PART 1

FOUNDATIONS

1 The nature of cosmology

1.1 The aims of cosmology

The physical universe is the maximal set of physical objects which are locally causally connected to each other and to the region of spacetime that is accessible to us by astronomical observation. The scientific theory of cosmology is concerned with the study of the large-scale structure of the observable region of the universe, and its relation to local physics on the one hand and to the rest of the universe on the other.

Thus cosmology deals with the distribution and motion of radiation and of galaxies, clusters of galaxies, radio sources, quasi-stellar objects, and other astronomical objects observable at large distances, and so – in response to the astronomical observations – contemplates the nature and history of the expanding universe. Following the evolution of matter back into the past, this inevitably leads to consideration of physical processes in the hot early universe (the 'Hot Big Bang', or HBB), and even contemplation of the origin of the universe itself. Such studies underlie our current – still incomplete – understanding of the origin of galaxies, and in particular of our own Galaxy, which is the environment in which the Solar System and the Earth developed. Hence, as well as providing an observationally based analysis of what we can see in distant regions and how it got to be as it is, cosmology provides important information on the environment in which life – including ourselves — could come to exist in the universe, and so sets the background against which any philosophy of life in the universe must be set.

Thus, when understood in the widest sense, cosmology has both narrow and broad aims. It has aspects similar to normal physics, at least in its role as an explanatory theory for astrophysical objects (even if laboratory experiments are impossible in this context); aspects peculiar to scientific theories dealing with unique observable objects (and in particular the universe itself, regarded as a physical object); and one can use it as a starting point when considering aspects that stretch beyond science to metaphysics and philosophy.

Sciences vary in their mix of explanatory power, verifiability and links with the rest of science. The relative value one puts on those different qualities of scientific theories affects one's view of the nature of cosmology as a science, and hence one's approach to cosmology. The importance of considering such issues arises from one of the fundamental limitations of cosmology: there is only one universe. We cannot compare it with similar objects, so neither repeatable nor statistical experiments are possible. Thus a prime problem in cosmology is *the uniqueness of the universe* (Bondi, 1960, Harré, 1962, North, 1965, McCrea, 1970, Ellis, 2007). This means we have to pay even more care and attention than in other sciences to extracting as much as possible from data and theory; and we have to be very aware of the limitations of what we can state with reasonable certainty. These issues will be developed in the analysis that follows.

1.1.1 Scientific cosmology

The starting point of cosmology is a *description* of what there is in the universe and how it is distributed and moving – the geography of matter in the large; this is sometimes called 'cosmography'. It inevitably involves a filter of theory through which the raw data has been passed. At this level, the main aim is descriptive and work of this type provides the most accurate representation of the actual universe. We can refer to it as *observational cosmology*. It often leads to unexpected discoveries: for example, the expansion of the universe – and its acceleration, the existence of dark matter, massive walls and voids in the large-scale distribution of matter and large-scale motions of matter.

However, the cosmologist also seeks to *explain* the observations, to give an understanding of what processes are occurring and how they have led to the structures we see – an explanation of the nature and operation of the universe in physical terms. This explores the dynamics of the expansion of the universe in the large, but can also be at the level of the structure and evolution of large-scale objects, e.g. the physics of galaxy formation, the evolution of radio sources and the clustering of galaxies, as well as considering micro-processes in the HBB epoch, such as nucleosynthesis and the decoupling of matter and radiation. These studies can be called *physical cosmology*. It is usual here to take as the background model of the universe on the largest scales one of the Friedmann–Lemaître–Robertson–Walker (FLRW) class, and study the inhomogeneities by considering perturbed FLRW models: the 'standard model' is such a perturbed FLRW model.

The great potential significance of quantum and particle physics for the evolution of the early universe in the big-bang picture has come to the fore in recent years; this field may be called *particle cosmology*. As with physical cosmology, the background model is usually assumed to be an FLRW universe. Aspects of particle cosmology, such as the concept of inflation – an extremely brief era of extraordinarily rapid expansion in the very early universe – are regarded by most cosmologists as part of the standard model of cosmology. This approach is extended by some to *quantum cosmology*, which attempts to describe the very origin of spacetime and of physics. That attempt is still speculative and controversial, inter alia because it involves an engagement with quantum gravity, an as yet speculative theory, and also necessarily raises profound questions about the nature of quantum theory itself.

Finally, this all takes place in the context of gravitational theories based on Einstein's General Relativity (GR) theory. Spacetime curvature – and hence the evolution of the universe – is determined by the matter present via the Einstein Field Equations (EFE). Both the motion of matter in the universe, and the paths of light rays by which we observe it, are determined by this curvature. Therefore an exploration of these features ultimately underlies understanding of the others. *Relativistic cosmology* puts emphasis on the curved spaces demanded by GR and related theories, and focuses on the spacetime geometry of the universe and its consequences for observational and physical cosmology. In order to situate our understanding fully, it considers wider families of universe models than the FLRW models. This is our main approach, and its importance has become apparent from subtleties in applying the standard framework to such issues as horizons, lensing, gauge invariance, chaotic inflation and the supernovae data. A further key issue, which we also explore, is whether GR itself is an adequate theory of gravity for explaining the universe on cosmological scales, or whether some generalization is required.

These approaches have to some extent developed as a historical sequence of new 'paradigms' for cosmology, each offering new depth in our understanding (Ellis, 1993). We believe each of them offers important insights, and that a full understanding of the universe can only come about from the interaction of these approaches, to their mutual enrichment. Thus while our own expertise and emphasis is on the relativistic approach, which is perhaps the most neglected at the present time, we shall endeavour to link this fully to the other views. The full depth of the subject of cosmology involves all of them.

1.1.2 Cosmology's wider implications

An investigation of the universe as a whole inevitably has implications for philosophy and the humanities. For example, we may seek some view on how the cosmos relates to humanity in general and our own individual lives in particular – some conceptualization of how cosmology relates to meaning. This necessarily takes one beyond purely scientific concerns to broader philosophical issues, constrained by the scientific data and theories but not encompassed by them. Science itself cannot resolve the metaphysical issues posed by seeking reasons for existence of the universe, the existence of any physical laws at all, or the nature of the specific physical laws that actually hold, because we cannot devise experimental tests that will answer such questions; they are inevitably philosophical and metaphysical. However such issues lie at the foundation of cosmology.

This book is concerned with the scientific and technical aspects of cosmology. It will not specifically deal with the wider concerns, except for some brief comments towards the end. However, it will contribute to these wider concerns by attempting to delineate carefully the boundaries of what can be reliably achieved in cosmology by use of the scientific method. This involves in particular a careful review of which aspects of cosmological theory are testable by presently possible observations, or by observations that will conceivably be possible some day. These limitations are not always taken seriously in writings on cosmology.

1.2 Observational evidence and its limitations

There are three broad ways in which we obtain the evidence used in cosmology (all of them discussed in more depth later).

1.2.1 Evidence from astronomical observations

By observing the sky with telescopes and other instruments – detecting electromagnetic radiation (infrared, radio, optical, ultraviolet, X-ray and γ-ray), neutrinos, and gravitational waves – we aim to determine the distribution of matter around us. We observe discrete objects and hydrogen clouds up to very large distances, and indirectly observe the total matter (dark plus baryonic) via weak lensing. We also observe background radiation of various kinds that does not come from identifiable discrete sources. The most important such radiation is the blackbody Cosmic Microwave Background (CMB) that we identify as being relic radiation from the HBB. Its study is a central part of present day cosmology. It has propagated freely through space since its emission by hot matter on the Last Scattering Surface (LSS) in the early universe at the time of decoupling of matter and radiation, as the universe cooled through its ionization temperature. The universe was opaque at earlier times.

All electromagnetic radiation travels to us at the speed of light, so, *via electromagnetic phenomena, we can only observe the universe on our past light cone*; hence, as we observe to greater distances, we also observe to earlier times: each object is seen when it emitted the radiation, at a 'look back time' determined by the speed of light. In addition, we can observe massive high-energy particles ('cosmic rays'), but because they are charged they are strongly affected by local magnetic fields, so only very high-energy cosmic rays could carry information across cosmological distances.

Although we have strong evidence for our estimates of distances to the nearer galaxies, determining the *distance* of objects further away is difficult and often controversial. The basic problem is that we have direct observational access only to a two-dimensional projection of a three-dimensional spacetime region: we have to de-convolve these data to recover a three-dimensional picture of what is out there. However, this problem is ameliorated because we can observe at many wavelengths, and so can obtain spectral information about the objects we observe. We can also separate out different polarizations of the radiation received.

Experimentally there are problems in measuring faint signals and excluding effects of intervening matter, theoretically we have to make assumptions about the physical laws and conditions at the sources, and from both together we have to try to establish the intrinsic properties of the sources. The essential idea is to determine some class of 'standard candles' whose intrinsic luminosity is known and whose measured luminosity therefore gives a well-defined distance (relationships such as the Tully–Fisher relation between luminosity and rotational velocity for spiral galaxies are used, as well as classes of objects, like Cepheid variable stars or brightest cluster galaxies).

In spite of these difficulties, we understand quite a lot about the broad nature of what lies around us, as we describe in the next section.

Size of the universe

Astronomical length scales are determined by a variety of methods. Perhaps the most important thing we learn from these scales is that *the universe is extremely large relative to our own size*; even the immensities of our own Galaxy are insignificant compared with the scale

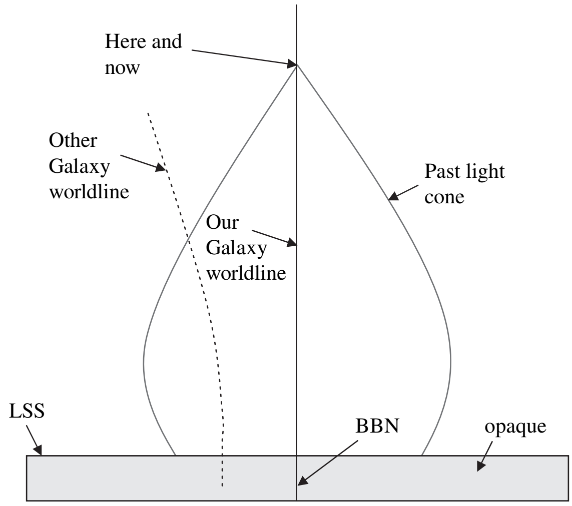

Fig. 1.1 Regions from which we have astronomical and 'geological' evidence, following Hoyle (1962).

of the observed region of the universe, which is of the order of 10^{10} light years (whereas the diameter of our Galaxy is of the order of $50,000$ light years, and the distance to the nearest other galaxy is about 10^6 light years).

This is the primary reason for our major observational problems in cosmology: in effect we can only observe the universe from one spacetime event, dubbed 'here and now', with all our direct observational information coming to us on a single light-cone (see Figure 1.1), supplemented by 'geological' data relating to the early history of our part of the universe (see below). Even a long-term astronomical data collection and analysis programme (say, collating data obtained over the next $10,000$ years by all available means including rocket probes able to travel at the speed of light) would not enable us to evade this restriction by observing the universe from an essentially different spatial or temporal vantage point, as, on cosmological scales, it would not move us from the point labelled 'here and now' in that spacetime diagram. Such a time scale is far too small to be detected relative to 10^{10} years, the scale of the universe itself.

1.2.2 Evidence of a geological nature

Additionally we obtain much useful information from evidence of a 'geological' nature, i.e. by careful study of the history of locally occurring objects as implied by their present-day structure and abundances. Particularly useful are measures of the abundances of elements, together with studies of the nature and hence the inferred ages of local astronomical objects such as star clusters. These observations test features of the early universe at times well before the earliest times accessible with telescopes (though only at points near our world line), thus enabling us to probe the physical evolution of matter in our vicinity at very early times (see Figure 1.1), for example testing Big Bang Nucleosynthesis (BBN) near our world line long before the LSS.

1.2.3 Evidence from local physics

Thirdly, a line of argument due to Mach, Olbers and others (see e.g. Bondi (1960)), argues that local physical conditions and even physical laws would be different if the universe were different; thus we can in principle use the nature of local physical conditions as evidence of the nature of the distant universe.

Mach raised this issue as regards the origin of inertia, and his proposal that inertia depends on the most distant matter in the universe had a profound influence on Einstein's cosmological thinking. Because the strength of the gravitational coupling constant G might be related to inertial properties, and so could depend on the state of the universe, this suggests there might be a *time-varying gravitational 'constant'*, $G = G(t)$ (Dirac, 1938). Two other examples are,

(a) The *dark night sky* ('Olbers' paradox') – the simplest static universe models suggest the entire sky at night (and, indeed, also during the day) should be as bright as the surface of the Sun. So why is the night sky dark?
(b) The *'arrow of time'* – the effects of the macroscopic laws of physics are dominated by irreversible processes with a unique arrow of time, despite the time reversibility of the fundamental local physical laws.

Plausibly, both may result from boundary conditions in the distant universe at very early times (Ellis and Sciama, 1972, Ellis, 2002), but they certainly have a profound effect on local physics. The essential point is that boundary conditions at the edge of the universe strongly affect the experienced nature of local physical laws, and conceivably affect the nature of the laws themselves – the distinction becomes blurred in the case of cosmology, where the boundary conditions are given and not open to change. We return to these issues in Sections 21.1 and 21.2.

1.2.4 Existence of horizons

Not only do signals fade with distance: if we live in an almost FLRW universe, as is commonly assumed, there is a series of horizons that limit what we can ever observationally or experimentally test in the cosmological context.

Firstly, the HBB era ends when the universe cools so much that matter and radiation, tightly coupled at earlier times, decouple from each other at the LSS in the early universe, which is the source of the CMB we detect today. The universe suddenly becomes transparent at this time: it was opaque to all electromagnetic radiation before then. Hence the earliest times we can access by electromagnetic experiments of all kinds are limited by a *visual horizon*: we can in principle have seen anything this side of the visual horizon, and cannot possibly have seen anything further out – and this will remain true, no matter how technology develops in the future.

There are two important provisos here. Firstly, there is no visual horizon if we live in a *small universe*, that is a universe spatially closed on such small scales that we have already seen around the entire universe more than once. This is a possibility we shall discuss below. Secondly, neutrino and gravitational wave detectors can in principle see to greater distances and earlier times. But they too will have their own horizons, limiting what they can ever see.

Because causal communication is limited by the speed of light, unless we live in a small universe, there exists outside the visual horizon a *particle horizon* limiting causality in the universe. We can have some kind of causal connection to any matter inside the particle horizon, but none whatever with matter outside it. This is a fundamental limitation on physical possibilities in the early universe. The proviso is that geometry at very early times may have been quite unlike that of an FLRW universe, and the situation may be different in FLRW models that collapse to a minimum radius, and then bounce to start a new expansion era. These possibilities also need investigation.

Because the energies we can attain in particle accelerators are limited by practical considerations (e.g. we cannot build a particle accelerator larger than the Solar System), there is a limit to our ability to experimentally determine the nature of the physical interactions that dominate what occurs at extremely early times, and in particular in the quantum gravity era. Hence there is a *physics horizon* preventing us from experimentally testing the relevant physics when we try to apply physical reasoning to earlier times (Section 20.5). Known, or at least potentially testable, physics applies at more recent times; what occurs at earlier times involves physics that cannot be directly observed or confirmed.

Unlike the other two horizons, this is technologically dependent, and the energies determining its location may change with time; nevertheless we may be certain that such a horizon exists. The ability of physical investigations to determine the nature of processes relevant to the very early universe is limited by technological and economic practicalities.

1.3 A summary of current observations

The current state of observations is discussed in detail in Chapter 13. The huge increase in available data and in accuracy of observations is a result of numerous technical developments such as space and balloon-borne telescopes, multi-mirror telescopes, interferometer techniques, adaptive optics, fibre optics, photon multipliers, CCDs, massive computing capabilities and so on, all coming together in an ability to do precision multi-wavelength observations (from radio through optical and infrared to gamma ray) across the entire sky. We shall not describe these developments in this book, but acknowledge that it is only through them that the era of data-based 'precision cosmology' has become possible (Bothun, 1998, Lena, Lebrun and Mignard, 2010). It is this solid grounding in observations and data that makes cosmology the exciting science that it is.

1.3.1 Expansion of the universe – and its acceleration

After Hubble determined the distance of other galaxies (and hence their nature) by observing Cepheid variables in them, the earliest observational result of modern cosmology was Hubble's 1929 law relating the magnitude and redshift of galaxies. The redshift z can be interpreted as due to the Doppler effect of a velocity of recession (since the measurements are made by comparing spectra and using known spectra of different elements, the interpretation depends on assuming that these were the same in the past). The flux received from a distant source depends on its distance, and may also be given as the source's apparent magnitude

m. For 'standard candles', such as supernovae, the flux is related to distance by the inverse square law; so magnitude is a proxy for distance. For relatively nearby sources believed to be intrinsically alike, the magnitude is related linearly to the redshift, as seen in Figure 1.2. This can be interpreted as a linear relation $v = H_0 d$ between velocity v and distance d, which is then in turn interpreted as due to *expansion of the universe*. The Hubble constant is $H_0 = 100h$ km/s/Mpc. For a long time there were uncertainties in its value of up to a factor 2, but recent observations have given much more accurately determined values. For example the Hubble Space Telescope Key Project gave $h = 0.73 \pm 0.06$ km/s/Mpc (Freedman and Madore, 2010). The constant H_0 gives a time scale $1/H_0$ for the present day expansion: using a linear extrapolation, this would be the time since a moment when all galaxies were in the same place, which gives an estimate of the age of the universe.

The fact that the universe is expanding does not necessarily imply it is evolving: it could conceivably be in a steady state, with the expansion rate always the same and a steady creation of matter keeping the density constant (Hoyle, 1948, Bondi, 1960). However there is a greater number density of radio sources at some distance than there is nearby, which disagrees with a 'steady state' picture. The initial rise and later fall in numbers as we go to fainter fluxes is consistent with a picture of an HBB universe in which radio sources form after the big bang, and their numbers rise to a peak and then start to decrease as their energy sources become exhausted. This is evidence of the crucial feature that *the universe*

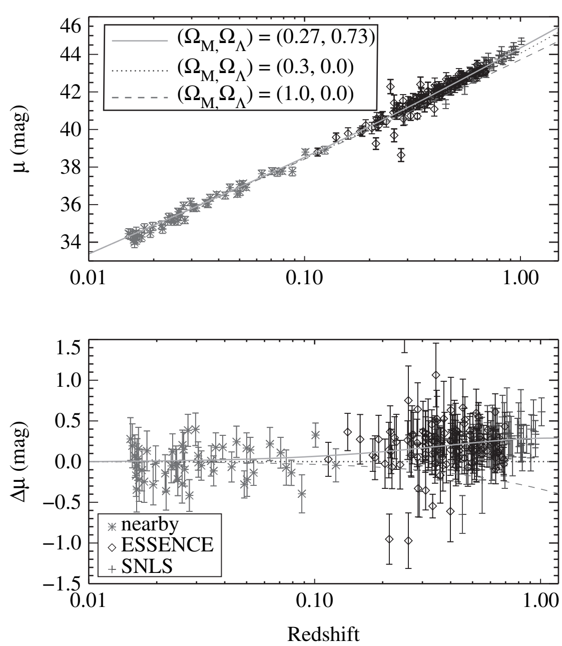

Fig. 1.2 *Top*: Magnitude-redshift diagram for SNIa from three surveys. The solid (green) curve gives the best fit. *Bottom*: Deviation from a model with no dark energy. (From Krisciunas (2008), courtesy of Kevin Krisciunas and the ESSENCE Supernova Search Team. Note that these are the preliminary results as of 2008.) A colour version of this figure is available online.