int_0^{r_*} \ell^{-1}\mathrm{e}^{-r/\ell}\mathrm{d}r = 1 - \mathrm{e}^{-r_*/\ell}. \tag{1.8}$$

Obviously, this fraction approaches 1 only in the limit of an infinite universe. In that case every point on the sky would be seen to be emitting photons, and the sky would indeed be as bright as the Sun at night. But since this is not the case, we must conclude that $r_*/\ell$ is small. Thus the reason why the whole field of vision is not filled with stars is that the volume of the presently observable Universe is not infinite, it is in fact too small to contain sufficiently many visible stars.

Lord Kelvin's original result follows in the limit of small $r_*/\ell$, in which case

$$f(r_*) \approx r/\ell.$$

The exponential effect in Equation (1.8) was neglected by Lord Kelvin.

We can also replace the mean free path in Equation (1.8) with the collision time [Equation (1.5)], and the distance $r_*$ with the age of the Universe $t_0$, to obtain the fraction

$$f(r_*) = g(t_0) = 1 - \mathrm{e}^{-t_0/\overline{\tau}}. \tag{1.9}$$

If $u_\odot$ is the average radiation density at the surface of the stars, then the radiation density $u_0$ measured by us is correspondingly reduced by the fraction $g(t_0)$:

$$u_0 = u_\odot(1 - \mathrm{e}^{-t_0/\overline{\tau}}). \tag{1.10}$$

In order to be able to observe a luminous night sky we must have $u_0 \approx u_\odot$, or the Universe must have an age of the order of the collision time, $t_0 \approx 10^{23}$ yr. However, this exceeds all estimates of the age of the Universe by 13 orders of magnitude! Thus the existing stars have not had time to radiate long enough.

What Olbers and many after him did not take into account is that even if the age of the Universe was infinite, the stars do have a finite age and they burn their fuel at well-understood rates.

If we replace 'stars' by 'galaxies' in the above argument, the problem changes quantitatively but not qualitatively. The intergalactic space is filled with radiation from the galaxies, but there is less of it than one would expect for an infinite Universe, at all wavelengths. There is still a problem to be solved, but it is not quite as paradoxical as in Olbers' case.

One explanation is the one we have already met: each star radiates only for a finite time, and each galaxy has existed only for a finite time, whether the age of the Universe is infinite or not. Thus when the time perspective grows, an increasing number of stars become visible because their light has had time to reach us, but at the same time stars which have burned their fuel disappear.

Another possible explanation evokes expansion and special relativity. If the Universe expands, starlight redshifts, so that each arriving photon carries less energy than when it was emitted. At the same time, the volume of the Universe grows, and thus the energy density decreases. The observation of the low level of radiation in the intergalactic space has in fact been evoked as a proof of the expansion.

Since both explanations certainly contribute, it is necessary to carry out detailed quantitative calculations to establish which of them is more important. Most of the existing literature on the subject supports the relativistic effect, but Harrison has shown (and P. S. Wesson [5] has further emphasized) that this is false: the finite lifetime of the stars and galaxies is the dominating effect. The relativistic effect is quantitatively so unimportant that one cannot use it to prove that the Universe is either expanding or contracting.

## 1.4 Hubble's Law

In the 1920s Hubble measured the spectra of 18 spiral galaxies with a reasonably well-known distance. For each galaxy he could identify a known pattern of atomic spectral lines (from their relative intensities and spacings) which all exhibited a common redward frequency shift by a factor $1 + z$. Using the relation in Equation (1.1) following from the assumption of homogeneity alone,

$$v = cz, \tag{1.11}$$

he could then obtain their velocities with reasonable precision.

**The Expanding Universe.** The expectation for a stationary universe was that galaxies would be found to be moving about randomly. However, some observations had already shown that most galaxies were redshifted, thus receding, although some of the nearby ones exhibited blueshift. For instance, the nearby Andromeda nebula M31 is approaching us, as its blueshift testifies. Hubble's fundamental discovery was that the velocities of the distant galaxies he had studied increased linearly with distance:

$$v = H_0 r. \tag{1.12}$$

This is called *Hubble's law* and $H_0$ is called the *Hubble parameter*. For the relatively nearby spiral galaxies he studied, he could only determine the linear, first-order approximation to this function. Although the linearity of this law has been verified since then by the observations of hundreds of galaxies, it is not excluded that the true function has terms of higher order in $r$. Later on we shall introduce a second-order correction.

The message of Hubble's law is that the Universe is expanding, and this general expansion is called the *Hubble flow*. At a scale of tens or hundreds of Mpc the distances to all astronomical objects are increasing regardless of the position of our observation point. It is true that we observe that the galaxies are receding *from us* as if we were at the center of the Universe. However, we learned from studying a homogeneous and isotropic Universe in Figure 1.1 that if observer A sees the Universe expanding with the factor $f(t)$ in Equation (1.1), any other observer B will also see it expanding with the same factor, and the triangle ABP in Figure 1.1 will preserve its form. Thus, taking the cosmological principle to be valid, every observer will have the impression that all astronomical objects are receding from him/her. A homogeneous and isotropic Universe does not have a center. Consequently, we shall usually talk about *expansion velocities* rather than *recession velocities*.

It is surprising that neither Newton nor later scientists, pondering about why the Universe avoided a gravitational collapse, came to realize the correct solution. An expanding universe would be slowed down by gravity, so the inevitable collapse would be postponed until later. It was probably the notion of an infinite scale of time, inherent in a stationary model, which blocked the way to the right conclusion.

**Hubble Time and Radius.** From Equations (1.11) and (1.12) one sees that the Hubble parameter has the dimension of inverse time. Thus a characteristic timescale for the expansion of the Universe is the *Hubble time*:

$$\tau_{\rm H} \equiv H_0^{-1} = 9.7778 h^{-1} \times 10^9 \text{ yr}. \tag{1.13}$$

Here $h$ is the commonly used dimensionless quantity

$$h = H_0/(100 \text{ km s}^{-1} \text{ Mpc}^{-1}).$$

The Hubble parameter also determines the size scale of the observable Universe. In time $\tau_{\rm H}$, radiation travelling with the speed of light $c$ has reached the *Hubble radius*:

$$r_{\rm H} \equiv \tau_{\rm H} c = 3000 h^{-1} \text{ Mpc}. \tag{1.14}$$

Or, to put it a different way, according to Hubble's nonrelativistic law, objects at this distance would be expected to attain the speed of light, which is an absolute limit in the theory of special relativity.

Combining Equation (1.12) with Equation (1.11), one obtains

$$z = H_0 \frac{r}{c}. \tag{1.15}$$

In the section on Special Relativity we will see limitations to this formula when $v$ approaches $c$. The redshift $z$ is in fact infinite for objects at distance $r_H$ receding with the speed of light and thus physically meaningless. Therefore no information can reach us from farther away, all radiation is redshifted to infinite wavelengths, and no particle emitted within the Universe can exceed this distance.

**The Cosmic Scale.** The size of the Universe is unknown and unmeasurable, but if it undergoes expansion or contraction it is convenient to express distances at different epochs in terms of a *cosmic scale* $R(t)$, and denote its present value $R_0 \equiv R(t_0)$. The value of $R(t)$ can be chosen arbitrarily, so it is often more convenient to normalized it to its present value, and thereby define a dimensionless quantity, the *cosmic scale factor*:

$$a(t) \equiv R(t)/R_0. \tag{1.16}$$

The cosmic scale factor affects all distances: for instance the wavelength $\lambda$ of light emitted at one time $t$ and observed as $\lambda_0$ at another time $t_0$:

$$\frac{\lambda_0}{R_0} = \frac{\lambda}{R(t)}. \tag{1.17}$$

Let us find an approximation for $a(t)$ at times $t < t_0$ by expanding it to first-order time differences,

$$a(t) \approx 1 - \dot{a}_0(t_0 - t), \tag{1.18}$$

using the notation $\dot{a}_0$ for $\dot{a}(t_0)$, and $r = c(t_0 - t)$ for the distance to the source. The *cosmological redshift* can be approximated by

$$z = \frac{\lambda_0}{\lambda} - 1 = a^{-1} - 1 \approx \dot{a}_0 \frac{r}{c}. \tag{1.19}$$

Thus $1/1 + z$ is a measure of the scale factor $a(t)$ at the time when a source emitted the now-redshifted radiation. Identifying the expressions for $z$ in Equations (1.18) and (1.15) we find the important relation

$$\dot{a}_0 = \frac{\dot{R}_0}{R_0} = H_0. \tag{1.20}$$

**The Hubble Constant.** The value of this constant initially found by Hubble was $H_0 = 550\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$: an order of magnitude too large because his distance measurements were badly wrong. To establish the linear law and to determine the global value of $H_0$ one needs to be able to measure distances and expansion velocities well and far out. Distances are precisely measured only to nearby stars which participate in the general rotation of the Galaxy, and which therefore do not tell us anything about cosmological expansion. Even at distances of several Mpc the expansion-independent, transversal *peculiar velocities* of galaxies are of the same magnitude as the Hubble flow. The measured expansion at the Virgo supercluster, 17 Mpc away, is about 1100 km s$^{-1}$, whereas the peculiar velocities attain 600 km s$^{-1}$. At much larger distances where the peculiar velocities do not contribute appreciably to the total velocity, for instance at the Coma cluster 100 Mpc away, the expansion velocity is 6900 km s$^{-1}$ and the Hubble flow can be measured quite reliably, but the imprecision in distance measurements becomes the problem. Every procedure is sensitive to small, subtle corrections and to systematic biases unless great care is taken in the reduction and analysis of data.

Notable contributions to our knowledge of $H_0$ come from *supernovae* observations with the Hubble Space Telescope (HST) [6, 7], from the measurements of the relic *cosmic microwave background* (CMB) radiation temperature and polarization by the (CMB) radiation temperature Planck satellite [9]. Also the observations WMAP9 [8] and the Baryonic Acoustic Oscillations (BAO) in the distribution of galaxies are important, but the values are reported combined with CMB.

The average of all these experiments [6, 8, 9] is

$$h \equiv H_0/(100 \text{ km s}^{-1} \text{ Mpc}^{-1}) = 0.696 \pm 0.007. \tag{1.21}$$

**Statistics.** Let us take the meaning of the term 'test' from the statistical literature, where it is accurately defined [10]. When the hypothesis under test concerns the value of a parameter, the problems of *parameter estimation* and *hypothesis testing* are related; for instance, good techniques for estimation often lead to analogous testing procedures. The two situations lead, however, to different conclusions, and should not be confused. If nothing is known *a priori* about the parameter involved, it is natural to use the data to estimate it. On the other hand, if a theoretical prediction has been made that the parameter should have a certain value, it may be more appropriate to formulate the problem as a test of whether the data are consistent with this value. In either case, the nature of the problem, estimation or test, must be clear from the beginning and consistent to the end. When two or more independent methods of parameter estimation are compared, one can talk about a *consistency test*.

A good example of this reasoning is offered by the discussion of Hubble's law. Hubble's empirical discovery tested the *null hypothesis* that the Universe (out to the probed redshifts) expands. The test is a valid proof of the hypothesis for any value of $H_0$ that differs from zero at a chosen confidence level, CL%. Thus the value of $H_0 = 0.673$ is unimportant for the test, only its precision 0.012 matters.

## 1.5 The Age of the Universe

One of the conclusions of Olbers' paradox was that the Universe could not be eternal, it must have an age much less than $10^{23}$ yr, or else the night sky would be bright. More recent proofs that the Universe indeed grows older and consequently has a finite lifetime comes from astronomical observations of many types of extragalactic objects at high redshifts and at different wavelengths: radio sources, X-ray sources, quasars, faint blue galaxies. High redshifts correspond to earlier times, and what are observed are clear changes in the populations and the characteristics as one looks toward earlier epochs. Let us therefore turn to determinations of the age of the Universe.

In Equation (1.13) we defined the Hubble time $\tau_{\mathrm{H}}$, and gave a value for it of the order of 10 billion years. However, $\tau_{\mathrm{H}}$ is not the same as the age $t_0$ of the Universe. The latter depends on the dynamics of the Universe, whether it is expanding forever or whether the expansion will turn into a collapse, and these scenarios depend on how much matter there is and what the geometry of the Universe is, all questions we shall come back to later.

All the large experiments [11] now agree with an average of

$$t_0 = 13.73 \text{ Gyr}. \tag{1.22}$$

**Cosmochronology by Radioactive Nuclei.** There are several independent techniques, *cosmochronometers*, for determining the age of the Universe. At this point we shall only describe determinations via the cosmochronology of long-lived radioactive nuclei, and via stellar modeling of the oldest stellar populations in our Galaxy and in some other galaxies. Note that the very existence of radioactive nuclides indicates that the Universe cannot be infinitely old and static.

Various nuclear processes have been used to date the age of the Galaxy, $t_{\mathrm{G}}$, for instance the 'Uranium clock'. Long-lived radioactive isotopes such as $^{232}$Th, $^{235}$U, $^{238}$U and $^{244}$Pu have been formed by fast neutrons from supernova explosions, captured in the envelopes of an early generation of stars. With each generation of star formation, burn-out and supernova explosion, the proportion of metals increases. Therefore the metal-poorest stars found in globular clusters are the oldest.

The proportions of heavy isotopes following a supernova explosion are calculable with some degree of confidence. Since then, they have decayed with their different natural half-lives so that their abundances in the Galaxy today have changed. For instance, calculations of the original ratio $K = {}^{235}\mathrm{U}/{}^{238}\mathrm{U}$ give values of about 1.3 with a precision of about 10%, whereas this ratio on Earth at the present time is $K_0 = 0.007\,23$.

To compute the age of the Galaxy by this method, we also need the decay constants $\lambda$ of $^{238}$U and $^{235}$U which are related to their half-lives:

$$\lambda_{238} = \ln\ 2/(4.46 \text{ Gyr}), \quad \lambda_{235} = \ln\ 2/(0.7038 \text{ Gyr}).$$

The relation between isotope proportions, decay constants, and time $t_{\mathrm{G}}$ is

$$K = K_0 \exp\ [(\lambda_{238} - \lambda_{235})t_{\mathrm{G}}]. \tag{1.23}$$

Inserting numerical values one finds $t_{\mathrm{G}} \approx 6.2$ Gyr. However, the Solar System is only 4.57 Gyr old, so the abundance of $^{232}$Th, $^{235}$U and $^{238}$U on Earth cannot be expected to furnish a very interesting limit to $t_{\mathrm{G}}$. Rather, one has to turn to the abundances on the oldest stars in the Galaxy.

The globular clusters (GCs) are roughly spherically distributed stellar systems in the spheroid of the Galaxy. During the majority of the life of a star, it converts hydrogen into helium in its core. Thus the most interesting stars for the determination of $t_G$ are those which have exhausted their supply of hydrogen, and which are located in old, metal-poor GCs, and to which the distance can be reliably determined. A recent age determination gives

$$t_{GC} = 14.61 \pm 0.8 \text{ Gyr.}$$

This includes an estimated age for the Universe when the clusters formed.

Of particular interest is the detection of a spectral line of $^{238}$U in the extremely metal-poor star CS 31082-001, which is overabundant in heavy elements. Theoretical nucleosynthesis models for the initial abundances predict that the ratios of neighboring stable and unstable elements should be similar in early stars as well as on Earth. Thus one compares the abundances of the radioactive $^{232}$Th and $^{238}$U with the neighboring stable elements Os and Ir ($^{235}$U is now useless, because it has already decayed away on the oldest stars). One result is

$$t_* = 13.5 \pm 2.9 \text{ Gyr.} \tag{1.24}$$

**Brightest Cluster Galaxies (BCGs).** Another cosmochronometer is offered by the study of elliptical galaxies in BCGs at very large distances. It has been found that BCG colors only depend on their star-forming histories, and if one can trust stellar population synthesis models, one has a cosmochronometer. From recent analyses of BCGs the result is

$$t_{BCG} \gtrsim 12 \text{ Gyr.} \tag{1.25}$$

Allowing 0.5–1.0 Gyr from the Big Bang until galaxies form stars and clusters, all the above estimates agree reasonably with the value in Equation (1.21) (This correction was already included in the value from globular clusters.).

There are many more cosmochronometers making use of well-understood stellar populations at various distances which we shall not refer to here, all yielding ages near those quoted. It is of interest to note that in the past, when the dynamics of the Universe was less well known, the calculated age $\tau_H$ was smaller than the value in Equation (1.21), and at the same time the age $t_*$ of the oldest stars was much higher than the value in Equation (1.23). Thus this historical conflict between cosmological and observational age estimates has now disappeared.

Later we will derive a general relativistic formula for $t_0$ which depends on a few measurable dynamical parameters determined in a combination of supernova analyses, cosmic microwave background analyses and a set of other data.

## 1.6 Matter in the Universe

Since antiquity the objects in the sky were known by the visible light they emit, absorb or reflect. Stars like the sun shine, planets and moons reflect sunlight, planets around distant stars reveal themselves by obscuration, and intergalactic dust by dimming absorption. However, there are other kinds of matter than these examples, and there is radiation at other wavelengths than visible light.

**Baryonic Matter.** Stable matter as we know it is composed of atoms, and the nuclei of atoms are composed of protons and neutrons which are called nucleons or baryons. The protons are stable particles, the neutrons in atomic nuclei are also stable because of the strong interactions between nucleons. Free neutrons are not stable, they decay dominantly into a proton, an electron and an antineutrino within about 885 s. There exist many more kinds of baryons, but they are unstable and do not form matter.

Stars form galaxies, galaxies form clusters and clusters form superclusters and other large-scale structures. Stars form in the regions of galaxies that are the hardest to observe with many of the common tools of astronomy—in dense, cool (10–100 K) clouds of molecular gas detected in relatively ordinary faraway galaxies. From this environment only a small fraction of visible light can escape. Once stars form, the pressure of their radiation expels the gas, and they can then be seen clearly at optical wavelengths. The results point to a continuous fuelling of gas into the star-forming guts of assembling galaxies.

The baryonic matter in stars and other collapsed objects is only a small fraction of the total baryonic content of the Universe. Much more baryonic matter exists in the form of interstellar dust, hot molecular gas and neutral gas within galaxies, mainly $^1$H and $^4$He. and in the form of intergalactic hot gas and hot diffuse ionized gas in the intergalactic medium (IGM). The amount of nonradiating diffuse components can be inferred from the absorption of radiation from a bright background source such as a quasar, a technique which is extremely sensitive. Most of the baryonic matter resides outside bound structures, in galaxy groups and in galactic halos.

Current observations of baryons extend from the present-day Solar System to the earliest and most distant galaxies which formed when their age was only 5% of the Universe's present age. About one-fifth of the large galaxies formed within the Universe's first four billion years; 50% of the galaxies had formed by the time the Universe was seven billion years old.

The electromagnetic radiation that stars emit covers all frequencies, not only as visible light but as infrared light, ultraviolet light, X-rays and gamma rays. The most extreme sources of radiation are the *Gamma Ray Bursts (GRB)* from Active Galactic Nuclei (AGN). The nuclear and atomic processes in stars also produce particle emissions: electrons, positrons, neutrinos, antineutrinos and cosmic rays.

There also exists baryonic antimatter, but not on Earth, and there is very little evidence for its presence elsewhere in the Galaxy. That does not mean that antibaryons are pure fiction: they are readily produced in particle accelerators and in violent astrophysical events. However, in an environment of matter, antibaryons rapidly meet baryons and annihilate each other. The asymmetry in the abundance of matter and antimatter is surprising and needs an explanation. We shall deal with that in a later section.

We shall also later see how the baryons came to be the stable end products of the Big Bang Nucleosynthesis and how the mean baryon density in the Universe today is determined from the same set of data as is the age of the Universe.

**Supernovae and Neutron Stars.** Occasionally, a very bright *supernova* explosion can be seen in some galaxy. These events are very brief (one month) and very rare: historical records show that in our Galaxy they have occurred only every 300 yr. The most recent nearby supernova occurred in 1987 (code name SN1987A), not exactly in our Galaxy but in our small satellite, the Large Magellanic Cloud (LMC). Since it has now become possible to observe supernovae in very distant galaxies, one does not have to wait 300 yr for the next one.

The physical reason for this type of explosion (a Type SNII supernova) is the accumulation of Fe group elements at the core of a massive red giant star of size 8–200$M_\odot$, which has already burned its hydrogen, helium and other light elements.

Another type of explosion (a Type SNIa supernova) occurs in binary star systems, composed of a heavy white dwarf and a red giant star. White dwarfs have masses of the order of the Sun, but sizes of the order of Earth, whereas red giants are very large but contain very little mass. The dwarf then accretes mass from the red giant due to its much stronger gravitational field.

As long as the fusion process in the dwarf continues to burn lighter elements to Fe group elements, first the gas pressure and subsequently the electron degeneracy pressure balance the gravitational attraction. But when a rapidly burning dwarf star reaches a mass of 1.44$M_\odot$, the so-called *Chandrasekhar mass*, or in the case of a red giant when the iron core reaches that mass, no force is sufficient to oppose the gravitational collapse. The electrons and protons in the core transform into neutrinos and neutrons, respectively, most of the gravitational energy escapes in the form of neutrinos, and the remainder is a *neutron star* which is stabilized against further gravitational collapse by the degeneracy pressure of the neutrons. As further matter falls in, it bounces against the extremely dense neutron star and travels outwards as energetic shock waves. In the collision between the shock waves and the outer mantle, violent nuclear reactions take place and extremely bright light is generated. This is the supernova explosion visible from very far away. The nuclear reactions in the mantle create all the elements; in particular, the elements heavier than Fe, Ni and Cr on Earth have all been created in supernova explosions in the distant past.

The released energy is always the same since the collapse always occurs at the Chandrasekhar mass, thus in particular the peak brightness of Type Ia supernovae can serve as remarkably precise standard candles visible from very far away. (The term *standard candle* is used for any class of astronomical objects whose intrinsic luminosity can be inferred independently of the observed flux.) Additional information is provided by the color, the spectrum and an empirical correlation observed between the timescale of the supernova light curve and the peak luminosity. The usefulness of supernovae of Type Ia as standard candles is that they can be seen out to great distances, $z \approx 1.0$, and that the internal precision of the method is quite high. At greater distances one can still find supernovae, but Hubble's linear law [Equation (1.15)] is no longer valid.

The SNeIa are the brightest and most homogeneous class of supernovae. (The plural of SN is abbreviated SNe.) Type II are fainter, and show a wider variation in luminosity. Thus they are not standard candles, but the time evolution of their expanding atmospheres provides an indirect distance indicator, useful out to some 200 Mpc.

The composition of neutron stars is not known. The density of their cores is a few times that of matter in terrestrial nuclei, but they contain far more neutrons than protons, and they are strongly degenerate, thus we have no similar baryonic matter to study in the laboratories. They could be dominated by *quark matter* or by excited forms of baryons such as hyperons which are unstable particles in terrestrial conditions.

**Dark components.** Nonbaryonic forms of matter or energy which are invisible in the electromagnetic spectrum are neutrinos, black holes, dark matter and dark energy. These components will be dedicated considerable space in later Chapters.

## 1.7 Expansion in a Newtonian World

In this Section we shall use Newtonian mechanics to derive a cosmology without recourse to Einstein's theory. Inversely, this formulation can also be derived from Einstein's theory in the limit of weak gravitational fields.

A system of massive bodies in an attractive Newtonian potential contracts rather than expands. The Solar System has contracted to a stable, gravitationally bound configuration from some form of hot gaseous cloud, and the same mechanism is likely to be true for larger systems such as the Milky Way, and perhaps also for clusters of galaxies. On yet larger scales the Universe expands, but this does not contradict Newton's law of gravitation.

The key question in cosmology is whether the Universe as a whole is a gravitationally bound system in which the expansion will be halted one day. We shall next derive a condition for this from Newtonian mechanics.

**Newtonian Mechanics.** Consider a galaxy of *gravitating mass* $m_G$ located at a radius $r$ from the center of a sphere of mean density $\rho$ and mass $M = 4\pi r^3\rho/3$. The gravitational potential of the galaxy is

$$U = -GMm_G/r = -\frac{4}{3}\pi G m_G \rho r^2, \tag{1.26}$$

where $G$ is the *Newtonian constant* expressing the strength of the gravitational interaction. Thus the galaxy falls towards the center of gravitation, acquiring a radial acceleration

$$\ddot{r} = -GM/r^2 = -\frac{4}{3}\pi G\rho r. \tag{1.27}$$

This is *Newton's law of gravitation*, usually written in the form

$$F = -\frac{GMm_G}{r^2}, \tag{1.28}$$

where $F$ (in old-fashioned parlance) is the force exerted by the mass $M$ on the mass $m_G$. The negative signs in Equations (1.28)–(1.30) express the attractive nature of gravitation: bodies are forced to move in the direction of decreasing $r$.

In a universe expanding linearly according to Hubble's law [Equation (1.12)], the kinetic energy $T$ of the galaxy receding with velocity $v$ is

$$T = \frac{1}{2}mv^2 = \frac{1}{2}mH_0^2r^2, \tag{1.29}$$

where $m$ is the *inertial mass* of the galaxy. Although there is no theoretical reason for the inertial mass to equal the gravitational mass (we shall come back to this question later), careful tests have verified the equality to a precision better than a few parts in $10^{13}$. Let us therefore set $m_G = m$. Thus the total energy is given by

$$E = T + U = \frac{1}{2}mH_0^2r^2 - \frac{4}{3}\pi Gm\rho r^2 = mr^2\left(\frac{1}{2}H_0^2 - \frac{4}{3}\pi G\rho\right). \tag{1.30}$$

If the mass density $\rho$ of the Universe is large enough, the expansion will halt. The condition for this to occur is $E = 0$, or from Equation (1.32) this *critical density* is

$$\rho_c = \frac{3H_0^2}{8\pi G} = 1.0539 \times 10^{10}h^2 \text{ eV m}^{-3}. \tag{1.31}$$

The value $h = 0.696$ from Equation (1.21) can be inserted here. A universe with density $\rho > \rho_c$ is called *closed*; with density $\rho < \rho_c$ it is called *open*.

**Expansion.** Note that $r$ and $\rho$ are time dependent: they scale with the expansion. Denoting their present values $r_0$ and $\rho_0$, one has

$$r(t) = r_0a(t), \quad \rho(t) = \rho_0a^{-3}(t). \tag{1.32}$$

The acceleration $\ddot{r}$ in Equation (1.27) can then be replaced by the acceleration of the scale

$$\ddot{a} = \ddot{r}/r_0 = -\frac{4}{3}\pi G\rho_0a^{-2}. \tag{1.33}$$

Let us use the identity

$$\ddot{a} = \frac{1}{2}\frac{d}{da}\dot{a}^2$$

in Equation (1.33) to obtain

$$d\dot{a}^2 = -\frac{8}{3}\pi G\rho_0\frac{da}{a^2}.$$

This can be integrated from the present time $t_0$ to an earlier time $t$ with the result

$$\dot{a}^2(t) - \dot{a}^2(t_0) = \frac{8}{3}\pi G\rho_0(a^{-1} - 1). \tag{1.34}$$

Let us now introduce the dimensionless *density parameter*:

$$\Omega_0 = \frac{\rho_0}{\rho_c} = \frac{8\pi G\rho_0}{3H_0^2}. \tag{1.35}$$

Substituting $\Omega_0$ into Equation (1.34) and making use of the relation in Equation (1.20), $\dot{a}(t_0) = H_0$, we find

$$\dot{a}^2 = H_0^2(\Omega_0a^{-1} - \Omega_0 + 1). \tag{1.36}$$

Thus it is clear that the presence of matter influences the dynamics of the Universe. Without matter, $\Omega_0 = 0$, Equation (1.36) just states that the expansion is constant, $\dot{a} = H_0$, and $H_0$ could well be zero as Einstein thought. During expansion $\dot{a}$ is positive; during contraction it is negative. In both cases the value of $\dot{a}^2$ is nonnegative, so it must always be true that

$$1 - \Omega_0 + \Omega_0/a \geqslant 0. \tag{1.37}$$

**Models of Cosmological Evolution.** Depending on the value of $\Omega_0$ the evolution of the Universe can take three courses.

(i) $\Omega_0 < 1$, the mass density is undercritical. As the cosmic scale factor $a(t)$ increases for times $t > t_0$ the term $\Omega_0/a$ decreases, but the expression (1.37) stays positive always. Thus this case corresponds to an open, ever-expanding universe, as a consequence of the fact that it is expanding now. In Figure 1.2 the expression in Equation (1.37) is plotted against $a$ as the long-dashed curve for the choice $\Omega_0 = 0.5$.

(ii) $\Omega_0 = 1$, the mass density is critical. As the scale factor $a(t)$ increases for times $t > t_0$ the expression in Equation (1.37) gradually approaches zero, and the expansion halts. However, this only occurs infinitely late, so it also corresponds to an ever-expanding universe. This case is plotted against $a$ as the short-dashed curve in Figure 1.2. Note that cases (i) and (ii) differ by having different asymptotes. Case (ii) is quite realistic because the observational value of $\Omega_0$ is very close to 1, as we shall see later.

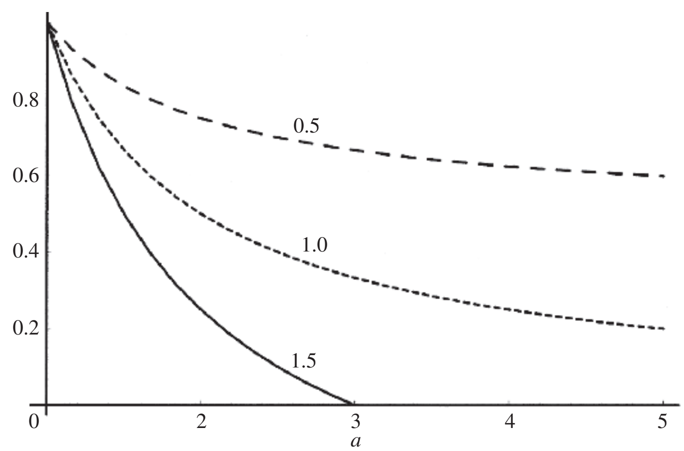


**Figure 1.2** Dependence of the expression in Equation (1.37) on the cosmic scale $a$ for an undercritical ($\Omega_0 = 0.5$), critical ($\Omega_0 = 1$) and overcritical ($\Omega_0 = 1.5$) universe. Time starts today at scale $a = 1$ in this picture and increases with $a$, except for the overcritical case where the Universe arrives at its maximum size, here $a = 3$, whereupon it reverses its direction and starts to shrink.

(iii) $\Omega_0 > 1$, the mass density is overcritical and the Universe is closed. As the scale factor $a(t)$ increases, it reaches a maximum value $a_{\rm mid}$ when the expression in Equation (1.37) vanishes, and where the rate of increase, $\dot{a}_{\rm mid}$, also vanishes. But the condition (1.37) must stay true, and therefore the expansion must turn into contraction at $a_{\rm mid}$. The solid line in Figure 1.2 describes this case for the choice $\Omega_0 = 1.5$, whence $a_{\rm mid} = 3$. For later times the Universe retraces the solid curve, ultimately reaching scale $a = 1$ again.

This is as far as we can go combining Newtonian mechanics with Hubble's law. We have seen that problems appear when the recession velocities exceed the speed of light, conflicting with special relativity. Another problem is that Newton's law of gravitation knows no delays: the gravitational potential is felt instantaneously over all distances. A third problem with Newtonian mechanics is that the Copernican world, which is assumed to be homogeneous and isotropic, extends up to a finite distance $r_0$, but outside that boundary there is nothing. Then the boundary region is characterized by violent inhomogeneity and anisotropy, which are not taken into account. To cope with these problems we must begin to construct a fully relativistic cosmology.

## Problems

1. How many revolutions has the Galaxy made since the formation of the Solar System if we take the solar velocity around the galactic center to be 365 km s$^{-1}$?
2. Use Equation (1.4) to estimate the mean free path $\ell$ of photons. What fraction of all photons emitted by stars up to the maximum observed redshift $z = 7$ arrive at Earth?
3. If Hubble had been right that the expansion is given by

$$H_0 = 550 \text{ km s}^{-1}\text{Mpc}^{-1},$$

how old would the Universe be then [see Equation (1.13)]?

4. What is the present ratio $K_0 = {}^{235}\text{U}/{}^{238}\text{U}$ on a star 10 Gyr old?
5. Prove Newton's theorem that the gravitational force at a radial distance $R$ from the center of a spherical distribution of matter acts as if all the mass inside $R$ were concentrated at a single point at the center. Show also that if the spherical distribution of matter extends beyond $R$, the force due to the mass outside $R$ vanishes.
6. Estimate the escape velocity from the Galaxy.

## References

[1] Ramella, M., Geller, M. J., Pisani, A. and da Costa, L. N. 2002 *Astron. J.* **123**, 2976.
[2] Fang Li Zhi and Li Shu Xian 1989 *Creation of the Universe*. World Scientific, Singapore.
[3] Peebles, P. J. E. 1993 *Principles of physical cosmology*. Princeton University Press, Princeton, NJ.
[4] Harrison, E. 1987 *Darkness at night*. Harvard University Press, Cambridge, MA.

[5] Wesson, P. S. 1991 *Astrophys. J.* **367**, 399.
[6] Efstathiou, G., 2013. Preprint arXiv:1311.3461 [astro-ph.CO].
[7] Freedman, W. L. *et al.* 2001 *Astrophys. J.* **553**, 47.
[8] Bennett, C. L. *et al.* 2013 *Astrophys. J. Suppl.* **208**, 20.
[9] Planck Collaboration: Ade, P. A. R. *et al.* 2014 *Astron. Astrophys* and preprint arXiv:1303.5076 [astro-ph.CO]
[10] Eadie, W. T., Drijard, D., James, F. E., Roos, M. and Sadoulet, B. 1971 *Statistical methods in experimental physics*. North-Holland, Amsterdam. Second Edition by Frederick James 2006. World Scientific Publishing Co.
[11] Bennett, C.L. *et al.* 2014, *Astrophys. J.* **794**, 135.