Bern University, Master of Physics

Cosmology

Julien Lesgourgues
CERN Theory Division
RWTH Aachen University

May 30, 2015

Contents

Chapter 1

Introduction to the universe expansion

1.1 Historical overview

1.1.1 The Doppler effect

At the beginning of the XX-th century, the understanding of the global structure of the universe beyond the scale of the solar system was still relying on pure speculation. In 1750, with a remarkable intuition, Thomas Wright understood that the luminous stripe observed in the night sky and called the Milky Way could be a consequence of stars being distributed along a thin plate. At that time, with the help of telescopes, many faint and diffuse objects had been already observed and listed, under the generic name of nebulae - in addition to the Andromeda nebula which is visible by eye, and has been known many centuries before the invention of telescopes. Soon after the proposal of Wright, the philosopher Emmanuel Kant suggested that some of these nebulae could be some other clusters of stars, far outside the Milky Way. So, the idea of a galactic structure appeared in the mind of astronomers during the XVIII-th century, but even in the following century there was no way to check it on an experimental basis.

At the beginning of the nineteenth century, some physicists observed the first spectral lines. In 1842, Johann Christian Doppler argued that if an observer receives a wave emitted by a body in motion, the wavelength that he will measure will be shifted proportionally to the speed of the emitting body with respect to the observer (projected along the line of sight):

$$\Delta\lambda/\lambda = \vec{v} \cdot \vec{n}/c \tag{1.1}$$

where c is the celerity of the wave (See figure 1.1). He suggested that this effect could be observable for both light and sound waves. The former assumption was checked experimentally in 1868 by Sir William Huggins, who found that the spectral lines of some neighboring stars were slightly shifted toward the red or blue ends of the spectrum. So, it was possible to know the projection of star velocities along the line of sight, v_r, using

$$z \equiv \Delta\lambda/\lambda = v_r/c \tag{1.2}$$

where z is called the redshift (it is negative in case of blue-shift) and c is the speed of light. Note that the redshift gives no indication concerning the distance of the star. At the beginning of the XX-th century, with increasingly good

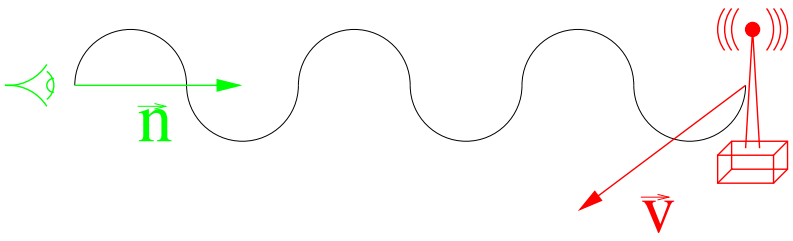


Figure 1.1: The Doppler effect.

instruments, people could also measure the redshift of some nebulae. The first measurements, performed on the brightest objects, indicated some arbitrary distribution of red and blue-shifts, like for stars. Then, with more observations, it appeared that the statistics was biased in favor of red-shifts, suggesting that a majority of nebulae were going away from us, unlike stars. This was raising new questions concerning the distance and the nature of nebulae.

1.1.2 The discovery of the galactic structure

In the 1920's, Leavitt and Shapley studied some particular stars, called the cepheids, known to have a periodic time-varying luminosity. They could show that the period of cepheids is proportional to their absolute luminosity L (the absolute luminosity is the total amount of light emitted by unit of time, i.e., the flux integrated on a closed surface around the star). This relation is well understood from current knowledge on stellar physics (it is due to the cycle of ionization of helium in the cepheids's athmosphere). Leavitt and Shapley were already able to measure the coefficient of proportionality (calibrated with the nearest cepheids, for which the parallax method can be employed; the parallax is half the angle under which a star appears to move when the earth makes one rotation around the sun). So, by measuring the apparent luminosity, i.e. the flux l per unit of surface through an instrument pointing to the star, it was easy to get the distance of the star r from

$$L = l \times \left(4\pi r^2\right) \ . \tag{1.3}$$

Using this technique, it became possible to measure the distance of various cepheids inside our galaxies, and to obtain the first estimate of the characteristic size of the stellar disk of the Milky Way (known today to be around 80.000 light-years).

But what about nebulae? In 1923, the 2.50m telescope of Mount Wilson (Los Angeles) allowed Edwin Hubble to make the first observation of individual stars inside the brightest nebula, Andromeda. Some of these were found to behave like cepheids, leading Hubble to give an estimate of the distance of Andromeda. He found approximately 900.000 light-years (but later, when cepheids were known better, this distance was established to be around 2 million light-years). That was the first confirmation of the galactic structure of the universe: some nebulae were likely to be some distant replicas of the Milky Way, and the galaxies were separated by large voids.

1.1.3 The Cosmological Principle

This observation, together with the fact that most nebulae are redshifted (excepted for some of the nearest ones like Andromeda), was an indication that on the largest observable scales, the universe was expanding. At the beginning,

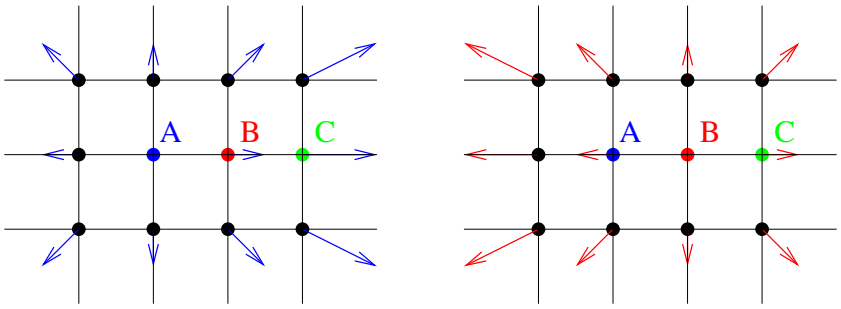


Figure 1.2: Homogeneous expansion on a two-dimensional grid. Some equally-spaced observers are located at each intersection. The grid is plotted twice. On the left, the arrays show the expansion flow measured by A; on the right, the expansion flow measured by B. If we assume that the expansion is homogeneous, we get that A sees B going away at the same velocity as B sees C going away. So, using the additivity of speeds, the velocity of C with respect to A must be twice the velocity of B with respect to A. This shows that there is a linear relation between speed and distance, valid for any observer.

this idea was not widely accepted. Indeed, in the most general case, a given dynamics of expansion takes place around a center. Seeing the universe in expansion around us seemed to be an evidence for the existence of a center in the universe, very close to our own galaxy.

Until the middle age, the Cosmos was thought to be organised around mankind, but the common wisdom of modern science suggests that there should be nothing special about the region or the galaxy in which we live. This intuitive idea was formulated by the astrophysicist Edward Arthur Milne as the "Cosmological Principle": the universe as a whole should be homogeneous, with no privileged point playing a particular role.

Was the apparently observed expansion of the universe a proof against the Cosmological Principle? Not necessarily. The homogeneity of the universe is compatible either with a static distribution of galaxies, or with a very special velocity field, obeying to a linear distribution:

$$\vec{v} = H\ \vec{r} \tag{1.4}$$

where $\vec{v}$ denotes the velocity of an arbitrary body with position $\vec{r}$, and H is a constant of proportionality. An expansion described by this law is still homogeneous because it is left unchanged by a change of origin. To see this, one can make an analogy with an infinitely large rubber grid, that would be stretched equally in all directions: it would expand, but with no center (see figure 1.2). This result is not true for any other velocity field. For instance, the expansion law

$$\vec{v} = H\ |\vec{r}|\ \vec{r} \tag{1.5}$$

is not invariant under a change of origin: so, it has a center.

1.2 The Hubble Law

1.2.1 Hubble's discovery

So, a condition for the universe to respect the Cosmological Principle is that the speed of galaxies along the line of sight, or equivalently, their redshift, should

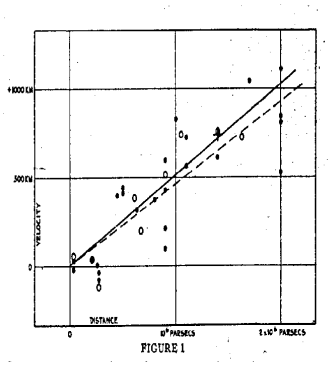


Figure 1.3: The diagram published by Hubble in 1929. The labels of the horizontal (resp. vertical) axis are 0, 1, 2 Mpc (resp. 0, 500, 1000 km.s^{-1}). Hubble estimated the expansion rate to be 500 km.s^{-1}Mpc^{-1}. Today, it is known to be around 70 km.s^{-1}Mpc^{-1}.

be proportional to their distance. Hubble tried to check this idea, still using the cepheid technique. He published in 1929 a study based on 18 galaxies (in which cepheids could be seen), for which he had measured both the redshift and the distance. His results were showing roughly a linear relation between redshift and distance (see figure 1.3). He concluded that the universe was in homogeneous expansion, and gave the first estimate of the coefficient of proportionality H, called the Hubble parameter.

Hubble's measurements were rather unprecise. It is now understood that his measurement were not based on regular cepheids. Moreover, at distances of the order of 1 Mpc probed by Hubble's experiment (Mpc denotes a Megaparsec, the unity of distance usually employed for cosmology; 1 Mpc $\simeq 3 \times 10^{22}$m $\simeq 3 \times 10^6$ light-years; the proper definition of a parsec is "the distance to an object with a parallax of one arcsecond"), peculiar velocities tend to dominate over the expansion flow. So, Hubble's conclusion was obviously quite biased. However, this experiment is generally considered as the starting point of experimental cosmology. Since then, many similar experiments have been performed with better and better techniques and instruments, using not only cepheids but also supernovae and other "standard candles" (i.e., objects which absolute magnitude can be inferred in some way, without knowing their distance) at larger and lager distances. Recent data (like that shown in figure 1.4) leave no doubt about the proportionality, but there is still an uncertainty concerning the exact value of H. The Hubble constant is generally parametrized as

$$H = 100\ h\ \text{km s}^{-1}\text{Mpc}^{-1} \tag{1.6}$$

where h is the dimensionless "reduced Hubble parameter", currently known to be in the range $h = 0.706 \pm 0.033$ (at the 68% confidence level) from astrophysical observations *(MNRAS 440 (2014) 1138)*. As we shall see later most cosmological observations confirm this range. So, for instance, a typical galaxy located at 10 Mpc goes away at a speed close to 700 km s^{-1}.

1.2.2 Homogeneity and inhomogeneities

Before leaving this section, we should clarify one point about the "Cosmological Principle", i.e. the assumption that the universe is homogeneous. Of course, nobody has ever claimed that the universe was homogeneous on small scales, since compact objects like planets or stars, or clusters of stars like galaxies

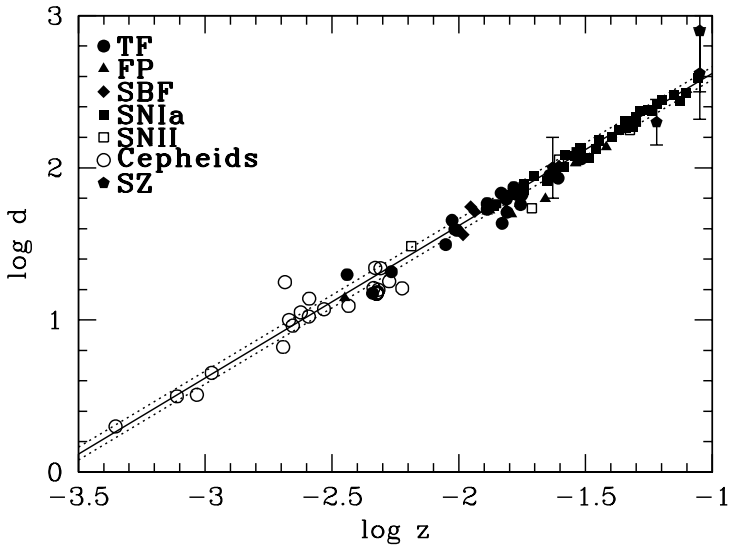


Figure 1.4: An example of Hubble diagram published by the Hubble Space Telescope Key Project in 2000 *(Astrophys.J. 553 (2001) 47-72)*, based on cepheids, supernovae and other standard candles till a distance of 400 Mpc. The horizontal axis gives the radial velocity, expressed as $\log_{10}[v/c] = \log_{10} z$ where z is redshift; the vertical axis shows the distance $\log_{10}[d/(1\text{Mpc})]$.

are inhomogeneities in themselves. The Cosmological Principle only assumes homogeneity after smoothing over some characteristic scale. By analogy, take a grid of step l (see figure 1.5), and put one object in each intersection, with a randomly distributed mass (with all masses obeying to the same distribution of probability). Then, make a random displacement of each object (again with all displacements obeying to the same distribution of probability). At small scales, the mass density is obviously inhomogeneous for three reasons: the objects are compact, they have different masses, and they are separated by different distances. However, since the distribution has been obtained by performing a random shift in mass and position, starting from an homogeneous structure, it is clear even intuitively that the mass density smoothed over some large scale will remain homogeneous again.

The Cosmological Principle should be understood in this sense. Let us suppose that the universe is almost homogeneous at a scale corresponding, say, to the typical intergalactic distance, multiplied by thirty or so. Then, the Hubble law doesn't have to be verified exactly for an individual galaxy, because of peculiar motions resulting from the fact that galaxies have slightly different masses, and are not in a perfectly ordered phase like a grid. But the Hubble law should be verified in average, provided that the maximum scale of the data is not smaller than the scale of homogeneity. The scattering of the data at a given scale reflects the level of inhomogeneity, and when using data on larger and larger scales, the scattering must be less and less significant. This is exactly what is observed in practice. An even better proof of the homogeneity of the universe on large scales comes from the Cosmic Microwave Background, as we shall see in section 6.6.

We will come back to these issues in section 6.6, and show how the formation of inhomogeneities on small scales are currently understood and quantified

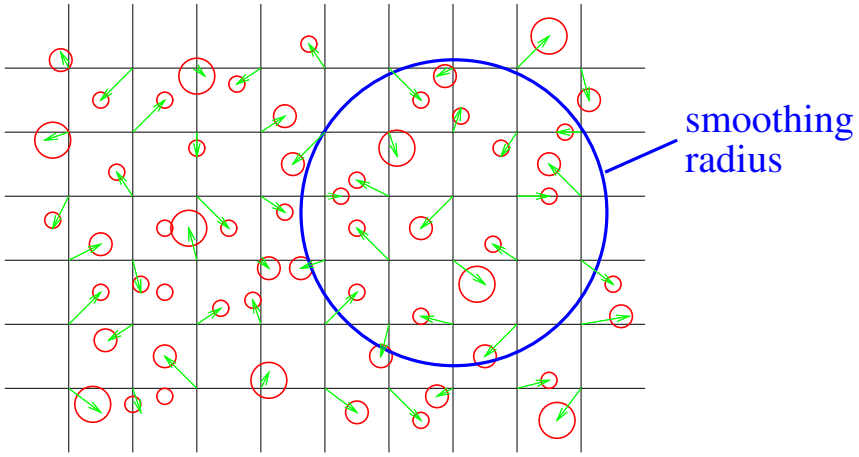


Figure 1.5: We build an inhomogeneous distribution of objects in the following way: starting from each intersection of the grid, we draw a random vector and put an object of random mass at the extremity of the vector. Provided that all random vectors and masses obey to the same distributions of probability, the mass density is still homogeneous when it is smoothed over a large enough smoothing radius (in our example, the typical length of the vectors is smaller than the step of the grid; but our conclusion would still apply if the vectors were larger than the grid step, provided that the smoothing radius is even larger). This illustrates the concept of homogeneity above a given scale, like in the universe.

within some precise physical models.

1.3 The universe Expansion from Newtonian Gravity

It is not enough to observe the galactic motions, one should also try to explain it with the laws of physics.

1.3.1 Newtonian Gravity versus General Relativity

On cosmic scales, the only force expected to be relevant is gravity. The first theory of gravitation, derived by Newton, was embedded later by Einstein into a more general theory: General Relativity (thereafter denoted GR). However, in simple words, GR is relevant only for describing gravitational forces between bodies which have relative motions comparable to the speed of light[1]. In most other cases, Newton's gravity gives a sufficiently accurate description.

The speed of neighboring galaxies is always much smaller than the speed of light. So, *a priori*, Newtonian gravity should be able to explain the Hubble flow. One could even think that historically, Newton's law led to the prediction of the universe expansion, or at least, to its first interpretation. Amazingly, and for reasons which are more mathematical than physical, it happened not to be the case: the first attempts to describe the global dynamics of the universe came with GR, in the 1910's. In this course, for pedagogical purposes, we will not follow the historical order, and start with the Newtonian approach.

[1]Going a little bit more into details, it is also relevant when an object is so heavy and so close that the speed of liberation from this object is comparable to the speed of light.

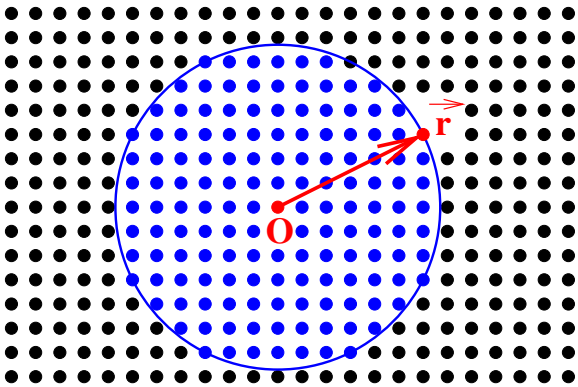


Figure 1.6: Gauss theorem applied to the local universe.

Newton himself did the first step in the argumentation. He noticed that if the universe was of finite size, and governed by the law of gravity, then all massive bodies would unavoidably concentrate into a single point, just because of gravitational attraction. If instead it was infinite, and with an approximately homogeneous distribution at initial time, it could concentrate into several points, like planets and stars, because there would be no center to fall in. In that case, the motion of each massive body would be driven by the sum of an infinite number of gravitational forces. Since the mathematical tools available at that time didn't allow to deal with this situation, Newton didn't proceed with his argument.

1.3.2 The rate of expansion from Gauss theorem

In fact, using Gauss theorem, this problem turns out to be quite simple. Suppose that the universe consists in many massive bodies distributed in an isotropic and homogeneous way (i.e., for any observer, the distribution looks the same in all directions). This should be a good modelling of the universe on sufficiently large scales. We wish to compute the motion of a particle located at a distance $r(t)$ away from us. Because the universe is assumed to be isotropic, the problem is spherically symmetric, and we can employ Gauss theorem on the sphere centered on us and attached to the particle (see figure 1.6). The acceleration of any particle on the surface of this sphere reads

$$\ddot{r}(t) = -\frac{GM(r(t))}{r^2(t)} \tag{1.7}$$

where G is Newton's constant and $M(r(t))$ is the mass contained inside the sphere of radius $r(t)$. In other words, the particle feels the same force as if it had a two-body interaction with the mass of the sphere concentrated at the center. Note that $r(t)$ varies with time, but $M(r(t))$ remains constant: because of spherical symmetry, no particle can enter or leave the sphere, which contains always the same mass.

Since Gauss theorem allows us to make completely abstraction of the mass outside the sphere[2], we can make an analogy with the motion e.g. of a satellite

[2]The argumentation that we present here is useful for guiding our intuition, but we should say that it is not fully self-consistent. Usually, when we have to deal with a spherically symmetric mass distribution, we apply Gauss theorem inside a sphere, and forget completely about the external mass. This is actually not correct when the mass distribution spreads out to infinity. Indeed, in our example, Newtonian gravity implies that a point inside the

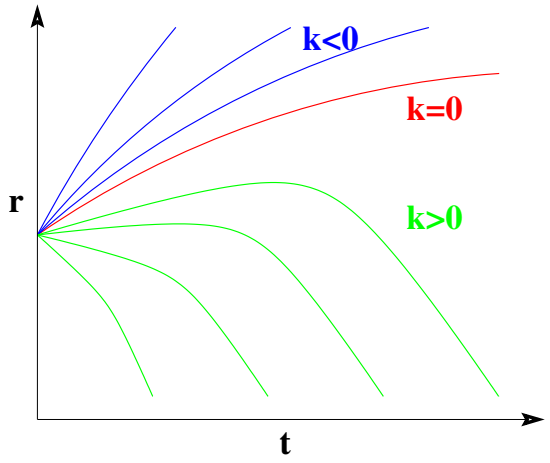


Figure 1.7: The motion of expansion in a Newtonian universe is equivalent to that of a body ejected from Earth. It depends on the initial rate of expansion compared with a critical density. When the parameter k is zero or negative, the expansion lasts forever, otherwise the universe re-collapses ($r \to 0$).

ejected vertically from the Earth. We know that this motion depends on the initial velocity, compared with the speed of liberation from the Earth: if the initial speed is large enough, the satellites goes away indefinitely, otherwise it stops and falls down. We can see this mathematically by multiplying equation (1.7) by $\dot{r}$, and integrating it over time:

$$\frac{\dot{r}^2(t)}{2} = \frac{GM(r(t))}{r(t)} - \frac{k}{2} \tag{1.8}$$

where k is a constant of integration. We can replace the mass $M(r(t))$ by the volume of the sphere multiplied by the homogeneous mass density $\rho_{\rm mass}(t)$, and rearrange the equation as

$$\left(\frac{\dot{r}(t)}{r(t)}\right)^2 = \frac{8\pi G}{3}\rho_{\rm mass}(t) - \frac{k}{r^2(t)} \ . \tag{1.9}$$

The quantity $\dot{r}/r$ is called the rate of expansion. Since $M(r(t))$ is time-independent, the mass density evolves as $\rho_{\rm mass}(t) \propto r^{-3}(t)$ (i.e., matter is simply diluted when the universe expands). The behavior of $r(t)$ depends on the sign of k. If k is positive, $r(t)$ can grow at early times but it always decreases at late times, like the altitude of the satellite falling back on Earth: this would correspond to a universe expanding first, and then collapsing. If k is zero or negative, the expansion lasts forever.

In the case of the satellite, the critical value, which is the speed of liberation (at a given altitude), depends on the mass of the Earth. By analogy, in the case

sphere would feel all the forces from all bodies inside and outside the sphere, which would exactly cancel out. Nevertheless, the present calculation based on Gauss theorem does lead to a correct prediction for the expansion of the universe. In fact, this can be rigorously justified only *a posteriori*, after a full general relativistic study. In GR, the Gauss theorem can be generalized thanks to the consequences of Birkhoff's theorem, which is valid also when the mass distribution spreads to infinity. In particular, for an infinite spherically symmetric matter distribution, Birkhoff's theorem implies that we can isolate a sphere as if there was nothing outside of it. Once this formal step has been performed, nothing prevents us from using Newtonian gravity and Gauss theorem inside a smaller sphere, as if the external matter distribution was finite. This argument justifies rigorously the calculation of this section.

of the universe, the important quantity that should be compared with some critical value is the homogeneous mass density. If at all times $\rho_{\rm mass}(t)$ is bigger than the critical value

$$\rho_{\rm mass}(t) = \frac{3(\dot{r}(t)/r(t))^2}{8\pi G} \tag{1.10}$$

then k is positive and the universe will re-collapse. Physically, it means that the gravitational force wins against inertial effects. In the other case, the universe expands forever, because the density is too small with respect to the expansion velocity, and gravitation never takes over inertia. The case $k = 0$ corresponds to a kind of equilibrium between gravitation and inertia in which the universe expands forever, following a power–law: $r(t) \propto t^{2/3}$.

1.3.3 The limitations of Newtonian predictions

In the previous calculation, we cheated a little bit: we assumed that the universe was isotropic around us, but we didn't check that it was isotropic everywhere (and therefore homogeneous). Following what we said before, homogeneous expansion requires proportionality between speed and distance at a given time. Looking at equation (1.9), we see immediately that this is true only when $k = 0$. So, it seems that the other solutions are not compatible with the Cosmological Principle. We can also say that if the universe was fully understandable in terms of Newtonian mechanics, then the observation of linear expansion would imply that k equals zero and that there is a precise relation between the density and the expansion rate at any time.

This argument shouldn't be taken seriously, because the link that we made between homogeneity and linear expansion was based on the additivity of speed (look for instance at the caption of figure 1.2), and therefore, on Newtonian mechanics. But Newtonian mechanics cannot be applied at large distances, where v becomes large and comparable to the speed of light. This occurs around a characteristic scale called the Hubble radius R_H :

$$R_H = cH^{-1}, \tag{1.11}$$

at which the Newtonian expansion law gives $v = HR_H = c$.

So, the full problem has to be formulated in relativistic terms. In the GR results, we will see again some solutions with $k \neq 0$, but they will remain compatible with the homogeneity of the universe.

Chapter 2

Homogeneous Cosmology

From now on, we will adopt units in which $c = \hbar = k_b = 1$ in most equations.

2.1 The Lemaître, Friedmann, Robertson & Walker metric

2.1.1 Cosmological background and perturbations

As already suggested in section 1.2.2, most calculations and predictions in cosmology are done under the assumption that the exact description of the universe can be decomposed in two problems: the background problem (which should be an independent, self-consistent problem) and the inhomogeneity problem (within a given imposed background). This is the usual approach in any theory of perturbations.

In the background problem, one assumes that in first approximation we can see the universe as a smooth distribution of matter, i.e. that one can average over small inhomogeneities like stars, galaxies and clusters, which are replaced by an idealized "cosmological fluid". The cosmological fluid can be thought to be a truly continuous distribution of matter, or equivalently, a regular distribution of compact objects, smoothed over a bigger scale than the smallest distance between these objects. The background problem consists in computing the evolution of the cosmological fluid (i.e., the distortions due to its own gravitational field, its possible transformations under phase transitions, etc.). The goal is to understand e.g. the average expansion rate as a function of time, the age of the universe, etc.

The perturbation problem consists in writing first-order (linear) perturbations in a given background and solve for their evolution. The goal is to understand, for instance, the large-scale structure of the universe or the Cosmic Microwave Background (CMB) anisotropies. The approach can even be pushed to second-order (quadratic) perturbations, but then equations become extremely complicated.

Of course, this approach cannot work for describing the formation of small scale structures. For instance, the merging of two galaxies is a fully non-linear gravitational problem which cannot be addressed by a perturbed expansion. On the other hand, it is not necessarily sensitive to General Relativity and to the expansion of the universe. The interesting question is to understand whether the cosmological perturbation theory is self-consistent on the largest scales today, and possibly on all scales in the remote past.

Today, all physicists agree that the cosmological perturbation theory provides an excellent description of the universe at early times on all scales (we

will quantify the statement "early time" later in the course), which can accurately explain e.g. observations of the CMB or of light element abundances. In addition, a large majority of cosmologists believes that cosmological perturbation theory is able to explain the structure and evolution of the universe on the largest observables scales until today. On small scales, the relativistic cosmological perturbation theory should be substituted by a Newtonian non-linear approach (involving N-body gravitational clustering simulations).

2.1.2 General Relativity in two words

Bern students following this course went through a detailed course on General relativity during the last semester. Hence *they may skip this subsection*, in which we try to summarise the main ideas of General Relativity, for people who never learnt anything about it.

Then, in the following sections, we will derive step by step the general relativistic laws governing the evolution universe, and stress the differences with their Newtonian counterparts.

When Einstein tried to build a theory of gravitation compatible with the invariance of the speed of light, the equivalence principle and Newton's law in some particular limit, he found that the minimal price to pay was :

- to abandon the idea of a gravitational potential, related to the distribution of matter, and whose gradient gives the gravitational field in any point.
- to assume that our four-dimensional space-time is curved, and that free-falling objects describe geodesics in this space-time.
- to relate the properties of curvature in a given point to the properties of matter in the same point.

What does that mean in simple words?

First, let's recall briefly what a curved space is, first with only two-dimensional surfaces. Consider a plane, a sphere and an hyperboloid. For us, it's obvious that the sphere and the hyperboloid are curved, because we can visualize them in our three-dimensional space: so, we have an intuitive notion of what is flat and what is curved. But if there were some two-dimensional aliens living on these surfaces, not being aware of the existence of a third dimension, how could they know whether they leave in a flat or in a curved space-time? There are several ways in which they could measure it. One would be to obey the following prescription: walk in straight line on a distance d; turn 90 degrees left; repeat this sequence three times again; see whether you are back at your initial position. The aliens on the three surfaces would find that they are back there as long as they walk along a small square, much smaller than the radius of curvature. But a good test is to repeat the operation on larger and larger distances. When the size of the square will be not so small compared the radius of curvature, the alien on the sphere will notice that before stopping, he crosses the first branch of his trajectory (see figure 2.1). The one on the hyperboloid will stop without closing his trajectory. Another way to specify the curvature of a two-dimensional surface is to map it with an arbitrary coordinate system (x, y), and to use a scaling law or *line element*, i.e. a function $dl(x, y, dx, dy)$ providing a measure of infinitesimal distances as a function of position and of infinitesimal coordinate differences. For example, on projected maps of the earth's surface, one should know the scaling law in order to correctly estimate distances between two points of given latitude and longitude. At the next level of precision, the surface of the earth is curved by mountains and valleys. In a given region, having under disposal a precise topological map with contour lines of constant

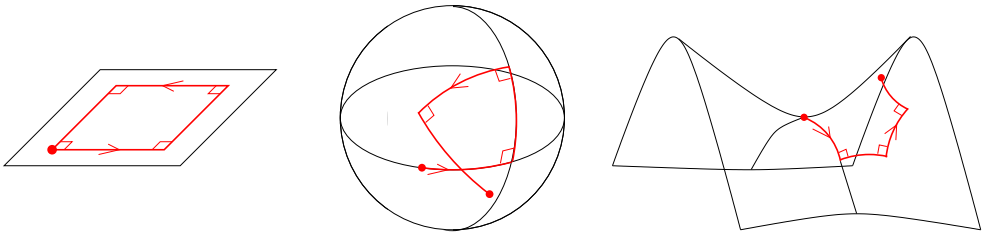


Figure 2.1: Measuring the curvature of some two-dimensional spaces. By walking four times in straight line along a distance d, and turning 90 degrees left between each walk, a small man on the plane would find that he is back at his initial point. Doing the same thing, a man on the sphere would walk across his own trajectory and stop away from his departure point. Instead, a man on the hyperboloid would not close his trajectory.

elevation, one can use a scaling law for estimating the physical distance between two neighboring points as a function of their latitude and longitude difference, and of the number of contour lines between the two points.

Getting an intuitive representation of a three-dimensional curved space is much more difficult. A 3-sphere and a 3-hyperboloid could be defined analytically as some 3-dimensional spaces obeying to the equation $a^2+b^2+c^2\pm d^2 = R^2$ inside a 4-dimensional Euclidean or Minkowski space with coordinates (a, b, c, d). If we wanted to define them by making use of only three dimensions, the problem would be exactly like for drawing a projected map of the Earth. We would need to specify the line element $dl(x, y, z, dx, dy, dz)$ everywhere, within a given (arbitrary) coordinate system. Of course, the coordinates can be defined arbitrarily, but the physical distances computed from dl are related to intrinsic properties of the curved space, invariant under a change of coordinate. The scaling law leads to the definition of a spatial metric tensor defined through $dl^2 = g_{ij}(x^i)\, dx^i dy^j$, and to the whole formalism of Riemannian geometry (curvature tensor, intrinsic curvature scalar, geodesics, etc.).

That was still for three dimensions. The curvature of a four-dimensional space-time is very difficult to visualize intuitively, first because it has even more dimensions, and second because in special and general relativity, there is a difference between time and space. For a given space-time manifold, one can choose an arbitrary system of coordinates (time x^0 and space x^1, x^2, x^3) and describe the space-time curvature by the line element ds (which represents the infinitesimal distance betwen two closeby *events* rather than two closeby spatial points). The 4×4 metric defined through $ds^2 = g_{\mu\nu}(x^\mu)\, dx^\mu dy^\nu$ must have a negative signature (i.e. negative determinant) in order to recover locally Lorentz invariance and special relativity[1].

Now, the definition of geodesics is the following. Take an initial point and an initial direction. They define a unique line, called a geodesic, such that any portion of the line gives the shortest trajectory between the two points (so, for instance, on a sphere of radius R, the geodesics are all the great circles of radius R, and nothing else). In general relativity (as in any theory of gravity respecting the equivalence principle and hence based on geometry and a metric tensor), the trajectories $x^\mu(\lambda)$ of free-falling bodies are geodesics of the space-time specified

[1]There are two sign conventions fulfilling this condition of negative signature: the $-+++$ convention in which $g_{00} < 0$, and the $+---$ convention in which $g_{00} > 0$. In this course we will use the $-+++$ convention.

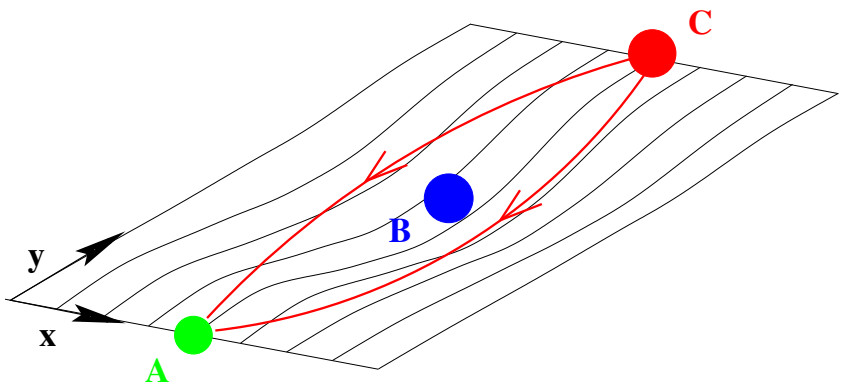


Figure 2.2: Gravitational lensing. Somewhere between an object C and an observer A, a massive object B - for instance, a galaxy - curves its surrounding space-time. Here, for simplicity, we only draw two spatial dimensions. In absence of gravity and curvature, the only possible trajectory of light between C and A would be a straight line. But because of curvature, the straight line is not anymore the shortest trajectory. Photons prefer to follow two geodesics, symmetrical around B. So, the observer will not see one image of C, but two distinct images. In fact, if we restore the third spatial dimension, and if the three points are perfectly aligned, the image of C will appear as a ring around B. This phenomenon is observed in practice.

by the metric $g_{\mu\nu}$. The geodesics obey to

$$\frac{d^2x^\alpha}{d\lambda^2} + \Gamma^\alpha_{\mu\nu}\frac{dx^\mu}{d\lambda}\frac{dx^\nu}{d\lambda} = 0 \tag{2.1}$$

and depend on curvature through the Christoffel symbols $\Gamma^\alpha_{\mu\nu}$, which in turn can be expressed as a function of the metric as

$$\Gamma^\alpha_{\mu\nu} = \frac{1}{2}g^{\alpha\beta}(g_{\mu\beta,\nu} + g_{\beta\nu,\mu} - g_{\mu\nu,\beta}) \ . \tag{2.2}$$

Note that the expression of $g_{\mu\nu}$ and $\Gamma^\alpha_{\mu\nu}$ are not invariant under a change of coordinate, while the curvature of the underlying manifold and the ensemble of geodesics on this manifold are intrinsic, coordinate-independent properties of the manifold.

All free-falling bodies follow geodesics, including light rays. This leads for instance to the phenomenon of gravitational lensing (see figure 2.2).

The Einstein theory of gravitation says that four-dimensional space-time is curved, and that the properties of curvature in each point (related to the metric and its derivatives) depends entirely on the matter distribution in that point. In simple words, this means that the metric tensor plays more or less the same role as the gravitational potential in Newtonian gravity.

So, in General Relativity, gravitation is not formulated as a force or a field, but as the curvature of space-time, sourced by matter. All isolated systems follow geodesics which are bent by the curvature. In this way, their trajectories are affected by the distribution of matter around them: this is precisely what gravitation means.

2.1.3 Frame comoving with an observer

Let us consider a free-falling observer M in an arbitrary curved space-time. The observer's trajectory (which is a geodesic) can be described parametrically by a

set of functions $\{x^1(t), x^2(t), x^3(t)\}$.

We can always perform a change of coordinates in such way that this particular observer has fixed spatial coordinates x^i_M: along this geodesic and in the new coordinates, $dx^i/dx^0 = 0$ and $x^i = x^i_M$. This frame is said to be comoving with the observer, and locally all terms g_{0i} vanish: in each point $x^\mu = (t, x^1_M, x^2_M, x^3_M)$ one has $g_{0i}(x^\mu) = 0$.

In addition, it is possible to define the time coordinate in such way that the coefficient $g_{00}^{1/2}$ (called the lapse function) is constant all over the geodesic of our particular observer, and equal to the speed of light c. If this is the case, the line element between two closeby events on the observer's trajectory is given by $ds^2 = -c^2(dx^0)^2 + g_{ij}dx^i dx^j = c^2(dx^0)^2$ (since $dx^i = 0$). Hence, the time coordinate $x^0 \equiv t$ obeys to the definition of proper time. It represents the physical time measured by our free-falling observer.

2.1.4 Building the first cosmological models

After obtaining the mathematical formulation of General Relativity around 1916, Einstein considered various testable consequences of his theory in the solar system (e.g., corrections to the trajectory of Mercury, or to the apparent diameter of the sun during an eclipse). But remarkably, he immediately understood that GR could also be applied to the universe as a whole, and published some first attempts in 1917. However, Hubble's results concerning the expansion were not known at that time, and most physicists had the prejudice that the background universe should be not only isotropic and homogeneous, but also static – or stationary. As a consequence, Einstein (and other people like De Sitter) found some interesting static cosmological solutions, but not the ones that really describe our universe.

A few years later, some other physicists tried to relax the assumption of stationarity. The first was the Russian physicist Alexander Friedmann (in 1922), followed independently by the Belgian physicist and priest Lemaître (in 1927), and then by some Americans, Robertson and Walker. When the Hubble flow was discovered in 1929, it became clear for a fraction of the scientific community that the universe could be described by the equations of Friedmann, Lemaître, Robertson and Walker. However, many people – including Hubble and Einstein themselves – remained reluctant to this idea for many years. Today, the Friedmann – Lemaître model is considered as one of the major achievements of modern physics.

2.1.5 Coordinate choice in the FLRW universe

The LFRW model is the most general solution of the GR equations under the assumption that the background universe is homogeneous and isotropic.

The fact that the universe is postulated to be homogeneous and isotropic (but not necessarily static) means that there exist a definition of time such that at each instant, all points and all directions are equivalent. For instance, the energy density should only be a function of this time, not of space: $\rho(x^\mu) = \rho(x^0)$.

We immediately notice that the fact of being "homogeneous, isotropic and non-stationary" cannot be a coordinate-independent property of a given universe, by construction: it privileges a particular definition of time, or more precisely, a particular time-slicing. A redefinition of time $t \longrightarrow t'(t)$ does not change the time-slicing. In the new time coordinate, at a given time t', all spatial points are still equivalent. A more general redefinition of time mixing time

and space, $t \longrightarrow t'(t, x^1, x^2, x^3)$, changes the time-slicing: 3D hypersurfaces of constant t' are no longer homogeneous.

The easiest way to build a system of coordinates in a homogeneous universe is to start from an initial homogeneous hypersurface, to assign it a time coordinate t_1 and some arbitrary spatial coordinates. In each point, we can place an observer at rest with respect to the coordinate system: for any of these observers, $dx^i/dx^0(t_1) = 0$. This is possible by assumption: since the hypersurface is assumed to be homogeneous, there is no "force" imposing some "bulk motion" to all observers. We then give a clock to each of our observers. These clocks indicate the proper time measured by each of them. We define a new hypersurface as the ensemble of all points in space-time such that the clocks indicate a common value t_2. We assign to this hypersurface the time coordinate t_2, and some spatial coordinates such that each of our observers keeps fixed spatial coordinates. This can be repeated in order to map the entire space-time with a set of coordinates such that: all our observers keep fixed spatial coordinates; and the time coordinate corresponds to the proper time measured by all observers. In other words, we have built a frame which is comoving not just with one observer (as in a previous subsection), but with an infinity of observers mapping the entire space. These particular observers are called "comoving observers", and any set of coordinates built in that way is called a comoving coordinate system.

In comoving coordinates and using proper time, the metric describing the whole space-time reads

$$ds^2 = -c^2 dt^2 + g_{ij} dx^i dx^j \tag{2.3}$$

in which t is the proper time, (x^i) are some spatial comoving coordinates, and g_{ij} must have a special form preserving the homogeneity and isotropy of three-dimensional space at any given time t. We will write down this form of g_{ij} in the next subsection. One is still free to perform some change of coordinates, and it is worth noticing that:

- a simple redefinition of time $t \longrightarrow t'(t)$ preserves the above form of the metric, excepted that $g_{00} \neq -c^2$. The new time coordinate does not represent the proper time of comoving observers anymore, but it still defines a time-slicing of space-time in homogeneous hypersurfaces. In the following, we will sometimes use different definitions of the time coordinate. Physical problems can be solved with any of these time coordinates, although observables involving physical periods of time or rates should always be computed with the proper time of the observer making the experiment.

- an internal redefinition of spatial coordinates $x^i \longrightarrow x^{i'}(x^i)$ preserves the above form, and the universe will still appear as homogeneous in the new system. Hence, there is an infinite number of possible comoving spatial coordinate systems. In the following we will use cartesian coordinates, spherical coordinates, etc.

- a general change of coordinates mixing space and time would not preserve the above form of the metric. In the new coordinate system, the universe would not appear as homogeneous, since quantities like e.g. the spatial curvature or the total energy density would depend on both time and space. The new frame could represent locally the comoving frame of an observer leaving in a homogeneous universe, but not being at rest with the ensemble of comoving observers (who see homogeneous and isotropic observables). Such an observer with a peculiar velocity should not perceive the universe as isotropic: for instance, if the universe is filled with a

homogeneous background of light, a non-comoving observer should see a Doppler effect affecting the color of this light (bluer in front of him, redder behind). It is important to understand that the FLRW assumption does not say that all possible observers see a homogeneous universe, but simply that there exists an ensemble of observers seeing a homogeneous universe, and hence, a global "comoving frame".

2.1.6 The curvature of the FLRW universe

So far, we have not specified the part g_{ij}. We only assumed that it preserves homogeneity and isotropy. So, the curvature should be the same everywhere at a given time. The list of possible three-dimensional spaces with constant curvature is very short: flat Euclidean space, 3-sphere and 3-hyperboloid.

In flat space, one can use e.g. Cartesian or polar coordinates and write the spatial line element as

$$dl^2 = dx^2 + dy^2 + dz^2 = dr^2 + r^2(d\theta^2 + \sin^2\theta\ d\phi^2)\ . \tag{2.4}$$

All possible changes of coordinate preserve this flatness. Let us rewrite the line element after the simplest possible change, namely an homothetic transformation with respect to the origin of coordinates:

$$dl^2 = a^2(dx^2 + dy^2 + dz^2) = a^2[dr^2 + r^2(d\theta^2 + \sin^2\theta\ d\phi^2)]\ . \tag{2.5}$$

Let's go back now to the full FLRW space-time. It is obvious that

$$ds^2 = -c^2dt^2+a^2(dx^2+dy^2+dz^2) = -c^2dt^2+a^2[dr^2+r^2(d\theta^2+\sin^2\theta\ d\phi^2)] \tag{2.6}$$

describes a flat, isotropic universe, but this universe is static. In fact we only want the universe to be homogeneous/isotropic *at any given time*, so

$$ds^2 = -c^2dt^2+a(t)^2(dx^2+dy^2+dz^2) = -c^2dt^2+a(t)^2[dr^2+r^2(d\theta^2+\sin^2\theta\ d\phi^2)] \tag{2.7}$$

(where we made the a factor time-dependent) is another obvious solution to the FLRW problem leading to a homogeneous, non-stationary and spatially flat universe. This is even the most general FLRW solution with zero spatial curvature (as usual, modulo trivial time redefinitions and spatial changes of coordinates)[2].

Again, three-dimensional spaces with constant non-zero curvature fall in two categories: 3-spheres and 3-hyperboloids. A convenient choice of polar coordinate leads to the following expression for the line elements in such spaces:

$$dl^2 = \left[\frac{dr^2}{1-kr^2} + r^2(d\theta^2 + \sin^2\theta\ d\phi^2)\right] \tag{2.8}$$

where k is a constant number, related to the spatial curvature: if $k = 0$, the universe is Euclidean (and called a "flat universe"), if $k > 0$, it is positively curved (and called a "closed universe"), and if $k < 0$, it is negatively curved (and called an "open universe"). In the last two cases, the radius of curvature is given by

$$r_c(t) = \frac{1}{\sqrt{|k|}}. \tag{2.9}$$

[2]Above, we performed a homothetic transformation of coordinates and allowed the factor appearing in the transformation to become a time-dependent function. We could have made a different transformation, generating other factors, and tried to make these other factors time-dependent. But in general this would break homogeneity and isotropy, unless the time-dependent factor can be factored out like in the above solution!

When $k > 0$, the universe has a finite volume, and the coordinate r is defined only in the range $0 \leq r < r_c$. This is the reason for which positively curved universes are usually called "closed". The terms "open universe" just refer to the opposite case.

The most general solution for an homogeneous, isotropic, non-stationary universe is obtained again by multiplying the above spatial line element by the square of a time-dependent factor $a(t)$ called the scale factor:

$$ds^2 = -c^2dt^2 + a(t)^2 \left[\frac{dr^2}{1-kr^2} + r^2(d\theta^2 + \sin^2\theta \; d\phi^2) \right] . \tag{2.10}$$

The corresponding metric is called the FLRW metric (in comoving spherical coordinates). So, in three-dimensional space, infinitesimal physical distances dl are always given by the scale factor $a(t)$ times the comoving line element computed from eq. (2.8). This is still true for a macroscopic length obtained by integrating dl over a given path in three-dimensional space: in the FRLW universe, the physical size of an object at a given time is always equal to its comoving size multiplied by the scale factor at that time[3]. In particular we can immediately notice that the physical size of the radius of curvature in the FLRW universe is

$$R_c^{\text{physical}}(t) = \frac{a(t)}{\sqrt{|k|}}. \tag{2.11}$$

The previous expression in eq. (2.9) provides only the comoving radius of curvature. Note that r and $a(t)$ can always be rescaled by $r \longrightarrow r\sqrt{|k|}$, $a(t) \longrightarrow a(t)/\sqrt{|k|}$. After the rescaling, the metric reads like in eq. (2.10), but with k restricted to the three possible values +1 (positive curvature), 0 (flat) or -1 (negative curvature) without loss of generality.

We know that observers at rest with the cosmological fluid have fixed comoving coordinates (it is trivial to check that all trajectories parametrized by ($x^i = x^i_M$=constant) are solutions of the geodesics equations in the FLRW metric). This doesn't mean that the universe is static, because all distances grow proportionally to $a(t)$: so, the scale factor accounts for the homogeneous expansion. An analogy helps in understanding this concept. Let us take a rubber balloon and draw some points on the surface. Then, we inflate the balloon. The distances between all the points grow proportionally to the radius of the balloon. This is not because the points have a proper motion on the surface, but because all the lengths on the surface of the balloon increase with time. In other words, in general relativity, the universe expansion is not described anymore through the velocity of objects like in Newtonian cosmology, but through the expansion of the background spacetime.

Intuitively, the FLRW metric describes a curved space-time with two types of curvature:

- the spatial curvature, described by $\pm a(t)/\sqrt{|k|}$ at each time.
- the space-time curvature described by the time evolution of $a(t)$.

The second is maybe more difficult to visualize as a curvature term, but we will see later that both terms contribute e.g. to the curvature of light ray trajectories in space-time. In a few sections, we will also see that the scale factor defines an actual radius of curvature, the Hubble radius $R_H(t) = ca(t)/\dot{a}(t)$.

If k was equal to zero and a was constant in time, we could redefine the coordinate system with $(r', \theta', \phi') = (ar, \theta, \phi)$, obtain the Minkowski metric and

[3]We will see however in the next sections that due to the finite speed of light, speaking of macroscopic distance in cosmology can be somewhat subtle and require more work and definitions.

go back to Newtonian gravity. So, we stress again that the curvature really manifests itself as $k \neq 0$ (for spatial curvature) and $\dot{a} \neq 0$ (for the remaining space-time curvature).

Note finally that in the rest of the course, some equations may take a simpler form after the time redefinition $dt = a(t)d\tau$. In this case, the time dependence factors out from the full FLRW line element:

$$ds^2 = a^2(\tau)\left(-c^2 d\tau^2 + \left[\frac{dr^2}{1-kr^2} + r^2(d\theta^2 + \sin^2\theta\ d\phi^2)\right]\right) \tag{2.12}$$

where the scalar factor $a(t)$ as been re-expressed as a function of the new time variable τ. This metric exhibits conformal symmetry; hence, τ is called *conformal time*, by opposition to the proper time t, also called *cosmological time*.

2.2 Curvature of light-rays in the FLRW universe

Our goal in this section is to understand the concrete consequences of the universe expansion for observers looking at the sky. Hence, we need to understand how light rays propagate in the universe.

2.2.1 Photon geodesics

Photon propagate in the vacuum at the speed of light along geodesics. Hence, over an infinitesimal time interval dt, they run over a distance $dl^2 = c^2dt^2$. On macroscopic scales, the relation between distance and time is given by integrating $dl = \pm cdt$ over the geodesic.

By definition, we are only interested in photons reaching us at some point, and allowing us to observe an object. Lets us consider that we are a comoving observer and choose the origin of spherical comoving coordinates to coincide with us (this choice is only made for getting simple calculations; it doesn't imply at all that we occupy some privileged point in space or anything like that). In the FLRW universe, a photon reaching us with a momentum aligned with a given direction (θ_e, ϕ_e) must have travelled along a straight line in space, starting from an unknown emission point (r_e, θ_e, ϕ_e). If its spatial trajectory was not a straight line, there would be a contradiction with the assumption of an isotropic universe. However the photon trajectory in space-time is curved, as can be checked by integrating over the infinitesimal distance between the emission point $(t_e, r_e, \theta_e, \phi_e)$ and a later point (t, r, θ_e, ϕ_e) with $t > t_e$, $r < r_e$:

$$\int_{r_e}^{r} -\frac{dr}{\sqrt{1-kr^2}} = \int_{t_e}^{t} \frac{c\,dt}{a(t)} \tag{2.13}$$

On can check that this trajectory is indeed a solution of the geodesic equation, and that it corresponds to a curved trajectory in space-time: if we draw this trajectory in two-dimensional (t, r) space, we see that the slope $dr/dt = -c\sqrt{1-kr^2}/a(t)$ changes along the trajectory. The photon is seen by the observer (at the origin of coordinates) at a reception time t_0 which can be deduced from r_e and t_e through the implicit relation:

$$\int_{r_e}^{0} -\frac{dr}{\sqrt{1-kr^2}} = \int_{t_e}^{t_0} \frac{c\,dt}{a(t)}\ . \tag{2.14}$$

The ensemble of all points (t_e, r_e, θ, ϕ) for which eq. (2.14) holds define our past light-cone at time t_0, as illustrated in figure 2.3. Note that the right-hand side

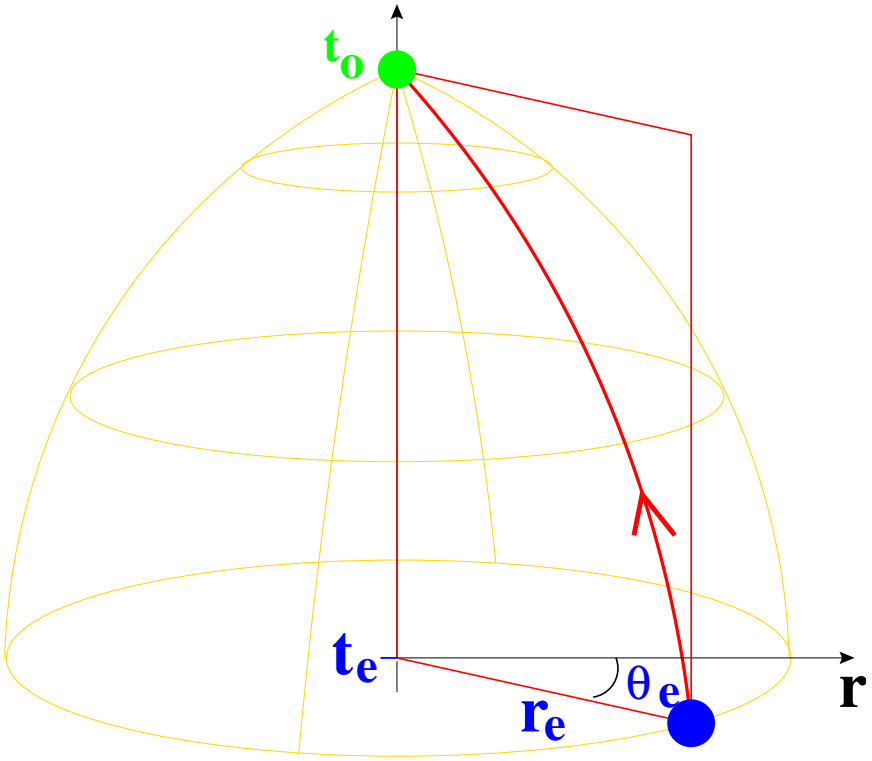


Figure 2.3: An illustration of the propagation of photons in our universe, skipping one spatial dimension. We are sitting at the origin, and at a time t_0 we can see the light of a galaxy emitted at (t_e, r_e, θ_e). Before reaching us, this light has travelled over a trajectory which is straight in three-dimensional space (constant angles), but curved in space-time. In any point, the slope dr/dt is given by equation (2.13). So, the relation between r_e, t_0 and t_e depends on the spatial curvature and on the scale factor evolution. The trajectory would be a straight line in space-time only if $k = 0$ and $a = constant$, i.e., in the limit of Newtonian mechanics in Euclidean space. The ensemble of all possible photon trajectories crossing $r = 0$ at $t = t_0$ is called our "past light cone", visible here in orange. Asymptotically, near the origin, it can be approximated by a linear cone with $dl = cdt$, showing that at small distance, the physics is approximately Newtonian. **Important remark: here, the past line cone has be drawn as a convex cone. Instead, for realistic cosmological scenarios, the cone is concave.**

corresponds exactly to the conformal time interval $(\tau_r - \tau_e)$ times the speed of light.

The equation (2.13) describing the propagation of light (more precisely, of radial incoming photons) is extremely useful - probably, one of the two most useful equations of cosmology, together with the Friedmann equation, that we will present soon. It is on the basis of this equation that we are able today to measure the curvature of the universe, its age, its acceleration, and other fundamental quantities.

2.2.2 A new definition of redshift

First, a simple calculation based on equation (2.13) gives the redshift associated with a given source of light. Let's still play the role of a comoving observer sitting at the origin of coordinates. We observe a galaxy located at (r_e, θ_e, ϕ_e), emitting light at a given frequency λ_e. The corresponding wave crests are emitted by the galaxy at a frequency $\nu_e = c/\lambda_e$ with a period $dt_e \equiv 1/\nu_e$. Each wave crests follows the trajectory described by eq. (2.13). We receive the light signal with a frequency $\nu_r = c/\lambda_r = 1/dt_r$ such that

$$\int_{r_e}^{0} -\frac{dr}{\sqrt{1-kr^2}} = \int_{t_e}^{t_r} \frac{dt}{a(t)} = \int_{t_e+dt_e}^{t_r+dt_r} \frac{dt}{a(t)} \ . \tag{2.15}$$

The second equality gives:

$$\int_{t_e}^{t_e+dt_e} \frac{dt}{a(t)} = \int_{t_r}^{t_r+dt_r} \frac{dt}{a(t)} \ . \tag{2.16}$$

Hence in very good approximation:

$$\frac{dt_e}{a(t_e)} = \frac{dt_r}{a(t_r)}. \tag{2.17}$$

We infer a simple relation between the emission and reception wavelengths:

$$\frac{\lambda_r}{\lambda_e} = \frac{dt_r}{dt_e} = \frac{a(t_r)}{a(t_e)}. \tag{2.18}$$

This result could have been easily guessed: a wavelength is a distance, subject to the same stretching as all physical distances when the scale factor increases. Hence, in the FLRW universe, the redshift imposed by the expansion is given by

$$z = \frac{\Delta\lambda}{\lambda} = \frac{\lambda_r - \lambda_e}{\lambda_e} = \frac{a(t_r)}{a(t_e)} - 1 \ . \tag{2.19}$$

In other words, if we observe an object now, at time t_0, its absorption lines are redshifted by a factor

$$z = \frac{a(t_0)}{a(t_e)} - 1 \ . \tag{2.20}$$

This is a crucial difference with respect to Newtonian mechanics, in which the redshift was defined as $z = v/c$, and seemed to be limited to $|z| < 1$. The true GR expression doesn't have such limitations, since the ratio of the scale factors can be arbitrarily large without violating any fundamental principle. And indeed, observations do show many objects - like quasars - at redshifts of $z \sim 4$ or even bigger. We'll see later that we also observe the Cosmic Microwave Background at a redshift of approximately $z = 1100$!

Note finally that in the real perturbed universe, objects are never exactly comoving, they have small peculiar velocities $\vec{v}_c$ with respect to the comoving frame. Hence, the observed redshift is given by the sum of a General Relativity contribution given by eq. (2.20), and a Doppler contribution given by $(\vec{v}_c \cdot \hat{n})/c$. The second term rarely exceeds $\mathcal{O}(10^{-3})$, while the first term grows from zero for nearby objects to infinity for remote objects. Hence, we expect that at very short distances, the Doppler contribution can dominate, while at larger distances the GR contribution takes over.

2.2.3 A new definition of the Hubble parameter

In the limit of small redshift, we expect to recover the Newtonian results, and to find a relation similar to $z = v/c = HL/c$ (where L is the physical distance to the object). To show this, let's assume again that t_0 is the present time, and that we are a comoving observer at $r = 0$. We want to compute the redshift of a nearby galaxy, which emitted the light that we receive today at a time $t_0 - dt$. In the limit of small dt, the equation of propagation of light shows that the physical distance L between the galaxy and us is simply

$$L \simeq dl = c\,dt \tag{2.21}$$

while the redshift of the galaxy is

$$z = \frac{a(t_0)}{a(t_0 - dt)} - 1 \simeq \frac{a(t_0)}{a(t_0) - \dot{a}(t_0)dt} - 1 = \frac{1}{1 - \frac{\dot{a}(t_0)}{a(t_0)}dt} - 1 \simeq \frac{\dot{a}(t_0)}{a(t_0)}dt \ . \tag{2.22}$$

By combining these two relations we obtain

$$z \simeq \frac{\dot{a}(t_0)}{a(t_0)}L/c \ . \tag{2.23}$$

So, at small redshift, we recover the Hubble law, and the role of the Hubble parameter is played by $\dot{a}(t_0)/a(t_0)$. In the Friedmann universe, we will define the Hubble parameter at any time as the expansion rate of the scale factor:

$$H(t) = \frac{\dot{a}(t)}{a(t)} \ . \tag{2.24}$$

The current value of the Hubble parameter (the one measured by Hubble himself) will be noted as H_0.

We have proved that in the FLRW universe, the proportionality between distance and velocity (or redshift) is recovered for small distances and redshifts. What happens at larger distance? This question actually raises a non-trivial problem: the definition of distances for objects which are so far from us that the (Euclidean) approximation $L = dl = dt$ becomes inaccurate.

2.2.4 The notion of distance to an object

Let us assume again that sitting at the origin of spherical coordinates at time t_0, we observe a remote comoving object emitting light from $(t_e, r_e, \theta_e, \phi_e)$. What is the physical distance to the object? This question is ambiguous in an expanding universe. Are we asking about the distance in units of today, i.e. the distance between us and the position of this object today? If it is a comoving object, it should be located now at coordinates $(t_0, r_e, \theta_e, \phi_e)$. Then, the distance computed on the constant-time hypersurface with $t = t_0$ is given by

$$d = \int_0^{r_e} dl = a(t_0) \int_0^{r_e} \frac{dr}{\sqrt{1 - kr^2}} \ . \tag{2.25}$$

Very often, the scale factor is defined in such way that $a(t_0) = 1$, and the above distance d coincides with the *comoving distance* $\chi(r_e)$:

$$\chi(r_e) \equiv \int_0^{r_e} \frac{dr}{\sqrt{1 - kr^2}} \ , \tag{2.26}$$

which can be integrated to

$$\chi(r) = \begin{cases} \sin^{-1}(r) & \text{if} \quad k = 1, \\ r & \text{if} \quad k = 0, \\ \sinh^{-1}(r) & \text{if} \quad k = -1. \end{cases} \tag{2.27}$$

Hence, it is useful to define the function

$$f_k(x) \equiv \begin{cases} \sin(x) & \text{if} \quad k=1, \\ x & \text{if} \quad k=0, \\ \sinh(x) & \text{if} \quad k=-1, \end{cases} \tag{2.28}$$

so that $r = f_k(\chi)$.

It follows from eq. (2.14) that $\chi(r)$ is equal to the conformal age of the object, $(\tau_0 - \tau_e)$, times the speed of light:

$$\chi(r) = \int_{t_e}^{t_0} \frac{c\,dt}{a(t)} = c(\tau_0 - \tau_e) \ . \tag{2.29}$$

At this point, conformal time takes a particular signification: it is a particular measure of time, which is equal to the comoving distance traveled by a light signal divided by c. In units in which $c = 1$ and assuming $a(t_0) = 1$, both χ and τ can be expressed in units of physical distances today, e.g. in Mega-parsecs. These are indeed the most comon units for comoving distance and conformal time.

Comoving distances are well-defined quantities, up to a choice of normalization for $a(t)$. They are used by observers in many circumstances. By construction, the comoving distance between two comoving objects does not depend on time, unlike the physical distance between them. However, this is a purely conventional and rather artificial definition of distances, since we can't see remote objects today - they might even have disappeared. Anyway, we should not argue about the definition of distances, because distances are not directly measurable quantities in cosmology. We should concentrate on experimental, indirect ways to probe them. Each experimental technique will lead to a particular definition of distance.

In astrophysics, distances are usually measured in three ways:

- *From the redshift.* In principle the observed redshift of objects measures the ratio $a(t_0)/a(t_e)$ plus corrections due to the local effects of small-scale inhomogeneities (peculiar velocity of the object, ...). On very large distances, one can neglect the impact of inhomogeneities and assume in first approximation that the observed redshift is really equal to $a(t_0)/a(t_e) - 1$. Then, if we know in advance the function $a(t)$, we can identify the time t_e and compute the comoving distance $\chi(t_e)$ by integrating $(c\,dt/a(t))$ from t_e to t_0. This method is (in first approximation) the one used by observers trying to infer the spatial distribution of galaxies from *galaxy redshift surveys.* The distance reported in pictures showing the distribution of galaxies in slices of our universe is obtained in that way. However, it assumes an *a priori* knowledge of the function $a(t)$. In many cases, this function is precisely what one wants to measure.

- *From the angular diameter of standard rulers.* Surprisingly, there exist a few objects in astrophysics and cosmology which physical size can be known in advance, given some physical properties of these objects. They are called *standard rulers.* In the next chapters we will introduce one example of standard ruler: the sound horizon at decoupling, "observed" in CMB anisotropies. In Euclidean space, the distance d to a spherical object can be inferred from its physical diameter dl and its angular diameter $d\theta$ through $dl = d \times d\theta$. In FLRW cosmology, although the geometry is not Euclidean, we will adopt exactly this relation as one of the possible definitions of distance. The corresponding quantity is called the angular diameter distance d_A,

$$d_A \equiv \frac{dl}{d\theta} \ . \tag{2.30}$$

In Euclidean space, d_A would be proportional to the usual Euclidean distance to the object and therefore to its redshift. In the FLRW universe, the relation between the angular diameter distance and the redshift is non-trivial and depends on the spacetime curvature, as we shall see in the next subsection.

- *From the luminosity of standard candles.* As we have seen already with Cepheids, there exists also objects called *standard candles* for which the absolute luminosity (i.e. the total luminous flux emitted per unit of time) can be estimated independently of its distance and apparent luminosity. In Euclidean space, the distance could be inferred from the absolute luminosity L and apparent one l through $l = L/(4\pi d^2)$. In cosmology, although the geometry is not Euclidean, we will adopt exactly this relation as one of the possible definitions of distance. The corresponding quantity is called the luminosity distance d_L,

$$d_L \equiv \sqrt{\frac{L}{4\pi l}} \ . \tag{2.31}$$

In Euclidean space, d_L would be again proportional to the usual Euclidean distance to the object and therefore to its redshift, while in the FLRW universe the relation between the luminosity distance and the redshift is as subtle as for the angular diameter distance.

2.2.5 Angular diameter distance – redshift relation

Recalling that in Euclidean space with Newtonian gravity and homogeneous (linear) expansion, one has $z = v/c$ and $v = H_0 d$, we easily find a trivial relation between the angular diameter distance and the redshift:

$$d_A = d = (c/H_0)\, z. \tag{2.32}$$

In General Relativity, because of the bending of light-rays by gravity, the steps of the calculation are different. Using the definition of infinitesimal distances (2.10), we see that the physical size dl (evaluated at time t_e) of an object orthogonal to the line of sight is related to its angular diameter $d\theta$ through

$$dl = a(t_e)\ r_e\ d\theta \tag{2.33}$$

where t_e is the time at which the galaxy emitted the light ray that we observe today on Earth, and r_e is the comoving coordinate of the object. Hence

$$d_A = a(t_e)\ r_e = a(t_0)\frac{r_e}{1+z_e} \ . \tag{2.34}$$

We can replace r_e using Eqs. (2.26) - (2.29):

$$\begin{aligned} d_A &= \frac{a(t_0)}{1+z_e}\ f_k(\chi) && (2.35)\\ &= \frac{a(t_0)}{1+z_e}\ f_k\left(\int_{t_e}^{t_0} \frac{c\,dt}{a(t)}\right) && (2.36)\\ &= \frac{a(t_0)}{1+z_e}\ f_k\left(\int_{a_e}^{a_0} \frac{c\,da}{a^2 H(a)}\right) && (2.37)\\ &= \frac{a(t_0)}{1+z_e}\ f_k\left(\int_{0}^{z_e} \frac{c\,dz}{a(t_0) H(z)}\right) && (2.38) \end{aligned}$$

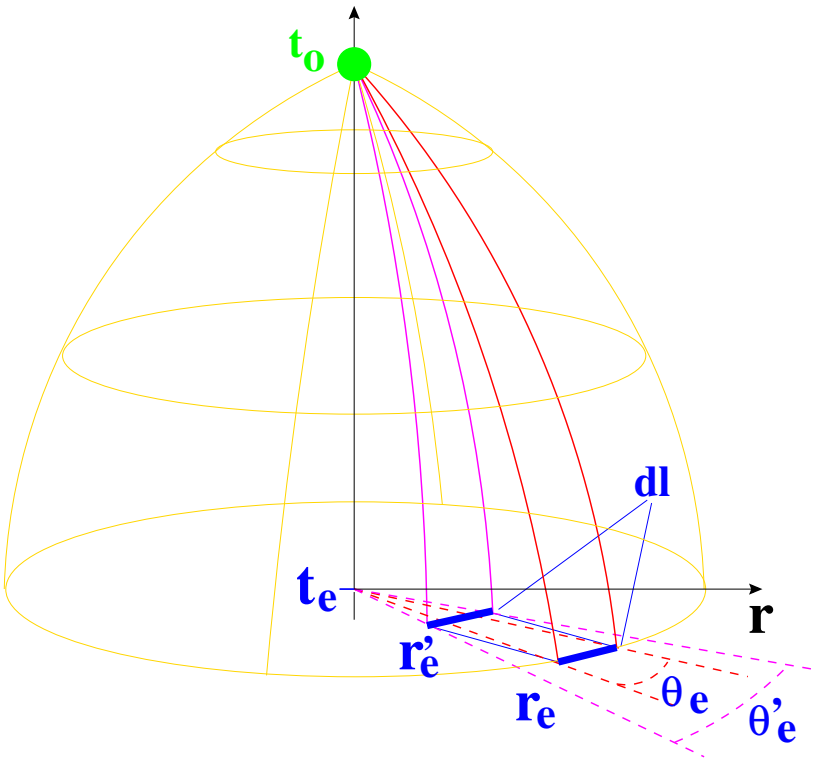


Figure 2.4: Angular diameter – redshift relation. We consider an object of fixed size dl and fixed redshift, sending a light signal at time t_e that we receive at present time t_0. All photons travel by definition with θ =constant. However, the bending of their trajectories in the (t, r) plane depends on the spatial curvature and on the scale factor evolution. So, for fixed t_e, the comoving coordinate of the object, r_e, depends on curvature. The red lines are supposed to illustrate the trajectory of light in a flat universe with $k = 0$. If we keep dl, $a(t)$ and t_e fixed, but choose a positive value $k > 0$, we infer from equation (2.13) that the new coordinate ${r_e}'$ has to be smaller. But dl is fixed, so the new angle $d\theta'$ has to be bigger, as easily seen on the figure for the purple lines. So, in a positively curved universe, objects are seen under a larger angle. Conversely, in a negatively curved universe, they are seen under a smaller angle. **Important remark: here, the past line cone has be drawn as a convex cone. Instead, for realistic cosmological scenarios, the cone is concave.**

If we know the the curvature sign k and the function $H(z)$ up to z_e, we can compute d_A as a function of z_e. The function $d_A(z_e)$ is called the "angular diameter distance – redshift relation".

A generic consequence is that in the Friedmann universe, for an object of fixed size and redshift, the angular diameter depends on the spatial curvature - as illustrated graphically in figure 2.4. Therefore, if we know in advance the physical size of an object, we can measure on the one hand its angular diameter, on the other hand its redshift z_e, and then look for cosmological models predicting the correct value for $d_A(z_e)$.

2.2.6 Luminosity distance – redshift relation

In absence of expansion and curvature, d_L would simply correspond to the Euclidean distance to the source. On the other hand, in general relativity, it is easy to understand that the apparent luminosity is given by

$$l = \frac{L}{4\pi\, a^2(t_0)\, r_e^2 (1+z_e)^2} \tag{2.39}$$

leading to

$$d_L = a(t_0)\, r_e (1+z_e)\ . \tag{2.40}$$

Let us explain this result. First, the reason for the presence of the factor $[4\pi\, a^2(t_0)\, r_e^2]$ in equation (2.39) is obvious. The photons emitted at a comoving coordinate r_e are distributed today on a sphere of comoving radius r_e surrounding the source. Following the expression for infinitesimal distances (2.10), the physical surface of this sphere is obtained by integrating over the infinitesimal surface element $dS^2 = a^2(t_0)\, r_e^2 \sin\theta\, d\theta\, d\phi$, which gives precisely $4\pi\, a^2(t_0)\, r_e^2$. In addition, we should keep in mind that L is a flux (i.e., an energy by unit of time) and l a flux density (energy per unit of time and surface). But the energy carried by each photon is inversely proportional to its physical wavelength, and therefore to $a(t)$. This implies that the energy of each photon has been divided by $(1+z_e)$ between the time of emission and now, and explains one of the two factors $(1+z_e)$ in (2.39). The other factor comes from the change in the rate at which photons are emitted and received (we have already seen in section 2.2.2 that since λ scales like $(1+z_e)$, both the energy and the frequence scale like $(1+z_e)^{-1}$).

We see that the luminosity distance is not indepent from the angular distance:

$$d_L = a(t_0)\ r_e\ (1+z_e) = a(t_e)\ r_e\ (1+z_e)^2 = (1+z_e)^2 d_A\ . \tag{2.41}$$

Like d_A, the luminosity distance can be written formally as a function of z_e:

$$d_L = a(t_0)\ (1+z_e)\ f_k\left(\int_0^{z_e} \frac{c\, dz}{a(t_0) H(z)}\right)\ . \tag{2.42}$$

Again, we would need to know the function $H(z)$ and the value of k in order to calculate explicitly the luminosity distance – redshift relation $d_L(z_e)$. In the limit $z \longrightarrow 0$, the three definition of distances given in the past sections (namely: $a(t_0)\chi$, d_A and d_L) are all equivalent and reduce to the usual definition of distance d in Euclidean space, related to the redshift through $d = z(c/H_0)$. Hence, the measurement of $d_A(z)$ and $d_L(z)$ at small redshift does not bring new information with respect to a Hubble diagram (i.e., it only allows to measure one number H_0), while measurements at high redshift depend on the spatial curvature and the dynamics of expansion. We will see in the next chapter that $d_L(z)$ has been measured for many supernovae of type Ia till roughly $z \sim 2$, leading to one of the most intriguing discovery of the past years.

In summary of this section, according to General Relativity, the homogeneous universe is curved by its own matter content, and the space–time curvature can be described by one number plus one function: the comoving spatial curvature k, and the scale factor $a(t)$. We should now be able to relate these two quantities with the source of curvature: the matter density.

2.3 The Friedmann law

In the rest of this course, we will use units such that $c = \hbar = k_B = 1$ for simplicity.

2.3.1 Einstein's equation

The relationship between the properties of matter in one point and those of curvature in the same point is given by the Einstein equation

$$G_{\mu\nu} = 8\pi G\, T_{\mu\nu} \ . \tag{2.43}$$

The Einstein tensor $G_{\mu\nu}$ can be computed for the FLRW metric using Christoffel's symbols. It is found to be diagonal ($G_{0i} = G_{i\neq j} = 0$) and isotropic ($G_{11} = G_{22} = G_{33}$). In fact, only diagonal and isotropic Einstein and energy-momentum tensors are compatible with the assumption of a homogeneous, isotropic universe with a comoving coordinate system. The most general energy-momentum tensor in such an idealized universe must be in the form

$$T^\mu_\nu = \begin{pmatrix} -\rho & 0 & 0 & 0 \\ 0 & p & 0 & 0 \\ 0 & 0 & p & 0 \\ 0 & 0 & 0 & p \end{pmatrix} \tag{2.44}$$

where ρ and p stand for the energy density and pressure of the cosmological fluid. The first component of the Einstein equation reads

$$G_{00} = 3\left[\frac{k}{a^2} + \left(\frac{\dot a}{a}\right)^2\right] \ . \tag{2.45}$$

This expression is interesting to discuss. In units with $c = 1$, G_{00} appears with the dimension of an inverse squared distance, representing intuitively the curvature of the space-time manifold. Here, indeed, G_{00} is the sum of the inverse squared spatial curvature radius, $R_c(t) = \pm a/\sqrt{|k|}$, and of the inverse squared Hubble radius, $R_H(t) = a/\dot a$, with a multiplicative factor 3 (coming from the number of spatial dimensions). We see that the Hubble radius really plays the role of a curvature radius for space-time. We can write now the first Einstein equation $G_{00} = 8\pi G\, T_{00}$ in the FLRW universe,

$$3\left[\frac{k}{a^2} + \left(\frac{\dot a}{a}\right)^2\right] = 8\pi G\,\rho \ , \tag{2.46}$$

or equivalently,

$$H^2 = \left(\frac{\dot a}{a}\right)^2 = \frac{8\pi G}{3}\,\rho - \frac{k}{a^2} \ . \tag{2.47}$$

The above relation between the scale factor $a(t)$, the comoving spatial curvature k and the homogeneous energy density of the universe $\rho(t)$ is called the Friedmann law. Together with the propagation of light equation, this law is the key ingredient of the Friedmann-Lemaître model.

In special/general relativity, the total energy of a particle is the sum of its rest energy $E_0 = mc^2$ (i.e. $E_0 = m$ in units $c = 1$), plus its momentum energy. So, if we consider only non-relativistic particles like those forming galaxies, we can neglect the momentum energy and write $\rho = \rho_{\rm mass}$. Then, the Friedmann equation looks exactly like the Newtonian expansion law (1.9), excepted that the function $r(t)$ (representing previously the position of objects) is replaced

by the scale factor $a(t)$. Of course, the two equations look the same, but they are far from being equivalent. First, we have already seen in section 2.2.2 that although the distinction between the scale factor $a(t)$ and the classical position $r(t)$ is irrelevant at short distance, the difference of interpretation between the two is crucial at large distances – of order of the Hubble radius (in particular, in one case the existence of objects with $d > R_H$ and $z > 1$ is violating the speed-of-light limit, in the other case it is not). Second, we have seen in section 1.3.3 that the term proportional to k seems to break the homogeneity of the universe in the Newtonian formalism, while in the Friedmann model, when it is correctly interpreted as the spatial curvature term, it is perfectly consistent with the Cosmological Principle.

The next crucial difference between the Friedmann law and the Newtonian expansion law is the possibility to account for a homogeneous, isotropic fluid of relativistic particles, as we shall see in the next subsection.

2.3.2 Energy conservation

The Einstein equation implies Bianchi identities of the form $G^{\nu}_{\mu;\nu} = T^{\nu}_{\mu;\nu} = 0$. The first Bianchi identity $T^{\nu}_{0;\nu} = 0$ is nothing but the energy conservation equation. In the FLRW universe it reduces to:

$$\dot{\rho} = -3\frac{\dot{a}}{a}(\rho + p) \ . \tag{2.48}$$

Hence, the relation between ρ and a (i.e. the way in which the energy gets diluted with the universe expansion) depends crucially on the pressure – or more precisely, on the equation of state $p(\rho)$. The most important limiting case in cosmology are:

- *non-relativistic matter.* In the limit of strongly non-relativistic matter, such as comobile objects, the negligible kinetic energy implies $p = 0$ (in absence of kinetic energy, a box enclosing the fluid would not feel any kind of pressure). If the comobile fluid represents a large-scale approximation for a homogeneous distribution of galaxies, then this approximation is fine. Hence:

  $$\dot{\rho} = -3\frac{\dot{a}}{a}\rho \qquad \Rightarrow \qquad \rho \propto a^{-3}. \tag{2.49}$$

  This result is obvious. For objects with negligible velocities, the energy density is equal to the mass density, which is conserved inside any given comoving volume, since the number of comobile objects in a comoving volume is by definition constant. Since a comoving volume V increases like $V \propto a^3$ in physical units, ρ decreases like a^{-3}.

- *ultra-relativistic matter.* In the limit of ultra-relativistic matter, such as photons or massless neutrinos, the particle velocity $v = c$ generates pressure. We know from statistical thermodynamics that an ultra-relativistic gas has an equation of state $p = \rho/3$. Hence:

  $$\dot{\rho} = -3\frac{\dot{a}}{a}(1 + \frac{1}{3})\rho = -4\frac{\dot{a}}{a}\rho \qquad \Rightarrow \qquad \rho \propto a^{-4}. \tag{2.50}$$

  We conclude that an ultra-relativistic fluid dilutes *faster* than a non-relativistic medium with the universe expansion. This can be understood in the following way. A homogeneous, ultra-relativistic fluid can be thought to be a gas of fast moving particles, each with $v = c$, either free-streaming or interacting with Brownian motions, such that at any time the density of particles is the same everywhere in the universe.

The cosmological fluid invoked in the FLRW model could include such a component. In this case, at a given time, a comoving volume V contains N ultra-relativistic particles of individual energy $E = \nu = 1/\lambda$ (still in units with $c = \hbar = 1$). As time passes by, V increases like a^3, N is fixed (the particles move in and out of the volume, but the number of particles remains constant, otherwise the assumption of homogeneity would be violated, since V would become an over dense or underdense region). Finally, E scales like a^{-1}. Hence the energy density in the volume scales like $\rho \propto E/V \propto a^{-4}$.

In the jargon of cosmology, the ultra-relativistic component of the cosmological fluid is usually called "radiation", while the word "matter" is reserved to the non-relativistic one. The Friedmann equation is true for any types of matter, relativistic or non-relativistic; if there are different species, the total energy density ρ is the sum over the density of all species.

2.3.3 Cosmological constant

When Einstein introduced its theory, he noticed that a simple geometrical term can be added to the left-hand side without violating any principle:

$$G_{\mu\nu} + \Lambda g_{\mu\nu} = 8\pi G\, T_{\mu\nu} \ . \tag{2.51}$$

The number Λ (which has the dimension of an inverse squared time, as can be seen when c is restored) should depend neither on space, nor on time. It is called the cosmological constant. At some point Einstein proposed that Λ could be non-zero and negative in order to allow for a static solution to the universe equations. Then he stepped back. Anyway, since we can write the above equation as

$$G_{\mu\nu} = 8\pi G\, T_{\mu\nu} - \Lambda g_{\mu\nu} \ , \tag{2.52}$$

or

$$G^\mu_\nu = 8\pi G\, T^\mu_\nu - \Lambda g^\mu_\nu \ , \tag{2.53}$$

with $g^\mu_\nu = g^{\mu\alpha} g_{\alpha\nu} = \delta^\mu_\nu$, we see that the cosmological constant above is rigorously equivalent to a homogeneous fluid with energy-momentum tensor

$$\tilde{T}^\mu_\nu = -\frac{\Lambda}{8\pi G}\delta^\mu_\nu = \begin{pmatrix} -\frac{\Lambda}{8\pi G} & 0 & 0 & 0 \\ 0 & -\frac{\Lambda}{8\pi G} & 0 & 0 \\ 0 & 0 & -\frac{\Lambda}{8\pi G} & 0 \\ 0 & 0 & 0 & -\frac{\Lambda}{8\pi G} \end{pmatrix} . \tag{2.54}$$

By comparison with eq. (2.44), we find that this fluid has $\rho = -p = \Lambda/8\pi G$. Looking at eq. (2.48), we see that the equation of state $p = -\rho$ implies $\dot{\rho} = 0$, consistently with the fact that Λ should not vary with time.

A priori, a cosmological constant could be present in the universe, either as a purely geometrical term (as in the Einstein proposal) or as some form of energy never being diluted. The vacuum energy which appears in quantum field theory (in particular, during a phase transition such as a spontaneous symmetry breaking) is of this kind: it does not dilute, and as long as the fundamental state of the theory is invariant, it remains indistinguishable from a cosmological constant. We will see that this term is probably playing an important role in our universe.

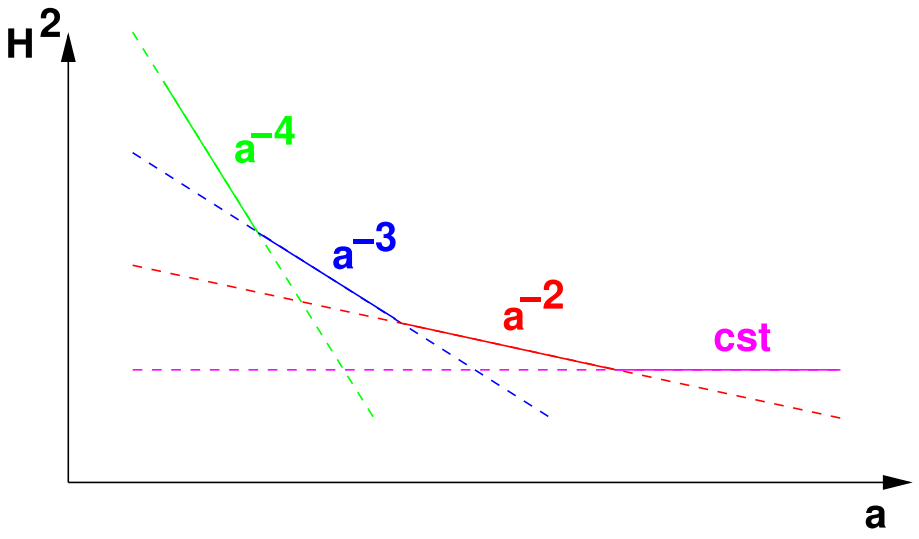


Figure 2.5: Evolution of the square of the Hubble parameter, in a scenario in which all typical contributions to the universe expansion (radiation, matter, curvature, cosmological constant) dominate one after each other.

2.3.4 Various possible scenarios for the history of the universe

Let us write the Friedmann law including all possible contributions to the homogeneous cosmological fluid mentioned so far:

$$H^2 = \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho_{\rm R} + \frac{8\pi G}{3}\rho_{\rm M} - \frac{k}{a^2} + \frac{\Lambda}{3} \tag{2.55}$$

where $\rho_{\rm R}$ is the radiation density and $\rho_{\rm M}$ the matter density. The order in which we wrote the four terms on the right–hand side – radiation, matter, spatial curvature, cosmological constant – is not arbitrary. Indeed, they evolve with respect to the scale factor as a^{-4}, a^{-3}, a^{-2} and a^0. So, if the scale factors keeps growing, and if these four terms are present in the universe, there is a chance that they all dominate the expansion of the universe one after each other (see figure 2.5). Of course, it is also possible that some of these terms do not exist at all, or are simply negligible. For instance, some possible scenarios would be:

- only matter domination, from the initial singularity until today (we'll come back to the notion of Big Bang later).
- radiation domination $\rightarrow$ matter domination today.
- radiation dom. $\rightarrow$ matter dom. $\rightarrow$ curvature dom. today
- radiation dom. $\rightarrow$ matter dom. $\rightarrow$ cosmological constant dom. today

But all the cases that do not respect the order (like for instance: curvature domination $\rightarrow$ matter domination) are impossible.

During each stage, if we assume that one component strongly dominates the others, the behavior of the scale factor, Hubble parameter and Hubble radius are given by:

1. Radiation domination:

$$\frac{\dot{a}^2}{a^2} \propto a^{-4}, \qquad a(t) \propto t^{1/2}, \qquad H(t) = \frac{1}{2t}, \qquad R_H(t) = 2t. \tag{2.56}$$

So, the universe is in decelerated power–law expansion.

2. Matter domination:

$$\frac{\dot{a}^2}{a^2} \propto a^{-3}, \qquad a(t) \propto t^{2/3}, \qquad H(t) = \frac{2}{3t}, \qquad R_H(t) = \frac{3}{2}t. \tag{2.57}$$

Again, the universe is in power–law expansion, but it decelerates more slowly than during radiation domination.

3. Negative curvature domination ($k < 0$):

$$\frac{\dot{a}^2}{a^2} \propto a^{-2}, \qquad a(t) \propto t, \qquad H(t) = \frac{1}{t}, \qquad R_H(t) = t. \tag{2.58}$$

A negatively curved universe dominated by its curvature is in linear expansion.

4. Positive curvature domination: if $k > 0$, and if there is no cosmological constant, the right–hand side finally goes to zero: expansion stops. After, the scale factor starts to decrease. H is negative, but the right–hand side of the Friedmann equation remains positive. The universe recollapses. We know that we are not in such a phase, because we observe the universe expansion. But *a priori*, we might be living in a positively curved universe, slightly before the expansion stops.

5. Cosmological constant domination:

$$\frac{\dot{a}^2}{a^2} \to \text{constant}, \qquad a(t) \propto \exp(\Lambda t/3), \qquad H = 1/R_H = \sqrt{\Lambda/3}. \tag{2.59}$$

The universe ends up in exponentially accelerated expansion.

So, in all cases, there seems to be a time in the past at which the scale factor goes to zero, called the initial singularity or the "Big Bang". The Friedmann description of the universe is not supposed to hold until $a(t) = 0$. At some time, when the density reaches a critical value called the Planck density, we believe that gravity has to be described by a quantum theory, where the classical notion of time and space disappears. Some proposals for such theories exist, mainly in the framework of "string theories". Sometimes, string theorists try to address the initial singularity problem, and to build various scenarios for the origin of the universe. Anyway, this field is still very speculative, and of course, our understanding of the origin of the universe will always break down at some point. A reasonable goal is just to go back as far as possible, on the basis of testable theories.

The future evolution of the universe heavily depends on the existence of a cosmological constant. If the latter is exactly zero, then the future evolution is dictated by the curvature (if $k > 0$, the universe will end up with a "Big Crunch", where quantum gravity will show up again, and if $k \leq 0$ there will be eternal decelerated expansion). If instead there is a positive cosmological term which never decays into matter or radiation, then the universe necessarily ends up in eternal accelerated expansion.

2.3.5 Cosmological parameters

In order to know the past and future evolution of the universe, it would be enough to measure the present density of radiation, matter and Λ, and also to measure H_0. Then, thanks to the Friedmann equation, it would be possible to

extrapolate $a(t)$ at any time[4]. Let us express this idea mathematically. We take the Friedmann equation, evaluated today, and divide it by H_0^2:

$$1 = \frac{8\pi G}{3H_0^2}\left(\rho_{\rm R0} + \rho_{\rm M0}\right) - \frac{k}{a_0^2 H_0^2} + \frac{\Lambda}{3H_0^2}. \tag{2.60}$$

where the subscript 0 means "evaluated today". Since by construction, the sum of these four terms is one, they represent the relative contributions to the present universe expansion. These terms are usually written

$$\Omega_{\rm R} = \frac{8\pi G}{3H_0^2}\rho_{\rm R0}, \tag{2.61}$$

$$\Omega_{\rm M} = \frac{8\pi G}{3H_0^2}\rho_{\rm M0}, \tag{2.62}$$

$$\Omega_k = -\frac{k}{a_0^2 H_0^2}, \tag{2.63}$$

$$\Omega_\Lambda = \frac{\Lambda}{3H_0^2}, \tag{2.64}$$

$$\tag{2.65}$$

and the "matter budget" equation is

$$\Omega_{\rm R} + \Omega_{\rm M} + \Omega_k + \Omega_\Lambda = 1. \tag{2.66}$$

The universe is flat provided that

$$\Omega_0 \equiv \Omega_{\rm R} + \Omega_{\rm M} + \Omega_\Lambda \tag{2.67}$$

is equal to one. In that case, as we already know, the total density of matter, radiation and Λ is equal at any time to the critical density

$$\rho_c(t) = \frac{3H^2(t)}{8\pi G}. \tag{2.68}$$

Note that the parameters Ω_x, where $x \in \{R, M, \Lambda\}$, could have been defined as the present density of each species divided by the present critical density:

$$\Omega_x = \frac{\rho_{x0}}{\rho_{c0}}. \tag{2.69}$$

The physical density today ρ_{x0} of a component can be expressed in standard units, e.g. g.cm^{-3}. Another alternative is to decompose it as:

$$\rho_{x0} = \Omega_x \frac{3H_0^2}{8\pi G} = \Omega_x h^2 \frac{3(100\,{\rm km.s^{-1}.Mpc^{-1}})^2}{8\pi G} \tag{2.70}$$

$$= \Omega_x h^2 \times 1.8788 \times 10^{-29} {\rm g.cm^{-3}}\ . \tag{2.71}$$

Hence, the physical density can be parametrized with the dimensionless number $\Omega_x h^2$. Later we will adopt the notation $\omega_x \equiv \Omega_x h^2$.

So far, we conclude that the evolution of the Friedmann universe can be described entirely in terms of four parameters, called the "cosmological parameters":

$$\Omega_{\rm R}, \Omega_{\rm M}, \Omega_\Lambda, H_0. \tag{2.72}$$

One of the main purposes of observational cosmology is to measure the value of these cosmological parameters.

[4]At least, this is true under the simplifying assumption that one component of one type does not decay into a component of another type: such decay processes actually take place in the early universe, and could possibly take place in the future.

Chapter 3

The Hot Big Bang cosmological model

3.1 Historical overview

Curiously, after the discovery of the Hubble expansion and of the Friedmann law, there were no significant progresses in cosmology for a few decades. The most likely explanation is that most physicists were not considering seriously the possibility of studying the universe in the far past, near the initial singularity, because they thought that it would always be impossible to test any cosmological model experimentally.

Nevertheless, a few pioneers tried to think about the origin of the universe. At the beginning, for simplicity, they assumed that the expansion of the universe was always dominated by a single component, the one forming galaxies, i.e., pressureless matter. Since going back in time, the density of matter increases as a^{-3}, matter had to be very dense at early times. This was formulated as the "Cold Big Bang" scenario.

According to Cold Big Bang, in the early universe, the density was so high that matter had to consist in a gas of nucleons and electrons. Then, when the density fell below a critical value, some nuclear reactions formed the first nuclei - this era was called Nucleosynthesis. But later, due to the expansion, the dilution of matter was such that nuclear reactions were suppressed (in general, the expansion freezes out all processes whose characteristic time–scale becomes smaller than the so–called Hubble time–scale H^{-1}). So, only a given number of nuclei had time to form, in some proportions which remained frozen afterward. After Nucleosynthesis, matter consisted in a gas of nuclei and electrons, with electromagnetic interactions. When the density became even smaller, they finally combined into atoms – this second transition is called recombination. At late time, any small density inhomogeneity in the gas of atoms was enhanced by gravitational interactions. The atoms started to accumulate into clumps like stars and planets - but this is a different story.

In the middle of the XX-th century, a few particle physicists tried to build the first models of Nucleosynthesis – the era of nuclei formation. In particular, four groups – each of them not being aware of the work of the others – reached approximately the same negative conclusion: in the Cold Big Bang scenario, Nucleosynthesis does not work properly, because the formation of hydrogen is strongly suppressed with respect to that of heavier elements. But this conclusion is at odds with observations: using spectrometry, astronomers know that there is a lot of hydrogen in stars and clouds of gas. The groups of the Russo-American Gamow in the 1940's, of the Russian Zel'dovitch (1964), of the British

Hoyle and Taylor (1964), and of Peebles in Princeton (1965) all reached this conclusion. They also proposed a possible way to reconcile Nucleosynthesis with observations. If one assumes that during Nucleosynthesis, the dominant energy density is that of photons, the expansion is driven by $\rho_{\rm R} \propto a^{-4}$, and the rate of expansion is different. This affects the kinematics of the nuclear reactions in such way that enough hydrogen can remain.

In that case, the universe would be described by a Hot Big Bang scenario, in which the radiation density dominated at early time. Before Nucleosynthesis and recombination, the mean free path of the photons was very small, because they were continuously interacting – first, with electrons and nucleons, and then, with electrons and nuclei. So, their motion could be compared with the Brownian motion in a gas of particles: they formed what is called a "black–body". In any black–body, the many interactions maintain the photons in thermal equilibrium, and their spectrum (i.e., the number density of photons as a function of wavelength) obeys to a law found by Planck in the 1890's. Any "Planck spectrum" is associated with a given temperature.

Following the Hot Big Bang scenario, after recombination, the photons did not see any more charged electrons and nuclei, but only neutral atoms. So, they stopped interacting significantly with matter. Their mean free path became infinite, and they simply traveled along geodesics – excepted a very small fraction of them which interacted accidentally with atoms, but since matter got diluted, this phenomenon remained subdominant. So, essentially, the photons traveled freely from recombination until now, keeping the same energy spectrum as they had before, i.e., a Planck spectrum, but with a temperature that decreased with the expansion. This is an effect of General Relativity: the wavelength of an individual photon is proportional to the scale factor; so the shape of the Planck spectrum is conserved, but the whole spectrum is shifted in wavelength. The temperature of a black–body is related to the energy of an average photon with average wavelength: $T \sim < E > \sim \hbar c / < \lambda >$. So, the temperature decreases like $1/ < \lambda >$, i.e., like $a^{-1}(t)$.

The physicists that we mentioned above noticed that these photons could still be observable today, in the form of a homogeneous background radiation with a Planck spectrum. Following their calculations – based on Nucleosynthesis – the present temperature of this cosmological black–body had to be around a few Kelvin degrees. This would correspond to typical wavelengths of the order of one millimeter, like microwaves.

These ideas concerning the Hot Big Bang scenario remained completely unknown, excepted from a small number of theorists.

In 1964, two American radio–astronomers, A. Penzias and R. Wilson, decided to use a radio antenna of unprecedented sensitivity – built initially for telecommunications – in order to make some radio observations of the Milky Way. They discovered a background signal, of equal intensity in all directions, that they attributed to instrumental noise. However, all their attempts to eliminate this noise failed.

By chance, it happened that Penzias phoned to a friend at MIT, Bernard Burke, for some unrelated reason. Luckily, Burke asked about the progresses of the experiment. But Burke had recently spoken with one of his colleagues, Ken Turner, who was just back from a visit in Princeton, during which he had followed a seminar by Peebles about Nucleosynthesis and possible relic radiation. Through this series of coincidences, Burke could put Penzias in contact with the Princeton group. After various checks, it became clear that Penzias and Wilson had made the first measurement of a homogeneous radiation with a Planck spectrum and a temperature close to 3 Kelvins: the Cosmic Microwave Background (CMB). Today, the CMB temperature has been measured with

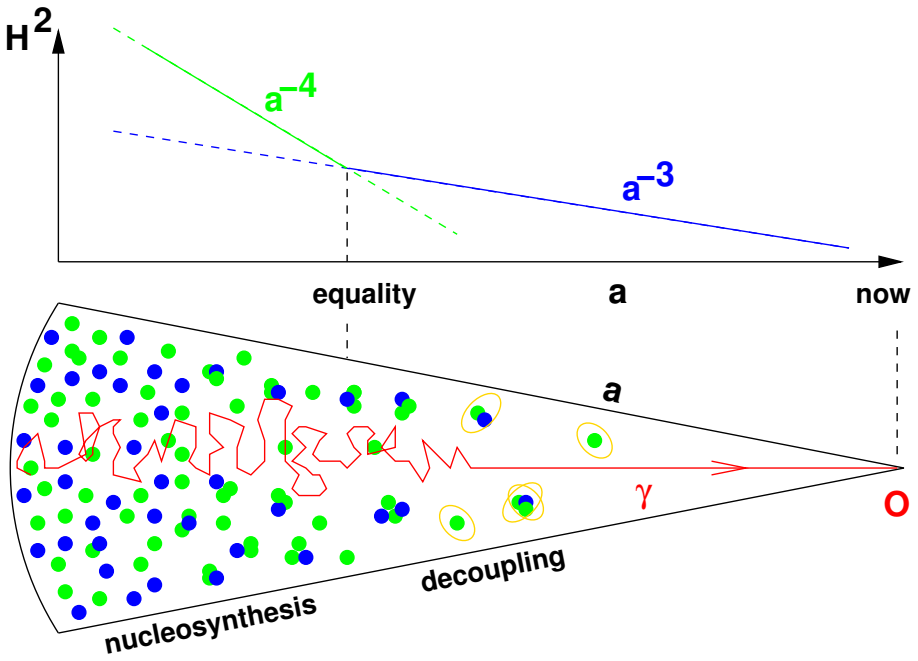


Figure 3.1: On the top, evolution of the square of the Hubble parameter as a function of the scale factor in the Hot Big Bang scenario. We see the two stages of radiation and matter domination. On the bottom, an idealization of a typical photon trajectory. Before decoupling, the mean free path is very small due to the many interactions with baryons and electrons. After decoupling, the universe becomes transparent, and the photon travels in straight line, indifferent to the surrounding distribution of electrically neutral matter.

great precision: $T_0 = 2.726$ K.

This fantastic observation was a very strong evidence in favor of the Hot Big Bang scenario. It was also the first time that a cosmological model was checked experimentally. So, after this discovery, more and more physicists realized that reconstructing the detailed history of the universe was not purely science fiction, and started to work in the field.

The CMB can be seen in our everyday life: fortunately, it is not as powerful as a microwave oven, but when we look at the background noise on the screen of a TV set, one fourth of the power comes from the CMB!

3.2 Relativistic quantum thermodynamics in the FLRW universe

We recall that we are using units such that $c = \hbar = k_B = 1$.

3.2.1 Momentum

Individual free-falling particles in the FLRW universe follow trajectories in space-time parametrised by a function $x^\mu(\lambda)$, where λ is a parameter monotonically increasing along the trajectory, and $x^\mu(\lambda)$ satisfies the geodesic equation:

$$\frac{d^2x^\alpha}{d\lambda^2} + \Gamma^\alpha_{\mu\nu}\frac{dx^\mu}{d\lambda}\frac{dx^\nu}{d\lambda} = 0 \ . \tag{3.1}$$

The energy-momentum 4-vector P^μ is defined as $P^\mu = m\frac{dx^\mu}{d\lambda}$, where λ is normalised in such a way that the energy-momentum vector satisfies everywhere the on-shell condition:

$$g_{\mu\nu}P^\mu P^\nu = -m^2 \ , \tag{3.2}$$

where m is the particle mass (zero for a photon). The geodesic equation can be used to show that P^i has a non-trivial evolution with time in the FLRW universe. Indeed we can write

$$\frac{d}{d\lambda}\left\{\frac{dx^i}{d\lambda}\right\} + \Gamma^i_{\mu\nu}\frac{dx^\mu}{d\lambda}\frac{dx^\nu}{d\lambda} = 0 \ . \tag{3.3}$$

After multiplying by m^2, this becomes

$$m\frac{dP^i}{d\lambda} + \Gamma^i_{\mu\nu}P^\mu P^\nu = 0 \ . \tag{3.4}$$

We use the fact that with the FLRW metric, $\Gamma^i_{00} = \Gamma^i_{jk} = 0$, while $\Gamma^i_{0j} = \Gamma^i_{j0} = \frac{\dot{a}}{a}$, where $\dot{a} = \frac{da}{dt}$ and t is proper time. So

$$m\frac{dP^i}{d\lambda} + 2\frac{\dot{a}}{a}P^0P^i = 0 \ . \tag{3.5}$$

We can write this relation as

$$m\frac{dt}{d\lambda}\frac{dP^i}{dt} + 2\frac{\dot{a}}{a}P^0P^i = 0 \tag{3.6}$$

However $m\frac{dt}{d\lambda} = m\frac{dx^0}{d\lambda} = P^0$. So P^0 can be simplified for the equation, and we are left with

$$\frac{dP^i}{dt} + 2\frac{\dot{a}}{a}P^i = 0 \ . \tag{3.7}$$

Finally, this can be written as

$$\frac{dP^i}{P^i} = -2\frac{da}{a} \ , \tag{3.8}$$

which shows that P^i scales like a^{-2} when the universe expands.

In special relativity, we would interpret P^0 as the energy and P^i as the physical momentum of the particle. In the FLRW universe, the latter is not correct, because P^i has been defined with respect to the comoving coordinates x^i rather than physical distances. In reality, one could show that the physical energy and physical momentum measured by comoving observers are

$$E = P^0 \ , \qquad p^i = aP^i \ . \tag{3.9}$$

We then learn that in the FLRW universe:

- The physical momentum scales like a^{-1}: in particular, its modulus $p = \sqrt{\delta_{ij}p^ip^j}$ scales like a^{-1}, or in other words (ap) is constant.
- The on-shell condition gives:

$$-(P^0)^2 + a^2\delta_{ij}P^iP^j = -m^2 \ , \tag{3.10}$$

  i.e.

$$-E^2 + p^2 = -m^2 \ . \tag{3.11}$$

  We conclude that the energy reads $E = \sqrt{m^2 + p^2}$, and is either constant for non-relativistic particles ($m^2 \gg p^2$), or redshifted like a^{-1} for ultra-relativistic particles ($m^2 \ll p^2$): this is a very intuitive result already used in the previous chapter, in section 2.3.2.

3.2.2 Phase-space distribution

Let us assume that the cosmological fluids is formed of many different species X (which can be either interacting with other species or free-streaming), each described by a phase-space distribution function $f_X(x^\mu, P^\nu)$. The number of arguments can be reduced in the FLRW universe by noticing that:

- Homogeneity implies that f_X should be the same everywhere and should not depend on x^i.
- Isotropy implies that f_X should not depend on the direction of the spatial part of the energy-momentum vector, P^i, or equivalently, it should not depend on the direction of the physical momentum p^i. However it could depend on the modulus p.
- P^0 is not an additional independent argument, since $P^0 = E = \sqrt{m_X^2 + p^2}$.

Hence the phase-space distribution can be written as a function of time and p only[1]: $f_X(t,p)$. The number density, energy density and pressure of each species read:

$$n_X(t) = \frac{g_X}{(2\pi)^3}\int d^3p\ f_X(t,p)\ , \tag{3.12}$$

$$\rho_X(t) = \frac{g_X}{(2\pi)^3}\int d^3p\ E_X\ f_X(t,p)\ , \tag{3.13}$$

$$P_X(t) = \frac{g_X}{(2\pi)^3}\int d^3p\ \frac{p^2}{3E}\ f_X(t,p)\ , \tag{3.14}$$

where g_X is the number of quantum degrees of freedom (spin or helicity states) of the considered species (e.g. $g_X = 2$ for photons γ, electrons e^-, positrons e^+, protons p, anti-protons $\bar{p}$, neutrons n, anti-neutrons $\bar{n}$, or $g_X = 1$ for neutrinos ν_i and anti-neutrons $\bar{\nu}_X$ where X is one of e, μ or τ).

Interactions can be represented by a set of reactions $1+2 \leftrightarrow 3+4$ (for elastic scattering, $1 = 3$ and $2 = 4$). In general the evolution of each species due to the above reaction is represented by a Boltzmann equation of the type:

$$\frac{df_X}{dt} = F[f_1, f_2, f_3, f_4] \tag{3.15}$$

where the right-hand side, which is quite complicated to write in the general case, is a function of the distribution of each species involved in the reaction.

3.2.3 Kinetic (or thermal) equilibrium

If two species X and Y have frequent interactions (like elastic scattering $X + Y \longrightarrow X + Y$), they exchange momentum in a random way and reach a kinetic equilibrium called "thermal equilibrium". Many species can be in thermal equilibrium, forming a so-called "thermal bath" or "thermal plasma". In thermal equilibrium, the distributions of each species depend on a common parameter, the temperature T. However the distributions f_X are not all equal to each other. They depend on:

- the mass m_X of each species (the mass appears in the energy of each particle, $E_X = \sqrt{m_X^2 + p^2}$) ;

[1] equivalently, we could choose to write it as a function of time and E only: this is just a matter of convention, and in some books you would often find $f_X(t,E)$ instead of $f_X(t,p)$.

- an additional parameter μ_X, the "chemical potential" of the species, which encodes the effect of the balance between the many reactions (inelastic scatterings) involved in the plasma;
- at the quantum level, the fact that each species should obey to the Bose-Einstein statistics for bosons (e.g. photons), or to the Fermi-Dirac statistics for fermions (in this chapter, apart from photons, we will only consider fermions).

Hence, a plasma of N species in thermal equilibrium with known masses m_X and known statistics (fermion or boson) can be entirely described in terms of a maximum of $N+1$ free parameter $(T, \mu_1, ...\mu_N)$, which values can be inferred from considerations e.g. on energy conservation, quantum number conservation, and on the the kinetic of the various reactions involved. Thermal distributions read

$$f_X = \begin{cases} \frac{1}{\exp[\frac{E_X - \mu_X}{T}]+1} & \text{(Fermi-Dirac)} , \\ \\ \frac{1}{\exp[\frac{E_X - \mu_X}{T}]-1} & \text{(Bose-Einstein)} . \end{cases} \tag{3.16}$$

The probability of interaction between individual particles depends on a cross-section σ and on their relative velocity v. In thermal equilibrium, the interaction between two species X and Y is characterized by a "thermally averaged cross-section – velocity product" $\langle \sigma v \rangle$. The interaction rate (or scattering rate) of X is given by $\Gamma_X = n_Y \langle \sigma v \rangle$, that of Y by $\Gamma_Y = n_X \langle \sigma v \rangle$. A detailed study would show that the scattering is efficient enough for maintaining X in thermal equilibrium with Y provided that the scattering rate Γ_X is larger than the inverse of the characteristic time set by the universe expansion: namely, $\Gamma_X > H$. Intuitively, when $\Gamma_X < H$, the cross-section is so low or the species Y is so diluted that the chance for X to scatter over Y within a time comparable to the age of the universe becomes negligible. When all possible scattering reactions which could maintain X in thermal equilibrium have $\Gamma_X < H$, the species X decouples from thermal equilibrium. In this case, assuming that the particles are stable and non-interacting, they can only free-stream with a frozen distribution (i.e., the distribution remains identical to the one at last scattering, apart from the effect of the universe expansion: $p \propto a$).

Let us review a few basic properties of thermal equilibrium which will be useful in the following sections.

- *Density of relativistic particles with negligible chemical potential.* Let us assume for simplicity that $|\mu_X| \ll T$. In this case,

$$f_X = \frac{1}{\exp[\sqrt{m_X^2 + p^2}/T] \pm 1} . \tag{3.17}$$

  From eq. (3.12), we see that in general the particles contributing mostly to the number density are those for which $p^2 f_X(p)$ is maximum. If $T \gg m_X$, the function $p^2 f_X(p)$ peaks at a value of p of the same order of magnitude as T, and hence for a huge majority of particles $p \gg m_X$. This corresponds to a gas of relativistic particles. The number density, energy density and pressure can be computed by integrating over the above distribution in the limit $m_X \longrightarrow 0$. The result is found to be:

$$n_X = \frac{\zeta(3)}{\pi^2} g_X T^3 \quad \left(\times \frac{3}{4} \quad \text{for fermions} \right) , \tag{3.18}$$

$$\rho_X = \frac{\pi^2}{30} g_X T^4 \quad \left(\times \frac{7}{8} \quad \text{for fermions} \right) , \tag{3.19}$$

$$P_X = \frac{1}{3} \rho_X , \tag{3.20}$$

where $\zeta(x)$ is the Riemann zeta function ($\zeta(3) \simeq 1.20206...$), and the extra factors for fermions come from the $+1$ term instead of -1 in the denominator of f_X. Note that the usual equation of state of a relativistic gas, $p = \sum_X p_X = \sum_X \rho_X/3 = \rho/3$, is recovered here. We conclude that boson and fermions in thermal equilibrium with each other and such that $m_X \ll T$ and $|\mu_X| \ll T$ share roughly the same number/energy density, apart from possible factors of order one.

- *Density of non-relativistic particles.* In the limit $m_X \gg T$, the function $p^2 f_X(p)$ peaks in between T and m_X, and most particles have $p \ll m_X$: hence this limits describes a gas of non-relativistic particles. Then, a detailed integration shows that for both fermions and bosons

$$n_X = g_X \left(\frac{m_X T}{2\pi}\right)^{3/2} \exp[-\frac{(m_X - \mu_X)}{T}] \ , \tag{3.21}$$

$$\rho_X = m_X n_X \ , \tag{3.22}$$

$$P_X = T n_X \ll \rho_X \ . \tag{3.23}$$

Let us compare the number density of these particles with that of relativistic ones still in thermal equilibrium with them:

$$\frac{n_X^{\rm NR}}{n_Y^{\rm R}} = e^{\frac{\mu_X}{T}} \left[\frac{g_X}{g_Y} \frac{\sqrt{\pi}}{2\sqrt{2}\zeta(3)}\right] \left(\frac{m_X}{T}\right)^{3/2} e^{-\frac{m_X}{T}} \ . \tag{3.24}$$

The factor between brackets is of order one. The part after the brackets is much smaller than one since we assumed $m_X \gg T$. Hence, unless the chemical potential is huge ($\mu_X \gg m_X \gg T$, a case that will never occur in the realistic situations considered later), the number density of non-relativistic species in thermal equilibrium is exponentially suppressed with respect to that of relativistic ones. The total number density in the thermal plasma is dominated by relativistic components.

3.2.4 Chemical equilibrium

Let's consider an inelastic scattering reaction of the type $1+2 \longleftrightarrow 3+4$. When this reaction is frequent enough, the relative number density of particles cannot be arbitrary, it must obey to the chemical equilibrium relation:

$$\mu_1 + \mu_2 = \mu_3 + \mu_4 \ . \tag{3.25}$$

When the reaction is not frequent, it is unable to maintain chemical equilibrium, and the kinetic of each particle production/annihilation must be followed using the Boltzmann equation. However, these particules can still be in thermal equilibrium (for instance, due to e.g. elastic scattering with photons). If all four species are still in thermal equilibrium, the Boltzmann equation describing e.g. the evolution of n_1 due to the above reaction takes a much simpler form than in the general case:

$$\dot{n}_1 + 3Hn_1 = n_1 n_2 \langle\sigma v\rangle \left[\exp\left(\frac{-\mu_1 - \mu_2 + \mu_3 + \mu_4}{T}\right) - 1\right] \ . \tag{3.26}$$

Here, we made two assumptions (apart for thermal equilibrium). First, we assumed that the cross section $\langle\sigma v\rangle$ is the same for the reactions $1+2 \longrightarrow 3+4$ and $3+4 \longrightarrow 1+2$. Otherwise, the right-hand side would split in two terms for creation and annihilation. However, for the realistic cases considered later, it is sufficient to consider a symmetric cross section. Second, we assumed that

$1 + 2 \longleftrightarrow 3 + 4$ is the only reaction leading to the creation or annihilation of type 1 particles. If there are other processes involved, the right-hand side should contain a sum over all possible creation and decay channels.

Note that the factor $n_2\langle\sigma v\rangle$ in the right-hand side is precisely the scattering rate Γ_1 for the scattering of type 1 particles. Hence, the second term on the left-hand side is of the order of Hn_1, while the right-hand side is of the order of $n_1\Gamma_1$ times the brackets. We see that if $\Gamma_1 \gg H$, the term involving H can be neglected; in this regime, the differential equation forces n_1 to reach an equilibrium value for which the brackets vanish, i.e. for which $\mu_1+\mu_2 = \mu_3+\mu_4$: chemical equilibirum will be maintained at any time. In the other limit, when $\Gamma_1 \ll H$, the right-hand side is negligible, and there is no reason for the relation $\mu_1 + \mu_2 = \mu_3 + \mu_4$ to be maintained; instead, $\dot{n}_1 = -3Hn_1$, which is equivalent to $n_1 \propto a^{-3}$: this simply corresponds to particle number conservation for a decoupled species. The intermediate regime can only be followed by integrating the above Boltzmann equation.

3.2.5 Conservation of quantum numbers

If the number of particles of a given type X was conserved in any comoving volume, we would have $n_X a^3$ =constant. This is usually *not* the case since in general, the particles X can be destroyed or created by various inelastic scatterings. So, conservation laws do not apply to the number density of individual particles, but to that of quantum numbers.

Let us consider for instance the conservation of electric charge. We can define n_+ as the sum over the number density of all particles with positive charge, weighted by the value of their charge; same for n_- (weighted by the absolute value of the charge so that $n_- > 0$). The total density of electric charge in the universe is then simply $n_Q \equiv n_+ - n_-$. Electric charge is a conserved number, so the charge in any comoving volume must be constant. Hence $n_Q a^3$ is constant. The same holds for other quantities such as baryon number ($n_B a^3$ =constant), lepton number ($n_L a^3$ =constant), etc. (except at very early times for which baryon or lepton number conservation can be violated in special circumstances, as we shall see later).

However, in the case of the electric charge, we have an even stronger constraint: since the electric charge is associated with Coulomb forces and the universe expansion is only governed by gravitational forces, the universe must be globally neutral: hence $n_Q = 0$ and $n_+ = n_-$.

Note that each conserved quantum number is *usually* associated with a non-zero chemical potential. When a particle X carries no conserved charge, nothing prevents reactions of the type $nX \to mX$ with $n \neq m$. This is the case for photons. For instance, as long as the universe contains electrons and positrons, the two reactions

$$3\gamma \longleftrightarrow e^+ + e^- \longleftrightarrow 2\gamma \tag{3.27}$$

are in chemical equilibrium, hence $2\mu_\gamma = 3\mu_\gamma$ and $\mu_\gamma = 0$. In addition, the above reactions tell us that electrons and positrons (which carry electric charges ± 1 and lepton numbers ± 1) have opposite chemical potentials, $\mu_{e^+} = -\mu_{e^-}$. It is not possible to find a reaction that would lead to the conclusion that $\mu_{e^+} = \mu_{e^-} = 0$ without violating charge or lepton number conservation. A species carrying a conserved charge can have a zero chemical potential, but only if we invoke external constraints on top of chemical equilibrium considerations.

3.2.6 Entropy conservation in the thermal bath

We just said that there is no reason for conserving the total number density of particles in a given comoving volume. However, it is possible to show that the total entropy (i.e. the number of possible states) in any comoving volume is conserved, and that the entropy density of a thermal plasma reads

$$s = \frac{\rho + P}{T} \tag{3.28}$$

where ρ and P are the total density and pressure of species in thermal equilibrium. The proofs of these results will be derived in the exercise sessions. Let us consider a thermal bath composed of a number of relativistic and non-relativistic species, and let us assume further that the density of non-relativistic particles is negligible with respect to that of relativistic ones (this assumption holds throughout the radiation dominated era in the early universe). The total density and pressure are then equal to

$$\rho_{\rm tot} = \frac{\pi^2}{30} g_* T^4 \ , \qquad P_{\rm tot} = \frac{1}{3}\rho_{\rm tot} \ , \tag{3.29}$$

where we have introduced the *effective number of bosonic relativistic degrees of freedom* g_* defined through

$$g_* = \sum_{\rm rel.bosons} g_X + \frac{7}{8} \sum_{\rm rel.fermions} g_X \ . \tag{3.30}$$

The entropy density is then

$$s = \frac{4}{3}\frac{\pi^2}{30} g_* T^3 \ , \tag{3.31}$$

and its conservation implies $g_* T^3 a^3$ =constant. We see that as long as g_* is constant, $T \propto a^{-1}$. However, when g_* varies (which can happen e.g. if one species becomes non-relativistic at some point), the temperature varies like $T \propto g_*^{-1/3} a^{-1}$.

Note that entropy conservation is really different from number density conservation. For instance, in the above example, the number density reads

$$n_{\rm tot} = \frac{\zeta(3)}{\pi^2} \left[\sum_{\rm rel.bosons} g_X + \frac{3}{4} \sum_{\rm rel.fermions} g_X \right] T^3 \ . \tag{3.32}$$

The term between brackets differs from g_* due to the factor 7/8. Hence, when g_* varies, the quantity $n_{\rm tot} a^3$ is *not* constant, since the entropy is *not* equivalent to the number density!

3.3 The Thermal history of the universe

3.3.1 Early stages

The earliest stages in the evolution of the universe are still partially unknown and subject to investigation, while the latest stages are very well modelled and constrained by observations. In summary, the epoch during which the energy scale $\rho_{\rm tot}^{1/4}$ of the universe was smaller than 100 MeV is rather well understood, while early stages are still quite uncertain. In this subsection, we will provide a very brief overview of what could have happened above 100 MeV. In the next

subsections, we will describe in more details the main events taking place below 100 MeV.

Following the most conventional picture, gravity became a classical theory (with well–defined time and space dimensions) at a time called the Planck time[2]: $t \sim 10^{-36}$s, $\rho \sim M_P^4 \sim (10^{18}\text{GeV})^4$ (where the Planck mass is defined by $M_P = G^{-1/2}$: the Friedmann equation can also be written as $3M_P^2H^2 = 8\pi\rho$, and the Planck time corresponds to $H = M_P$, i.e. to a Hubble radius equal to the Planck length $R_H = 1/M_P = \lambda_P$; all these relations are written as usual for $c = \hbar = k_B = 1$ units). Later, there was most probably a stage of accelerated expansion called *inflation*. Current observations provide some indirect, but precise information on inflation, which is quite extraordinary since this stage took place at extremely high energy. Inflation might be related to the spontaneous symmetry breaking of the GUT (Grand Unified Theory) symmetry around $t \sim 10^{-32}$s, $\rho \sim (10^{15}$ to $10^{16}\text{GeV})^4$. However, it could also take place at much lower energy. Besides, we are not even sure that Grand Unification ever occurred. We will describe the motivations and predictions of inflation in the last chapter.

After inflation, during a stage called reheating, the scalar field responsible for inflation decayed into the particles of the standard model (three families of quarks, anti-quarks, leptons and anti-leptons; Higgs boson(s); gauge bosons), and possibly also some particles belonging to extensions of the standard model, like maybe supersymmetric particles, although recent LHC results bring no evidence for such an extension, at least until now. It is likely that all these particles reached thermal equilibrium after some time. At such high energy, most (if not all) particles were ultra-relativistic ($T > m_X$), and the total energy and pressure were given by eq. (3.29). The end of reheating marks the beginning of the radiation dominated era assumed by Gamow, Peebles and others. Note that during this era, $T \propto a^{-1}$ and $\rho \propto a^{-4}$ in good approximation, although these scaling laws are slightly violated each time that g_* varies (this occurs from time to time e.g. when some particles become non-relativistic). Around $t \sim 10^{-6}$s, $\rho \sim (100\text{ GeV})^4$, the EW (Electro Weak) symmetry is spontaneously broken and the quarks acquire a mass through the Higgs mechanism. Later, at $t \sim 10^{-4}$s, $\rho \sim (100\text{ MeV})^4$, the QCD (Quantum Chromo Dynamics) transition forces quarks to get confined into hadrons: baryons and mesons.

All these stages are quite complicated and extremely interesting to investigate in details (here we will not address them). Let us mention that a particularly fascinating and important issue is the evolution of the baryon and lepton number.

Let us focus first on the baryon number. Before reheating, there is no baryon number. Hence, if the baryon number is always conserved, each time that a particle is created during reheating with a given baryon number, its antiparticle with opposite baryon number will also be created. The pairs of particle-antiparticles will not annihilate in the relativistic regime. For simplicity, let us do as if there was only one type of particle with a baryon number, say b with baryon number $B = 1$ and its antiparticle $\bar{b}$ with $B = -1$. These particles could in principle annihilate through e.g.

$$b + \bar{b} \leftrightarrow n\gamma \tag{3.33}$$

(n being the number of produced photons). Note that a particle and its antiparticle should share the same mass m_b. Intuitively, as long as $T \gg m_b$, the photons carry enough energy for creating pairs of b and $\bar{b}$, so they will coexist in the thermal plasma: annihilation and creation compensate each other. However,

[2]By convention, the origin of time is chosen by extrapolating the scale-factor to $a(0) = 0$. Of course, this is only a convention, it has no physical meaning.

when $T < m_b$, the photons do not carry enough energy for creating pairs, and only annihilation can occur: so, b and $\bar{b}$ annihilate. If we assume that the baryon number is always conserved, then the annihilation will be total and we will be left with no baryons at all today. This is not the case since the nuclei of atoms are made of protons and neutrons. Hence, the baryon number conservation has to be weakly violated at some point between reheating and $T \sim m_b$. When the violation occurs, an excess of particles with positive B can be created. This is called baryogenesis. When $T \sim m_b$, all baryons annihilate with antibaryons, excepted the few ones in excess, which remain around until today.

Let us give a very simplified mathematical description of this phenomenon: after baryogenesis, the universe contains relativistic baryons and anti-baryons in thermal and kinetic equilibrium. The reaction

$$b + \bar{b} \leftrightarrow n\gamma \tag{3.34}$$

with different possible values of n guaranties that $\mu_\gamma = 0$ and $\mu_b = -\mu_{\bar{b}}$. If $\mu_b = 0$, then n_b is exactly equal to $n_{\bar{b}}$. The outcome of baryogenesis should be a small excess of baryons, hence $\mu_b > 0$. The conserved baryon number $n_B a^3$ is non-zero and obtained from

$$n_B = n_b - n_{\bar{b}} = \frac{g_b}{(2\pi)^3} \int d^3p \left[\frac{1}{\exp(\frac{E-\mu_b}{T}) + 1} - \frac{1}{\exp(\frac{E+\mu_b}{T}) + 1} \right] . \tag{3.35}$$

In the relativistic limit $E = p$ this gives

$$n_B = \frac{g_b T^3}{6\pi^2} \left[\pi^2 \left(\frac{\mu_b}{T} \right) + \left(\frac{\mu_b}{T} \right)^3 \right] , \tag{3.36}$$

which is positive for $\mu_b > 0$. As long as Ta =constant (i.e. as long as g_* is constant in the thermal bath), the conservation of $n_B a^3$ implies that μ_b/T is also constant. The baryon asymmetry can be parametrized by

$$\frac{n_B}{n_b + n_{\bar{b}}} \simeq n_B / \left[2 \times \frac{3}{4} \frac{\zeta(3)}{\pi^2} g_b T^3 \right] \sim \left[\pi^2 \left(\frac{\mu_b}{T} \right) + \left(\frac{\mu_b}{T} \right)^3 \right] \tag{3.37}$$

but this is not a conserved number. Usually, the asymmetry is parameterized by n_B/s, which is really a conserved number since both the baryon number $n_B a^3$ and entropy sa^3 are conserved. We will see later that in order to obtain the correct baryon density today, we must assume that n_B/s is of the order of 10^{-10}.

Note that when the universe is filled with a thermal plasma, s is of the order of $g_* T^3$, while n_γ is of the order of $g_\gamma T^3$ with $g_\gamma = 2$. So, instead of n_B/s, we will often use the ratio n_B/n_γ, although strictly speaking the second number is not conserved and differs from the first one by a factor of the order of g_* (which can vary between ~ 3 and ~ 10 during the period that we will study in the next sections). In the recent universe we will see that

$$\eta_b \equiv \frac{n_B}{n_\gamma} \sim 5 \times 10^{-10} . \tag{3.38}$$

When $T \sim m_B$, the number density of both n_b and $n_{\bar{b}}$ drops down very quickly due to the $\exp(-m_b/T)$ factor. Intuitively, this means that a smaller and smaller fraction of photons have enough energy for producing $b+\bar{b}$ pairs. The assumption of thermal and kinetic equilibrium and the conservation of entropy and baryon number provide enough equations for following $\mu_b(T)$ and $T(a)$ until $n_{\bar{b}}$ becomes really negligible. We don't even need to do that: it is enough to know that

when $n_{\bar{b}} = 0$, baryon number conservation simply implies that $n_b a^3 = n_B a^3$ is constant. Note that at that time

$$n_b = g_b \left(\frac{m_b T}{2\pi} \right)^{3/2} e^{-\frac{(m_b - \mu_b)}{T}} , \tag{3.39}$$

so the quantity μ_b/T now varies with time, in order to maintain a constant $n_b a^3$.

This description of the matter-antimatter asymmetry in the early universe was quite simplistic with respect to reality. Actually, baryogenesis and baryon-antibaryon annihilation are two active topics of research. Baryogenesis could be associated with B-violating processes during GUT symmetry breaking or EW symmetry breaking, or could also be induced by leptogenesis, for which a similar discussion can hold. The baryon-antibaryon annihilation is expected to take place roughly around $T \sim 1000$ MeV, which is the order of magnitude of the proton mass; it is intimately related to the quark-hadron transition.

3.3.2 Content of the universe around $T \sim$10 MeV

In the next sections, we will describe a list of phenomena induced by the fact that the weak interactions become inefficient around 1 MeV, and also that the MeV is the order of magnitude of binding energies in light nuclei. Before these sections, we should look at initial conditions before $T \sim$MeV.

Let us list the species present after the quark-hadron transition. A species can be present at a given time if it satisfies one of two conditions:

- either it is relativistic: $m \ll T$. In this case the particle can be easily produced by other species in the thermal bath (annihilation and creation compensate each other).

- or it is stable thanks to the conservation of a quantum number. In this case, the particle may have $m \gg T$, but cannot decay with violating the conservation of this number. Typically, the particles in the category are the lightest ones carrying a given quantum number. For instance, the proton is the lightest baryon.

Generally speaking, hadrons consist of baryons, mesons and their antiparticles. Mesons carry zero baryon number and quickly annihilate. Antibaryons annihilate well before $T \sim$ 10MeV, as described above. Baryons made of heavy quarks are unstable at the temperature considered here since they can decay into lighter baryons (protons and neutrons). Protons are perfectly stable in the limit of no B violation since they are the lightest baryons. Neutrons can decay into protons through beta decay ($n \longrightarrow p + e^- + \bar{\nu}_e$) but it is possible to show that at the temperature considered here, the inverse process is still efficient (electrons and neutrons carry enough energy for converting a proton into a neutron: this only requires $m_n - m_p = 1.203$ MeV). So, around ∼10 MeV, both protons and neutrons are present. They are still maintained in thermal and kinetic equilibrium by weak and electromagnetic interactions. They are of course both non-relativistic since $m_n \sim m_p \sim$ GeV. They have approximately the same density $n_n = n_p$, as will be shown explicitly in the section on Nucleosynthesis.

In the lepton sector, μ, $\bar{\mu}$, τ and $\bar{\tau}$ are so heavy that they decay into electrons and positrons. The mass of electrons and positrons is close to 0.5 MeV, so they are still relativistic at that time. Electric neutrality implies $n_{e^-} - n_{e^+} = n_p$. Does this imply a large asymmetry for electrons versus positrons? Remember that n_B/s is conserved and of the order of 10^{-10}. At the temperature considered here, we can consider that $n_B = n_p + n_n \simeq 2n_p$ and that $s \sim n_\gamma \sim n_{e^-}$ modulo

factors of order at most ten. Hence, speaking only of orders of magnitude,

$$\frac{n_{e^-} - n_{e^+}}{n_{e^-} + n_{e^+}} \sim \frac{n_{e^-} - n_{e^+}}{s} \sim \frac{n_B}{s} \sim 10^{-10} \ . \tag{3.40}$$

We see that electric neutrality implies that the electron-positron asymmetry is as tiny as the initial baryon asymmetry.

Besides, the universe contains all six neutrinos: ν_e, ν_μ, ν_τ and their antiparticles, maintain in thermal and kinetic equilibrium by weak interactions. Their mass is at most of the order of eV, so they have no reason to annihilate, and they contribute to the thermal plasma as ultra-relativistic components. They could in principle carry some asymmetry associated to chemical potentials μ_e, μ_μ and μ_τ (each antineutrino would then have an opposite chemical potential due to the chemical equilibrium of the reactions $\nu_e + \bar{\nu}_e \longleftrightarrow e^- + e^+ \longleftrightarrow \gamma$). Due to the large mixing angles in the neutrino mass matrix, the three potentials should share a unique value at this epoch. This issue is still a topic of research, but since such an asymmetry is difficult to motivate and has not been observed so far, we will assume throughout this course that neutrino chemical potentials are null, and hence that at the time considered here all six neutrino species share exactly the same number density.

Finally, the universe should contain photons. All other particles are expected to have decayed by that time, excepted one or more stable "dark matter particle" that will be discussed in Chapter 4. In summary, around $T \sim 10$ MeV, the universe should contain: p, n, e^-, e^+, six neutrino species, γ and possibly dark matter particles. The latter, if they exist, are expect to be non-relativistic at that time. So the number of relativistic degrees of freedom is given by photons, electrons, positrons and six neutrinos:

$$g_*(\sim 10\text{MeV}) = 2 + \frac{7}{8}(2 + 2 + 6) = 10.75 \ . \tag{3.41}$$

3.3.3 Neutrino decoupling

Weak interactions maintain neutrinos in thermal equilibrium through elastic and inelastic interactions like e.g.

$$\begin{aligned} \nu_e + e^- &\longleftrightarrow \nu_e + e^- \qquad\qquad (3.42) \\ \nu_e + \bar{\nu}_e &\longleftrightarrow \nu_i + \bar{\nu}_i \qquad (i = \mu \ \text{ or } \ \tau) \\ \nu_e + \nu_i &\longleftrightarrow \nu_e + \nu_i \\ \nu_e + \bar{\nu}_i &\longleftrightarrow \nu_e + \bar{\nu}_i \\ &\text{etc.} \qquad\qquad (3.43) \end{aligned}$$

which are all of the weak interaction type (they involve exchanges of weak bosons Z^0, $W^\pm$). The thermally averaged cross sections of these reactions are of the order of $\langle \sigma v \rangle \sim G_F^2 T^2$, where $G_F \sim 10^{-5}\text{GeV}^{-2}$ is the Fermi constant (which characterizes the magnitude of weak interactions). Hence the relevant scattering rates are of the order of $\Gamma = n_{e^-} \langle \sigma v \rangle \sim G_F^2 T^5$. Let us compare the evolution of Γ with that of the Hubble rate $H^2 = (8\pi G/3)\rho \sim M_P^{-2} T^4$. We find that

$$\frac{\Gamma}{H} \sim M_P G_F^2 T^3 \sim \left(\frac{T}{1 \text{ MeV}} \right)^3 \ . \tag{3.44}$$

Hence, when the temperature of the plasma drops below $T \sim$ MeV, the neutrinos leave thermal equilibrium, and their distribution remains frozen, with

$$f_i(p) = \frac{1}{\exp[p/T_\nu] + 1} \ . \tag{3.45}$$

By "frozen", one means that f_i varies only due to the universe expansion, which imposes a very trivial evolution. Each decoupled particle is free-falling in the FLRW universe. The geodesic equation shows that for such particles $p \propto a^{-1}$ (we already used this result many times for photons). Hence each individual particle has a momentum redshifting like $p(t) = p(t_D)a(t_D)/a(t)$ where t_D is the time of decoupling. For particles which decoupled when they were relativistic (like the neutrinos considered in this section), the distribution $f_i(p)$ depends on p only through the ratio p/T_ν. So, saying that all momenta shift like a^{-1} is strictly equivalent to saying that T_ν shifts like a^{-1}. Hence, after neutrino decoupling and for each of the six species i, the product $(T_\nu a)$ remains *exactly* constant at all times. Besides, as long as they remain relativistic with $T_\nu \gg m_{\nu_i}$, they obey:

$$n_{\nu_i} = \frac{3}{4}\frac{\zeta(3)}{\pi^2} T_\nu^3 \propto a^{-3} , \tag{3.46}$$

$$\rho_{\nu_i} = \frac{7}{8}\frac{\pi^2}{30} g_i T_\nu^4 \propto a^{-4} , \tag{3.47}$$

$$p_{\nu_i} = \frac{1}{3}\rho_{\nu_i} . \tag{3.48}$$

Neutrino decoupling is a very smooth process because before decoupling (and as long as the number of relativistic degrees of freedom g_* was conserved), we already had $T = T_\nu \propto a^{-1}$, $n_{\nu_i} \propto a^{-3}$, $\rho_{\nu_i} \propto a^{-4}$ and $p_{\nu_i} = \rho_{\nu_i}/3$. Hence, from the point of view of the universe expansion, one could say that "nothing particular happens" when neutrinos decouple. The temperature of neutrinos and of the thermal bath remain equal, both scaling like a^{-1}. The entropy density before decoupling reads:

$$s = \left.\frac{\rho + p}{T}\right|_{\text{plasma}} = \frac{4}{3}\frac{\pi^2}{30} g_* T^3 \qquad \text{with} \quad g_* = 2 + \frac{7}{8}(2+2+6) = 10.75 . \tag{3.49}$$

After decoupling, the entropy receives contribution from the plasma and from neutrinos. We have not derived the expression of entropy for a decoupled relativistic species, but it is simple: it reads like the entropy of relativistic species in equilibrium, with the appropriate value of the temperature:

$$s = \left.\frac{\rho + p}{T}\right|_{\text{plasma}} + \left.\frac{\rho_\nu + p_\nu}{T_\nu}\right|_{\text{neutrinos}} \tag{3.50}$$

$$= \frac{4}{3}\frac{\pi^2}{30}\left(2 + \frac{7}{8}(2+2)\right) T^3 + \frac{4}{3}\frac{\pi^2}{30}\left(\frac{7}{8}\times 6\right) T_\nu^3 . \tag{3.51}$$

Since both T and T_ν scale like a^{-1} around the time of neutrino decoupling, they remain equal to each other, and the expression of the entropy is absolutely unchanged.

3.3.4 Positron annihilation

The electron and positron mass is close to 0.5 MeV. Hence, when the temperature of the plasma drops below this value, electrons and positron become gradually non-relativistic. This is the same situation as the one described previously for b and $\bar{b}$: the number density of e^- and e^+ drops down very quickly with respect to that of photons, due to the suppression factor $\exp[-m_e/T]$. Basically, this means that electrons and positrons annihilate each other without being regenerated, until positrons disappear completely; a small number of electrons survives, in equal proportion to protons in order to ensure electric neutrality. After this process, $n_{e^-} = n_p \sim n_B \sim 10^{-10} n_\gamma$.

It is particularly interesting to follow the evolution of entropy during electron-positron annihilation. Intuitively, entropy conservation implies that when electrons and positrons annihilate each other, their entropy has to go into other species, namely: photons, which are the only remaining relativistic species in the plasma. In other words, the reaction $e^- + e^+ \longrightarrow n\gamma$ generates an excess of photons; since photons are in thermal equilibrium, any excess in the number density must be described in terms of an increase in the product (Ta). Let us check this explicitly. Before positron annihilation, the expression of entropy is given by eq. (3.51). After annihilation, it reads:

$$s = \left.\frac{\rho+p}{T}\right|_{\rm plasma} + \left.\frac{\rho_\nu+p_\nu}{T_\nu}\right|_{\rm neutrinos} \tag{3.52}$$

$$= \frac{4}{3}\frac{\pi^2}{30}(2)\,T^3 + \frac{4}{3}\frac{\pi^2}{30}\left(\frac{7}{8}\times 6\right)T_\nu^3\ . \tag{3.53}$$

Note that the total entropy in a comoving volume sa^3 is conserved, but the separate entropy of neutrinos is also conserved since they are decoupled and $(T_\nu a)$ is exactly constant. This implies that $s_{\rm plasma}a^3$ is also conserved separately. Hence:

$$\frac{11}{2}(Ta)^3_{\rm before} = 2(Ta)^3_{\rm after}\ . \tag{3.54}$$

We conclude that the temperature of the plasma does not scale like a^{-1} during electron positron annihilation: this is a typical example in which it is rescaled according to $g_*^{-1/3}$. In fact, Ta increases in order to compensate the loss of the electron and positron degrees of freedom. But the most interesting outcome of this is that the temperature of photons and neutrinos after annihilation differs by:

$$\frac{(T_\nu a)_{\rm after}}{(Ta)_{\rm after}} = \frac{(T_\nu a)_{\rm before}}{(11/4)^{1/3}(Ta)_{\rm before}} = \left(\frac{4}{11}\right)^{1/3}\ . \tag{3.55}$$

After positron annihilation, the photons are the only remaining species in thermal equilibrium, hence $g_* = 2$ and (Ta) is exactly constant. Finally, we will see that photons decouple around $T \sim 0.3$ eV. Like for neutrinos, the distribution of photons remains frozen after decoupling, with $T(t) = T(t_D)a(t_D)/a(t)$ until today. We conclude that between $T \sim 0.5$ MeV and today, the relation $T_\nu = (4/11)^{1/3}T$ holds at any time, with the photon temperature given by $T = T_0(a_0/a)$. Here, T_0 is the CMB temperature measured today, $T_0 = 2.726$K. So $T_{\nu 0} = 1.946$K. Knowing the photon and neutrino temperature today, we can infer their number densities:

$$n_\gamma^0 = \frac{\zeta(3)}{\pi^2}\times 2\ T_0^3 = 137\ {\rm cm}^{-3}\ , \tag{3.56}$$

$$n_\nu^0 = \frac{\zeta(3)}{\pi^2}\times\frac{3}{4}\times 6\times\frac{4}{11}\,T_0^3 = 112\ {\rm cm}^{-3}\ , \tag{3.57}$$

(the second number being the total density summed over the six neutrinos).

3.3.5 Nucleosynthesis

A nucleus X containing Z protons can have various isotopes ${}^AX$ of mass number A (hence containing $A-Z$ neutrons). The following reactions can increase Z by one unit, starting from a simple proton (i.e. ionized hydrogen nucleus $H^+ = p$; in the following we will omit to write the charge of the various ions):

$$p + n \longrightarrow D + \gamma \tag{3.58}$$

$$D + D \longrightarrow {}^3He + n \tag{3.59}$$

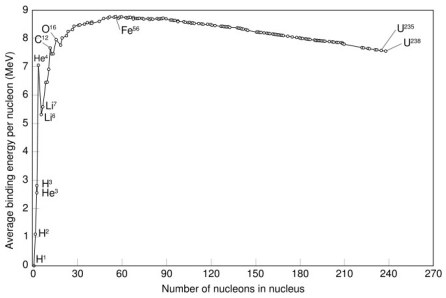


Figure 3.2: Average binding energy per nucleon B/A as a function of A.

$$^3He + D \longrightarrow {}^4He + p \tag{3.60}$$

$$... \tag{3.61}$$

In order to know whether these reactions are favored or not from the point of view of energetics, we should know the binding energy B of each element. We recall that the binding energy is the minimal amount of energy which must be furnished in order to break a nucleus X in Z protons and $A-Z$ neutrons. Hence the rest energy of X reads:

$$E_0(X) = m_X = Zm_p + (A-Z)m_n - B \ . \tag{3.62}$$

For instance, the binding energy of deuterium is $B_D = 2.22$ MeV, since $m_p = 938.27$ MeV, $m_n = 939.57$ MeV, $m_p+m_n = 1877.84$ MeV and $m_D = 1875.62$ MeV. Hence, from a purely energetic point of view, protons and neutrons should combine and form the isotope with the largest possible binding energy per nucleon B/A: once this isotope exists, any nuclear reaction destroying it would cost energy. Figure 3.2 shows the average binding energy per nucleon B/A as a function of A. Starting from zero for hydrogen ^1H (p), the curve raises for deuterium ^2H (pn), helium ^3He (ppn), tritium ^3H (pnn), and reaches a local maximum for ^4He $(ppnn)$. The first isotope with a ratio B/A larger than that of ^4He is ^{12}C. The global maximum is reached at $A = 56$ for iron ^{56}Fe.

Preliminary overview of Nucleosynthesis. From a purely energetic point of view, we could expect the following picture. The reaction

$$D + \gamma \longrightarrow p + n \tag{3.63}$$

requires an energy of at least $B_D = 2.22$ MeV. For $T > B_D$, photons carry enough energy for breaking any deuterium nucleus into pairs $p + n$. Hence, protons and neutrons can be significantly converted into deuterium only when the temperature drops below B_D. Once deuterium forms, it is energetically more favorable to convert it in ^3He, and so on and so on, until the universe contains only heavy elements like iron.

In the above reasoning, we forgot that the kinetic of the various reactions involved does not depend only on initial and final energies, but also on number densities and cross sections. In fact, the previous reasoning is more or less correct in the frame of the Cold Big Bang scenario, which was rejected on this basis: far from stars, the real universe seems dominated by hydrogen rather than heavy elements. In the Hot Big Bang scenario, a key feature is that baryons are considerably suppressed with respect to photons, $n_B \sim 10^{-10} n_\gamma$. So, our argument that when $T < B_D$ the reaction (3.63) cannot occur is wrong. There

are so many photons that even if the average photon energy is much less than T_D, but a tiny fraction of them (of order 10^{-10}) have a momentum larger than B_D (which is possible if they are in the high-momentum tail of the Fermi-Dirac distribution), then the reaction is still very efficient. So, in the Hot Big Bang scenario, neutrons and protons start forming deuterium at a significantly *smaller* temperature than B_D. The formation of heavier elements is also suppressed by consideration on number densities. Once deuterium is formed, most of it is efficiently converted into ^4He as could be expected from energetics, but then the gap between ^4He and ^{12}C is very difficult to cross: it requires a three-body reaction $3 \times {}^4He \longrightarrow {}^{12}C$. When ^4He forms, the temperature is far too low for the scattering rate of the above reaction to be comparable with H. Hence the chain will stop at ^4He. Let us now check these qualitative expectations using our knowledge of thermal and chemical equilibrium. The discussion can be carried out in two steps.

Formation of Deuterium. We first study the reaction of deuterium formation:

$$n + p \longleftrightarrow D + \gamma \ . \tag{3.64}$$

The cross-section of this reaction is large enough for ensuring chemical equilibrium in the temperature range considered here. Hence $\mu_D = \mu_n + \mu_p$. At $T \ll$ GeV, neutrons, protons and deuterium are all non-relativistic with densities given by eq. (3.21). Hence

$$\frac{n_D}{n_p n_n} = \exp\left(\frac{\mu_D - \mu_p - \mu_n}{T}\right) \ \frac{3}{4}\left(\frac{2\pi m_D}{m_p m_n T}\right)^{3/2} \exp\left(\frac{m_p + m_n - m_D}{T}\right) \ , \tag{3.65}$$

where we used the number of spin states: $g = 2$ for p and n, $g = 3$ for deuterium. The argument of the first exponential cancels because of chemical equilibrium. The argument of the second one involves the binding energy B_D of deuterium:

$$\frac{n_D}{n_p n_n} = \frac{3}{4}\left(\frac{2\pi m_D}{m_p m_n T}\right)^{3/2} \exp\left(\frac{B_D}{T}\right) \ . \tag{3.66}$$

We will now use this equation for getting a rough estimate of the order of magnitude of the deuterium density to baryon number density ratio. We know that roughly, $n_p \sim n_n \sim n_B \sim 10^{-10} n_\gamma \sim 10^{-10} T^3$. Hence we obtain

$$\begin{aligned} \frac{n_D}{n_B} &\sim 10^{-10}\left(\frac{T}{m_p}\right)^{3/2} \exp\left(\frac{B_D}{T}\right) \\ &\sim 10^{-10}\left(\frac{T}{0.94 \text{ GeV}}\right)^{3/2} \exp\left(\frac{2.22 \text{ MeV}}{T}\right) \ . \end{aligned} \tag{3.67}$$

As long as $T > B_D$, it is clear that the ratio remains tiny. As expected, there is no significant deuterium abundance above that scale; all baryons are in the form of neutrons and protons. The first terms can be compensated only if the argument of the exponential is large enough. A quick estimate shows that for $T \sim 0.06$ MeV, the above ratio reaches the order of one. A more careful estimate shows that the deuterium abundance becomes sizable around 0.07 MeV. We will retain 0.07 MeV as the temperature of Nucleosynthesis.

Once deuterium forms, one can show that it is efficiently converted to ^3He and ^4He, since the scattering rate of the relevant reactions exceeds the Hubble rate, and ^4He is the most stable configuration. However, at $T \sim 0.07$ MeV, the scattering rate of the three-body reaction $3 \times {}^4He \longrightarrow {}^{12}C$ is considerably suppressed and the chain stops. We conclude that below $T \sim 0.07$ MeV,

nucleons combine into ^4He, which is formed of two protons and two neutrons. However, protons and neutrons are not necessarily in exactly equal proportions before this temperature is reached. Hence, together with ^4He, there might be a relic density of protons or neutrons. We see that it is crucial to compute the neutron over proton ratio for $T \geq 0.07$ MeV.

Neutron versus proton density above $T \sim 0.07$ MeV. The balance between neutrons and protons depends essentially on the reaction (called β-decay):

$$p + e^- \longleftrightarrow n + \nu_e \ . \tag{3.68}$$

At high energy ($T >$ MeV), this reaction is in chemical equilibrium, with $\mu_p + \mu_e = \mu_n + \mu_{\nu_e}$. The chemical potential of neutrinos is zero in the simplest cosmological model considered in this course. The one of electrons is non-zero, but before electron-positron annihilation the asymmetry between electrons and positrons is so small ($\mu_e/T \sim 10^{-10}$) that we can work in the approximation $\mu_e \simeq 0$. Hence:

$$\begin{aligned} 1 &= \exp\left(\frac{\mu_p + \mu_e - \mu_n - \mu_{\nu_e}}{T}\right) \\ &\simeq \exp\left(\frac{\mu_p - \mu_n}{T}\right) \\ &= \frac{n_p}{n_n}\left(\frac{m_n}{m_p}\right)^{3/2} \exp\left(\frac{m_p - m_n}{T}\right) \ . \end{aligned} \tag{3.69}$$

(for the last equality, we used eq. (3.21) for the number density of non-relativistic species). The difference between the neutron and proton mass is $Q \equiv m_n - m_p = 1.203$ MeV. Hence, for $T \gg 1$ MeV, the neutron to proton ratio is given by:

$$\left.\frac{n_n}{n_p}\right|_{T \gg 1 \text{ MeV}} = (m_n/m_p)^{3/2} = 1.002 \ , \tag{3.70}$$

i.e. the density of neutrons and protons is essentially the same. When $T \sim 1$ MeV, chemical equilibrium would force the neutron to proton ratio to drop exponentially like $\exp(-Q/T)$. If this was true, at 0.07 MeV there would be essentially no neutron left, and Nucleosynthesis would not happen: the primordial universe would contain only hydrogen.

However, the above reaction is mediated by weak interactions. Hence, it becomes quite weak around $T \sim$MeV, and we are forced to consider its departure from chemical equilibrium. In fact we will see that the reaction freezes out with a significant leftover of neutrons. The neutron density obeys to the Boltzmann equation:

$$\dot{n}_n + 3Hn_n = n_n[n_{\nu_e}\langle\sigma v\rangle]\left\{\exp\left(\frac{\mu_e + \mu_p - \mu_n - \mu_{\nu_e}}{T}\right) - 1\right\} \ . \tag{3.71}$$

The term between square brackets is the scattering rate Γ_{np} for neutron to proton conversion, and the exponential can be approximated using eq. (3.69). Hence

$$\dot{n}_n + 3Hn_n = n_n\Gamma_{np}\left\{\frac{n_p}{n_n}\left(\frac{m_n}{m_p}\right)^{3/2} e^{-Q/T} - 1\right\} \ . \tag{3.72}$$

This equation can be written in terms of a dimensionless variable, the neutron fraction $X_n = n_n/(n_n + n_p)$. We have

$$n_n = X_n(n_n + n_p) = X_n n_B, \qquad n_p = (1 - X_n)n_B \ . \tag{3.73}$$

The conservation of the baryon number implies $n_B \propto a^{-3}$, so

$$\dot{n}_n = \dot{X}_n n_B - 3HX_n n_B \ . \tag{3.74}$$

Replacing n_n and n_p in eq. (3.72) and dividing by n_B, we get

$$\dot{X}_n = \Gamma_{np} \left[(1 - X_n)\, e^{-Q/T} - X_n \right] \ . \tag{3.75}$$

The dependence of Γ_{np} with respect to T can be computed using nuclear physics. Still, in order to integrate the equation, we need to know the relation between time t and temperature T. This relation can be inferred from the Friedmann equation. In first approximation, $T \propto a^{-1}$ (neglecting the effect of the electron-positron annihilation on Ta) and $dT/T = -da/a$. So,

$$\frac{dT}{dt} = -T\,\frac{da}{a\,dt} = -TH \tag{3.76}$$

$$= -\sqrt{\frac{8\pi G}{3}\rho T^2} \tag{3.77}$$

$$= -\sqrt{\frac{8\pi^3 G}{90} g_* T^6} \tag{3.78}$$

with $g_* = 10.75$ before electron-positron annihilation. Hence the reaction reads

$$\frac{dX_n}{dT} = -\sqrt{\frac{90}{8\pi^3 g_*}}\frac{M_P}{T^3}\Gamma_{np}(T) \left[(1 - X_n) e^{-Q/T} - X_n \right] \ . \tag{3.79}$$

Knowing $\Gamma_{np}(T)$, this equation can be integrated. The result is that around $T \sim 0.1$ MeV, X_n gets close to an asymptotic value of 0.15, corresponding to the freeze-out of the neutron to proton ratio.

Equation (3.79) is just a first-order approximation. The precise calculation includes two additional effects: the change in g_* and Ta due to the electron-positron annihilation, and the neutron beta-decay ($n \longrightarrow p + e^- + \bar{\nu}_e$) which should be included in the right-hand side of the Boltzmann equation since it represents another decay channel. Altogether, these effects lead to a slightly different neutron to proton ratio at freeze-out, $X_n(T < 0.1 \text{ MeV}) \sim 0.11$, while at the time of Deuterium creation, $n_n = 0.124\, n_B$ and $n_p = 0.876\, n_B$. Then, all available neutrons will combine into deuterium, ^{3}He and finally ^{4}He nuclei, together with the same number of protons. The final ^{4}He density should be $n_{^4\text{He}} = 0.062\, n_B$, with a leftover of $n_H = 0.752\, n_B$ protons. The helium fraction, usually defined as:

$$Y_P \equiv \frac{4n_{^4He}}{n_B} \ , \tag{3.80}$$

is predicted to be 0.228 at any time after Nucleosynthesis, in every region of the universe not affected by the ejection of particles from stars (since inside stars, nuclear reactions can form other elements in very different proportions).

Exact results from a full calculation. The above calculation was rather simplistic. A full simulation of Nucleosynthesis can be performed using numerical codes (a few Nucleosynthesis codes are even publicly available). Instead of studying the kinetics of just two reactions, these codes follow of the order of one hundred possible reactions between neutrons, protons and heavier nuclei (typically, till ^{12}C). The main differences between the outcome of a full simulation and the results of the above section are:

- when reactions freeze-out, the density n_i of other elements than ^{4}He is nonzero - but still very small: the number density of D and ^{3}He is smaller than that of ^{4}He by a factor $\sim 10^5$, the density of ^7Li is smaller by $\sim 10^9$, and all other species are even more suppressed.

- the final helium fraction depends slightly on the free parameter of this problem, namely $n_B/s \sim 10^{-10}$, which controls mainly the temperature at which deuterium starts forming (see eq. (3.67)). Hence the neutron-to-proton ratio at the beginning of deuterium formation depends on n_B/s, as well as the final helium abundance. The ratio n_B/s is easy to relate today to $(n_p + n_n)/n_\gamma$, and for fixed CMB temperature, this ratio can finally be expressed as a function of ω_b.

3.3.6 Recombination

After Nucleosynthesis, the universe contains a thermal plasma composed essentially of relativistic photons and non-relativistic electrons, hydrogen nuclei and helium nuclei; plus decoupled relativistic neutrinos. At $T \ll$ MeV, weak interactions are inefficient, but electromagnetic interactions ensure equilibrium between electrons, nuclei and photons. More precisely, photons remain tightly coupled to electrons via Compton scattering ($e^- + \gamma \longrightarrow e^- + \gamma$) and electrons to nuclei via Coulomb scattering ($e^- + p \longrightarrow e^- + p$ or $e^- + {}^4\mathrm{He} \longrightarrow e^- + {}^4\mathrm{He}$). These interactions are efficient at least as long as hydrogen and helium remain ionized.

The formation of neutral hydrogen depends on the reaction:

$$e^- + p \longleftrightarrow H + \gamma \ . \tag{3.81}$$

The exact description of recombination is considerably complicated by the fact that hydrogen can form in various excited states, and then relax to its fundamental state while emitting photons: so, there are many states and reactions to follow. Here we will neglect this issue and do as if hydrogen could only be in its fundamental state.

Like for Nucleosynthesis, let us start from purely energetic considerations. The binding energy of hydrogen, defined through:

$$m_H = m_p + m_e - \epsilon_0 \ , \tag{3.82}$$

is equal to $\epsilon_0 = 13.6$ eV. Hence we expect that for $T \gg 13.6$ eV hydrogen is fully ionized: any neutral hydrogen atom would immediately interact with an energetic photon and get ionized. This does not mean that neutral hydrogen forms immediately below $T \sim 13.6$ eV. Just like for the formation of deuterium during Nucleosynthesis, the balance of the above reaction depends on relative abundances. We know that the density of electrons and protons is 10^{10} times smaller than that of photons. So, much below $T \sim 13.6$ eV, there should still be enough energetic photons for preventing recombination.

In the exercise sessions, you will find that this expectation is confirmed by the actual equations. You will define the hydrogen ionization fraction:

$$X_e \equiv \frac{n_e}{n_e + n_H} = \frac{n_p}{n_p + n_H} \ . \tag{3.83}$$

Assuming thermal equilibrium, you will derive the Saha equation

$$\frac{X_e^2}{1 - X_e} = \frac{1}{n_e + n_H} \left(\frac{m_e T}{2\pi} \right)^{3/2} e^{-\epsilon_0/T} \ , \tag{3.84}$$

which gives an approximation of the temperature of recombination, found to be close to $T_{\mathrm{rec}} \sim 0.254$ eV. Below this temperature, the reaction leaves thermal equilibrium, implying that the ionisation fraction freezes out. You will write the Boltzmann equation governing the evolution of X_e. By integrating this equation, one would find that the ionization fraction X_e becomes significantly smaller than one around $z \sim 1080$, and tends to an asymptotic freeze-out value of order $X_e \to 5 \times 10^{-4}$ for $z < 100$.

3.3.7 Photon decoupling

Till the time of recombination, photons are maintained in thermal equilibrium mainly through Compton scattering off electrons:

$$\gamma + e^- \longrightarrow \gamma + e^- \ . \tag{3.85}$$

The cross section $\langle \sigma v \rangle$ of the above reaction is the Thomson cross section, equal to $\langle \sigma_T v \rangle = 0.665 \times 10^{-24} \mathrm{cm}^2$. Compton scattering of photons off electrons becomes inefficient roughly when the scattering rate $\Gamma = n_e \langle \sigma_T v \rangle$ equals the Hubble parameter. In order to evaluate this characteristic time, we can write $n_e = n_p = X_e n_B$ (like in the previous section, we neglect helium) and $n_B \sim \rho_b / m_p$. We obtain:

$$\frac{\Gamma}{H} = 0.07 (a_0/a)^3 X_e \Omega_b h \frac{H_0}{H} \ . \tag{3.86}$$

The Hubble rate in units of Hubble rate today can be estimated to be $H/H_0 = \Omega_M^{1/2} (a_0/a)^{3/2}$ during matter domination. Taking a to be of the order of $a_{\rm dec}$, $\Omega_b \simeq 0.4$, $\Omega_M \simeq 0.25$ and $h \simeq 0.7$, we see that photon decoupling occurs when X_e drops below $\sim 10^{-2}$ during recombination. Hence, recombination directly triggers photon decoupling. This is in fact the main reason for which recombination is important to study: it controls the decoupling of the CMB photon that we observe today. The details of recombination affect CMB anisotropy patterns. However, the temperature evolution of photons is completely unaffected by their decoupling, exactly like for neutrinos. When photons decouple, their relativistic Bose-Einstein distribution freezes-out, and only evolves at later times due to the universe expansion, which induces $p \propto a^{-1}$ and hence $T \propto a^{-1}$.

A precise calculation shows that photon decoupling takes place near the redhsift of recombination, $z_{\rm dec} = z_{\rm rec} = 1080$. Translating this redshift in terms of proper time, one finds photons decouple approximately 380,000 years after the initial singularity.

3.3.8 Very recent stages

From the point of view of the thermal history of the universe, very few phenomena occur after photon decoupling. Each neutrino family i becomes non-relativistic when $T_\nu < m_i$, but since they are already decoupled, this has no effect on the temperature and number density evolution ($T_\nu \propto a^{-1}$ and $n_\nu \propto a^{-3}$). Only the energy density and pressure of neutrinos are affected by the non-relativistic transition. The consequences of this transition on structure formation are interesting, but not discussed in this course.

There is however another important phenomenon occurring at low redshift, than we just mention here briefly. When the first stars form, they emit a new population of photons which partially reionize hydrogen and heavier elements. However, this reionization is not sufficient for "re-coupling" photons to electrons and ionized matter: only a small fraction of CMB photons have a chance to experience Compton scattering between the time of decoupling and today. This can be understood from eq. (3.86): when X_e goes back to one at small redshifts $z \sim 10$, the ratio $(a/a_0)^3$ is much smaller than at the time of recombination (100^3 smaller), so Γ/H remains smaller than one.

In figure 3.3, we summarize qualitatively the main results of this section.

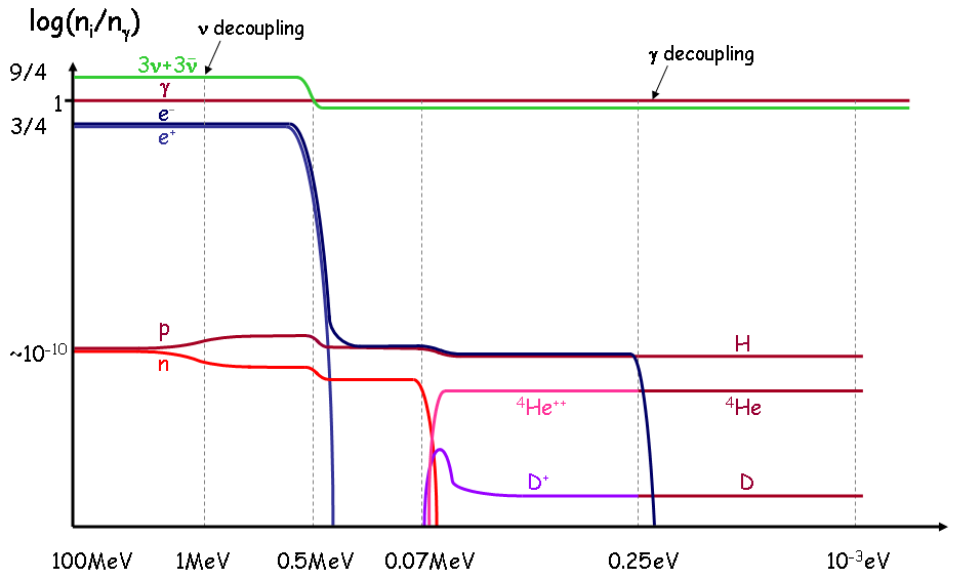


Figure 3.3: As a summary of Chapter 3, we show the qualitative evolution of n_i for each species, normalized in terms of n_γ.

Chapter 4

Dark Matter

4.1 Historical arguments

There are many strong reasons to believe that in the recent universe, the non-relativistic matter is of two kinds: ordinary matter and dark matter. One of the well-known evidences for dark matter arises from galaxy rotation curves.

Inside galaxies, the stars orbit around the center. If we can measure the redshift in different points inside a given galaxy, we can reconstruct the distribution of velocity $v(r)$ as a function of the distance r to the center. It is also possible to measure the distribution of luminosity $I(r)$ in the same galaxy. What is not directly observable is the mass distribution $\rho(r)$. However, it is reasonable to assume that the mass distribution of the *observed luminous matter* is proportional to the luminosity distribution: $\rho_{\rm lum}(r) = b\ I(r)$, where b is an unknown coefficient of proportionality called the bias. From this, we can compute the gravitational potential $\Phi_{\rm lum}$ generated by the luminous matter, and the corresponding orbital velocity, given by ordinary Newtonian mechanics:

$$\rho_{\rm lum}(r) = b\ I(r), \tag{4.1}$$

$$\Delta\Phi_{\rm lum}(r) = 4\pi G\ \rho_{\rm lum}(r), \tag{4.2}$$

$$v^2_{\rm lum}(r) = r\frac{\partial}{\partial r}\Phi_{\rm lum}(r). \tag{4.3}$$

So, $v_{\rm lum}(r)$ is known up to an arbitrary normalisation factor $\sqrt{b}$. However, for many galaxies, even by varying b, it is impossible to obtain a rough agreement between $v(r)$ and $v_{\rm lum}(r)$ (see figure 2.3). The stars rotate faster than expected at large radius. We conclude that there is some non–luminous matter, which deepens the potential well of the galaxy.

We can explain the same result in slightly different words. Assuming that stars have a circular orbit (this is just an approximation), the relation between force and accelerations gives us

$$\frac{v^2(r)}{r} = \frac{\partial}{\partial r}\Phi(r) \tag{4.4}$$

while the Poisson equation of newtonian mechanics gives us

$$\Delta\Phi(r) = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\Phi\right) = 4\pi G\rho(r) \tag{4.5}$$

Finally, the mass of objects enclosed in a radius r is just

$$M(r) = 4\pi\int_0^r dr'(r')^2\rho(r')\ . \tag{4.6}$$

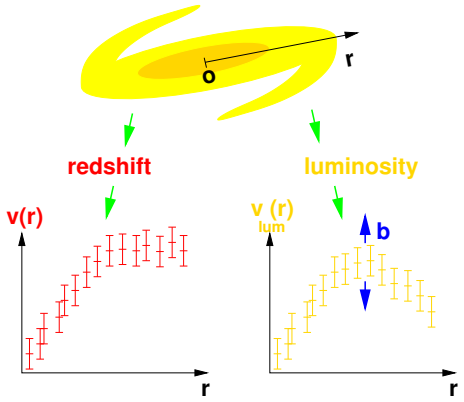


Figure 4.1: A sketchy view of the galaxy rotation curve issue. The genuine orbital velocity of the stars is measured directly from the redshift. From the luminosity distribution, we can reconstruct the orbital velocity under the assumption that all the mass in the galaxy arises form of the observed luminous matter. Even by varying the unknown normalisation parameter b, it is impossible to obtain an agreement between the two curves: their shapes are different, with the reconstructed velocity decreasing faster with r than the genuine velocity. So, there has to be some non–luminous matter around, deepening the potential well of the galaxy.

These relations and a simple integration by part give the exact relation

$$v^2(r) = \frac{GM(r)}{r} \ . \tag{4.7}$$

If we assume that all the mass is in the form of visible matter, there is a mismatch between measurements of $v(r)$ and estimates of $M(r)$. In particular, when we see that most of the mass is located within a radius r_v (where v stands for visible), we expect that above r_v, $M(r)$ reaches a constant asymptote. Then $v(r > r_v)$ should decrease nearly like $1/r$ (this is obvious from the last relation, and such a decrease is called a Keplerian decrease). Instead, in many galaxies, beyond such a radius r_v, the few remaining starts tend to orbit much too fast. An obvious solution is to assume that there is another type of non-visible matter contributing to $M(r)$, and even dominating it. If the non-visible matter is spread over a larger radius than visible matter, then the most distant observable galaxies are still orbiting in the gravitational potential created by dark matter. This supports the notion of a dark matter halo.

A qualitatively similar argument applies to the dynamics of galaxies within galaxy clusters. Actually, the hypothesis of dark matter was formulated for the first time by Franz Zwicky in 1933, following the observation of surprisingly large galaxy velocities inside the Coma galaxy cluster.

4.2 Other evidences for dark matter

Apart from galactic rotation curves, there are many arguments – of more cosmological nature – which imply the presence of a large amount of non–luminous matter in the universe, called dark matter.

The observation of *CMB anisotropies* is the strongest one. It requires the presence of a component *not interacting with ordinary electromagnetic forces*. This component had to be present in the universe at least at early times, between

$z \sim 10^6$ and $z \sim 10^3$. Assuming that this dark matter is stable and that its number density is conserved until today, we obtain from the CMB an estimate of the dark matter density today corresponding to 25% of the total energy density of the universe (and 85% of the total non-relativistic matter density). Before CMB observations, one could have thought that dark matter is just ordinary matter that we cannot see, because it is not luminous, and it does not absorb the light of other objects. If this was the case, from the point of view of CMB physics, non-luminous ordinary matter would count as baryonic matter, not as dark matter. Thanks to CMB data, we are now sure that dark matter is truly different from ordinary atomic matter.

Strong lensing. Gravitational lensing is by definition sensitive to the total gravitational potential created by both ordinary matter and dark matter. The study of arclets and of strong lensing patterns allows to reconstruct the shape of dark matter halos around some particular galaxy clusters, and to bring further proofs of the existence of such halos.

Weak lensing. By looking at the statistics of the apparent orientation of galaxies in different regions of the sky, one can average out the random distribution of intrinsic shapes, and estimate the effect of the weak lensing of galaxy images by dark matter (since this effect is coherent over many galaxies in a given region). Weak lensing works surprisingly well and gives us a very good map of the gravitational potential projected along the line-of-sight in each direction around us. By studying the weak lensing of galaxies located at a given redshift, one can even do *tomography* and reconstruct the 3D distribution of dark matter (always with a poorer resolution along the 3rd dimension, i.e. along the line-of-sight). This technique brings further evidence for dark matter, and allows to estimate its abundance. It gives consistent results with other techniques, and in particular with CMB observations.

The analysis of the *bullet cluster* shows very well the presence of two halos that have crossed each other in the recent past without being deformed, unlike the two associated clouds of gas, now displaced from the center of halos, and shaped like a shock wave. This is another way to find that dark matter is very weakly interacting - possibly only gravitationally.

There are other ways to prove the existence and the properties of dark matter that we do not have time to summarise here. However, we should stress one important fact: dark matter must be *cold* rather than *hot*. What do we mean by this?

Within the standard model of particle physics, a good candidate for non-baryonic dark matter would be a neutrino with a small mass. Then, dark matter would become non-relativistic only recently. Today it would still possess large velocities, just a few orders of magnitude smaller than the speed of light (this hypothesis is called Hot Dark Matter or HDM). Because of these large velocities, neutrinos could not remain confined in small gravitational potential wells. Even in presence of gravitational clustering, they would form very large and not-so-dense halos (while particles with negligible velocities, called Cold Dark Matter or CDM, can cluster much better: they can form much smaller and much denser halos).

If halos were huge and not very dense like in the HDM case, the number and the distribution of galaxies (which depends on the gravitational effects of dark matter) would be very different from what we observe. For that reason, HDM is strongly excluded by several types of observations. Dark matter particles have to be strongly non-relativistic, otherwise galaxy could not form with a sufficient abundance during matter domination.

4.3 Thermal WIMP model

Here we will review only one out of many possible models for explaining the dark matter problem: the case of so-called WIMPs. WIMP means *Weakly Interacting Massive Particle*.

Here *Interacting* refers to the fact that these particles are assumed to have interactions with standard model particles (otherwise no interesting calculations or predictions could be made...). More specifically, WIMP interactions are supposed to be sufficiently efficient in the early universe for bringing WIMPs in thermal equilibrium with other particles.

Weakly refers to the fact that we assume specifically that these interactions are of the weak type (i.e. mediated by Z and $W^{\pm}$ bosons). Hence we expect that these particles decouple when the temperature is roughly of the order of the MeV, like for neutrinos. After decoupling, WIMPs interact only gravitationally with other species.

Massive refers to the fact that we impose a sufficient mass for WIMPs to decouple when they are non-relativistic: i.e. we assume a mass $m_\chi \gg O(\text{MeV})$. If we did not make such an assumption, i.e. if the WIMPs decoupled when they are still relativistic, they would share the same number density as neutrinos until today. Then the correct value of the relic density ρ^0_{dm} (or equivalently ω_{dm}) could only be obtained for $m_\chi \sim O(10)$ eV, and the typical velocity of WIMPs today would be of the order of $v \sim \langle p \rangle / m_\chi \sim T/m_\chi \sim 10^{-4}c$. Such velocities are still very large, and this dark matter candidate would fall in the category of Hot Dark Matter. To avoid this, we impose non-relativistic decoupling: we will see that this leads to a number density n^0_χ very suppressed with respect to that of neutrinos, and because the mass is large, to very small dark matter velocities: in that case WIMPs fall in the category of Cold Dark Matter.

WIMPs are usually assumed to be neutral and to be their own anti-particle, because they are Majorana fermions, that we will denote χ. Moreover, one usually imposes a Z_2 symmetry on these particles. This means that the Lagrangian can only feature even powers of χ. For instance, an interaction term χAB is forbidden by the Z_2 symmetry, while a term $\chi^2 AB$ respects the symmetry. As a consequence, WIMPs cannot decay ($\chi \longrightarrow A + B + ...$ does not respect the symmetry) but they can annihilate ($\chi + \chi \longrightarrow A + B + ...$ does respect the symmetry. In the latter reaction, the total charge on the left-hand side is zero, since χ is its own antiparticle. Hence the total charge must be zero also on the right-hand side. It means that pairs of WIMPs can only annihilate into neutral particles, or into pairs of particles and anti-particles. For instance, they can annihilate into higgs bosons, Z^0 bosons, quark-antiquark pairs, lepton-anti-lepton pairs, etc. This is model-dependent. "Popular" annihilation channels are, for instance, electron-positron, or muon-antimuon, but there are many other possibilities. For the calculations in the rest of this section, it is not necessary to specify explicitly what the dominant annihilation channel is.

Our goal is to compute the relic density of WIMPs and see whether it can be related to the WIMP mass and/or annihilation cross-section. First, we will assume that the annihilation reaction

$$\chi + \chi \longrightarrow A(+B) \tag{4.8}$$

remains in chemical equilibrium at every time. In that case, we can write $2\mu_\chi = \mu_A(+\mu_B)$. But since the right-hand side must be a particle carrying no conserved charge (so $\mu_A = 0$) or a particle-antiparticle pair (so $\mu_A = -\mu_B$), this simplifies to $\mu_\chi = 0$. In that case, we know that for $T \gg m_\chi$

$$n_\chi = \frac{\zeta(3)}{\pi^2}\frac{3}{4} g_\chi T^3 \ , \tag{4.9}$$

while for $T \ll m_\chi$

$$n_\chi = g_\chi \left(\frac{m_\chi T}{2\pi}\right)^{3/2} e^{-\frac{m_\chi}{T}} \ . \tag{4.10}$$

In fact, one could integrate the Fermi-Dirac phase-space distribution in order to get n_χ for any value of temperature. The exact relation has the two asymptotes written above. This means that $n_\chi a^3$ evolves with T like: $(aT)^3$ for $T \gg m_\chi$, i.e. like a constant when the effective number of effective degrees of freedom g_* is constant; and like an exponentially decaying function for $T \ll m_\chi$. In the rest of this section, we will call this particular solution $n_\chi^{\text{ch.eq.}}(T)$ (the "density assuming chemical equilibrium").

In reality, we expect that WIMPs will not remain in chemical equilibrium for a long time after the non-relativistic transition. Indeed, the mass of WIMPs is usually assumed to be slightly above the order of the MeV (a typical range for the most popular models is 100 MeV $< m_\chi < 10^3$ GeV). But we know that weak interactions become inefficient around $T \sim O(\text{MeV})$ for neutrinos, because $\Gamma = n_\nu \langle \sigma_\nu v \rangle$ becomes smaller than H. For WIMPs, we expect a cross section of the same order of magnitude (because it still depends on the Fermi constant), and an annihilation rate $\Gamma = n_\chi \langle \sigma_A v \rangle$ even smaller than for neutrinos, because n_χ gets exponentially suppressed after the non-relativistic transition (here $\langle \sigma_A v \rangle$ is the WIMP thermally averaged annihilation cross section). Hence we expect WIMPs to leave thermal equilibrium even before neutrinos, and very soon after their non-relativistic transition, when n_χ becomes very small.

We know that "leaving chemical equilibrium" means, concretely, that the WIMP number density will freeze out, and that $n_\chi a^3$ will be conserved after freeze-out. The evolution of n_χ is given by the Boltzmann equation

$$\dot{n}_\chi + 3Hn_\chi = n_\chi^2 \langle \sigma_A v \rangle \left[e^{-\frac{2\mu_\chi}{T}} - 1 \right] \ . \tag{4.11}$$

Using our definition of $n_\chi^{\text{ch.eq.}}(T)$, and the fact that for $T \ll m_\chi$ it is given by eq. (4.10), we can write the Boltzmann equation in the following form for any time *after* the WIMP non-relativistic transition:

$$\dot{n}_\chi + 3Hn_\chi = n_\chi^2 \langle \sigma_A v \rangle \left[\left(\frac{n_\chi^{\text{ch.eq.}}}{n_\chi} \right)^2 - 1 \right] = \langle \sigma_A v \rangle \left[\left(n_\chi^{\text{ch.eq.}}\right)^2 - (n_\chi)^2 \right] \ . \tag{4.12}$$

We now want to estimate the relic density of WIMPs at freeze-out. Let us use a superscript f to denote quanties at that time (for instance, the freeze-out temperature will be T^f). At freeze-out, the real solution n_χ and the chemical equilibrium solution $n_\chi^{\text{ch.eq.}}$ start to depart significantly from each other, but they must still be of the same order of magnitude:

$$n_\chi^{\text{ch.eq. f}} \sim n_\chi^f \ . \tag{4.13}$$

Given our previous discussion on the Boltzmann equation in Chapter 3, we also know that at freeze-out, the term $3Hn_\chi$ must be of the same order of magnitude as the term on the right-hand side of eq. (4.12), so we can write

$$3H^f n_\chi^f \sim \langle \sigma_A v \rangle (n_\chi^{\text{ch.eq. f}})^2 \sim \langle \sigma_A v \rangle (n_\chi^f)^2 \ . \tag{4.14}$$

Using these relations, we get a very useful result:

$$n_\chi^{\text{ch.eq. f}} \sim \frac{3H^f}{\langle \sigma_A v \rangle} \ . \tag{4.15}$$

We can estimate H^f as a function of temperature using what we have learned in Chapter 2 (Friedmann equation) and 3 (density of the thermal bath and effective number of relativistic degrees of freedom g_*):

$$(H^f)^2 = \frac{8\pi G}{3}\rho^f = \frac{8\pi G}{3}\frac{\pi^2}{30}g_*^f(T^f)^4 \sim \frac{g_*^f(T^f)^4}{M_P^2} \tag{4.16}$$

Hence

$$n_\chi^{\text{ch.eq. f}} \sim \frac{(g_*^f)^{1/2}(T^f)^2}{M_P\langle\sigma_A v\rangle} \tag{4.17}$$

Moreover, we know that after freeze-out, $n_\chi a^3$ is exactly conserved. At the same time the entropy conservation law says that $g_*(Ta)^3$ is constant, so n_χ evolves proportionally to g_*T^3. Finally we can evaluate the WIMP relic density today:

$$n_\chi^{\text{ch.eq. 0}} = n_\chi^{\text{ch.eq. f}}\ \frac{g_*^0}{g_*^f}\ \left(\frac{T^0}{T^f}\right)^3 \ . \tag{4.18}$$

Putting everything together, this gives

$$n_\chi^{\text{ch.eq. 0}} \sim \frac{g_*^0}{(g_*^f)^{1/2}}\frac{(T^0)^3}{T^f}\frac{1}{M_P\langle\sigma_A v\rangle} \ . \tag{4.19}$$

Since WIMPs are strongly non-relativistic today, we can infer the energy density by multiplying the number density by the WIMP mass. This gives:

$$\rho_\chi^{\text{ch.eq. 0}} \sim \frac{m_\chi}{T^f}\frac{g_*^0}{(g_*^f)^{1/2}}\frac{(T^0)^3}{M_P\langle\sigma_A v\rangle} \ . \tag{4.20}$$

In order to get a definite prediction, and assuming that in a given model we know g_*^0, g_*^f and $\langle\sigma_A v\rangle$, the only remaining task is to evaluate the ratio $\frac{m_\chi}{T^f}$.

We know that $\frac{m_\chi}{T}$ crosses one at the time of the non-relativistic transition. A crude approximation would consist in saying that freeze-out takes place very soon after the non-relativistic transition, because it is triggered by the fast exponential decay of n_χ as a function of $\frac{m_\chi}{T}$ (as indicated by eq. (4.10)). In this approximation we could use $\frac{m_\chi}{T^f} \sim 1$.

This approximation is actually not so bad, because the true value of the freeze-out temperature would be given by solving the equation

$$\frac{\Gamma}{H} = \frac{n_\chi^{\text{ch.eq.}}\langle\sigma_A v\rangle}{H} \sim 1 \ , \tag{4.21}$$

with n_χ given by eq. (4.10). Due to the factor $e^{-\frac{m_\chi}{T}}$, the solution for $\frac{m_\chi}{T^f}$ depends *logarithmically* on the parameters of the problem (mass, cross-section): hence it is true that $\frac{m_\chi}{T^f}$ is of order one, and can be considered as nearly independent of the mass and cross-section in first approximation.

We are led to our final result: the relic density of WIMPs is governed by the inverse annihilation cross-section, with

$$\rho_\chi^{\text{ch.eq. 0}} \sim \frac{g_*^0}{(g_*^f)^{1/2}}\frac{(T^0)^3}{M_P\langle\sigma_A v\rangle} \ . \tag{4.22}$$

This is a very well-known result. Intuitively, WIMPs with a larger cross-section remain in thermal equilibrium for a longer time after their non-relativistic transition. Hence they are more Boltzmann-suppressed at freeze-out, and a smaller fraction of them survives until today. By matching $\rho_\chi^{\text{ch.eq. 0}}$ to observations, one

gets a prediction for $\langle \sigma_A v \rangle \sim 10^{-26} cm^3 s^{-1}$. This is precisely the order of magnitude that one would expect for a particle interacting only with weak forces. Many experiments of so-called *dark matter direct detection* have been built for probing such particles. They are usually located in underground laboratories, to filter out as many cosmic rays, ordinary electromagnetic radiation and terrestrial radio-activity as possible. Most of them try to measure the small heating of the detector caused by elastic interactions between detector particles and the WIMPs crossing them. This is of course very difficult, due to the very small interaction rate. But at least, people know what to search for: we have seen that we can estimate the WIMP annihilation cross-section, and hence, also the typical WIMP interaction cross sections. These detectors have not found any significant signal so far.

There are many dark matter candidates falling in the category of WIMPs: it is not so difficult to build a reasonable extension of the standard model of particle physics featuring new particles, some of them having the basic properties that we mentioned in this section. A famous examples is the *neutralino* of supersymmetric models. However, not only WIMPs have not been discovered, but the LHC is currently bringing no evidence in favour of supersymmetry.

Search for WIMPs will continue in the next years. In parallel, people are thinking about other types of dark matter candidates (axions, sterile neutrinos, ...), and are working on other types of experiments to probe them. We do not have time to describe these alternative scenarios in this chapter.

Chapter 5

Cosmological perturbations

In all this chapter, we will study the evolution of cosmological perturbations under the assumption that the universe is flat: this simplifies all equations considerably.

5.1 Linear cosmological perturbations

5.1.1 Classification

We decompose the metric and stress-energy tensor of the universe into spatial averages and linear perturbations,

$$g_{\mu\nu}(t,\vec{x}) = \bar{g}_{\mu\nu}(t) + \delta g_{\mu\nu}(t,\vec{x}) \ , \tag{5.1}$$
$$T_{\mu\nu}(t,\vec{x}) = \bar{T}_{\mu\nu}(t) + \delta T_{\mu\nu}(t,\vec{x}) \ , \tag{5.2}$$

where $\bar{g}_{\mu\nu}$ stands for the metric of the homogeneous and isotropic Friedmann–Lemaître (FL) model. Being symmetric, the two perturbed tensors contain ten degrees of freedom each, describing different aspects of gravity. Bardeen showed in 1980 that they can be decomposed on the basis of scalars, vectors and tensors under spatial rotations (spatial rotations play a special role because they leave the FL background invariant). These three sectors are decoupled at first order in perturbation theory.

In the vacuum, scalar and vector perturbations vanish, while tensor perturbations can propagate if they have been excited: they account for gravitational waves, the only "real" (propagating) gravitational degrees of freedom in General Relativity (GR). In the presence of matter, scalars represent the response of the metric to an irrotational distribution of matter, and generalize Newton's theory of gravitation. Vectors represent the response of the metric to vorticity, and describe phenomena with no equivalent in Newton's theory, called "gravitomagnetism".

In minimal cosmological models, the vorticity of the various matter components decays with time, and vectors can be neglected. Tensors may play a small role in CMB anisotropies, that we will mention briefly in Sec. 5.2.8. They can be studied separately, since they decouple from the scalar sector at first order in perturbations. Hence this course will be essentially focused on scalar perturbations.

The four scalar components of both the metric and stress-energy perturbed tensors are contained in:

1. the (00) term,
2. the trace of the (ij) matrix,

3. the irrotational part of the $(0i)$ vector,
4. the traceless longitudinal part of the (ij) tensor.

For the perturbed metric $\delta g_{\mu\nu}$, these components correspond (in the same order) to:

1. the generalized gravitational potential ψ,
2. the local distortion ϕ of the average scale factor $a(t)$: the "local scale factor" is given by $(1-\phi)a$,
3. the potential b such that $\delta g_{0i} = \partial_i b$,
4. the potential μ of the metric shear: $\delta g_{ij} = (\partial_i\partial_j - \frac{1}{3}\delta_{ij}\Delta)\mu$.

For the perturbed stress-energy tensor $\delta T_{\mu\nu}$, these components (still in the same order) represent:

1. the energy density perturbations $\delta\rho$ (we will usually refer to the relative perturbations $\delta \equiv \delta\rho/\bar{\rho}$),
2. the pressure perturbations δp,
3. the potential v of the irrotational component of the flux of energy, $\delta T^0_i = \partial_i v$ (v is sometimes called the velocity potential, since in the case of a fluid it is related to the bulk velocity),
4. the potential s of the shear stress or anisotropic stress: $\delta T^i_j = (\partial_i\partial_j - \frac{1}{3}\delta_{ij}\Delta)s$.

It is equivalent to use as a variable the "velocity potential" v or the "velocity divergence" θ defined as

$$\partial_i \delta T^0_i \equiv (\bar{\rho}+\bar{p})\theta = \Delta v \ . \tag{5.3}$$

Similarly, we can use the function σ instead of s, with the definition

$$(\partial_i\partial_j - \frac{1}{3}\delta_{ij}\Delta)\delta T^i_j \equiv (\bar{\rho}+\bar{p})\sigma = \Delta(\Delta s) \ . \tag{5.4}$$

The function σ is usually called the anisotropic stress, although the true anisotropic stress is the component of δT^i_j derived from the potential s. The factors $(\bar{\rho}+\bar{p})$ in the two previous equations are introduced in the definitions in order to obtain simple equations (some authors use alternative notations without these factors or with different ones). To summarize, we see that we can manipulate four degrees of freedom representing the scalar perturbations of matter fields, that can be chosen to be the density fluctuation, pressure perturbation, velocity divergence and anisotropic stress: $\{\delta, \delta p, \theta, \sigma\}$.

5.1.2 Gauges

In an idealised FL universe, there is only one time slicing compatible with the assumption of homogeneity. Instead, in a perturbed universe, there is an infinity of time slicings compatible with perturbation theory (i.e. such that on each slice, all quantities remain close to their average value).

The perturbation of any quantity in a given point is the difference between the true and the average quantity in this point. For instance, for the total energy density ρ,

$$\delta\rho(t,\vec{x}) = \rho(t,\vec{x}) - \bar{\rho}(t) \ . \tag{5.5}$$

While $\rho(t,\vec{x})$ is a locally, unambiguously defined quantity, $\bar{\rho}(t)$ depends on the choice of equal-time hypersurface going through the point $(t,\vec{x})$. With a different choice, $\rho(t,\vec{x})$ would be compared to the average performed on a different sheet, that would take a different value. Hence $\delta\rho(t,\vec{x})$ also depends on the choice of time slicing.

A gauge is a choice of time slicing. Gauge transformations are induced by coordinate transformations $x_\mu \mapsto x_\mu + \epsilon_\mu$ mapping the points of one time slicing to those of another time slicing. All coordinate transformations do not induce a valid gauge transformation: the condition that perturbations must still be linear after the transformation restricts ϵ_μ to be small in every point.

A naive study of the equations of motion of perturbed quantities would be plagued by the freedom to change the gauge without changing physical results: some solutions of the full equations would be "gauge modes" with no observable consequences. To deal with this issue, one can adopt one of two point of views:

- one can derive gauge-invariant quantities (i.e., non-trivial integro-differential combinations of the metric and stress-energy tensor components left invariant by a gauge transformation), and gauge-invariant equations of motions for these quantities. Note that there are four scalar degrees of freedom in $\delta g_{\mu\nu}$ and two scalar degrees of freedom in the four-vector field ϵ_μ inducing gauge transformations: namely, ϵ_0 and the potential e such that $\epsilon_i = \partial_i e$. Hence we can use gauge transformations to cancel two scalar degrees of freedom, and build up two independent gauge-invariant scalars. One way to define them is through the two Bardeen potentials $\Phi_{\rm A}$ and Φ_H, defined by Bardeen (1980) as two integro-differential combinations of ψ, ϕ, b, μ.

- one can fix the gauge, i.e. introduce a condition such that the time slicing is unique. Then the number of independent solutions to the equations will be the same as in the gauge-invariant formalism. In any case, one can show that truly observable quantities are always independent of the gauge. Obtaining them after solving equations in the gauge-invariant formalism or in one particular gauge should not make any difference in practice.

A convenient gauge choice for a pedagogical introduction to CMB and matter scalar perturbations is the so-called Newtonian gauge or longitudinal gauge, in which one imposes that non-diagonal scalar perturbations of the metric vanish: $b = \mu = 0$. This prescription can be showed to fix a unique time slicing. In this gauge, adopting units such that $c = 1$ and using proper time t, the line element reads:

$$ds^2 = -(1+2\psi)dt^2 + (1-2\phi)a^2 d\vec{l}^{\,2} \ , \tag{5.6}$$

where $d\vec{l}^{\,2}$ stands for the cartesian measure $(dx^2 + dy^2 + dz^2)$ for a flat FL model, or for $(1 \pm kr^2)^{-1}dr^2 + r^2(\sin\theta^2 d\theta^2 + d\varphi^2)$ for an open/closed FL model in spherical coordinates. We are still free to redefine time (by definition, a time redefinition leaves the time slicing invariant). Lots of results in cosmological perturbation theory look simpler when using conformal time η, defined up to a constant by $d\eta = dt/a$. In this course we fix the constant in such way that $\eta \longrightarrow 0$ at the vicinity of the initial singularity, when $\rho \longrightarrow \infty$ (this prescription would not work if we were studying cosmological inflation, but in this course, we are not). Conformal time is convenient because photons traveling in a flat unperturbed FL universe along geodesics crossing the origin of the system of coordinates obey to $dr = d\eta$ (this comes from $a\,dl = dt$, i.e. from $ds = 0$ with the restriction $d\theta = d\phi = 0$). Hence conformal time is a measure of time based on the comoving distance travelled by a photon[1], and the comoving

[1] In a universe with non-zero spatial curvature, this remains true, provided that the comoving distance is defined not like r, but like $\chi \equiv \int (1 \pm kr^2)^{-1/2} dr$.

distance to a given object coincides with its "look-back conformal time". In this course we will use dots for derivatives with respect to proper time and primes for derivatives with respect to conformal time. The Hubble parameter (or expansion rate parameter) reads

$$H = \frac{\dot{a}}{a} = \frac{a'}{a^2} \ , \tag{5.7}$$

the Hubble radius is $R_H = 1/H$, and the condition that a Fourier mode of physical wavelength λ crosses the Hubble radius is

$$\lambda = R_H \quad \Leftrightarrow \quad \frac{2\pi}{k}a = \frac{1}{H} \quad \Leftrightarrow \quad k \sim aH = \frac{a'}{a} \ . \tag{5.8}$$

One advantage of the Newtonian gauge is that the gauge-invariant Bardeen potentials $\{\Phi_{\rm A}, \Phi_H\}$ reduce in this gauge to the metric perturbations $\{\phi, \psi\}$: the evolution of the Newtonian metric perturbations informs us directly on that of two gauge-invariant quantities. Other interesting properties of $\{\phi, \psi\}$ appear when writing the Einstein equations in the Newtonian gauge. The full Einstein equations linearized at first order in perturbations feature four equations relating scalar degrees of freedom. One of them, associated to the traceless longitudinal part of $\delta G^i_j = 8\pi G T^i_j$, gives (assuming a flat FL background)

$$\frac{2}{3}(k/a)^2(\phi - \psi) = 8\pi G \sum_x (\bar{\rho}_x + \bar{p}_x)\sigma_x \ , \tag{5.9}$$

where the index x runs over all the species contributing to the total stress-energy tensor. This means that when the universe contains only shearless components with $\sigma_x = 0$ (as would be the case in the presence of perfect fluids), the two metric perturbations are equal. Next, the Einstein equation $\delta G^0_0 = 8\pi G T^0_0$ gives (still assuming a flat FL background)

$$2a^{-2}\left[k^2\phi + 3\frac{a'}{a}\left(\phi' + \frac{a'}{a}\psi\right)\right] = -8\pi G \sum_x \bar{\rho}_x \delta_x \ . \tag{5.10}$$

The term on the right-hand side involves the total energy perturbation $\delta\rho_{\rm tot} = \sum_x \bar{\rho}_x \delta_x$. In the short scale (more precisely, sub-Hubble) limit, the term containing k^2 dominates the other terms in the square brackets, and we recover the Poisson equation

$$-\frac{k^2}{a^2}\phi = 4\pi G \, \delta\rho_{\rm tot} \ , \tag{5.11}$$

where the factor $-k^2/a^2$ represents the Fourier transform of the physical Laplace operator in an expanding universe. Note that in the Poisson equation one may have expected to see the generalized gravitational potential ψ instead of ϕ: however, in the sub-Hubble limit, the shear of individual components is usually either null or negligible, so that $\phi = \psi$.

5.1.3 Equations of motion

In the minimal cosmological scenario, the universe features several species with spatial fluctuations, described with different equations because of their distinct properties: cold dark matter (CDM) is non-relativistic and collisionless, neutrinos are ultra-relativistic and collisionless at the times of interest, baryons are non-relativistic and smoothly interpolating from a strongly coupled to decoupled regime, and finally photons are ultra-relativistic and interpolating between the same two regimes.

The equation of conservation of the total stress-energy tensor, $D_\mu T^\mu_\nu = 0$ (deriving from Bianchi identities), yields two scalar and two vector equations. The scalar ones are the conservation of energy equation and the Euler equation.

For any single component experiencing no interaction with other species (other than gravitational), the equation $D_\mu T^\mu_\nu = 0$ applies to the individual stress-energy tensor: it gives one continuity and one Euler equation for that component. The evolution of single component experiencing interactions is also given by the continuity and Euler equation, but with an extra source term accounting for stress-energy injection/leak caused by the interaction.

We have seen that the perturbations of each component x can be described by four variables $\{\delta_x, \delta p_x, \theta_x, \sigma_x\}$. Hence, in general, two equations of motion are not sufficient for closing the system. However:

- for a perfect fluid, microscopic interactions impose local thermodynamical equilibrium. The pressure is then isotropic[2], with $\sigma_x = 0$. In addition, pressure perturbations obey $\delta p_x = c_{\rm a}^2 \delta\rho_x$, where $c_{\rm a}$ is the adiabatic sound speed inferred from the equation of state of the fluid. If σ_x vanishes and δp_x is a function of $\delta\rho_x$, perturbations in the fluid are described by only two independent functions $\delta_x \equiv \delta\rho_x/\bar{\rho}_x$ and θ_x. If the collision term also vanishes or is specified, the two equations of motion (continuity and Euler) are sufficient for closing the system and computing the evolution of perturbations.

- for a decoupled or weakly interacting species, there are no such simplifications concerning the anisotropic stress and pressure perturbation. Hence, in general, the two equations inferred from stress-energy conservation are not sufficient. For such species, one has to use the more general Boltzmann equation, giving the evolution of each phase-space distribution:

$$\frac{d}{d\eta} f_x = \sum_y C_{xy}[f_x, f_y] \ , \tag{5.12}$$

  where the sum holds over the species y coupled with x. Each phase-space distribution can be decomposed into a background and perturbation part:

$$f_x(\eta, \vec{x}, \vec{p}) = \bar{f}_x(\eta, |\vec{p}|) + \delta f_x(\eta, \vec{x}, \vec{p}) \ , \tag{5.13}$$

  where $\vec{p}$ stands for momentum, $|\vec{p}|$ for its modulus, and the background part does not depend on the direction of $\vec{p}$ by assumption of isotropy.

- fully decoupled CDM is a particular case of a collisionless species with negligible velocity dispersion (the word "cold" refers precisely to this last assumption). Hence, it behaves in the same way as a pressureless perfect fluid, although in reality it has no interactions and should not be called a fluid. Since the velocity dispersion is negligible, in a given point, all particles share the same velocity, imposed by gravitational flows (while for non-cold collisionless species, the velocity would get two contributions, one from gravitational flows, and one from the phase-space distribution function). Hence the anisotropic pressure vanishes (see the previous footnote). The pressure perturbation δp_x is also related to the velocity dispersion in a given point, and can be neglected with respect to the density perturbation in the CDM case. Hence CDM is formally equivalent to a perfect

[2] An anisotropic shear stress $\sigma \neq 0$ reflects the fact that in each given point, particles travel with different velocities (due to some intrinsic velocity dispersion and/or a superposition of several flows in phase space), leading to anisotropic pressure. This contradicts the assumption of a perfect fluid, in which local interactions result in a unique bulk velocity (after coarse-graining over microscopic scales), and erase anisotropic pressure.

fluid with no anisotropic stress and no pressure perturbation (or in other words, with a sound speed $c_s = 0$). In that case, the two equations of motion inferred from $D_\mu T^\mu_\nu = 0$ are sufficient, like for a fluid.

Gravitational interactions between species are accounted by the presence of metric perturbations in each equation of motion (more specifically, terms in $k^2\psi$ accounting for gravitational forces, and terms in ϕ' accounting for dilation effects, i.e. for local distortions of the scale factor with respect to a). Hence, in order to close the full system of equations, we still need two independent relations, to be chosen among the four scalar Einstein equations: they provide the value of $k^2\psi$ and ϕ' at each time, as a function of all matter fields.

5.1.4 Initial conditions

We wish to study the evolution of matter perturbations, starting from some early time at which all Fourier modes of interest (those which are observable in the CMB spectrum and in the matter power spectrum on linear or mildly non-scales) are still outside the Hubble radius. Indeed, super-Hubble modes experience a trivial evolution, unaffected by small scale interactions (Thomson or Coulomb scattering, usual gravitational force $\vec{\nabla}\phi$, etc.). Hence, the perturbations evaluated at some arbitrary time but on super-Hubble scales reflect directly the mechanism responsible for the formation of perturbations in the very early universe. In the standard cosmological model, these initial conditions can be inferred from inflation.

Typically, a good time for setting initial conditions is when the redshift $z = a_0/a - 1$ is of the order of 10^5: at this time, all comoving scales that are observable in the CMB and linear matter power spectrum still verify $k \ll aH$.

It is crucial to understand that, as long as the background cosmology is assumed to be of the FL type, the perturbed stress-energy momentum tensor δT^ν_μ must be diagonal on super-Hubble scales. Indeed, the background tensor $\bar{T}^\nu_\mu$ is diagonal, and of the form: $\mathrm{diag}(-\bar{\rho}, \bar{p}, \bar{p}, \bar{p})$. This can be showed to be the most general assumption compatible with homogeneity and isotropy. Let us assume that we Taylor-expand δT^ν_μ in powers of the variable $(k\eta)$. For any power-law scale factor, aH is given by $1/\eta$ times a factor of order one. Hence the limit $k\eta \ll 1$ represents precisely the super-Hubble limit. In the Taylor expansion, the zero mode should share the same properties as the background solution, and be diagonal. Higher order terms account for contributions to δT^ν_μ growing with time, and possibly becoming important around the time of Hubble crossing.

The total scalar perturbations $\delta\rho$ and δp are the only ones preserving the diagonal form of δT^ν_μ: we conclude that they are the only ones that do not vanish at order zero in the $(k\eta)$ expansion. Using stress-energy conservation equations, one can show that the part of δT^ν_μ associated with the velocity divergence is of order one in $(k\eta)$, while the part associated to the total anisotropic stress is of order two.

Suppose that the universe contains initially N uncoupled perfect fluids[3], with N known sound speeds $c_x^2 = \delta p_x/\delta\rho_x$. There are $2N$ independent initial conditions, corresponding to possible initial values of each δ_x and each δ'_x. Importantly, in the $2N$-dimensional basis of IC's, one basis vector is very special, as we shall see below.

Before studying perturbations, one should have solved for the background evolution: all background quantities should be known, including for instance

[3]The discussion presented in this section could be generalized to N coupled species, not all of them being perfect fluids: the conclusions would not change qualitatively, and we restrict here to N uncoupled fluids for simplicity.

the density $\bar{\rho}_x(\eta)$ and pressure $\bar{p}_x(\eta)$ of each species x. Now, let us assume that the real universe is perturbed initially by a single degree of freedom (one may say, by a single initial time shifting function). This is the case in single-field inflationary cosmology: during inflation, there is a single clock (the inflaton), and perturbations arise from a single time shifting function (the inflaton perturbation).

As long as we are dealing with super-Hubble modes, we can neglect microscopic interactions and say that the evolution in each point (in fact, in each Hubble patch) is still given by homogeneous cosmology, taking this shift function $\delta\eta(\vec{x})$ into account:

$$\begin{aligned} \forall x, \qquad \rho_x(\eta,\vec{x}) &= \bar{\rho}_x\left(\eta+\delta\eta(\vec{x})\right) = \bar{\rho}_x(\eta) + \bar{\rho}_x'(\eta)\,\delta\eta(\vec{x}) \\ p_x(\eta,\vec{x}) &= \bar{p}_x\left(\eta+\delta\eta(\vec{x})\right) = \bar{p}_x(\eta) + \bar{p}_x'(\eta)\,\delta\eta(\vec{x}) \end{aligned} \tag{5.14}$$

where in the last equalities, terms of order two or higher in $\delta\eta$ have been neglected. The above ansatz restricts a lot the choice of possible initial conditions. Indeed, you will show in the exercise sessions that it implies a relation between all density fluctuations,

$$\forall x, y, \qquad \frac{\delta\rho_x}{\bar{\rho}_x+\bar{p}_x} = \frac{\delta\rho_y}{\bar{\rho}_y+\bar{p}_y}\ . \tag{5.15}$$

It shows that in presence of such initial conditions, everything is fixed up to a single function of $\vec{x}$. Let us take the example of a universe containing only photons, baryons, cold dark matter and neutrinos. We can use the fact that for non-relativistic species $\bar{p}_x \ll \bar{\rho}_x$, while for ultra-relativistic ones $\bar{p}_x = \bar{\rho}_x/3$. Hence, if one function is known — for instance, $\delta_\gamma(\vec{x})$ at initial time — the others can be derived from

$$\delta_{\rm b} = \delta_{\rm cdm} = \frac{3}{4}\delta_\nu = \frac{3}{4}\delta_\gamma\ . \tag{5.16}$$

In the exercises, you will also show that the ansatz of Eqs. (5.14) implies

$$\delta p_{\rm tot}(\eta,\vec{x}) = c_{\rm s}^2(\eta)\,\delta\rho_{\rm tot}(\eta,\vec{x})\ , \tag{5.17}$$

where the squared total sound speed $c_{\rm s}^2$ can be easily expressed as a weighted average over the squared adiabatic sound speed of individual components. Hence, the total matter content resulting from the sum of all components also features an adiabatic sound speed. For that reason, such initial conditions are usually called *adiabatic initial conditions*.

More general initial conditions not obeying Eqs. (5.14) can be expanded on different bases. A famous basis is formed by (i) the adiabatic mode; (ii) $(N-1)$ non-decaying isocurvature modes, getting their name from the property that for each of them, the total density (and spatial curvature) perturbations vanish asymptotically in the super-Hubble limit, while two species have opposite density perturbations compensating each other; (iii) N decaying modes that are irrelevant for most purposes.

It is clear from the previous discussion that non-adiabatic initial conditions should only be considered when assuming that primordial perturbations are generated by more than one degree of freedom (for instance, two inflaton fields, or one inflaton and one axion, etc.). The assumption of several degrees of freedom is necessary but not sufficient. The primordial universe may contain a mixture of adiabatic and isocurvature modes until a time at which all species are brought in thermal equilibrium. At that time, if we further assume that all chemical potentials vanish, the perturbations of each species can be inferred from those of temperature, $T(\eta,\vec{x}) = \bar{T}(\eta) + \delta T(\eta,\vec{x})$. Then, temperature plays

precisely the role of a single time-shifting function. Any non-adiabatic initial condition is washed out and becomes irrelevant. Hence, isocurvature modes can be observable — for instance in the CMB — only under additional assumptions. For instance, one species carrying isocurvature perturbations might remain out of equilibrium at all times, or might feature a chemical potential with spatial fluctuations. There exist a few non-minimal, but still reasonable scenarios featuring isocurvature perturbations (with axions, curvatons etc.). Ultimately, the presence of isocurvature modes is to be tested with observations. Current CMB observations put strong limits on the amplitude of such modes, and prefer purely adiabatic initial conditions. We will restrict to this case in the rest of this course.

For adiabatic initial conditions, we found the relation (5.15) holding between density fluctuations. Also, if we do not consider the case of species with a non-zero anisotropic stress, we can assume that $\phi = \psi$ at initial time. By plugging Eqs. (5.16), (5.17) and $\phi = \psi$ into the Einstein equations, one is led to a second-order differential equation for ψ only. During the radiation dominated era, one can show that this equation has two solutions, one constant in time, and one decaying. The (00) Einstein equation then shows that the constant solution is related to density fluctuations through

$$-2\psi = -2\phi = \delta_{\rm tot} \simeq \delta_\gamma = \text{constant} \ . \tag{5.18}$$

Hence, in the Newtonian gauge and for adiabatic initial conditions, super-Hubble metric fluctuations and density fluctuations are static. In fact, on super-Hubble scales, there can only be some time evolution when the universe changes of total equation of state (or equivalently, of expansion law). This is the case at the time of equality between radiation and matter. During matter domination and on super-Hubble scales, they freeze out again. Then, the relation $-2\psi = -2\phi = \delta_{\rm tot}$ and eq. (5.16) are still satisfied, but now $\delta_{\rm tot} \simeq \delta_{\rm b} = \delta_{\rm cdm} = \frac{3}{4}\delta_\gamma$. This detail will become important when studying the Sachs-Wolfe effect in section 5.2.3.

5.1.5 Power spectra and transfer functions

The theory of cosmological perturbations is a stochastic theory: the fluctuations of a given quantity in a given point, $A(\eta, \vec{x})$, obey a distribution of probability. As long as we stick to linear perturbation theory, there is a "linear transport of probability" from one time to another time. Let the probability of A in a given point at time η_1 be $\mathcal{P}_1(A)$. In the same point and at η_2, the linear evolution would have transformed each value A into αA, where α is the linear growth (or decrease) factor of A between η_1 and η_2. So the probability of A at time η_2 is given by $\mathcal{P}_2(A) = \mathcal{P}_1(A/\alpha)$. This "linear transport of probability" implies that the linear evolution respects the shape of the probability distribution, and rescales all its statistical moments of order n by α^n. In particular, if the initial probability is Gaussian, the statistics will remain Gaussian at any later time, and the evolution of the system can be formulated as an evolution of the root mean square of all quantities. In summary, as long as we assume linear perturbations with Gaussian initial conditions, our goal is to solve for the evolution of the root mean squares of the fluctuations.

The equal-time 2-point correlation function of any quantity A in real space is given by

$$\langle A(\eta, \vec{x}) A(\eta, \vec{x}') \rangle \equiv \xi(\eta, \vec{x}, \vec{x}') \stackrel{SHI}{=} \xi(\eta, |\vec{x}' - \vec{x}|) \ , \tag{5.19}$$

where the last equality holds as a consequence of Statistical Homogeneity and Isotropy (SHI, assumed to hold in a perturbed FL universe). Indeed, the correlation function should be invariant under spatial translations and rotations.

We can go to Fourier space and use the same letter to denote the Fourier transform of A. If A is real, $A(\eta, -\vec{k}) = A^*(\eta, \vec{k})$. It is easy to show that the previous equation (using the assumption of SHI) implies that the two-point correlation function in Fourier space reads

$$\langle A(\eta, \vec{k}) A^*(\eta, \vec{k}') \rangle \stackrel{SHI}{=} \delta_D(\vec{k}' - \vec{k}) P_{\rm A}(k) \ , \tag{5.20}$$

where δ_D is the Dirac distribution. Here, statistical homogeneity is responsible for the fact that the two-point correlation vanishes for $\vec{k} \neq \vec{k}'$, and statistical isotropy for the fact that $P_{\rm A}$ only depends on the modulus $k = |\vec{k}|$. The function $P_{\rm A}(k)$ is usually called the power spectrum of A. Some authors prefer to refer to the "dimensionless power spectrum", defined as

$$\mathcal{P}_{\rm A}(k) \equiv \frac{k^3}{2\pi^2} P_{\rm A}(k) \ . \tag{5.21}$$

The reason is that typical expressions for the average of various quantitites in real space consist in the convolution of the power spectrum with some window function $f(k)$, like in

$$\int \frac{d^3\vec{k}}{(2\pi)^3} P_{\rm A}(k) f(k) = \int \frac{4\pi k^2 dk}{(2\pi)^3} P_{\rm A}(k) f(k) = \int d{\rm log}k \ \mathcal{P}_{\rm A}(k) f(k) \ . \tag{5.22}$$

Hence the dimensionless spectrum $\mathcal{P}_{\rm A}(k)$ stands for the weight of each logarithmic interval in the integral. The term "scale-invariant spectrum" refers to $\mathcal{P}_{\rm A}(k)$ being independent of k, i.e. $P_{\rm A}(k) \propto k^{-3}$.

We know that for adiabatic initial conditions, all perturbations are related to each other through Eqs. (5.16, 5.18). Hence, with Gaussian adiabatic initial conditions, specifying the primordial power spectrum of one quantity is sufficient for knowing everything about the system. For instance, if we assume that the power spectrum of the metric perturbation ψ is a given function $P_\psi(k)$, then we infer from Eqs. (5.16, 5.18) that the photon and baryon primordial spectra are given by $P_\gamma(k) = 4P_\psi(k)$ and $P_{\rm b}(k) = \frac{9}{16} P_\gamma(k) = \frac{9}{4} P_\psi(k)$.

By convention, the primordial spectrum is usually given for the variable $\mathcal{R}$, which represents the spatial curvature perturbation on one initial comoving hypersurface (i.e. an hypersurface orthogonal to the energy flux of the total cosmic fluid in each point). The advantage of using this quantity is that it is conserved on super-Hubble scales for adiabatic initial conditions (while ϕ and ψ get rescaled when the equation of state of the universe changes, e.g. at radiation-matter equality). In the Newtonian gauge, $\mathcal{R}$ relates to ψ and to the total density perturbation through

$$\mathcal{R} = \psi - \frac{1}{3} \frac{\delta\rho_{\rm tot}}{\bar{\rho}_{\rm tot} + \bar{p}_{\rm tot}} \ . \tag{5.23}$$

The power spectrum of a given quantity at some arbitrary time can be decomposed into two parts, one accounting for initial conditions, and one accounting for linear evolution with time:

$$\langle A(\eta, \vec{k}) A^*(\eta, \vec{k}') \rangle = \delta_D(\vec{k}' - \vec{k}) \left[\frac{A(\eta, \vec{k})}{\mathcal{R}(\vec{k})} \right]^2 P_{\mathcal{R}}(k) \ . \tag{5.24}$$

In the above equation, there is no time argument for $\mathcal{R}(\vec{k})$ since this quantity is conserved on super-Hubble scales. We just assume that $\mathcal{R}$ is evaluated at a time such that $k \ll aH$ for all modes of interest. In a FL universe, the equations of motion of all perturbations respect isotropy, and do not depend

on the direction of the wavevector $\vec{k}$. Hence the ratio between brackets in the last equation is a function of k, not $\vec{k}$. This function accounts for the linear evolution, independently of initial conditions. It is called the "transfer function" of A. In this course, we will denote transfer functions with the same letter as the perturbations themselves, but with the modulus of k as an argument, i.e.

$$A(\eta, k) \equiv \frac{A(\eta, \vec{k})}{\mathcal{R}(\vec{k})} \ . \tag{5.25}$$

In summary of this section, solving for cosmological perturbations (in a model with adiabatic Gaussian IC's) amounts in

- postulating a primordial spectrum, or calculating it within the framework of a model, for instance of inflation;
- solving the equations of motion of all perturbations, with quantities normalized initially to $\mathcal{R}(\vec{k}) = 1$, in such a way to obtain all transfer functions $A(\eta, k)$ and to infer the evolution of the various root mean squares.

5.2 CMB temperature anisotropies

From now on, we will assume for simplicity that the universe is flat.

5.2.1 Photon scattering rate

CMB physics consists is the study of electron, baryon and photon perturbations on cosmological scales, taking into account their gravitational coupling with collisionless species such as decoupled neutrinos and CDM. Electrons and baryons carry opposite electric charges and are coupled to each other through very efficient Coulomb scattering processes. Electrons and photons are coupled through Thomson scattering, which is the limit of Compton scattering when electrons are non-relativistic and photons carry a smaller energy than the rest mass of the electron. Then, the scattering process results mainly in a deflection of the photon, with a negligible transfer of energy between the two particles. The leading interaction between baryons and photons is the gravitational one.

In units such that $c = 1$, the Thomson scattering rate (with respect to conformal time) is given by $\Gamma = \sigma_T a n_e x_e$, where σ_T is the Thomson scattering rate, a is the scale factor, n_e is the total electron number density (scaling like a^{-3} due to dilution), and x_e is the ionized electron fraction. The product an_e scales like a^{-2}. The ionized fraction is close to one at high energy. Then, at the time of recombination between electrons and nuclei (around $z \sim 1080$), which falls at the beginning of matter domination, n_e drops abruptly to very small values. This causes Thomson scattering to become suddenly very inefficient, and photons to decouple from electrons.

Hence photon decoupling is the story of Thomson scattering becoming inefficient, while Coulomb scattering remains very strong. For that reason, one can describe electrons and baryons as a single tightly-coupled fluid. People often refer only to baryons for simplicity. At early time, the full system of (electrons)-baryons-photons is also tightly coupled, but later on, it splits progressively into two collisonless species, (electrons)-baryons on the one hand, photons on the other hand.

The thermodynamical description of recombination is very technical, due to the different energy level of atoms (in particular, of hydrogen). In order to understand CMB anisotropies, we only need to describe the main results of recombination studies at a very qualitative level.

- The free electron fraction $x_e(\eta)$ starts from one at high redshift. It decreases sharply at the recombination time (corresponding to $z \sim 1080$ or $T \sim 0.3$ eV), and freezes at a very small value (due to departure from thermal equilibrium). Around $z \sim 10$, star formation causes a reionization of the universe, and x_e goes up again to one.

- The Thomson scattering rate $\Gamma = \sigma_T a n_e x_e$ evolves like $a^{-2} x_e$. Before recombination, $\Gamma \gg \frac{a'}{a}$, and the universe is opaque. The sudden drop of x_e at recombination renders the universe transparent: $\Gamma \ll \frac{a'}{a}$. Due to the dilution factor coming from n_e, Γ remains much smaller than $\frac{a'}{a}$ even during reionization, and the universe keeps being transparent (this is why despite of reionization, most photons emitted at recombination do not interact anymore, and allow us to observe anisotropies on the last scattering surface).

- The optical depth $\tau(\eta) \equiv \int_\eta^{\eta_0} d\eta \Gamma(\eta)$ represents the opacity of the universe at a given time, when seen from today (when $\eta = \eta_0$). It tends to infinity when $\eta \longrightarrow 0$, falls below one at recombination and stabilize at a value of the order of 0.1 between recombination and reionization. After reionization it decreases smoothly and reaches zero today by definition.

- The visibility function $g(\eta) \equiv -\tau' e^{-\tau}$ gives the probability that a CMB photon seen today experienced its last scattering at time η. It starts from negligible values at high redshift (suppressed by the $e^{-\tau}$ factor). It has a narrow spike around the time of recombination, and then it falls again to negligible values due to the smallness of τ' between recombination and reionization. It develops a second smaller and wider spike around reionization. This function shows that most CMB photons did not interact between the last scattering surface and today, while a minority rescattered at reionization. The width of the recombination spike gives an indication on the thickness of the last scattering "surface".

- The diffusion length $\lambda_{\rm d}(\eta)$ is an important quantity for understanding the damping of temperature anisotropies on small scales. At any given time, the comoving mean free-path (mfp) of photons is given by $r_{\rm mfp} = 1/\Gamma(\eta)$ (still in units where $c = 1$). If the photons experience a random walk analogous to Brownian motion in a gas, the comoving distance over which they travel between time $\eta_{\rm ini}$ and η can be approximated by

$$r_{\rm d} \sim \left[\int_{\eta_{\rm ini}}^{\eta_0} d\eta \; \Gamma \; r_{\rm mfp}^2\right]^{1/2} = \left[\int_{\eta_{\rm ini}}^{\eta_0} d\eta \; \Gamma^{-1}\right]^{1/2} . \tag{5.26}$$

  This physical diffusion length of photons is given in this approximation by $\lambda_{\rm d} \simeq a r_{\rm d}$. It grows quickly from very small to very large scales (comparable to the Hubble scale) around the time of recombination.

5.2.2 Boltzmann equation

Since photons decouple from baryons near the recombination time, we cannot describe them with fluid equations, and need to solve the Boltzmann equation

$$\frac{d}{d\eta} f = C\left[f, f_e\right] \tag{5.27}$$

at order one in perturbations. The right-hand side stands for the photon-electron coupling due to Thomson scattering. As explained in the previous

section, electrons and baryons are so tightly coupled that it makes no difference to think of this term as a photon-electron or photon-baryon coupling term. Solving this equation is involved because the photon phase-space distribution f involves many arguments, $f(\eta, \vec{x}, \vec{p})$. Fortunately one can reduce the dimensionality of the problem. First, we notice that as long as photons are in thermal equilibrium with electrons (and hence with baryons), they are entirely described in any point by the local value of the equilibrium temperature $T(\eta, \vec{x})$. The phase space distribution is then of the Bose-Einstein form:

$$f(\eta, \vec{x}, \vec{p}) = \frac{1}{e^{\frac{p}{T(\eta,\vec{x})}} - 1} \ . \tag{5.28}$$

It can be expanded into a background part and a first-order perturbation, $f = \bar{f} + \delta f$, with:

$$\bar{f}(\eta, p) = \frac{1}{e^{\frac{p}{\bar{T}(\eta)}} - 1} \ , \qquad \delta f(\eta, \vec{x}, \vec{p}) = \frac{d\bar{f}}{d\log p} \ \frac{\delta T(\eta, \vec{x})}{\bar{T}(\eta)} \ . \tag{5.29}$$

We see that in the tightly-coupled regime we could replace the variable $f(\eta, \vec{x}, \vec{p})$ by the lower-dimensional variable $\Theta(\eta, \vec{x}) \equiv [\delta T(\eta, \vec{x})/\bar{T}(\eta)]$. The Boltzmann equation leads to an equation of motion for $\Theta(\eta, \vec{x})$.

At later times, when photons decouple, the shape of the Bose-Eisntein distribution of photons is preserved. This can be infered from the geodesic equation. This equation tells how the individual momentum p of photons evolve when they are decoupled and they travel in the perturbed universe. In the Newtonian gauge, it reads

$$\frac{d(a\,p)}{d\eta} = -a\,p\,\phi' - a\,\epsilon\,\hat{n}\cdot\vec{\nabla}\psi \ . \tag{5.30}$$

The left-hand side represents the time evolution of the product (ap) for a photon of momentum p traveling over a geodesic. In a perfectly homogeneous universe, the photon would only experience the average cosmological redshifting, $p \propto a^{-1}$, and the product (ap) would be conserved. Due to presence of perturbations in the universe, photons experience gravitational interactions and the product (ap) varies. The first term $-a\,p\,\phi'$ accounts for dilation, i.e. for the fact that locally, the expansion of the universe is a bit more advanced or delayed than the average (remember that $a(1+\phi)$ can be seen as the "local scale factor"). This means that the redshifting of the photon is also a bit more advanced or delayed locally. The second term accounts for the gravitational blueshifting of photons falling in gravitational potential wells (or redshifting of those leaving potential wells). The energy $\epsilon \equiv \sqrt{p^2 + m^2}$ can be simply replaced by p for massless photons. In that case we can rewrite the geodesic equation as

$$\frac{d\ln(a\,p)}{d\eta} = -\phi' - \hat{n}\cdot\vec{\nabla}\psi \ . \tag{5.31}$$

The fact that p disappeared from the right-hand side is crucial: it shows that photons of different momenta traveling in the perturbed universe along a given geodesic all experience the same relative momentum variation. The consequence is that there can be no distortions of the Bose-Einstein shape of f. However, photons traveling along different geodesics and in different directions experience different redshifting. Hence the distribution function acquires a dependence on one extra argument, the direction $\hat{n}$ of propagation ($\hat{n} \equiv \vec{p}/p$):

$$f(\eta, \vec{x}, \vec{p}) = \frac{1}{e^{\frac{p}{T(\eta,\vec{x},\hat{n})}} - 1} \ . \tag{5.32}$$

We can perform the same decomposition in terms of background and perturbations as in eq. (5.29). The difference is that $\Theta = \frac{\delta T}{T}$ is now a function of $(\eta, \vec{x}, \hat{n})$. The Boltzmann equation can now be used to derive an equation of motion for $\Theta(\eta, \vec{x}, \hat{n})$.

If we work in Fourier space, we can derive the equation of motion for the function $\Theta(\eta, \vec{k}, \hat{n})$. Because of the statistical isotropy of the FL universe, this equation does not depend explicitly on $\vec{k}$ nor $\hat{n}$, since there is no preferred direction: it depends only on the direction of propagation relatively to the considered wavenumber , i.e. on the product $(\vec{k} \cdot \hat{n})$. Hence the equation of motion can be written in terms of k and of the angle θ such that $(\vec{k} \cdot \hat{n}) = k \cos\theta$. The initial conditions for Θ do depend on the wavevector $\vec{k}$ (since each mode gets random initial conditions), but they also depend only on θ rather than $\hat{n}$, for a reason that will become clear in the paragraphs below. Hence we can entirely eliminate $\hat{n}$ from the problem, and solve the equation of motion for $\Theta(\eta, \vec{k}, \theta)$.

Finally, we can expand the temperature anisotropy with respect to θ using a Legendre transformation:

$$\Theta(\eta, \vec{k}, \theta) = \sum_l (-i)^l (2l+1) \Theta_l(\eta, \vec{k}) P_l(\cos\theta) \ . \tag{5.33}$$

Here the Θ_l's are the temperature anisotropy multipoles, and P_l the Legendre polynomials. It can be shown that the monopole Θ_0 is related to the photon density fluctuation δ_γ in a given point, the dipole Θ_1 to its velocity divergence θ_γ, and the quadrupole Θ_2 to its anisotropic stress σ_γ. The Boltzmann equation can be written as an infinite hierarchy of equations of motion for the coupled multipoles Θ_l.

Actually, the equation of motion for Θ takes a striking form when written in real space, before Fourier and Legendre expansions:

$$\Theta' + \hat{n} \cdot \vec{\nabla}\Theta - \phi' + \hat{n} \cdot \vec{\nabla}\psi = -\Gamma \left(\Theta - \Theta_0 - \hat{n} \cdot \vec{v}_e \right) \ . \tag{5.34}$$

We recall that Γ is the conformal Thomson scattering rate, Θ_0 the temperature monopole (i.e. the average of $\Theta(\eta, \vec{x}, \hat{n})$ over all directions $\hat{n}$), and $\vec{v}_e$ the bulk velocity of electrons, equal to that of baryons due to tight Coulomb interactions. We recall that the variable θ_b is defined as the divergence of $\vec{v}_\mathrm{b} = \vec{v}_e$.

This equation is illuminating, since it shows that at early times, when photons, electrons and baryons form a tightly-coupled fluid, the fact that Γ is huge forces Θ to evolve in such way that the parenthesis on the right-hand side vanishes. When this is the case, Θ can only have two non-zero component: a monopole Θ_0, and a dipole equal to $\hat{n} \cdot \vec{v}_\mathrm{b}$. The condition on the dipole can be written equivalently as $\theta_\gamma = \theta_\mathrm{b}$.

This should remind us of the discussion of Sec. 5.1.3, when we noticed that a perfect fluid can be described in terms of δ_x, c_s^2 and θ_x only, with a vanishing anisotropic stress σ_x. Here, we see concretely how this conclusion emerges from the Boltzmann equation in the tightly-coupled regime: Θ_2 (related to σ_γ) and all higher multipoles vanish; the photon perturbations can be described in terms of only two independent variables Θ_0 and Θ_1 (or δ_γ and θ_γ), like in a perfect fluid.

The fact that in the tightly-coupled limit the system is driven towards $\theta_\gamma = \theta_\mathrm{b}$ simply shows that the interaction imposes a common bulk velocity to photons, baryons and electrons, as it should be the case in any tightly-coupled fluid (note that we are referring to bulk velocities, not to individual velocities of particles, which are still equal to c for interacting photons).

In the tightly-coupled regime, the dipole component of $\Theta(\eta, \vec{x}, \hat{n})$ is given by $\hat{n} \cdot \vec{v}_\mathrm{b}$, with $\theta_\mathrm{b} \equiv \vec{\nabla} \cdot \vec{v}_\mathrm{b}$. Hence in Fourier space this component reads

$i(\hat{n}\cdot\vec{k})k^{-2}\theta_{\rm b} = i(\cos\theta)k^{-1}\theta_{\rm b}$. This justifies the fact that initial conditions for Θ in Fourier space depend on θ, but not on the two degrees of freedom of $\hat{n}$. As mentioned above, this property is preserved by the isotropic equations of motion, so that at all times we can study Θ as a function of the arguments $(\eta, \vec{k}, \theta)$ instead of $(\eta, \vec{k}, \hat{n})$.

5.2.3 Temperature anisotropy in a given direction

The map of temperature anisotropies that we observe today ($\eta = \eta_0$) in our location of the universe ($\vec{x} = \vec{o}$ with a proper choice of origin) when looking in a direction $\hat{n}$ is represented mathematically by

$$\frac{\delta T}{T}(\hat{n}) = \Theta(\eta, \vec{o}, -\hat{n}) \tag{5.35}$$

(since in a direction $\hat{n}$, we see photons traveling towards $-\hat{n}$). Our scope is now to relate this quantity to perturbations on the point of the last scattering surface seen in the same direction $\hat{n}$. This can be done by integrating the Boltzmann equation along the corresponding line-of-sight.

A good starting point consists in computing the total derivative of the product $e^{\tau}(\Theta+\psi)$ along the trajectory of photons between the last scattering surface and the observer. The reason for choosing this product rather than just Θ will become clear in a few lines: the derivative of this term will be easy to simplify, using the Boltzmann equation.

The total derivative of an arbitrary function $\mathcal{F}$ of $(\eta, \vec{x}, \hat{n})$ along the trajectory of photons going in a direction $\hat{n}$ reads

$$\frac{d}{d\eta}\mathcal{F}(\eta, \vec{x}, \hat{n}) = \mathcal{F}' + \frac{dx_i}{d\eta}\frac{\partial \mathcal{F}}{\partial x_i} + \frac{dn_i}{d\eta}\frac{\partial \mathcal{F}}{\partial n_i} \ . \tag{5.36}$$

If $\mathcal{F}$ is of order one in perturbations, the first two terms on the right-hand side are also of order one. Instead the last term is of order two, since $\frac{dn_i}{d\eta}$ is of order one (this is clear from the fact that in an unperturbed universe, photons would travel in straight line with $\frac{dn_i}{d\eta} = 0$). Hence we can drop this term in first-order perturbation theory, and for $\frac{dx_i}{d\eta}$ we only need to keep the zero-th order contribution, that can be computed assuming a homogeneous universe:

$$\frac{dx_i}{d\eta} = \hat{n} \ . \tag{5.37}$$

The previous relation is just telling that photons are traveling in the direction $\hat{n}$ and at the velocity of light: hence, in units where $c = 1$, $d\vec{x}^2 = d\eta^2$. In summary, the total derivative of $\mathcal{F}$ is given at first order by

$$\frac{d}{d\eta}\mathcal{F}(\eta, \vec{x}, \hat{n}) = \mathcal{F}' + \hat{n}\cdot\vec{\nabla}\mathcal{F} \ . \tag{5.38}$$

Let us now replace the generic function $\mathcal{F}$ by

$$\mathcal{F}(\eta, \vec{x}, \hat{n}) = e^{-\tau(\eta)}\left(\Theta(\eta, \vec{x}, \hat{n}) + \psi(\eta, \vec{x})\right) \ . \tag{5.39}$$

The total derivative of this function reads

$$\frac{d}{d\eta}\left[e^{-\tau}(\Theta+\psi)\right] = e^{-\tau}\left(\Theta' + \psi' + \hat{n}\cdot\vec{\nabla}(\Theta+\psi)\right) - \tau' e^{-\tau}(\Theta+\psi) \ . \tag{5.40}$$

We now use the linearized Boltzmann equation (5.34) and the fact that $\tau' = -\Gamma$ to write the result as

$$\frac{d}{d\eta}\left[e^{-\tau}(\Theta+\psi)\right] = -e^{-\tau}\tau'(\Theta_0 + \psi + \hat{n}\cdot\vec{v}_{\rm b}) + e^{-\tau}(\phi' + \psi') \ . \tag{5.41}$$

Finally, using the definition of the visibility function g given in Sec. 5.2.1, we get

$$\frac{d}{d\eta}\left[e^{-\tau}(\Theta+\psi)\right] = g\left(\Theta_0+\psi+\hat{n}\cdot\vec{v}_{\rm b}\right)+e^{-\tau}(\phi'+\psi') \ . \tag{5.42}$$

We can integrate this relation along the line of sight, i.e. along a straight line seen by the observer in a given direction $-\hat{n}$ (since the photons go in the direction $\hat{n}$), starting from an early time *before* recombination (such that $e^{-\tau(\eta_{\rm ini})}\simeq 0$) until the present time at which photons reach the observer (with by definition $e^{-\tau(\eta_0)}=1$). The result reads

$$(\Theta+\psi)|_{\rm obs} = \int_{\eta_{\rm ini}}^{\eta_0} d\eta\left[g\left(\Theta_0+\psi+\hat{n}\cdot\vec{v}_{\rm b}\right)+e^{-\tau}(\phi'+\psi')\right] \ , \tag{5.43}$$

where the notation $|_{\rm obs}$ means "evaluated at the observer location, along this line of sight", i.e. at the coordinate $(\eta_0,\vec{o},\hat{n})$.

We can gain further intuition from this equation if we use the instantaneous decoupling approximation, in which all photons are assumed to decouple precisely at the time $\eta_{\rm dec}$. In this limit, we can replace the visibility function g by the Dirac function $\delta_D(\eta-\eta_{\rm dec})$, and $e^{-\tau}$ by the Heaviside function $H(\eta-\eta_{\rm dec})$. Note that the approximation $g(\eta)=\delta_D(\eta-\eta_{\rm dec})$ is correctly normalized, since the definition of g implies $\int d\eta\, g(\eta)=1$. In the instantaneous decoupling limit, eq. (5.43) reads:

$$(\Theta+\psi)|_{\rm obs} = \left(\Theta_0+\psi+\hat{n}\cdot\vec{v}_{\rm b})\right)|_{\rm dec}+\int_{\eta_{\rm dec}}^{\eta_0} d\eta\,(\phi'+\psi') \ , \tag{5.44}$$

where the notation $|_{\rm dec}$ means "evaluated on the last scattering surface, along this line of sight", i.e. at the coordinate $(\eta_{\rm dec},-r_{\rm dec}\hat{n},\hat{n})$ (here $r_{\rm dec}$ is the comoving radius of the last scattering surface). Let us now give the interpretation of each term in this crucial equation.

First, $\Theta|_{\rm obs}$ is the temperature anisotropy measured by the observer in the direction $-\hat{n}$, while $\Theta_0|_{\rm dec}$ is the temperature anisotropy in the point of the last scattering surface seen in the same direction. If only these two terms where present, this relation would simply tell us that that the temperature anisotropy seen today in a given direction is equal to the intrinsic anisotropy in the point where the observed photons last scattered.

Second, the term $\hat{n}\cdot\vec{v}_{\rm b}|_{\rm dec}$ stands for the correction to this temperature coming from by the usual Doppler. Indeed, this correction is caused by the velocity of the baryon-photon fluid (that we assumed to be tightly coupled until $\eta_{\rm dec}$) projected along the line of sight.

Next, we expect a correction from gravitational effects. The redshifting and blueshifting of the photons traveling along gravitational potential fluctuations should affect the observed temperature anisotropy relatively to the intrinsic one. It turns out that if the gravitational potential was constant in time (but, of course, not in space), this effect would be conservative, and would only depend on $\psi|_{\rm obs}-\psi|_{\rm dec}$. This explains the second term on the left-hand and right-hand sides. But if ψ varies in time, the effect is not conservative anymore: intuitively, the amount of blueshifting and redshifting experienced by photons traveling across a potential well do not compensate each other if the gravitational well gets deeper between the time at which the photon enters and leaves the well. This explains the addition term $\int d\eta\,\psi$. A similar effect is caused by dilation effects along the line-of-sight, and contributes like $\int d\eta\,\phi$.

Finally, we can drop the second term on the left-hand side, because this term represents only a tiny *isotropic* correction to the observed anisotropies. It is impossible to measure it with the CMB map only, because it is formally

equivalent to a redefinition of the average temperature $\bar{T}$, but only by a tiny amount of the order of $10^{-5}\bar{T}$.

Let us write once more our result, dropping this unobservable correction, and grouping the terms in a particular way:

$$\Theta|_{\rm obs} = \underbrace{(\Theta_0 + \psi)|_{\rm dec}}_{\rm SW} + \underbrace{\hat{n}\cdot\vec{v}_{\rm b}|_{\rm dec}}_{\rm Doppler} + \underbrace{\int_{\eta_{\rm dec}}^{\eta_0} d\eta\,(\phi' + \psi')}_{\rm ISW} \ . \tag{5.45}$$

The first term is conventionally called the Sachs-Wolfe (SW) term, and includes the intrinsic temperature term Θ_0 and the "gravitational Doppler shift" term ψ at one point on the last scattering surface. The second term is the conventional Doppler term. The last term is called the Integrated Sachs-Wolfe (ISW) term and contains all non-conservative gravitational effects occurring in a universe with non-static metric fluctuations.

We can gain further insight on the Sachs-Wolfe term. We will try to find a simpler expression for $(\Theta_0 + \psi)|_{\rm dec}$, that applies at least for describing large angular patterns on CMB maps, i.e. maps smoothed over small scales. For instance, this expression would describe very well the map of the COBE satellite, which had limited angular resolution. In more precise terms, we wish to calculate the contribution to the Sachs-Wolfe term of large wavelengths, which are bigger than the Hubble radius at the time of recombination.

Let us first focus on the term $\Theta_0|_{\rm dec}$. We have seen that on super-Hubble scales, temperature anisotropies only have a monopole and a dipole component, related respectively to δ_γ and θ_γ. We can be more precise now. We know from thermodynamics that the local value of the photon density is proportional to the temperature to the power four. Taking the derivative of $\log\rho_\gamma = \log T^4$, we get $\delta_\gamma = 4\,\delta T/\bar{T}$, where on the right-hand side the temperature anisotropy is averaged over all directions $\hat{n}$: hence $\delta_\gamma = 4\Theta_0$.

Let us now focus on the term $\psi|_{\rm dec}$. We have seen in Sec. 5.1.4 that on super-Hubble scales and for adiabatic initial conditions, $\delta_\gamma = \frac{4}{3}\delta_{\rm b}$. Also, we know that decoupling takes place at the beginning of matter domination, and we mentioned at the very end of section 5.1.4 that for super-Hubble scales and during matter domination, one has $-2\phi = -2\psi = \delta_{\rm tot} = \delta_{\rm b} = \delta_{\rm cdm}$. Hence, $-2\psi = \frac{3}{4}\delta_\gamma$. Putting all these equalities together, we conclude that on the last scattering surface,

$$\Theta_0 + \psi = \frac{1}{4}\delta_\gamma + \psi = (-\frac{1}{4}\,2\,\frac{4}{3} + 1)\,\psi = (-\frac{2}{3} + 1)\,\psi = \frac{1}{3}\psi \ . \tag{5.46}$$

Moreover, on super-Hubble scales, we can neglect the Doppler term, which can be shown to be important only on sub-Hubble scales. We will see later that the integrated Sachs-Wolfe term plays a role on large scales, but only a small role, because ϕ and ψ are static during most of the evolution after decoupling. Hence we get an approximation for CMB anisotropies smoothed over small scales (that was first derived by Sachs and Wolfe in 1967):

$$\Theta|_{\rm obs,\,large\,scales} \simeq \frac{1}{3}\psi|_{\rm dec} = -\frac{1}{8}\delta_\gamma|_{\rm dec} \ . \tag{5.47}$$

In this calculation, we have seen that the term ψ wins over the term Θ_0, leading to a minus sign in front of δ_γ in the above relation. This means that an over-density on the last scattering surface ($\delta_\gamma > 0$), corresponding to a potential well ($\psi < 0$), leads to a cold spot in the observed map ($\Theta < 0$). Conversely, a hot spot corresponds to an under-density. Hence, due to the "gravitational Doppler shift" effect accounted by the term ψ (often called the Sachs-Wolfe effect), the

patterns that we observe on CMB map are inverted with respect to intrinsic fluctuations on the last scattering surface.

Equation (5.45) (and its large-scale approximation (5.47)) are important for pedagogical purposes, but they have no practical application: indeed, what we wish to calculate and to compare to observations is a theoretical prediction for the statistical properties of CMB anisotropies. Hence, we need to compute at least the CMB two-point correlation function.

5.2.4 Spectrum of temperature anisotropies

Definition. The map of CMB temperature anisotropies can be expanded in spherical harmonics:

$$\frac{\delta T}{\bar{T}}(\hat{n}) = \Theta(\eta_0, \vec{o}, -\hat{n}) = \sum_{lm} a_{lm} Y_{lm}(\hat{n}) \ . \tag{5.48}$$

Using the Legendre expansion of Θ introduced in eq. (5.33), and some basic relations between Legendre polynomials and spherical harmonics, it is easy to express a_{lm} as a function of Θ_l:

$$a_{lm} = (-i)^l \int \frac{d^3\vec{k}}{2\pi^2} Y_{lm}(\hat{k}) \Theta_l(\eta_0, \vec{k}) \ , \tag{5.49}$$

where we recall that hats denote unit vectors: $\hat{k} \equiv \vec{k}/k$. In linear perturbation theory and assuming Gaussian initial conditions, both Θ_l and a_{lm} are Gaussian random variables. Using the orthogonality relation of spherical harmonics and the definitions given in Sec. 5.1.5, we can infer the two-point correlation function of the a_{lm}'s as a function of the power spectrum of Θ_l, or even better, of the primordial curvature power spectrum:

$$\langle a_{lm} a^*_{l'm'} \rangle = \delta^K_{ll'} \delta^K_{mm'} \left[\frac{1}{2\pi^2} \int \frac{dk}{k} \Theta^2_l(\eta_0, k) \mathcal{P}_\mathcal{R}(k) \right] \ . \tag{5.50}$$

Here $\delta^K_{ll'}$ represents the Kronecker symbol. The fact that $\langle a_{lm} a^*_{l'm'} \rangle$ vanishes for $l \neq l'$ or $m \neq m'$ comes out of the algebra, but physically, it is a consequence of the homogeneity of the universe — just like the fact that all power spectra in Fourier space are proportional to $\delta_D(\vec{k}' - \vec{k})$. Similarily, the fact that the quantity between brackets is a function of l but not of m is a consequence of isotropy — like the fact that in Fourier space, power spectra are functions of k but not $\vec{k}$.

The quantity between brackets is usually denoted by C_l, and is called the power spectrum of temperature anisotropies in harmonic space, or the temperature harmonic power spectrum:

$$C_l \equiv \frac{1}{2\pi^2} \int \frac{dk}{k} \Theta^2_l(\eta_0, k) \mathcal{P}_\mathcal{R}(k) \ . \tag{5.51}$$

In a universe with linear and Gaussian perturbations, the C_l's encode all the information concerning the cosmological model describing our universe that is contained in the CMB temperature map.

It is worth coming back on the meaning of the averaging symbols $\langle ... \rangle$ in eq. (5.50). Since the theory of cosmological perturbation is stochastic, the a_{lm}'s should be seen as random numbers, and the average is meant over many realizations of the theory. In a sense, "a given realization" means "a given universe", and the average holds over many universes, all obeying the same cosmological model, which is encoded in the spectrum C_l.

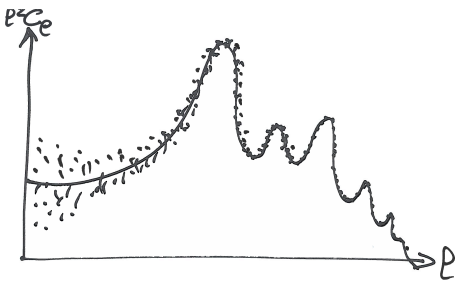


Figure 5.1: Observed values of the temperature harmonic spectrum, C_l^{obs}, are expected to be scattered around the true underlying C_l's. The scattering, called cosmic variance, decreases with increasing l.

However, we observe CMB anisotropies in our universe, i.e. in only one realization. CMB maps allow us to measure a definite value of each a_{lm}. Hence the squares $|a_{lm}|^2$ are not expected to be equal to C_l, even if we postulated the right model: there should be some scattering around C_l. This scattering limits the possibility to find the best theory matching the observations. However we can reduce considerably the scattering by noticing that for any fixed l, the statistical distribution of $|a_{lm}|^2$ is independent of m (as a consequence of isotropy, and as expressed by eq. (5.51)). Hence, for an ideal full-sky CMB experiment, the best estimator of the underlying C_l's is the average between all observed coefficients $|a_{lm}|^2$ with fixed l,

$$C_l^{obs} \equiv \frac{1}{2l+1} \sum_{-l \leq m \leq l} |a_{lm}^{obs}|^2 \ . \tag{5.52}$$

In a typical universe, this quantity should be closer to the underlying C_l than a single $|a_{lm}^{obs}|^2$. The way to see this mathematically is to consider again the theoretical (stochastic) a_{lm}'s, and to define

$$\hat{C}_l \equiv \frac{1}{2l+1} \sum_{-l \leq m \leq l} |a_{lm}|^2 \ . \tag{5.53}$$

By performing averages in the same sense as in equation (5.50), one can easily show that for Gaussian a_{lm}'s,

$$\langle \hat{C}_l \rangle = C_l \qquad \text{and} \qquad \langle (\hat{C}_l - C_l)^2 \rangle = \frac{2}{2l+1} C_l^2 \ . \tag{5.54}$$

The first equality shows that the observed C_l's defined in eq. (5.52) are unbiased estimators of the true underlying C_l's. The second equality gives the typical scattering between the theory and the observations. Since for larger l we perform an average over more values of m, the relative scattering decreases. This is the case with true data points, which are distributed qualitatively like in Fig. 5.1. This scattering is called cosmic variance. It can be seen as a theoretical error: because of cosmic variance, we cannot reconstruct the underlying model with infinite precision, even if we have infinitely precise observations. Cosmic variance is large for small l's, meaning that the shape of the true underlying C_l's will always be poorly known at low l.

Line-of-sight integral in Fourier space. According to equation (5.51), for a given primordial spectrum, the shape of the CMB spectrum C_l depends on the square of the transfer function $\Theta_l(\eta_0, k)$. We would like to understand this shape at

least qualitatively. In real space, we did learn a lot on the behavior of $\Theta(\eta, \vec{x}, \hat{n})$ by using the line-of-sight integral approach presented in section 5.2.3. A similar approach can be worked out in Fourier and harmonic space, i.e. for the variable $\Theta_l(\eta, k)$. We do not present here the intermediate steps. The final result shows many similarity with its real space counterpart, eq. (5.43). It can be decomposed into:

$$\begin{aligned} \Theta_l(\eta_0, k) &= \int_{\eta_{\rm ini}}^{\eta_0} d\eta \; S_T(\eta, k) \; j_l(k(\eta_0 - \eta)) \; , \\ S_T(\eta, k) &\equiv \underbrace{g \, (\Theta_0 + \psi)}_{\rm SW} + \underbrace{\left(g \, k^{-2} \theta_{\rm b}\right)'}_{\rm Doppler} + \underbrace{e^{-\tau} (\phi' + \psi')}_{\rm ISW} \; . \end{aligned} \tag{5.55}$$

We see that $\Theta_l(\eta_0, k)$ is given by the convolution of spherical Bessel functions $j_l(x)$ with a function $S_T(\eta, k)$, called the temperature source function, which contains the usual three terms: Sachs-Wolfe, Doppler, and Integrated Sachs-Wolfe. Like in the previous section, we can use the instantaneous decoupling approximation, integrate the Doppler term by part, and write $\Theta_l(\eta_0, k)$ as:

$$\begin{aligned} \Theta_l(\eta_0, k) \simeq & \left[\Theta_0(\eta_{\rm dec}, k) + \psi(\eta_{\rm dec}, k)\right] j_l(k(\eta_0 - \eta_{\rm dec})) \\ & + k^{-1} \theta_{\rm b}(\eta_{\rm dec}, k) \; j_l'(k(\eta_0 - \eta_{\rm dec})) \\ & + \int_{\eta_{\rm dec}}^{\eta_0} d\eta \; \left[\phi'(\eta, k) + \psi'(\eta, k)\right] j_l(k(\eta_0 - \eta)) \end{aligned} \tag{5.56}$$

(note that in the second line, i.e. in the Doppler term, the prime stands for the derivative of the function $j_l(x)$ with respect to its argument, not with respect to conformal time). This approximate result can be plugged into eq. (5.51) to obtain the final spectrum C_l. We see that each C_l can be decomposed into six terms: the power spectrum $C_l^{\rm SW}$ of the SW term, coming from the first line of eq. (5.56) squared, that of the Doppler term, coming from the second line of eq.(5.56) squared, that of the Integrated Sachs-Wolfe term, coming from the third line of eq. (5.56) squared, and finally the three cross-spectra involving each pair of terms.

For large values of l, the spherical Bessel functions $j_l(x)$ and $j_l'(x)$ are very peaked near $x \simeq l$. Hence, for the Sachs-Wolfe and Doppler contributions to the spectrum C_l, the integral over k in eq. (5.51) will pick up mainly modes with $k(\eta_0 - \eta_{\rm dec}) \simeq l$. This shows that the SW contribution to C_l is given by the product of the primordial spectrum with the squared transfer function $(\Theta_0 + \psi)$ at a given value of η and k:

$$C_l^{\rm SW} \sim \left[\Theta_0(\eta_{\rm dec}, k) + \psi(\eta_{\rm dec}, k)\right]^2 \mathcal{P}_{\mathcal{R}}(k) \; , \qquad k \simeq \frac{l}{(\eta_0 - \eta_{\rm dec})} \tag{5.57}$$

(for simplicity, we did not write numerical factors and powers of l or k in front of this expression). In other words, $C_l^{\rm SW}$ depends on the power spectrum of the perturbation $(\Theta_0 + \psi)$, evaluated at the time of decoupling, and for wavenumbers in the vicinity of $k = \frac{l}{(\eta_0 - \eta_{\rm dec})}$,

$$C_l^{\rm SW} \sim \langle |\Theta_0 + \psi|^2 \rangle_{(\eta, k) \simeq (\eta_{\rm dec}, \, l/(\eta_0 - \eta_{\rm dec}))} \; . \tag{5.58}$$

We reached this result with mathematical arguments, but it has a very simple geometrical interpretation, illustrated in Fig. 5.2. The spectrum C_l encodes the correlation between structures on CMB maps seen under an angle $\theta = \pi/l$. This angle subtends a given physical scale on the last scattering surface, namely $\theta \times d_{\rm a}(z_{\rm dec})$, where $d_{\rm a}(z)$ is the angular diameter distance to objects of redshift z. Since the Sachs-Wolfe term $(\Theta_0 + \psi)$ contributes to the temperature map only

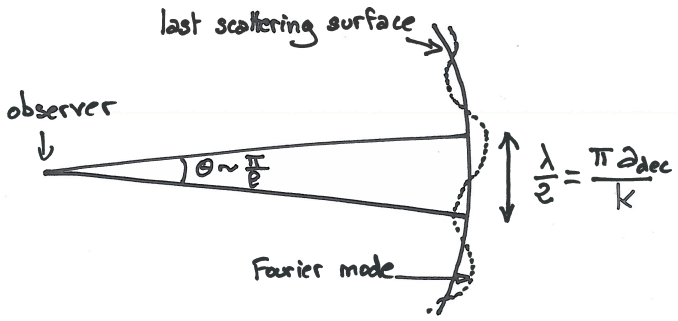


Figure 5.2: A multipole l refers to structures seen on the last scattering surface under an angle $\theta = \pi/l$. These structures are seeded by Fourier modes with a half wavelength $\frac{\lambda}{2} = \frac{\pi a_{\rm dec}}{k}$.

at the time $\eta \simeq \eta_{\rm dec}$, the spectrum $C_l^{\rm SW}$ should depend on the power spectrum of the Sachs-Wolfe term $\langle|\Theta_0 + \psi|^2\rangle$ at that time, and for a wavenumber such that

$$\frac{\lambda}{2} = \frac{\pi a(\eta_{\rm dec})}{k} = \theta\, d_{\rm a}(z_{\rm dec}) \ . \tag{5.59}$$

The reason for which λ has been divided by two is that for spherical harmonics, $\theta = \pi/l$ is the angle between a maximum and a minimum, while for a Fourier mode the distance between a maximum and a minimum is one half of the wavelength, $\frac{\lambda}{2} = \frac{\pi a}{k}$. eq. (5.59) leads to

$$\frac{a(\eta_{\rm dec})}{k} = \frac{d_{\rm a}(z_{\rm dec})}{l} \ . \tag{5.60}$$

We recall that in a flat FL universe the angular diameter distance is given by

$$d_{\rm a}(z) = a(t(z)) \int_{t(z)}^{t_0} \frac{dt}{a} \ , \tag{5.61}$$

$t(z)$ being the proper time at which an object seen today with a redshift z emitted light. The conformal time $\eta(z)$ is defined similarly. In terms of conformal time,

$$d_{\rm a}(z) = a(\eta(z)) \int_{\eta(z)}^{\eta_0} d\eta = a(\eta(z))\ [\eta_0 - \eta(z)] \ , \tag{5.62}$$

and for the case of a point located on the last scattering surface,

$$d_{\rm a}(z_{\rm dec}) = a(\eta_{\rm dec})\ (\eta_0 - \eta_{\rm dec}) \ . \tag{5.63}$$

Hence, eq. (5.60) can be written as

$$\frac{1}{k} = \frac{(\eta_0 - \eta_{\rm dec})}{l} \ , \tag{5.64}$$

which is the same relation between k and l as in Eqs. (5.57, 5.58). In those equations, we implicitly performed a small-angle approximation. For large angles (small l's), it is inaccurate to say that a given angle/mutipole corresponds to a single Fourier mode on the last-scattering surface, and it is important to keep the spherical Bessel function of eq. (5.55) and the integral over k of eq.(5.51). In summary, Eqs. (5.57, 5.58) represent the instantaneous decoupling *and* small-angle limit of the true power spectrum $C_l^{\rm SW}$.

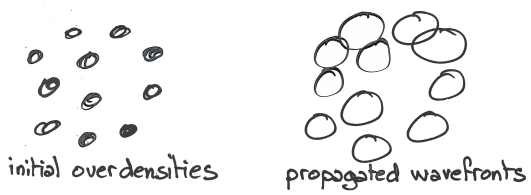


Figure 5.3: Initial over-densities in the early universe propagate in the form of wavefronts. The distance travelled by any wavefront at a given time is given by the sound horizon in the photon-baryon fluid. Here, we represent initial over-densities as spherical patterns at a given scale, while in the real universe primordial over-densities result from a superposition of structures on all scales.

Using the same two limits, a similar discussion can be carried for the Doppler and Integrated Sachs-Wolfe power spectra. The Doppler term depends on the power spectrum of the baryon velocity divergence evaluated roughly at the same time and scale,

$$C_l^{\rm Doppler} \sim \langle |\theta_{\rm b}|^2 \rangle_{(\eta,k)\simeq(\eta_{\rm dec},\, l/(\eta_0-\eta_{\rm dec}))} \ , \tag{5.65}$$

while the ISW term can be written approximately in terms of the integral

$$C_l^{\rm ISW} \sim \int_{\eta_{\rm dec}}^{\eta_0} d\eta\, (\eta_0-\eta)\ \langle |\phi'+\psi'|^2 \rangle_{(\eta,k)\simeq(\eta,\, l/(\eta_0-\eta))} \ . \tag{5.66}$$

Again, for simplicity, we did not write numerical factors and powers of l and k in front of these expressions.

In the next sections, we will infer the shape of the full spectrum C_l from that of the three power spectra appearing in Eqs. (5.58, 5.65, 5.66):

$$\begin{array}{lcll} \text{SW} & : & \langle |\Theta_0+\psi|^2 \rangle & \text{at } (\eta,k) \simeq (\eta_{\rm dec},\ l/(\eta_0-\eta_{\rm dec})) \ , \\ \text{Doppler} & : & \langle |\theta_{\rm b}|^2 \rangle & \text{at } (\eta,k) \simeq (\eta_{\rm dec},\ l/(\eta_0-\eta_{\rm dec})) \ , \\ \text{ISW} & : & \langle |\phi'+\psi'|^2 \rangle & \text{for all } (\eta,k) \simeq (\eta,\ l/(\eta_0-\eta)) \ . \end{array}$$

Before entering into details, we can make a guess. Even if the primordial spectrum $\mathcal{P}_{\mathcal{R}}(k)$ is smooth, the temperature spectrum C_l should contain some structure. Indeed, as illustrated by Fig. 5.3, any primordial over-density is expected to propagate. In Fig. 5.3, we represented naively the primordial over-densities with little dots, giving rise to spherical wavefronts at later time. In the real universe, wavefront patterns are less visible, because primordial over-densities result from a superposition of structures on all scales. However, there is always a characteristic scale in this problem: namely, the distance by which a wavefront travels between some time in the primordial universe and the time of photon decoupling. This distance, called the sound horizon at decoupling $d_{\rm s}(\eta_{\rm dec})$, obeys

$$d_{\rm s} \equiv a \int_{t_{\rm ini}}^{t} \frac{c_{\rm s} dt}{a} = a \int_{\eta_{\rm ini}}^{\eta} c_{\rm s} d\eta \ , \tag{5.67}$$

where $c_{\rm s}$ is the sound speed in the photon-baryon fluid (in units of the speed of light). Two points on the last scattering surface separated by this distance should be partially correlated, since density waves have propagated from one point to the other. Hence, in angular space, the two-point correlation function of CMB anisotropies should exhibit a characteristic feature for angular scales corresponding to the sound horizon at decoupling, $\theta \sim d_{\rm s}(\eta_{\rm dec})/d_{\rm a}(z_{\rm dec})$. Similarly,

the harmonic power spectrum C_l should exhibit a feature at the corresponding scale, $l \sim \pi/\theta \sim \pi\ d_{\rm a}(z_{\rm dec})/d_{\rm s}(\eta_{\rm dec})$, and also for all the harmonics of this scale. We will get a confirmation of this in the next section.

5.2.5 Acoustic oscillations

As long as electrons, baryons and photons are tightly coupled, they form an effective single fluid in which density waves propagate at the sound speed

$$c_{\rm s}^2 = \frac{\delta p_\gamma + \delta p_{\rm b}}{\delta\rho_\gamma + \delta\rho_{\rm b}} \tag{5.68}$$

(the density and pressure of electrons is always negligible with respect to that of photons). The density fluctuation δ_x of each species $x = \gamma, b$ can be inferred from the local value of the equilibrium temperature. The fact that $\rho_{\rm b} \propto T^3$ and $\rho_\gamma \propto T^4$ implies $\delta_\gamma = \frac{4}{3}\delta_{\rm b}$, and tight coupling imposes $\theta_\gamma = \theta_{\rm b}$, as we already saw in Sec. 5.2.2. We can simplify the expression of the sound speed, using also the fact that $|\delta p_{\rm b}| \ll |\delta p_\gamma|$. The result reads

$$c_{\rm s}^2 = \frac{1}{3(1+R)}\ , \quad R \equiv \frac{4\bar\rho_{\rm b}}{3\bar\rho_\gamma} \propto a\ . \tag{5.69}$$

It is possible to derive a simple equation of motion for the photon temperature fluctuation $\Theta_0(\eta, \vec{x})$ in the tightly-coupled regime:

$$\Theta_0'' + \frac{R'}{1+R}\Theta_0' + k^2 c_{\rm s}^2 \Theta_0 = -\frac{k^2}{3}\psi + \frac{R'}{1+R}\phi' + \phi''\ . \tag{5.70}$$

This equation follows from the combination of the continuity and Euler equations for photons and baryons. Given that R is proportional to the scale factor, we could replace R' by $(a'/a)R = aHR$. The second term on the left-hand side is a damping term, increasing with the contribution of baryons to the total energy of the fluid. The third term accounts for pressure forces in the effective fluid. The first term on the right-hand side accounts for the gravitational force, and the last two terms for dilation effects.

This equation would be that of a simple harmonic oscillator if R was a constant (no friction term, constant sound speed) and in absence of gravitational source terms. Then, the solution would be of the form

$$\Theta_0 = \Theta_{\rm ini}\ \cos(kc_{\rm s}\eta + \varphi)\ , \tag{5.71}$$

with two constants of integration $(\Theta_{\rm ini}, \varphi)$. We know that for adiabatic initial conditions and in the Newtonian gauge, photon density/temperature fluctuations should be constant in the super-Hubble limit, $k\eta \ll 1$: this fixes the phase to $\varphi = 0$. In the opposite limit, this solution corresponds to the propagation of acoustic oscillations. Actually, the limit between the constant and oscillatory regime is not set by the value of $k\eta$, but by that of $kc_{\rm s}\eta$. In fact, the condition $kc_{\rm s}\eta \ll 1$ is equivalent to $\lambda \gg d_{\rm s}$, where λ is a physical wavelength ($\lambda = 2\pi a/k$), and $d_{\rm s}$ is the physical sound horizon, given in the case of a constant sound speed by:

$$d_{\rm s} = a\int_{\eta_{\rm ini}}^{\eta} c_{\rm s} d\eta \simeq a c_{\rm s}\eta \tag{5.72}$$

(assuming $\eta \gg \eta_{\rm ini}$). Hence, the phase $kc_{\rm s}\eta$ of the cosine stands for the ratio $2\pi d_{\rm s}/\lambda$. Modes start oscillating when their wavelength becomes smaller than the sound horizon, and later on, the number of oscillations is given by the ratio between these two scales.

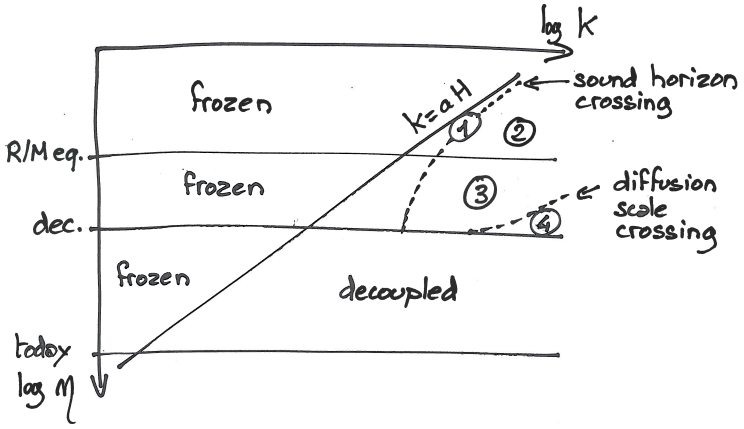


Figure 5.4: Different regions in (k, η) space, corresponding to qualitatively different behaviors for photon (and baryon) perturbations.

In reality, R grows with time (and crosses one roughly around the time of decoupling). In addition, the gravitational source terms in eq. (5.70) can play a role in some regimes. Let us describe qualitatively the evolution of Θ_0 in the different regions in (k, η) space shown in Fig. 5.4. In this figure, the horizontal axis corresponds to wavenumbers k (large wavelengths are on the left), and the vertical axis to conformal time, flowing from top to bottom. The super-Hubble and sub-Hubble regions are separated by the solid diagonal line corresponding to $k = aH$ (equivalent to $k\eta = 1$ during radiation domination). From top to bottom, the horizontal lines correspond to the time of equality between radiation and matter, to the time of photon decoupling, and to the time today. The upper dashed line separates wavelengths bigger/smaller than the sound horizon in the baryon-photon fluid before decoupling (after decoupling, this notion does not make sense anymore). As we have just seen, this limit corresponds to $\lambda = d_{\rm s}$, or (up to a factor 2π) to $k = (a/d_{\rm s})$. At early times, $c_{\rm s} = 1/\sqrt{3}$, and this condition reads $k\eta = \sqrt{3}$. Just before decoupling, R becomes large, $c_{\rm s}$ goes to zero, and the comoving sound horizon $(d_{\rm s}/a)$ becomes asymptotically constant, explaining the shape of the upper dashed line. Finally, the lower dashed line separates wavelengths bigger/smaller than the diffusion length defined in section 5.2.1: this line corresponds to $k = 1/r_{\rm d}$, where $r_{\rm d}$ is given in first approximation by eq. (5.26).

The evolution of Θ_0 in the super-Hubble region is trivial: as long as $k\eta \ll 1$ — and *a fortiori* $kc_{\rm s}\eta \ll 1$ — the fluctuation Θ_0 is frozen, and remains approximately equal to its initial value.

The region marked with a ① in the figure corresponds to modes that are crossing the sound horizon before decoupling. This is precisely the region in which gravitational source terms are important. They shift the zero point of oscillations, and boost their amplitude (due to gravitational forces and dilution effects). This happens during a limited amount of time, because the metric fluctuations quickly decay inside the sound horizon during radiation domination, making the gravitational source terms negligible. An approximation for the zero-point of oscillations can be found by setting Θ_0'' and Θ_0' to zero in equation (5.70), and by keeping only the first gravitational term:

$$\Theta_0^{equilibrium} = -\frac{1}{3c_{\rm s}^2}\psi = -(1+R)\psi \ . \tag{5.73}$$

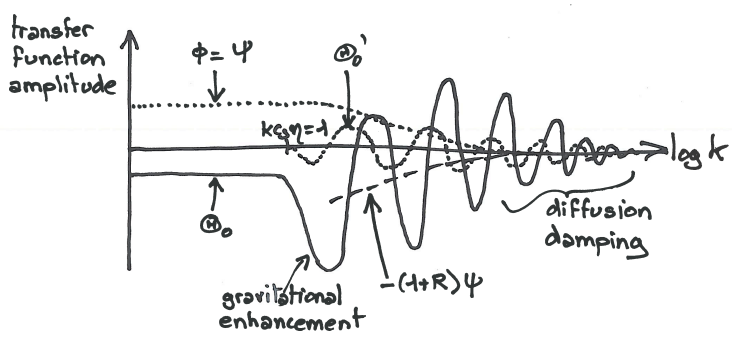


Figure 5.5: Transfer functions at the time of decoupling.

Since the gravitational potential is non-zero on super-Hubble scales (we have seen that for adiabatic initial conditions $-2\psi = \delta_{\rm tot}$), the equilibrium point is shifted away from zero on those scales. It reaches asymptotically zero on sub-sound-horizon scales.

Region ② corresponds to wavelengths smaller than the sound horizon during radiation domination. In this regime, the metric fluctuations have decayed, so the source term in eq. (5.70) can be neglected. The friction term can also be neglected, because during radiation domination, $R \ll 1$. Finally the effective mass $k^2 c_{\rm s}^2$ is constant in time because $R \ll 1$ implies $c_{\rm s}^2 = 1/3$. Hence we are in the simple case discussed before, and the solution is proportional to $\cos(kc_{\rm s}\eta)$, corresponding to stationary oscillations, symmetric around $\Theta_0 = 0$.

Region ③ refers to wavelengths smaller than the sound horizon during the intermediate stage between the time of equality and that of photon decoupling. In this region, the metric perturbations have decayed, but R cannot be neglected (baryons and photons contribute to the total energy density with the same order of magnitude). Hence the oscillator equation has a non-negligible friction term (increasing with time), and a time-varying effective mass (decreasing with time). The solution of the equation corresponds to damped oscillations. Physically, this damping is caused by the increasing inertia and decreasing pressure of the baryon-photon fluid when the energy density of non-relativistic baryons takes over. Using the WKB formalism, one can find a good analytic approximation to these damped oscillations:

$$\Theta_0 = \Theta_{\rm ini}\,(1+R)^{-1/4}\cos(kc_{\rm s}\eta)\ , \tag{5.74}$$

Region ④ refers to modes with smaller wavelength than the diffusion scale in the photon-baryon fluid. We have defined this scale in section 5.2.1. At early times, in the tightly-coupled limit, the mean free path of particles in the fluid is negligible, and cosmologically interesting scales are all well above the diffusion length. At the approach of decoupling, the diffusion length suddenly increases, and encompasses most of sub-sound-horizon wavelengths. In this regime, the oscillator equation (5.70) does not apply anymore, because we cannot describe baryons and photons in terms of a perfect fluid. Perturbations are then strongly damped, since diffusion tends to average out any small-scale fluctuation.

After describing these different regions, we are ready for understanding the qualitative behavior of the various relevant transfer functions, evaluated at the time of decoupling. Figure 5.5 shows the transfer functions $\Theta_0(\eta_{\rm dec}, k)$ (solid line), $\psi(\eta_{\rm dec}, k)$ (upper dotted line) and $\Theta_0'(\eta_{\rm dec}, k)$ (middle dotted line) as a function of $\log k$. For simplicity, we neglect the role of the anisotropic stress generated by neutrinos (and to a lesser extent by photons near decoupling time).

Hence we can assume $\phi = \psi$ at all times. For scales above the sound horizon ($kc_s\eta \ll 1$), the transfer functions are constant: indeed we know that, on those scales and in the Newtonian gauge, density and metric fluctuations are frozen. Adiabatic initial conditions impose $-2\psi = \delta_{\rm tot} \simeq \delta_{\rm b} \simeq \frac{3}{4}\delta_\gamma \simeq 3\Theta_0$. The opposite sign of Θ_0 and ψ reflects the fact that an over-density corresponds to a temperature excess and a gravitational potential well (and vice-versa). We remember that transfer functions are all normalized to $\mathcal{R}(\vec{k}) = 1$ (see section 5.1.5): this corresponds to a negative δ_γ (and Θ_0) and to a positive ψ, like in the figure.

Because of the decay of metric fluctuations inside the Hubble radius, the ψ curve smoothly decreases and tends towards zero in the small wavelength limit. The behavior of Θ_0 is more complicated. Modes which are just crossing the sound horizon near $\eta = \eta_{\rm dec}$ are experiencing the boost caused by gravitational source terms in eq. (5.70): this explains the first bump in the solid line. For smaller wavelengths, we see oscillatory patterns corresponding to acoustic oscillations. Smaller wavelengths crossed the sound horizon earlier, and had more time for oscillating before decoupling. The maxima observed at the time of decoupling correspond to modes that could experience 0.5, 1, 1.5, 2, ..., periods of oscillations before that time. The zero point of oscillations follows $-(1+R)\psi$, represented on the figure with a dashed line: this zero point reaches zero well inside the sound horizon. The amplitude of the oscillations is maximal for the first oscillatory pattern, i.e. for modes that crossed the sound horizon very recently. The second oscillatory pattern is reduced by the fact that those modes stayed for a longer time inside the sound horizon during the matter dominated regime, and experienced more damping due to baryons. The third and higher oscillatory patterns are reduced even more by diffusion damping just before photon decoupling. Temperature fluctuations on very small wavelengths are completely suppressed by photon diffusion.

Figure 5.5 also shows qualitatively the behavior of the time derivative $\Theta_0'(\eta_{\rm dec}, k)$, which exhibits oscillations that are out of phase with respect to those of $\Theta_0(\eta_{\rm dec}, k)$. This will be important in a few paragraphs, when discussing the Doppler effect.

Now that we understand qualitatively the behavior of the metric and photon transfer functions, we can go back to the decomposition of the CMB temperature spectrum C_l in three terms (Sachs-Wolfe, Doppler and integrated Sachs-Wolfe) discussed in section 5.2.5.

Sachs-Wolfe contribution. We have seen that the Sachs-Wolfe contribution to C_l is approximately given by the power spectrum of the combination $(\Theta_0 + \psi)$ at $\eta = \eta_{\rm dec}$, with a correspondence between k and l given by eq. (5.64). We know that this power spectrum is given by the product of the primordial spectrum $\mathcal{P}_\mathcal{R}(k)$ (that we can choose to be scale-invariant in first approximation) by the square of the transfer function $(\Theta_0 + \psi)^2$. The qualitative behavior of the latter has no more secrets for us. We can pick up the solid and upper dotted lines in figure 5.5, add them up, and square the result. The result is shown in figure 5.6 as a function of $\log k$. For modes with $kc_s^2\eta \ll 1$, we see a flat plateau: there, our previous calculation of the Sachs-Wolfe effect would apply (this part of the curve is actually called the Sachs-Wolfe plateau). We then observe a series of peaks. Due to the shift of the zero-point of oscillations given by $-(1+R)\psi$ for Θ_0, and hence by $-R\psi$ for the Sachs-Wolfe term $(\Theta_0 + \psi)$, there is an asymmetry between the first few odd and even peaks, with odd peaks being enhanced. Moreover the overall amplitude of the peaks is suppressed in the large k limit by diffusion damping. A bit of algebra would show us that the envelope of the peaks is given in first approximation by the function $\exp[-(k/k_{\rm d})^2]$, where $k_{\rm d}$ is the diffusion wavenumber, related to the diffusion comoving scale of eq. (5.26) by $k_{\rm d} r_{\rm d} = 1$.

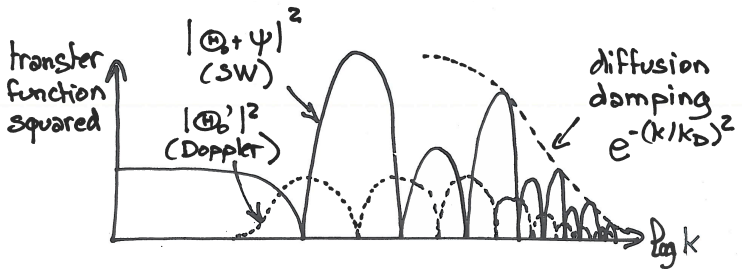


Figure 5.6: Squared transfer functions at the time of decoupling.

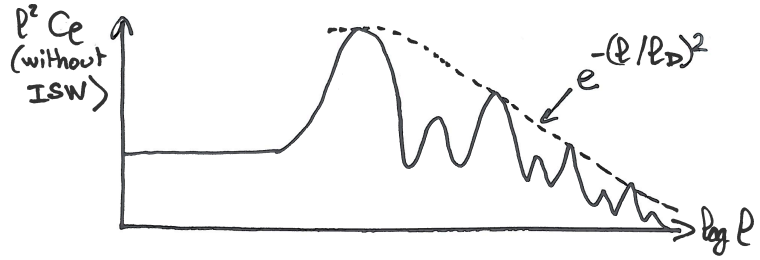


Figure 5.7: Contribution to the C_l's from the SW and Doppler terms.

Doppler contribution. Next, we know that the C_l's receive a second contribution from the Doppler effect, related to the power spectrum of $\theta_{\rm b}$, which is equal to θ_γ until baryon and photons decouple from each other. It turns out that the photon velocity divergence θ_γ is itself related to the time derivative of the temperature fluctuation Θ_0. At a very qualitative level, we can infer the Doppler contribution from the shape of the transfer function $\Theta_0'(\eta_{\rm dec}, k)$. This contribution is null for scales above the sound horizon, since in this regime there are no oscillations and no significant dynamics in the fluid. On smaller scales, the Doppler contribution has oscillatory patterns, that are out of phase with respect to those of the Sachs-Wolfe term. The Doppler contribution is represented schematically as a dotted line in figure 5.6.

We can now sum up the Sachs-Wolfe and Doppler contributions. Also, we can transpose our results for power spectra as a function of k in terms of C_l's as a function of l. We have already seen in section 5.2.5 that there is a mapping betwwen the two, at least in the small-angle and instantaneous decoupling approximation, with a correspondence between values of k and l given by eq. (5.64). The result is shows in figure 5.7. Note that the vertical axis stands for $l^2 C_l$, for the following reason. If the primordial spectrum was scale invariant ($\mathcal{P}_\mathcal{R}(k) =$ constant) and the transfer functions were flat (as it is the case for large wavelengths/small l's), the quantity $l(l+1)C_l$ would also be flat and independent of l. This would follow from eq. (5.57) if we had been more carefull in keeping all the factors. It is convenient to plot $l^2 C_l$ or $l(l+1)C_l$ instead of C_l, in order to display a roughly constant curve, just modulated by acoustic oscillations and diffusion damping.

Diffusion damping effect. In Fig. 5.7, we can identify all the features mentioned before: the flat Sachs-Wolfe plateau, the series of oscillations with enhanced odd peaks, and the exponentially decaying envelope of the peaks for large l. The envelope is now given by $\exp[-(l/l_{\rm d})^2]$, where $l_{\rm d}$ is the diffusion multipole, related to the diffusion angle by $l_{\rm d} = \pi/\theta_{\rm d}$. The diffusion angle is

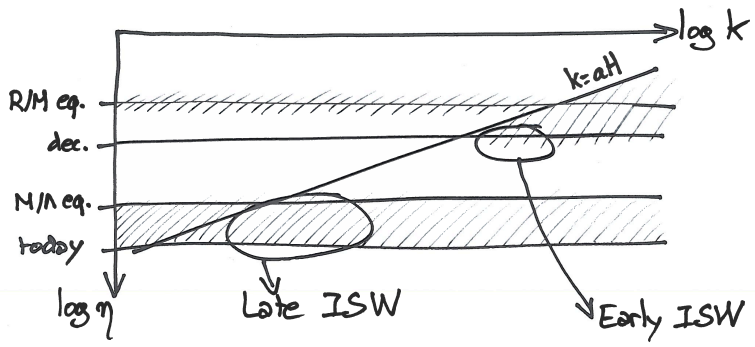


Figure 5.8: Regions in (k, η) space where metric fluctuations are expected to vary with time (giving rise potentially to an ISW effect after decoupling).

related to the diffusion scale $r_{\rm d}$ by the usual angular diameter distance relation,

$$\theta_{\rm d}\, d_{\rm a}(\eta_{\rm dec}) = \lambda_{\rm d} = a(\eta_{\rm dec}) r_{\rm d}(\eta_{\rm dec}). \tag{5.75}$$

At this point, we are almost done with the qualitative description of the CMB temperature spectrum C_l. We only missed the integrated Sachs-Wolfe contribution, and the effect of reionization.

Integrated Sachs-Wolfe contribution. We have seen that the integrated Sachs-Wolfe effect is given by an integral over $(\psi' + \phi')$ between photon decoupling and today. In fact, $(\psi' + \phi')$ remains vanishingly small in a large part of the space (k, η). In Figure 5.8, we hatched all the regions in (k, η) space where metric fluctuations are expected to vary with time. Let us discuss these different regions.

Before photon decoupling, we know that metric fluctuations decay inside the sound horizon. Instead, in the Newtonian gauge, they remain frozen outside the Hubble radius, except near times at which the equation of state of the universe changes: namely, at the time of equality between radiation and matter.

Let us now discuss the variation of metric perturbations after photon decoupling (this is the relevant epoch for the ISW effect). Deep inside the matter dominated regime, one can show that metric fluctuations are static, even inside the Hubble radius (at least within linear perturbation theory). We will justify this result in section 5.3.2. Hence in figure 5.8 there are no hatches during matter domination on whatever scales. Note however that at the beginning of matter domination, it takes some time for sub-sound-horizon metric fluctuations to freeze around a constant value: hence the hatches continue below the line corresponding to the time of equality, and extend till the line corresponding to photon decoupling.

Later on, during the Λ (or dark energy) dominated regime, the equation of state of the universe changes again, so metric fluctuations vary on all scales, like at the time of equality. A simple calculation based on Einstein equations would show that metric fluctuations are damped during this stage.

In summary, contribution to the integrated Sachs-Wolfe effect can only come from two regions:

- just after photon decoupling, on sub-sound-horizon scales, and
- during Λ domination, on all scales.

These two distinct contributions are usually called the Early Integrated Sachs-Wolfe (EISW) and Late Integrated Sachs-Wolfe (LISW) effects. One can show

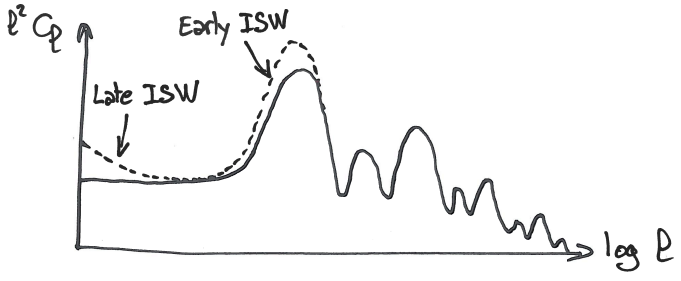


Figure 5.9: ISW contribution to the temperature spectrum.

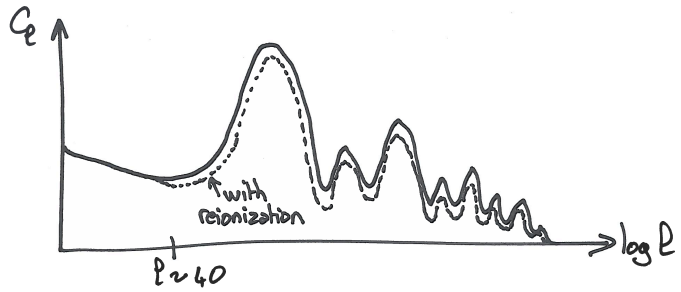


Figure 5.10: Impact of reionization on the temperature spectrum.

that both effects decrease with wavelength, for geometrical reasons. Hence the EISW effect is maximal for scales crossing the sound horizon just at the time of photon decoupling, while the LISW effect is maximal for the largest observable scales today. In multipole space, this means that the EISW effect contributes mainly to the scale of the first peak, i.e. to $l \sim 200$, while the LISW contributes mainly to the smallest multipoles $l = 2, 3, 4$, etc. The two ISW contributions are drawn on figure 5.9. The EISW effect enhances the first peak, while the LISW effect tilts the Sachs-Wolfe plateau even if the primordial spectrum is exactly scale invariant.

Reionization effect. The last effect that we omitted to describe is that of reionization. We have seen in section 5.2.1 that at small redshift ($z \simeq 10$), the reionization of the universe produces a small secondary bump in the visibility function, corresponding physically to a small probability for CMB photons to rescatter at late times. This rescattering will tend to smooth out any temperature anisotropy pattern. Hence, reionization lowers the overall amplitude of the C_l', but only a small amount (by approximately 15%). Note that the suppression of power is not uniform over the whole multipole range: smoothing effects cannot reach the largest observable scales (corresponding to the smallest values of l). Hence the effect of reionization is step-like shaped, and saturates for l of the order of 40 or so (as illustrated on figure 5.10).

5.2.6 Parameter dependence of the temperature spectrum

We summarized in the last section the various physical effect contributing to the shape of the CMB temperature spectrum C_l. We will now recapitulate the effect of the various cosmological parameters on the C_l's, within the framework of the minimal ΛCDM model.

This model assumes zero spatial curvature, and a power-law primordial spectrum of scalar perturbations:

$$\mathcal{P}_{\mathcal{R}}(k) = A_{\rm s}(k/k_*)^{n_{\rm s}-1} \ , \tag{5.76}$$

where k_* is an arbitrary fixed pivot scale, $A_{\rm s}$ is the spectrum amplitude at this scale, and $n_{\rm s}$ is called the scalar tilt (the exponent is chosen to be $n_{\rm s}-1$ rather than just $n_{\rm s}$ for historical reasons; with such notations, a scale-invariant spectrum corresponds to $n_{\rm s}=1$).

We recall that the Hubble parameter today, H_0, can be expressed in terms of a dimensionless reduced Hubble parameter h:

$$H_0 \equiv 100\,h\,{\rm km.s^{-1}.Mpc^{-1}} \tag{5.77}$$

The physical energy density of a given component x today can be expressed in terms of a dimensionless parameter $\omega_x \equiv \Omega_x h^2$:

$$\rho_x^0 = \Omega_x \rho_{\rm crit}^0 = \Omega_x \frac{3H_0^2}{8\pi G} = \beta\,\omega_x \tag{5.78}$$

where $\beta = \frac{3(H_0/h)^2}{8\pi G}$ is a fixed number, with the dimension of an energy per volume.

The six free parameters of the minimal ΛCDM model can be chosen to be

$$\{A_{\rm s}, n_{\rm s}, \omega_{\rm b}, \omega_{\rm m}, \Omega_\Lambda, \tau_{\rm reio}\}, \tag{5.79}$$

where $\tau_{\rm reio}$ is the optical depth to recombination, which is non-zero because of reionization in the recent universe. At first order, this is the only parameter that one needs to introduce for describing reionization. For a very accurate description, one should introduce other parameters specifying the full reionization history, but these extra parameters are very difficult to probe experimentally. Hence, they are usually not specified.

The above parameter basis specifies the baryon density $\omega_{\rm b}$, the total non-relativistic (baryons + CDM) matter density $\omega_{\rm m}$, and the fractional density of cosmological constant Ω_Λ. The photon density is implicitly assumed to match the measured value of the CMB temperature ($T = 2.725$ K implies $\omega_\gamma \sim 2.10^{-5}$). Neutrinos are assumed to be still relativistic today, with an abundance relative to photons given by the prediction of the standard neutrino decoupling model[4]. Finally, we did not include the parameter H_0 (or h) in the parameter basis (5.79). Given that we are assuming a flat universe, h can be inferred from other parameters in (5.79): $h = \sqrt{\omega_{\rm m}/(1-\Omega_\Lambda)}$.

The parameters basis (5.79) is just one particular choice. Many other bases would be valid, for instance, $\{A_{\rm s}, n_{\rm s}, \omega_{\rm b}, \omega_{\rm cdm}, H_0, \tau_{\rm reio}\}$. The choice of (5.79) is dictated by purely pedagogical considerations: we will show that the shape of the CMB spectrum can easily be related to these parameters.

We summarize in Fig. 5.11 the evolution of background densities in the minimal ΛCDM model. The normalization of the radiation density ρ_r is fixed by the CMB temperature and by the standard neutrino decoupling model. The normalization of the matter density $\rho_{\rm m}$ and of the cosmological constant density ρ_Λ is given respectively by the parameters $\omega_{\rm m}$ and Ω_Λ. The scale factor at radiation/matter equality is given by

$$\rho_r = \rho_{\rm m} \quad \Longrightarrow \quad \omega_r \left(\frac{a_0}{a_{\rm eq}}\right)^4 = \omega_{\rm m}\left(\frac{a_0}{a_{\rm eq}}\right)^3 \quad \Longrightarrow \quad \frac{a_{\rm eq}}{a_0} = \frac{\omega_r}{\omega_{\rm m}} \ . \tag{5.80}$$

[4] In this course, for simplicity, we do not discuss the effect of neutrinos; this effect is far from being negligible, but since we consider the abundance of neutrinos as fixed, and their mass as irrelevant, we can explain the effect of other free parameters without taking neutrinos into account.

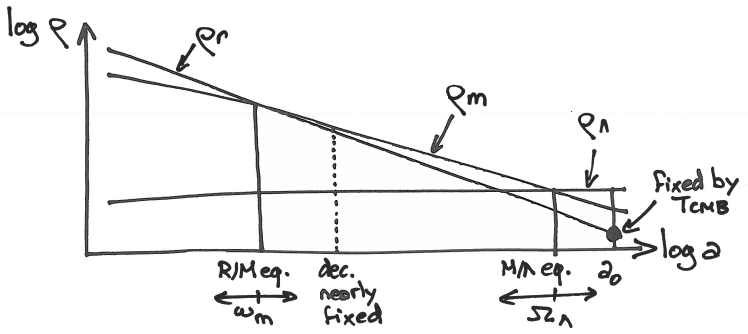


Figure 5.11: In the minimal ΛCDM model, evolution of the background density of radiation, matter and cosmological constant as a function of the scale factor.

Since ω_r is considered as fixed, the redshift of equality is controlled by $\omega_{\rm m}$ only. The scale factor at matter/Λ equality is given by

$$\rho_{\rm m} = \rho_\Lambda \quad \Longrightarrow \quad \omega_{\rm m}\left(\frac{a_0}{a_\Lambda}\right)^3 = \Omega_\Lambda h^2 \quad \Longrightarrow \quad \frac{a_\Lambda}{a_0} = \left(\frac{1-\Omega_\Lambda}{\Omega_\Lambda}\right)^{1/3} , \quad (5.81)$$

so it is controlled by Ω_Λ only. The scale factor at photon decoupling has a small (logarithmic) dependence on $\omega_{\rm b}$, that we can neglect — we will consider that decoupling takes place at a fixed temperature, and hence a fixed scale factor and redshift.

Among the various physical effects described in the previous section, we can identify eight independent leading effect. We summarize them in Table 5.1, and show by which parameters they are governed. Below, we give more details on

Table 5.1: Independent leading effects controlling the shape of the CMB temperature power spectrum C_l in the minimal ΛCDM model.

	Effect	Relevant quantity	Parameter
(C1)	Peak scale	$\theta_{\rm peak} = \dfrac{\pi}{l_{\rm peak}} \sim \dfrac{d_{\rm s}\|_{\rm dec}}{d_{\rm a}\|_{\rm dec}}$	$\leftarrow \omega_{\rm m}, \omega_{\rm b}$ $\leftarrow \Omega_\Lambda, \omega_{\rm m}$
(C2)	Odd/even peak amplitude ratio	$R\|_{\rm dec}$	$\omega_{\rm b}$
(C3)	Amplitude of rirst peaks	$\frac{a_{\rm dec}}{a_0}$	$\omega_{\rm m}$
(C4)	Damping envelope	$\theta_{\rm d} = \dfrac{\pi}{l_{\rm d}} = \dfrac{a_{\rm dec} r_{\rm d}\|_{\rm dec}}{d_{\rm a}\|_{\rm dec}}$	$\leftarrow \omega_{\rm m}, \omega_{\rm b}$ $\leftarrow \Omega_\Lambda, \omega_{\rm m}$
(C5)	Global amplitude	$\mathcal{P}_\mathcal{R}(k_*)$	$A_{\rm s}$
(C6)	Global tilt	$\frac{d\log \mathcal{P}_\mathcal{R}}{d\log k}$	$n_{\rm s}$
(C7)	Additional plateau tilting (LISW)	$\frac{a_\Lambda}{a_0}$	Ω_Λ
(C8)	Amplitude for $l \geq 40$ only	$\tau_{\rm reio}$	$\tau_{\rm reio}$

these effects.

(C1) Since all peaks in multipole space correspond to the harmonics of a single correlation length in real space, the scale of the peak is controlled (in good approximation) by a single number $l_{\rm peak}$. It depends on the ratio of the sound horizon at decoupling by the angular diameter distance to decoupling. The first quantity depends on the evolution prior to decoupling, and in particular on the expansion history and sound speed. Hence it depends on $\omega_{\rm m}$ (governing the time of equality) and $\omega_{\rm b}$ (governing $c_{\rm s}^2$ as a function of a). The second quantity depends on the expansion and geometry of the universe after decoupling, i.e. on Ω_Λ and H_0, or in our parameter basis Ω_Λ and $\omega_{\rm m}$.

(C2) When studying the SW contribution to the C_l's, we have seen that the asymmetry between the amplitude of odd and even peaks depends on the shift of the zero-point of acoustic oscillations by a term $-R\psi$. The value of the ratio R at decoupling is governed in our parameter basis by $\omega_{\rm b}$.

(C3) A shift in the time of radiation/matter equality affects the amplitude of all peaks for two reasons: it controls the duration of the intermediate stage between equality and decoupling, during which: (i) modes crossing the sound horizon are not enhanced by a gravitational boosting effects as much as during radiation domination (since $\psi = \phi$ does not strongly decay after sound crossing during MD), and (ii) the EISW effect can take place. Both effects go in the same direction. If equality takes place later, there is less time between equality and decoupling. Then, more modes experience gravitational boosting, and some peaks are higher (typically the 2nd and 3rd). The metric fluctuations are also less stabilized at decoupling, and the EISW is larger (hence the first peak is increased even more).

(C4) Diffusion damping near the time of recombination controls the envelope of the peaks (the function $\exp[-(l/l_{\rm d})^2]$ suppresses them, starting essentially from the third one). It depends on the ratio of the damping scale at decoupling by the angular diameter distance to decoupling. The first quantity depends on the Thomson scattering rate prior to decoupling, and in particular on $\omega_{\rm m}$ (governing the value of conformal time at equality) and $\omega_{\rm b}$ (governing the ionization fraction as a function of a). The second quantity depends on the expansion and geometry of the universe after decoupling, i.e. on Ω_Λ and H_0, or in our parameter basis Ω_Λ and $\omega_{\rm m}$. The parameter dependence of effects (C1) and (C4) could be thought to be similar: in fact, it is not, because the sound horizon and the diffusion scale depend on very different combinations of $\omega_{\rm m}$ and $\omega_{\rm b}$.

(C5) The global amplitude of the C_l's depends trivially on that of the primordial spectrum, fixed by $A_{\rm s}$.

(C6) The global slope of the C_l's depends trivially on the tilt of the primordial spectrum, fixed by $n_{\rm s}$.

(C7) The LISW effect tilts the Sachs-Wolfe plateau (on top of the effect of $n_{\rm s}$). The plateau is more lifted at small l's if Λ domination is longer, i.e. if metric fluctuations decay during a larger amount of time. Hence this effect is enhanced by large values of Ω_Λ.

(C8) The C_l amplitude is suppressed if reionization takes place early, i.e. if $\tau_{\rm reio}$ is large, but without affecting the largest scales (small l's).

We see that in the framework of the minimal ΛCDM model, six parameters control eight distinct physical effects with different impacts on the C_l's. This

suggests that an accurate enough measurement of the temperature spectrum is sufficient for fixing the six parameters of the cosmological model describing our universe, at least if the data is consistent with ΛCDM. This conclusion is roughly correct, but must be refined with some words of caution. Indeed, we remember that for small l's, cosmic variance is large, so that the average C_l's of the "true model" describing our universe cannot be measured precisely, even in the case of an ideal experiment. However, two of the previous effects (C1) - (C8) can only affect the smallest multipoles:

- effect (C7) affects only the Sachs-Wolfe plateau,
- a combination of effects (C5) and (C8), corresponding to the product $e^{-2\tau_{\rm reio}}A_{\rm s}$, controls the global amplitude for $l \gg 40$, but a variation of both $\tau_{\rm reio}$ and $A_{\rm s}$ with the previous product being kept fixed would only affect the smallest multipoles.

Hence, the measurement of the CMB temperature spectrum is sufficient for constraining the six parameters of the ΛCDM model, but with a relatively large error bar for Ω_Λ and for a particular combination of $\tau_{\rm reio}$ and $A_{\rm s}$. In the next section, we will say a few words on the measurement of CMB polarisation, which allows to better constrain reionization: polarisation data allow to remove the degeneracy between $\tau_{\rm reio}$ and $A_{\rm s}$. Instead, the error bar on Ω_Λ can only be improved by combining CMB data with other cosmological probes (e.g. supernovae luminosity or large scale structure data).

5.2.7 A quick word on polarisation *(not treated)*

The Bolztmann equation presented in eq. (5.34) was a bit over-simplified. We did as if the only degree of freedom describing photons with a blackbody spectrum was their temperature. In fact, photons are described by more degrees of freedom, called the Stokes parameters, involving also their polarisation. In eq. (5.34), the Thomson scattering rate was integrated over polarisation parameters, but in reality some polarized correction terms are present.

Well before decoupling, photons remain unpolarized on average. Indeed, we have seen in section 5.2.2 that, throughout the tight-coupling regime and in the frame comoving with the fluid (i.e. such that $\theta_{\rm b} = \theta_\gamma = 0$), the photon temperature is isotropic in every point. This isotropy (resulting from frequent interactions) implies that photons acquire no net polarisation patterns when they scatter over electrons.

Instead, at the approach of decoupling, when Thomson scattering becomes inefficient, the photon temperature is no longer isotropic in the frame comoving with the electrons. This means that a given electron will scatter simultaneously some hotter photons coming from one direction, and some colder photons coming from another direction. What is important for polarisation is the quadrupolar component of the the temperature distribution in each point. This component starts from zero and grows at the approach of decoupling. When it becomes significant, photon scattering leads to a net linear polarisation. Hence, today, CMB photons have a different polarisation amplitude and orientation in each direction of the sky.

The map of CMB temperature anisotropies is a scalar map: it can be represented with a one-dimensional color code. The map of CMB polarisation can be represented with sticks of different size and orientation in different points of the map. Roughly speaking, the size and orientation of the sticks can be related to the magnitude and orientation of the quadrupole anisotropy in each point of the last scattering surface. Since these quadrupolar patterns reflect variations of temperature in the region of the last scattering surface, it is clear that there is

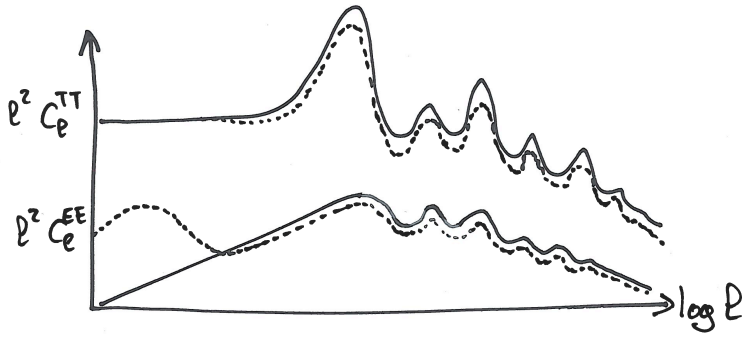


Figure 5.12: Typical shape of the temperature and E-type polarisation power spectra in a ΛCDM model without (solid lines) or with (dotted lines) reionization of the universe at small redshift.

a non-zero correlation between temperature and polarisation maps. Still, both maps contain some independent information.

In general, a vector field can be decomposed into a gradient and a curl component, like the electric and magnetic fields. Similarly, CMB polarisation maps can be expanded in two scalar maps called, by analogy, E-type and B-type polarisation maps.

One can show that primordial scalar perturbations can produce both temperature anisotropies and E-type polarisation anisotropies. If the early universe only features Gaussian scalar perturbations on cosmological scales, all the information contained in CMB maps is encoded in the temperature power spectrum C_l^{TT}, the E-type polarisation power spectrum C_l^{EE}, and the cross-correlation power spectrum C_l^{TE}. The later really contains additional independent information, because polarisation patterns are only partially correlated with temperature patterns (in mathematical terms, $[C_l^{TE}]^2 \leq C_l^{TT} C_l^{EE}$).

We will mention in the next section the possibility that a significant amount of tensor perturbations are produced in the primordial universe. Such perturbations can also generate B-type polarisation. If this is the case, all the information is encoded in C_l^{TT}, C_l^{EE}, C_l^{TE} and C_l^{BB}. We did not include in this list the cross-correlation spectra C_l^{TB} and C_l^{EB}, because the parity symmetry imposes that they should vanish (at the level of primary anisotropies).

Mapping the CMB polarisation is interesting because it contains additional information with respect to temperature anisotropies.

First, reionization has a very distinct effect on the polarisation spectrum. The small fraction of photons rescattered around the time of reionization acquire additional polarisation that will show up in the form of a peak in the small-l branch of C_l^{EE} (see figure 5.12) and C_l^{TE}. We have seen in the previous section that with temperature data only, cosmic variance limits our ability to measure $\tau_{\rm reio}$, and *a fortiori* to constrain additional parameters describing the reionization history. The effect of reionization in the small-l branch of C_l^{EE} and C_l^{TE} is so big that despite of cosmic variance, $\tau_{\rm reio}$ can be well measured using polarisation data.

So the effect of reionization, described before as (C8), is very different for temperature and polarisation. Instead, the other effects (C1) - (C7) have a qualitatively similar impact on C_l^{EE} and C_l^{TE}. Still, measuring polarisation is interesting, because temperature anisotropies feature a particular combination of the SW, Doppler and ISW effects, while polarisation provides information on the quadrupole Θ_2 on the last scattering surface, which is correlated with the SW

and Doppler terms, but probing a different combination of them. Hence the role of polarisation measurements is to remove some of the parameter degeneracies appearing in the analysis of temperature data.

Note that in the full Boltzmann equation, the equation of motion of temperature and polarisation degrees of freedom are coupled. In the last sections, we neglected such a coupling. In fact, the impact of polarisation on the evolution of temperature anisotropies is very small. Hence our qualitative discussion of the various effects affecting CMB temperature remains valid.

5.2.8 A quick word on tensors *(not treated)*

We have seen in section 5.1.1 that tensor modes are related to the traceless divergenceless components of the metric and stress-energy tensor, obeying

$$\sum_i \delta T_{ii} = 0 \qquad \text{and} \qquad \forall j, \quad \sum_i \partial_i \delta T_{ij} = 0 \ . \tag{5.82}$$

Since these relations impose four constraints on the six degrees of freedom of the 3×3 symmetric matrix δT_{ij} (or δg_{ij}), there are two tensor degrees of freedom in $g_{\mu\nu}$ and $T_{\mu\nu}$.

The two degrees of freedom in δg_{ij} are the two degrees of polarisation of gravitational waves, that can propagate even in the vacuum. In the presence of matter with a non-diagonal stress tensor δT_{ij}, gravitational waves can be seeded by the tensor components δT_{ij}.

If tensor perturbations are sufficiently large at recombination or later, they can generate CMB temperature and polarisation anisotropies: one can show that they interact with CMB photons, and produce effects similar to the SW and ISW ones. As mentioned in the previous section, gravitational waves can even seed B-type polarisation.

The stress-energy tensor is diagonal for perfect fluids, and negligible for pressureless components like CDM. Hence, during radiation and matter domination, tensor perturbations can only be seeded by decoupled neutrinos and/or decoupled photons. However, neutrino and photon tensor modes are far too small for generating a detectable amount of CMB anisotropies. However, in the early universe, other mechanisms may produce gravitational waves on scales sufficiently large to be observable in the CMB. The most famous one is related to inflation. During inflation, quantum fluctuations of the metric can excite primordial tensor perturbations at a level that will produce a detectable signature in CMB maps. The amplitude of this signal is directly proportional to the energy scale of inflation. If this scale is large enough, tensors can contribute to C_l^{TT}, but only on small l's, because gravitational waves decay quickly inside the Hubble radius. Hence, one way to detect tensors would be through a small distortion of the C_l^{TT} spectrum shape for $l \leq 100$ (i.e. for scales that are equal or larger than the Hubble radius at the time of decoupling).

However, if the tensor signal is small with respect to the error bars associated to cosmic variance, it will remain undetectable. In that case, further sensitivity could be obtained by measuring C_l^{BB}: in absence of tensors, C_l^{BB} would vanish, so even if the tensor amplitude is very small it can still dominate the C_l^{BB} signal. Unfortunately, this is true only up to some extent, because secondary anisotropies (in particular, those generated by weak lensing) produce a non-zero C_l^{BB} that could mask the primary tensor anisotropy spectrum.

It is very challenging for CMB experiments to reach the sensitivity level required for detecting a C_l^{BB} signal, even for that generated by weak lensing. Current limits on the tensor primordial spectrum (and on the energy scale of inflation) mainly come from the observation of C_l^{TT} at small l by WMAP.

The sensitivity of Planck to tensors will also mainly come from temperature. Future experiments dedicated to CMB polarisation will improve the sensitivity to B-type polarisation and obtain more precise bounds, until they reach the theoretical limit set by the lensing contamination.

5.3 Matter power spectrum *(not treated)*

5.3.1 Definition

The total energy perturbation in the universe can be expanded as

$$\delta\rho_{\rm tot} = \delta\rho_\gamma + \delta\rho_{\rm b} + \delta_{\rm cdm} + \delta\rho_\nu \,(+\delta\rho_{\rm de} + ...) \tag{5.83}$$

where $\delta\rho_{\rm de}$ refers to possible Dark Energy (DE) perturbations, and the three dots for extra relics. In the minimal ΛCDM model, only the first four components are present.

Many Large-Scale Structure (LSS) observables are related to the power spectrum of $\delta\rho_{\rm tot}$, at different wavenumbers and redshifts. This is the case of the galaxy and of the halo correlation function, of the cluster mass function, of CMB lensing, of the cosmic shear spectrum, etc.

All these observations probe the power spectrum during matter or Λ/DE domination, when photons are subdominant and $\delta\rho_\gamma$ can be neglected. If neutrinos are still ultra-relativistic today, $\delta\rho_\nu$ can also be neglected (in this course, we do not have time to discuss the impact of small neutrino masses). If the acceleration of the universe is caused by a cosmological constant, there is no term $\delta\rho_{de}$. More generally, most Dark Energy models would predict a negligible amount of DE perturbations. In summary, in a wide category of cosmological scenarios including ΛCDM, we can use $\delta\rho_{\rm tot} \simeq \delta_{\rm m} \equiv \delta\rho_{\rm b} + \delta_{\rm cdm}$. Hence, in the context of LSS observations, it is customary to refer to the power spectrum of the non-relativistic matter fluctuation $\delta_{\rm m}$, defined as

$$\delta_{\rm m} = \frac{\delta\rho_{\rm m}}{\bar\rho_{\rm m}} = \frac{\delta\rho_{\rm b} + \delta\rho_{\rm cdm}}{\bar\rho_{\rm b} + \bar\rho_{\rm cdm}} \ , \tag{5.84}$$

which is indistinguishable from the ratio $\delta\rho_{\rm tot}/\bar\rho_{\rm m}$. Only in models with large dark energy perturbations or modifications of Einstein gravity, the two quantities might be different, and special care about the definition of the matter power spectrum is needed.

The matter power spectrum $P(z,k)$ of $\delta_{\rm m}$ is defined like in section 5.1.5:

$$\langle \delta_{\rm m}(z,\vec{k}) \delta^*_{\rm m}(z,\vec{k}') \rangle = \delta_D(\vec{k}-\vec{k}') \ P(z,k) \ . \tag{5.85}$$

Here we used the redshift as a time variable, but we could have indifferently used proper or conformal time. We have seen in section 5.1.5 that for Gaussian initial conditions and as long as perturbations are linear, the power spectrum at a given time can be written as the product of the primordial spectrum by the square of the relevant transfer function, in our case $\delta_{\rm m}(z,k)$. Sticking to the same conventions as in section 5.1.5 and assuming a power-law primordial spectrum like in section 5.2.6, this gives

$$P(z,k) = \frac{2\pi^2}{k^3} A_{\rm s} \left(\frac{k}{k_*}\right)^{n_{\rm s}-1} \delta^2_{\rm m}(z,k) \ . \tag{5.86}$$

Hence, by studying qualitatively the evolution of the transfer function $\delta_{\rm m}(z,k)$, we will get some insight on the cosmological information encoded in the matter power spectrum.

5.3.2 Transfer function evolution

CDM dominated universe. In order to simplify the presentation, let us first assume that we live in a ΛCDM universe with a negligible amount of baryons: $\Omega_{\rm b} \ll \Omega_{\rm cdm}$ and $\delta_{\rm m} \simeq \delta_{\rm cdm}$. In section 5.2.5, we wrote the master equation governing the evolution of photon perturbations during the tightly-coupled regime. Similarly, by combining the continuity and Euler equation of CDM perturbations, we can write here a master equation for $\delta_{\rm cdm}$, actually valid in all regimes:

$$\delta''_{\rm cdm} + \frac{a'}{a}\delta'_{\rm cdm} = -k^2\psi + 3\phi'' + 3\frac{a'}{a}\phi' \ . \tag{5.87}$$

In an expanding universe, the clustering rate depends on the expansion rate: expansion increases distances, weakens gravitational forces, and slows down clustering processes. In the above equation, this is accounted by the second term, often called the Hubble friction term. On the right-hand side, the first term represents gravitational forces, and the last two terms account for dilation effects.

On super-Hubble scales, we have seen that in the Newtonian gauge, adiabatic ICs predict constant density fluctuations $\delta_{\rm cdm}$. To be precise, ϕ and $\delta_{\rm cdm}$ vary on super-Hubble scales only when the total equation of state of the universe changes (i.e. around the time of radiation/matter equality, and during Λ domination). Instead, they remain constant on those scale during the radiation and matter dominated regime.

Inside the Hubble radius, we can neglect dilation terms, and replace the gravitational potential term:

$$\delta''_{\rm cdm} + \frac{a'}{a}\delta'_{\rm cdm} - \frac{3}{2}\left(\frac{a'}{a}\right)^2 \Omega_{\rm cdm}(a)\ \delta_{\rm cdm} = 0 \ , \tag{5.88}$$

where $\Omega_{\rm cdm}(a)$ is the fraction of the critical density coming from CDM at a given value of time (or of the scale factor). This equation is often called the Mészáros equation. It can be obtained by combining eq. (5.87) with the (00) component of the Einstein equation (or its Poisson limit) and the Friedmann equation. The careful reader might have noticed something suspicious: the gravitational force term $k^2\psi$ has been eliminated in favor of $\delta_{\rm cdm}$, while the first Einstein equation (or its Poisson limit) involves the total density fluctuation. Hence, shouldn't eq. (5.88) feature also δ_γ, $\delta_{\rm b}$ and δ_ν? Actually, it turns out that the above equation is a really good approximation for the CDM equation of evolution in all regimes, under our assumption $\Omega_{\rm b} \ll \Omega_{\rm cdm}$. It applies even during radiation domination, when photon fluctuations are potentially large. The reason is subtle and we don't have time to study it in this course. In few words, it has to do with the fact that the photons feel mainly pressure forces and CDM gravitational forces; these interactions have very different time scales, and it is possible to show mathematically that they almost decouple; for details, see *Weinberg*'s cosmology textbook, or *Lesgourgues, Tram and Voruz 2013*.

During radiation domination, the Friedmann equation gives $a \propto \eta$, and $\Omega_{\rm cdm}$ is much smaller than one. Hence we can neglect the last term in the Mészáros equation, and find that the two solutions are $\delta_{\rm cdm}$ = constant and $\delta_{\rm cdm} \propto \log\eta$. Hence CDM fluctuations grow logarithmically. During matter domination, $a \propto \eta^2$ and $\Omega_{\rm cdm} \simeq 1$, so the solutions are $\delta_{\rm cdm} \propto \eta^2$ and $\delta_{\rm cdm} \propto \eta^{-3}$. Then, CDM fluctuations grow quadratically with η. (These are only asymptotic solutions, but the Mészáros equation can actually be solved analytically at all times). During Λ domination, the function $a(\eta)$ is more complicated, and $\Omega_{\rm cdm}$ decreases. With a bit of work, one can show that $\delta_{\rm cdm}$ grows at a smaller rate than during matter domination (i.e. slower than η^2), and that the reduction of the growth rate does not depend on k.

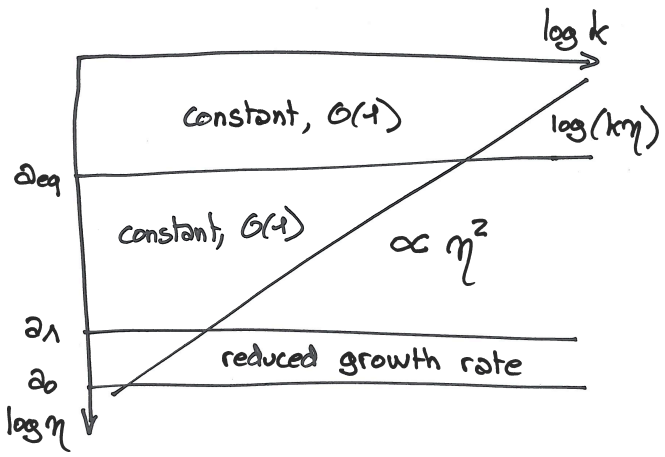


Figure 5.13: Qualitative evolution of the transfer function $\delta_{\rm cdm}(\eta, k)$ (normalized as usual to $\mathcal{R}(k) = 1$ at initial time) in different regimes: during radiation domination, matter domination, Λ domination, and on super/sub-Hubble scales.

In summary, during radiation domination, $\delta_{\rm cdm}(\eta, k)$ is constant on super-Hubble scales and grows logarithmically on sub-Hubble scales. A more precise calculation would show that up to a numerical factor of order one, $\delta_{\rm cdm}$ is given on sub-Hubble scales by $\delta_{\rm cdm}(\eta, k) = \log(k\eta)$. During matter domination, $\delta_{\rm cdm}(\eta, k)$ is still constant on super-Hubble scales, and grows like η^2 on sub-Hubble scales. Finally, during Λ domination, it grows more slowly. These different behaviors are reported in Fig. 5.13.

This simple discussion is sufficient for understanding the shape of the matter power spectrum at different time. Like in a cartoon, Fig. 5.14 shows this shape at four different times: at some initial time when all relevant modes are super-Hubble; at radiation/matter equality; at matter/Λ equality; and today. Let us comment these plots. We first need to define the comoving wavenumbers corresponding to wavelengths crossing the Hubble radius at the time of radiation/matter equality, of matter/Λ equality, and today:

$$k_{\rm eq} = a_{\rm eq} H_{\rm eq} \ , \qquad k_\Lambda = a_\Lambda H_\Lambda \ , \qquad k_0 = a_0 H_0 \ . \tag{5.89}$$

We can now review the evolution of $P(k)$ with respect to time, following the same steps as in Fig. 5.14.

1. At initial time, if we assume a scale-invariant spectrum with $n_{\rm s} = 1$, we see from eq. (5.86) that $P(k) \propto k^{-3}$, with an amplitude given by $A_{\rm s}$.

2. During radiation domination, modes grow logarithmically inside the Hubble radius, like $\log(k\eta)$. At equality, super-Hubble modes ($k_{\rm eq}\eta \ll 1$) are still shaped like at initial time, while sub-Hubble modes ($k_{\rm eq}\eta \gg 1$) have been enhanced by a factor $[\delta_{\rm cdm}(\eta_{\rm eq}, k)/\delta_{\rm cdm}(\eta_{\rm ini}, k)]^2 \simeq [\log(k\eta_{\rm eq})]^2$. The two asymptots of $P(k)$ are then given by k^{-3} for $k \ll k_{\rm eq}$ and $k^{-3}[\log(k)]^2$ for $k \gg k_{\rm eq}$.

3. At the end of matter domination, when $a = a_\Lambda$, modes still outside the Hubble radius keep being shaped like at initial time. This concerns all modes with $k \ll k_\Lambda$. Modes $k \gg k_{\rm eq}$ have been amplified during matter

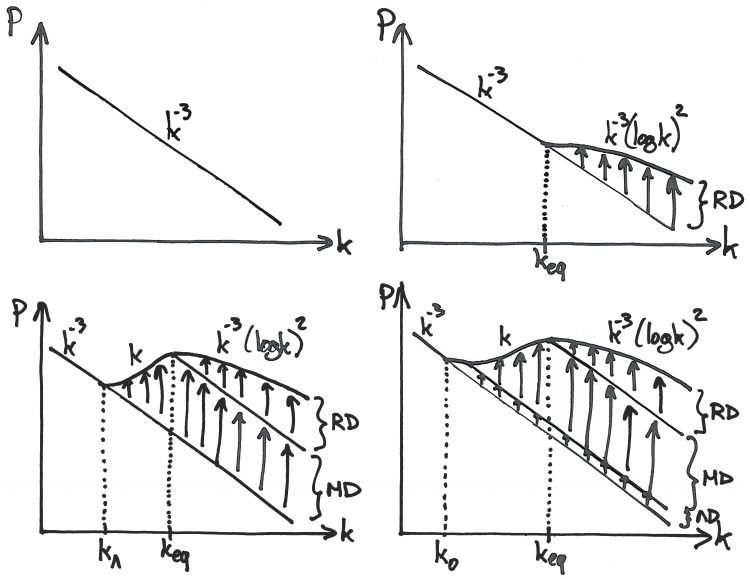


Figure 5.14: Shape of the matter power spectrum $P(k)$ (log-log scale) at four different times: *(upper left)* when initial conditions are imposed (and all wavenumbers are super-Hubble); *(upper right)* at radiation/matter equality (arrows show the logarithmic growth during radiation domination); *(lower left)* at matter/Λ equality (lower set of arrows show the growth during matter domination); *(lower right)* today (lower set of arrows show the growth during Λ domination).

domination by a factor $[\delta_{\rm cdm}(\eta_\Lambda, k)/\delta_{\rm cdm}(\eta_{\rm eq}, k)]^2 \simeq (\eta_\Lambda/\eta_{\rm eq})^4$. This factor does not depend on k and preserves the shape of the power spectrum on those scales. Finally, intermediate modes entering the Hubble scale during matter domination have been amplified by $(\eta_\Lambda/\eta_*)^4$, where η_* is their time of Hubble crossing, given approximately by $\eta_* = 1/k$. Hence they have been amplified by $(k\eta_\Lambda)^4$. Putting all these informations together, we see that the spectrum has three branches, scaling respectively like:

- $P(k) \propto k^{-3}$ for $k < k_\Lambda$,
- $P(k) \propto k^{-3}k^4 = k$ for $k_\Lambda < k < k_{\rm eq}$,
- $P(k) \propto k^{-3}(\log k)^2$ for $k > k_{\rm eq}$.

4. During Λ domination, $\delta_{\rm cdm}(k, \eta)$ grows more slowly than η^2, but it still grows at the same rate for all sub-Hubble modes. So the shape of the power spectrum today is unaltered by this stage, and given by:

   - $P(k) \propto k^{-3}$ for $k < k_0$,
   - $P(k) \propto k^{-3}k^4 = k$ for $k_\Lambda < k < k_{\rm eq}$,
   - $P(k) \propto k^{-3}(\log k)^2$ for $k > k_{\rm eq}$.

   We do not enter into details for the small range of modes obeying $k_0 < k < k_\Lambda$.

All this discussion was carried under the assumption of a scale-invariant primordial spectrum. If $n_{\rm s} \neq 1$, the above shape should simply be rescaled by

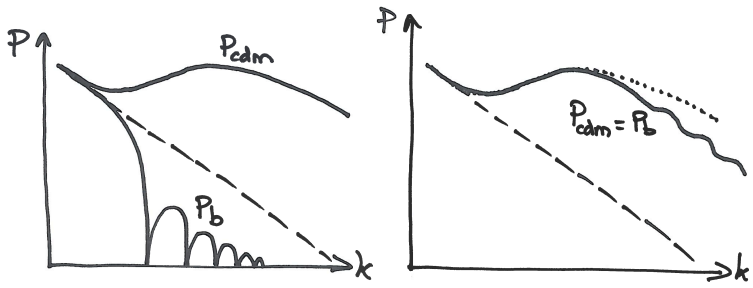


Figure 5.15: *(Left)* Shape of the baryon and CDM power spectra just before the baryon drag time, compared to the shape of the primordial spectrum (dashed line). *(Right)* Shape of the matter power spectrum well after the baryon drag time, when baryons and CDM perturbations have reached gravitational equilibrium, compared to the shape of the primordial spectrum (dashed lines) and to the spectrum of a CDM-dominated universe (dotted line).

$k^{n_\mathrm{s}-1}$, and the three branches of the power spectrum are given by:

- $P(k) \propto k^{n_\mathrm{s}-4}$ for $k < k_0$,
- $P(k) \propto k^{n_\mathrm{s}}$ for $k_\Lambda < k < k_\mathrm{eq}$,
- $P(k) \propto k^{n_\mathrm{s}-4}(\log k)^2$ for $k > k_\mathrm{eq}$.

This closes the presentation of the shape of $P(z,k)$ in the limit $\Omega_\mathrm{b} \ll \Omega_\mathrm{cdm}$. What remains to be seen is the impact of a non-negligible baryon fraction on the power spectrum.

Baryon corrections. We must define a new important time in the evolution of the universe: the baryon drag time. The photon decoupling time was defined through the maximum of the photon visibility function $g(\eta)$. But there are many more photons than baryons in the universe. Hence, for some amount of time after photon decoupling, baryons keep tracking photon perturbations: in other words, the baryon population decouples later than the photon population (seen as a whole).

Until the baryon drag time, we know that $\delta_\mathrm{b} = \frac{3}{4}\delta_\gamma$. After that time, baryons do not experience significant Thomson scattering anymore. They only feel gravity, and collapse in gravitational potential wells.

We know qualitatively the behavior of the baryon transfer function $\delta_\mathrm{b}(\eta,k)$ just before baryon drag, since in the CMB section we have studied the behavior of the photon transfer function. We know that $\delta_\mathrm{b} = \frac{3}{4}\delta_\gamma$ is constant on super-sound-horizon scales, experiences stationary oscillations on sub-sound-horizon scales during radiation domination, and finally damped oscillations on sub-sound-horizon scales during matter domination. Moreover, we expect that at any time — and particularly at the approach of decoupling — the relation $\delta_\mathrm{b} = \frac{3}{4}\delta_\gamma$ breaks on very small wavelengths comparable to the mean free path of baryons in the imperfect baryon-photon fluid, but here we will not discuss such small scales.

The behavior of the CDM transfer function $\delta_\mathrm{cdm}(\eta,k)$ before baryon drag is still given by the Mészáros equation: deep inside the sound horizon, baryons are much less clustered than CDM, so that CDM is almost self-gravitating.

In the left plot of Fig. 5.15, we sketch the qualitative behavior of the individual power spectrum of baryons and CDM just before baryon drag ($P_\mathrm{b}(k)$ and $P_\mathrm{cdm}(k)$ are defined as in eq. (5.86), with δ_m replaced either by δ_b or δ_cdm).

After baryon drag, baryons collapse in potential gravitational wells. Since

$\delta_{\rm b}$ grows, CDM will start to feel the gravity of baryons. Finally, because CDM and baryons are two collisionless species feeling the same gravitational forces, they will equilibrate with $\delta_{\rm b} = \delta_{\rm cdm}$. The gravitational potential is then related to this common value by the (00) Einstein equation (or its Poisson sub-Hubble limit). This is not true however on very small scales, for which the baryon pressure cannot be neglected, but here we do not discuss such small scales.

Hence, after a quick relaxation period, the two power spectra of baryons and CDM are equal to each other, $P_{\rm b}(k) = P_{\rm cdm}(k) = P(k)$. In order to relate this common power spectrum to the individual power spectrum of baryons and CDM before baryon drag, one must follow a matching process described in details in *Eisenstein and Hu 1997*. Intuitively, the matching depends on the relative weight of baryons and CDM, i.e. on the ratio $\Omega_{\rm b}/\Omega_{\rm cdm}$. In the limit $\Omega_{\rm b} \ll \Omega_{\rm cdm}$, what we wrote at the beginning of this section applies. In the limit $\Omega_{\rm b} \gg \Omega_{\rm cdm}$, the power spectrum is very suppressed with respect to the CDM-dominated case, with a much more negative slope on average, and some large oscillations corresponding to photon-baryon acoustic waves before decoupling. Finally, if $\Omega_{\rm b}$ is a bit smaller than $\Omega_{\rm cdm}$ but not negligible (as it is the case in our universe), the power spectrum departs slightly from a pure CDM one, with a smooth step-like suppression plus small oscillatory patterns, which are the smoking gun in the recent universe of baryon-photon acoustic oscillations happening before photon decoupling. This is illustrated on the right plot of Fig. 5.15. These oscillations are called Baryon Acoustic Oscillations (BAO).

LSS observations can only probe the power spectrum on scales much smaller than the radius of the observable universe, i.e. than the current Hubble radius. For this reason, the first of the three branches described above is unobservable[5]. We can only measure the second and third branches, behaving respectively like $P(k \ll k_{\rm eq}) \propto k^{n_{\rm s}}$ and, in first approximation, $P(k \gg k_{\rm eq}) \propto [k^{n_{\rm s}-4}(\log k)^2]$, plus the titling and superimposed oscillations coming from baryons.

5.3.3 Parameter dependence

The discussion presented in the previous section allows us to understand which effects and which parameters impact the shape of the matter power spectrum in the minimal (flat) ΛCDM model, parametrized by

$$\{A_{\rm s}, n_{\rm s}, \omega_{\rm b}, \omega_{\rm m}, \Omega_\Lambda, \tau_{\rm reio}\} \tag{5.90}$$

(see Sec. 5.2.6 for details on this parametrization). We can already notice that the reionization optical depth is relevant for the CMB spectrum but not for $P(k)$, since it only impacts the scattering rate of photons in the recent universe. Other effects are described in Table 5.2. Below, we give more details on these effects.

(P1) The time of equality determines the scale $k_{\rm eq}$ of the power spectrum peak. More precisely, if this scale is expressed in units of $h\,{\rm Mpc}^{-1}$, which is the usual convention, then one can show that the scale of the maximum depends on both $z_{\rm eq}$ and $\Omega_{\rm m}$ (i.e. on our parameter basis on $\omega_{\rm m}$ and $\Omega_\Lambda = 1 - \Omega_{\rm m}$).

[5]The first unobservable branch is actually gauge-dependent: it behaves like $k^{n_{\rm s}-4}$ in the Newtonian gauge, but not in other gauges. Note that truly observable quantities are always gauge-invariant. We have been a bit uncareful when saying that LSS data probes the power spectrum of $\delta_{\rm m}$. In fact, different LSS observations probe different quantities, each of them being gauge-invariant. However, well inside the Hubble region, all these quantities coincide with each other and with the spectrum of $\delta_{\rm m}$ computed in an arbitrary gauge (up to small corrections that may become important in the future, but not with current experimental sensitivities).

Table 5.2: Independent leading effects controlling the shape of the matter power spectrum $P(k)$ in the minimal ΛCDM model.

	Effect	Relevant quantity	Parameter
(P1)	scale of the maximum	$k_{\rm eq}$	$\omega_{\rm m}, \Omega_\Lambda$
(P2)	slope for $k \gg k_{\rm eq}$ and BAO amplitude	$\Omega_{\rm b}/\Omega_{\rm cdm}$	$\omega_{\rm b}, \omega_{\rm m}$
(P3)	BAO scale	$r_{\rm s}(\eta_{\rm drag})$	$\omega_{\rm b}, \omega_{\rm m}$
(P4)	Global amplitude	amplitude of primordial spectrum and duration of ΛD	$A_{\rm s}, \Omega_\Lambda$
(P5)	Global tilt	tilt of primordial spectrum	$n_{\rm s}$

(P2) The baryon abundance (relative to CDM) is crucial at the matching time: a high baryon abundance leads to more suppression of the power spectrum for $k \geq k_{\rm eq}$, and to more pronounced BAOs.

(P3) We have seen that the scale of acoustic oscillations is set by the sound horizon at a given time. Since photons decouple at $\eta_{\rm dec}$, the scale of oscillations on the last scattering surface is set by $d_s(\eta_{\rm dec})$, corresponding to the comoving scale $r_s(\eta_{\rm dec}) = d_s(\eta_{\rm dec})/a(\eta_{\rm dec})$. Similarly, since baryons decoupled at $\eta_{\rm drag}$, the scale of BAOs depends on $d_s(\eta_{\rm drag})$ at baryon drag, corresponding to the comoving scale $r_s(\eta_{\rm drag}) = d_s(\eta_{\rm drag})/a(\eta_{\rm drag})$. As explained in Sec. 5.2.6, the sound horizon at a given time depends on $\omega_{\rm b}$ and $\omega_{\rm m}$. In the case of $d_s(\eta_{\rm drag})$ the dependence on $\omega_{\rm b}$ is even stronger because the time of baryon drag itself depends strongly on the baryon abundance.

(P4) The global amplitude depends of course on the primordial spectrum amplitude, i.e. on $A_{\rm s}$. Also, we have seen that during Λ domination, the growth rate of fluctuations is reduced with respect to matter domination, but is independent of k for all sub-Hubble scales. Hence, Ω_Λ also affects the global amplitude of the power spectrum.

(P5) The global tilt depends of course on the primordial spectrum tilt, i.e. on $n_{\rm s}$.

This discussion shows that in principle, a precise measurement of the matter power spectrum today (or at a given redhsift) would allow to measure independently $\omega_{\rm b}$, $\omega_{\rm m}$, $n_{\rm s}$, $A_{\rm s}$ and Ω_Λ (assuming a flat ΛCDM universe). In practice, given the limited precision of current data sets, some of the effects described above are degenerate with each other, and matter power spectrum observations are mainly useful in combination with CMB observations, since they bring independent and complementary information.

Future experiments will use all the discriminating power of the matter power spectrum. They will perform accurate measurements of $P(k, z)$ at different redshifts. This is crucial for at least two reasons:

- First, the effect of Ω_Λ in (P5) is redshift dependent: at different redshifts, fluctuations have spent more or less time during the Λ dominated regime, and the power spectrum amplitude has been more or less affected. Hence, the comparison of the amplitude at different redshifts allows to measure

Ω_Λ, or more generally to test the compatibility of the data with a cosmological constant (rather than some dynamical Dark Energy model).

- Second, BAOs appear at a fixed comoving wavenumber $k_{\rm BAO}$ in Fourier space (related to $r_s(\eta_{\rm drag})$), but if we measure it in different LSS datasets at different redshifts (corresponding to different shells in real space), this scale will be seen under different angles[6]. By comparing the BAO angle at different redshifts, one can reconstruct the angular diameter distance, and therefore the expansion history at different redshifts. This is another way of measuring Ω_Λ or testing the Λ model versus DE models.

[6]Here we are assuming that the BAO scale is measured in each redshift shell transversally, i.e. in the direction orthogonal to the line of sight. Real experiments probe the BAO scale both transversally and longitudinally, so the situation is a bit more subtle than in this simplified discussion.

Chapter 6

Cosmological observations

6.1 Minimal set of parameters

One can build arbitrarily complicated cosmological models with an arbitrary number of physical ingredients and free parameters. But given what we have seen in the course, a number of parameters are unavoidable:

- by postulating that the Universe is described by the Friedmann-Lemaître-Robertson-Walker metric, we are forced to introduce one parameter: the Hubble rate today H_0 (or the reduced Hubble rate h), and to wonder about the value of two more parameters: the fractional density of cosmological constant Ω_Λ and the effective fractional density of curvature Ω_k.
- the presence of a photon background is confirmed by the observation of Penzias & Wilson (and their successors), and the presence of a neutrino background is strongly suggested by the assumption of thermal equilibrium in the early universe. In principle this should lead to two new parameters: the physical density parameters ω_γ and ω_ν. However we have seen in the course that given the precise measurement of the CMB temperature since many decades, ω_γ can be considered as a fixed parameter:

$$\begin{aligned} \omega_\gamma &\equiv \Omega_\gamma h^2 \\ &= \frac{\bar{\rho}_\gamma^0}{\bar{\rho}_c^0} h^2 \\ &= \left(\frac{\pi^2}{15} T_0^4\right) \left(\frac{8\pi G}{3H_0^2}\right) h^2 \\ &= \frac{8\pi^3 T_0^4}{45 (H_0/h)^2 M_P^2} \ . \end{aligned} \tag{6.1}$$

  Today we know that $T_0 = 2.7255 \pm 0.0006$ K (68% CL, Fixsen et al. 2009, ApJ, 707, 916), leading to $\omega_r = (2.472 \pm 0.002) \times 10^{-5}$. Moreover, if nothing strange occurs in the universe near the time of neutrino decoupling and if the sequence of events after neutrino decoupling is similar to what we described in Chapter 3, the neutrino abundance is fixed relative to that of photons, and the total radiation density reads:

$$\begin{aligned} \omega_{\rm r} &\equiv \omega_\gamma + \omega_\nu \\ &= \left[\frac{\pi^2}{15} T_0^4 + N_{\rm eff} \times \frac{7}{8} \times \frac{\pi^2}{15} T_{\nu 0}^4\right] \left(\frac{8\pi G}{3H_0^2}\right) h^2 \\ &= \left[1 + N_{\rm eff} \times \frac{7}{8} \times \left(\frac{4}{11}\right)^{4/3}\right] \omega_\gamma \end{aligned} \tag{6.2}$$

Here, according to the simple calculations of Chapter 3, $N_{\rm eff}$ should be equal to the number of neutrino species, $N_{\rm eff} = 3$ (and not 6: the abundance of neutrinos plus anti-neutrinos is already counted in the previous formula). However, refined numerical calculations taking into account various effects (non-instantaneous decoupling near positron annihilation, neutrino oscillations, ...) gives a small correction to this result, usually absorbed in a redefinition of the parameter $N_{\rm eff}$, such that $N_{\rm eff} = 3.046$ (Mangano et al., Nucl.Phys. B729 (2005) 221-234). With $N_{\rm eff} = 3.046$ and $T_0 = 2.7255$ K one gets $\omega_{\rm r} = 4.183 \times 10^{-5}$.

- we know that there must be baryons around because we see them, and also because they are crucial for Nucleosynthesis and CMB physics. We know as well that there must be cold dark matter, given the observations mentioned in Chapter 6. This leads to two new parameters $\omega_{\rm b}$ and $\omega_{\rm cdm}$ summing up to the total matter density $\omega_{\rm m}$.

- to describe the primordial spectrum of scalar perturbations, and more precisely of curvature fluctuations, we need at least an amplitude A_s and a tilt n_s, where s refers to "scalar" (the tilt will be better motivated in Chapter 7).

- we briefly mentioned reionisation. It should be described by at least one parameters, the optical depth to reionisation $\tau_{\rm reio}$, giving the amplitude of the plateau in the optical depth evolution curve, stretching from the end of recombination to the beginning of reionisation.

At this point we are left with a set of 8 parameters:

$$\{h, \Omega_\Lambda, \Omega_k, \omega_b, \omega_m, A_s, n_s, \tau_{\rm reio}\} \ . \tag{6.3}$$

However they are not all independent because the relation $\sum \Omega_i = 1$ gives

$$\frac{\omega_m}{h^2} + \Omega_\Lambda + \Omega_k = 1 \ . \tag{6.4}$$

So we need to drop one parameter in the basis. Usually one drops either h or Ω_Λ. Here we will usually drop h.

6.2 Brief history of the minimal cosmological model(s)

We gave strong motivations for the presence of photons, neutrinos, baryons and CDM, but not for Λ and curvature. For the latter parameters, we don't understand well enough the fundamental theory describing the universe at early times and high energies for being able to make a neat prediction, like we do, e.g., for neutrinos. So the answer can only come from observations.

Several decades ago, when observations were not very accurate, people tried to stick to the simplest possible model in terms of number parameters, assuming $\Omega_\Lambda = \Omega_k = 0$. The minimal model was then called "standard Cold Dark Matter" (sCDM), with 5 free independent parameters $\{\omega_b, \omega_m, A_s, n_s, \tau_{\rm reio}\}$. In the 70's-80's, this model became better and better established, thanks to observations of the CMB background, on the theory of Nucleosynthesis combined with the first good observations of primordial element abundances, and on observations of the distribution and of the dynamics of clusters and galaxies proving the existence of dark matter. There were still some doubts concerning the fact that dark matter is cold (heavy particles with small velocities) or hot (light neutrinos). So sCDM

had a competitor, called sHDM. In the early nineties it became clear that Hot Dark Matter cannot cluster enough to explain the abundance of clusters today, and sHDM was disqualified in favour of sCDM.

In the early and mid-nineties, several observations started to be in tension with sCDM, and in particular, those of the age of the most distant quasars. Given observational bounds on h, the sCDM model predicted a universe too young to accommodate such objects. There were two obvious solutions: assuming a cosmological constant ($\Omega_\Lambda > 0$), or a negatively curved/open universe ($k < 0$, $\Omega_k > 0$).

The question was settled with the first accurate measurements of the supernovae luminosity–redshift relation in 1998: fitting the data required *at least* $\Omega_\Lambda > 0$. The data could be fitted with $\Omega_\Lambda > 0$ and $\Omega_k = 0$, so the minimal cosmological model became ΛCDM with 6 free independent parameters: $\{\omega_b, \omega_m, \Omega_\Lambda, A_s, n_s, \tau_{\rm reio}\}$.

Since 1998 observations have done amazing progress, in particular, as far as CMB anisotropies are concerned. Many cosmologists thought that with such progress we would detect some new effects, and switch to a minimal model with more ingredients and more parameters. In fact this was not the case. The new data has confirmed ΛCDM with incredible precision, and not detected any effect requiring more ingredients.

The next sections will review the main categories of observations described in this short summary, following the same order (referring to their chronological importance): abundance of primordial elements, age of the universe, supernovae, CMB anisotropies and galaxy correlation function.

6.3 Abundance of primordial elements

In sections 3.3.5, we have seen that the theory of Nucleosynthesis can predict the abundance of light elements formed in the early universe, when the energy density was of order $\rho \sim (0.07\,\mathrm{MeV})^4$. After Nucleosynthesis, there are no more nuclear reactions in the universe, excepted in the core of stars. So, today, in regions of the universe which were never filled by matter ejected from stars, the proportion of light elements is still the same as it was just after Nucleosynthesis. Fortunately, the universe contains clouds of gas fullfilling this criteria, and the abundance of deuterium, helium, etc. can be measured in such regions (e.g. by spectroscopy). The results can be directly compared with theoretical predictions.

The predictions presented in this course were based on a very simplistic description of Nucleosynthesis. Precise predictions arise from codes simulating the evolution of a system of many different reactions. Table 6.1 shows, for instance, the first 40 reactions used in the public code PARTHENOPE[1]. In section 3.3.5, we only studied the reactions called 1 and 12 in this table.

Numerical simulation of Nucleosynthesis accurately predict all relative abundances as a function of the only free parameter in the theory, the baryon density. We remember that the temperature at which light elements start forming is fixed by equation (3.67) and depends on $\eta_b \equiv n_b/n_\gamma \sim 10^{-10}$, which precise value is given by $\eta_b = 5.5 \times 10^{-10}(\omega_b/0.020)$ (note that η_b is defined at any time between positron annihilation and today: it is constant in this range). Hence, relative abundances depend on ω_b. Figure 6.1 shows the dependence of the abundance of ^{4}He, D, ^{3}He and ^{7}Li on η_b. Using this dependence, Nucleosynthesis experts are able to convert measurements of the primordial Helium and Deuterium abundance into a prediction for the baryon density.

[1] `http://parthenope.na.infn.it/`

No.	Reaction	Type	No.	Reaction	Type
1	$n \longrightarrow p$	*weak*	22	$^6Li + p \longrightarrow \gamma + {}^7Be$	(p,γ)
2	$^3H \rightarrow \bar{\nu}_e + e^- + {}^3He$	*weak*	23	$^6Li + p \longrightarrow {}^3He + {}^4He$	^{3}He Pickup
3	$^8Li \rightarrow \bar{\nu}_e + e^- + 2\,{}^4He$	*weak*	24	$^7Li + p \longrightarrow {}^4He + {}^4He$	^{4}He Pickup
4	$^{12}B \rightarrow \bar{\nu}_e + e^- + {}^{12}C$	*weak*	24 bis	$^7Li + p \longrightarrow \gamma + {}^4He + {}^4He$	(p,γ)
5	$^{14}C \rightarrow \bar{\nu}_e + e^- + {}^{14}N$	*weak*	25	$^4He + {}^2H \longrightarrow \gamma + {}^6Li$	(d,γ)
6	$^8B \rightarrow \nu_e + e^+ + 2\,{}^4He$	*weak*	26	$^4He + {}^3H \longrightarrow \gamma + {}^7Li$	(t,γ)
7	$^{11}C \rightarrow \nu_e + e^+ + {}^{11}B$	*weak*	27	$^4He + {}^3He \longrightarrow \gamma + {}^7Be$	(^{3}He,γ)
8	$^{12}N \rightarrow \nu_e + e^+ + {}^{12}C$	*weak*	28	$^2H + {}^2H \longrightarrow n + {}^3He$	^{2}H Strip.
9	$^{13}N \rightarrow \nu_e + e^+ + {}^{13}C$	*weak*	29	$^2H + {}^2H \longrightarrow p + {}^3H$	^{2}H Strip.
10	$^{14}O \rightarrow \nu_e + e^+ + {}^{14}N$	*weak*	30	$^3H + {}^2H \longrightarrow n + {}^4He$	^{2}H Strip.
11	$^{15}O \rightarrow \nu_e + e^+ + {}^{15}N$	*weak*	31	$^3He + {}^2H \longrightarrow p + {}^4He$	^{2}H Strip.
12	$p + n \longrightarrow \gamma + {}^2H$	(n,γ)	32	$^3He + {}^3He \longrightarrow p + p + {}^4He$	(^{3}He,2p)
13	$^2H + n \longrightarrow \gamma + {}^3H$	(n,γ)	33	$^7Li + {}^2H \longrightarrow n + {}^4He + {}^4He$	(d,n α)
14	$^3He + n \longrightarrow \gamma + {}^4He$	(n,γ)	34	$^7Be + {}^2H \longrightarrow p + {}^4He + {}^4He$	(d,p α)
15	$^6Li + n \longrightarrow \gamma + {}^7Li$	(n,γ)	35	$^3He + {}^3H \longrightarrow \gamma + {}^6Li$	(t,γ)
16	$^3He + n \longrightarrow p + {}^3H$	charge ex.	36	$^6Li + {}^2H \longrightarrow n + {}^7Be$	^{2}H Strip.
17	$^7Be + n \longrightarrow p + {}^7Li$	charge ex.	37	$^6Li + {}^2H \longrightarrow p + {}^7Li$	^{2}H Strip.
18	$^6Li + n \longrightarrow {}^3H + {}^4He$	^{3}H Pickup	38	$^3He + {}^3H \longrightarrow {}^2H + {}^4He$	(^{3}H,d)
19	$^7Be + n \longrightarrow {}^4He + {}^4He$	^{4}He Pickup	39	$^3H + {}^3H \longrightarrow n + n + {}^4He$	(t,n n)
20	$^2H + p \longrightarrow \gamma + {}^3He$	(p,γ)	40	$^3He + {}^3H \longrightarrow p + n + {}^4He$	(t,n p)
21	$^3H + p \longrightarrow \gamma + {}^4He$	(p,γ)			

Table 6.1: The first forty reactions used in the Nucleosynthesis code PARTENOPE. *Table taken from [arXiv:0705.0290] by Ofelia Pisanti et al.*

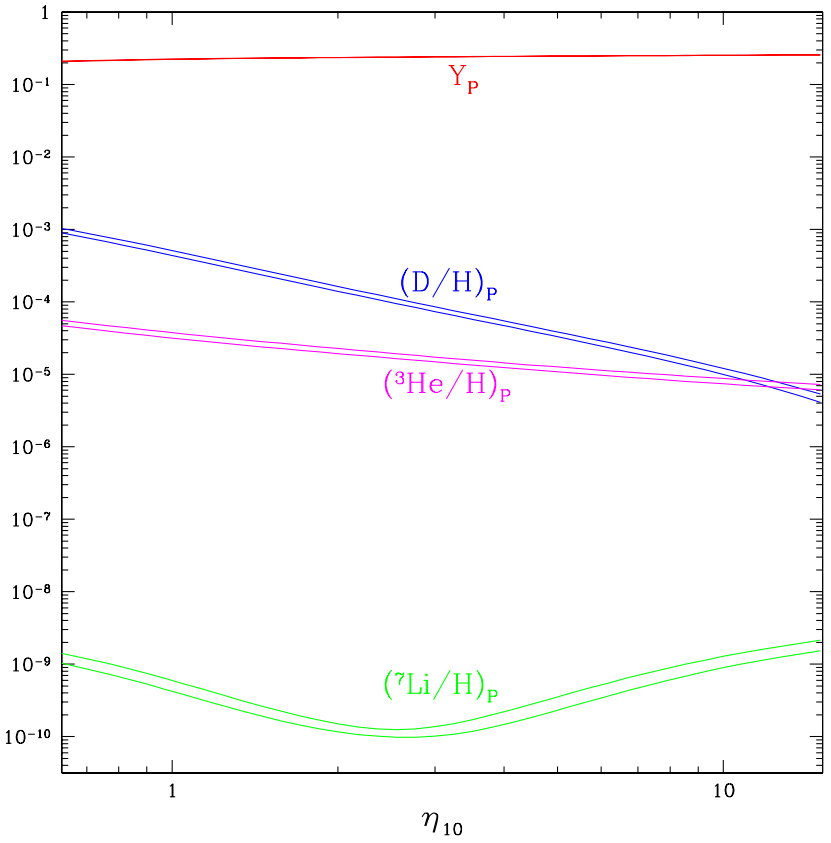


Figure 6.1: The Nucleosynthesis-predicted primordial abundances of D, ^{3}He, ^{7}Li (relative to hydrogen by number), and the ^{4}He mass fraction (Y_P), as functions of the baryon abundance parameter $\eta_{10} \equiv 10^{10}\eta_b$. The widths of the bands reflect the uncertainties in the nuclear and weak interaction rates. *Plot taken from Int.J.Mod.Phys. E15 (2006) 1-36 [arXiv:astro-ph/0511534v1] by Gary Steigman.*

Currently, the agreement between Nucleosynthesis and other types of cosmological observations is impressive. The latest CMB observations from the planck satellite indicate a baryon density

$$\omega_b \equiv \Omega_b h^2 = 0.0223 \pm 0.0002 \quad (68\% CL). \tag{6.5}$$

When reported in Nucleosynthesis calculations, this number leads to predictions for the ^{4}He and D abundance that are in very good agreement with observations. It is considered as a huge success for cosmology that two very different techniques (Nucleosynthesis, which relies on nuclear physics when $T \sim (0.01-1)$ MeV, and CMB, relaying on relativistic hydrodynamics and QED when $T \sim (0.1-100)$ eV) give compatible results for ω_b.

Hence, for $h = 0.67$ (the current best-fit value), the baryon fraction is of the order of $\Omega_b \sim 0.05$: approximately five percent of the universe density comes from ordinary matter. This is already more than the sum of all luminous matter, which represents only one per cent: so, 80% of ordinary matter is not even visible.

Note that if $\omega_{\rm r}$ was a free parameter, the outcome of Nucleosynthesis would also depend crucially on $\omega_{\rm r}$. So, Nucleosynthesis can also be used as a tool for testing the fact that eq. (6.2) with $N_{\rm eff} \simeq 3$ is correct. It turns out to be the case: primordial element abundances provide a measurement of $\omega_{\rm r}$ precise at the 10% level, and perfectly compatible with eq. (6.2).

6.4 Age of the universe

The age of the universe can be conveniently computed once the function $H(a)/H_0$ or $H(z)/H_0$ is known. This function follows from the Friedmann equation divided by H_0^2:

$$\begin{aligned}
\frac{H^2}{H_0^2} &= \frac{\bar{\rho}_{tot}}{\bar{\rho}_c} - \frac{k}{a^2 H_0^2} \\
&= \Omega_r \left(\frac{a_0}{a}\right)^4 + \Omega_m \left(\frac{a_0}{a}\right)^3 + \Omega_k \left(\frac{a_0}{a}\right)^2 + \Omega_\Lambda \qquad (6.6) \\
&= \Omega_r \left(1+z\right)^4 + \Omega_m \left(1+z\right)^3 + \Omega_k \left(1+z\right)^2 + \Omega_\Lambda \ , \qquad (6.7)
\end{aligned}$$

with the constraint that $\Omega_r + \Omega_m + \Omega_k + \Omega_\Lambda = 1$ by construction. Since $H = da/(adt)$, we can write:

$$dt = \frac{da}{aH} = -\frac{dz}{(1+z)H} \ . \tag{6.8}$$

Hence, the age of the universe can be computed from the integral

$$t = \int_0^{a_0} \frac{da}{aH} = H_0^{-1} \int_0^{a_0} \frac{da}{a} \left(\frac{H_0}{H(a)}\right) \ , \tag{6.9}$$

or equivalently from

$$t = \int_0^{\infty} \frac{dz}{(1+z)H} = H_0^{-1} \int_0^{\infty} \frac{dz}{1+z} \left(\frac{H_0}{H(z)}\right) \ . \tag{6.10}$$

This integral converges with respect to the boundary corresponding to the initial singularity, $a \longrightarrow 0$ or $z \longrightarrow \infty$. Actually, it is easy to show that the radiation dominated period gives a negligible contribution to the age of the universe, hence the term proportional to Ω_r can be omitted in the integral. If the universe is

matter-dominated today ($\Omega_\Lambda = \Omega_k = 0$), then $\Omega_m = 1$ and the age of the universe is simply given by:

$$t = H_0^{-1} \int_0^\infty dz\,(1+z)^{-5/2} = \frac{2}{3H_0} = 6.52 h^{-1}\text{Gyr} \ , \tag{6.11}$$

where 1 Gyr $\equiv$ 1 billion years. If $\Omega_\Lambda > 0$ and/or $\Omega_k < 0$ (negatively curved universe), the ratio $H(z)/H_0$ decreases with respect to the $\Omega_\Lambda = \Omega_k = 0$ case for all values of z corresponding to Λ or curvature domination. For $\Omega_k > 0$ (closed universe), it increases. Hence, the age of the universe increases with respect to $6.52h^{-1}$Gyr if $\Omega_\Lambda > 0$ and/or $\Omega_k < 0$, and decreases if $\Omega_k > 0$.

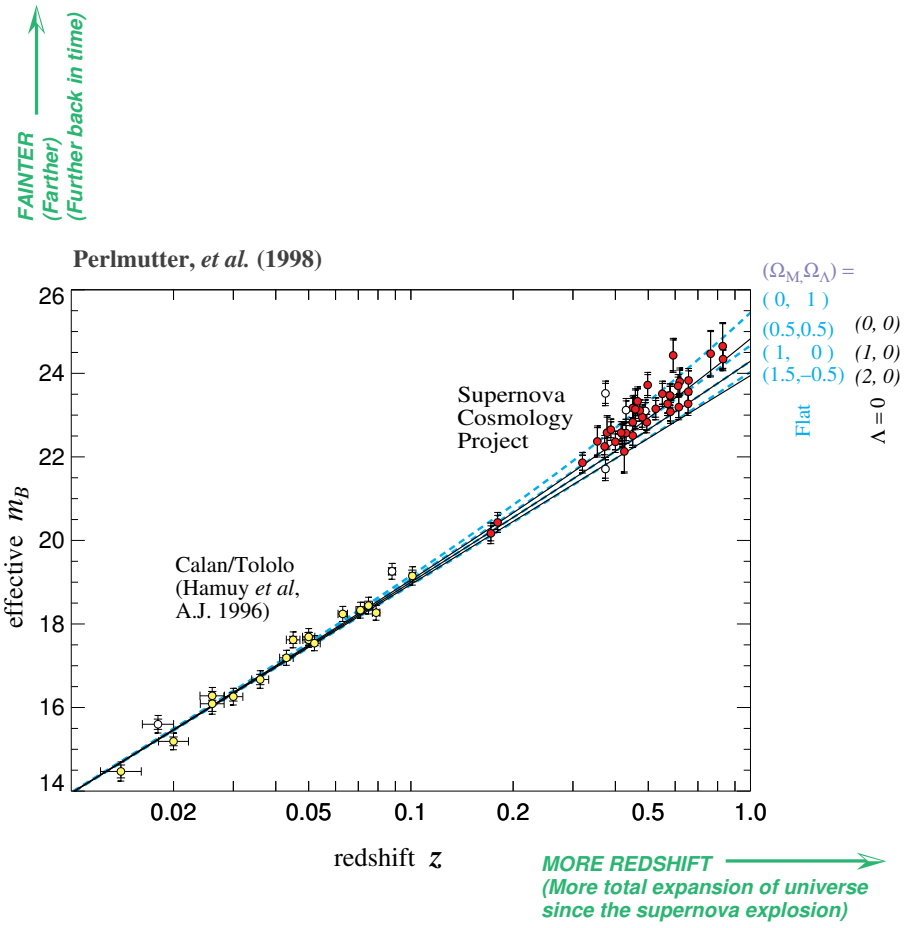
The age of of a few specific object in the universe can be evaluated with a number of techniques, e.g. by nucleochronology (studying the radioactive decay of isotopes inside an object, exactly like in the ^{14}C method used in archeology); or by measuring the cooling of stars in their final state, called "white dwarfs", and comparing with the mean evolution curve of white dwarfs; etc. If the age of an object is found to be extremely large, it provides a lower bound on the age of the universe itself. This set of observations sets a reliable lower bound on the age of the universe: $t > 11$Gyr. This is incompatible with the matter-dominated universe of eq. (6.11) unless $h < 0.59$. But different ways to measure the Hubble parameter point at $h \sim 0.67$. Hence, these observations provide a strong hint that that the universe is either negatively curved or Λ–dominated today. This "age problem" was already known in the 90's.

6.5 Luminosity of Type Ia supernovae

The evidence for a non–zero cosmological constant has increased considerably in 1998, when two independent groups studied the apparent luminosity of distant type Ia supernovae (SNIa). For this type of supernovae, astronomers believe that there is a simple relation between the absolute magnitude and the luminosity decay rate. In other words, by studying the rise and fall of the luminosity curve during a few weeks, one can deduce the absolute magnitude of a given SNIa. Therefore, it can be used in the same way as cepheids, as a probe of the luminosity distance – redshift relation. In addition, supernovae are much brighter than cepheids, and can be observed at much larger distances (until redshifts of order one or two). While observable cepheids only probe short distances, where the luminosity distance – redshift relation only gives the Hubble law (the proportionality between distance and redshift), the most distant observable SNIa's are in the region where general relativity corrections are important: so, they can provide a measurement of the scale factor evolution (see section 2.2.2).

On figure 6.2, the various curves represent the effective magnitude–redshift relation, computed for various choices of Ω_M and Ω_Λ. The effective magnitude m_B plotted here is essentially equivalent to the luminosity distance d_L, since it is proportional to $\log[d_L]$ plus a constant. For a given value of H_0, all the curves are asymptotically equal at short distance. Significant differences show up only at redshifts $z > 0.2$. Each red data point corresponds to a single supernovae in the first precise data set: that of the "Supernovae Cosmology Project", released in 1998. Even if it is not very clear visually from the figure, a detailed statistical analysis of this data revealed that a flat matter–dominated universe (with $\Omega_m = 1$, $\Omega_\Lambda = 0$) was excluded. This result has been confirmed by various more recent data sets. The top panel of figure 6.3 shows the luminosity distance – redshift diagram for the SNLS data set, released in 2005 (this is not the recent one). The corresponding constraints on Ω_m and Ω_Λ are displayed in Figure 6.4, and summarized by:

$$(\Omega_m - \Omega_\Lambda, \Omega_m + \Omega_\Lambda) = (-0.49 \pm 0.12, 1.11 \pm 0.52) \ . \tag{6.12}$$



In flat universe: $\Omega_M = 0.28$ [± 0.085 statistical] [± 0.05 systematic]

Prob. of fit to $\Lambda = 0$ universe: 1%

Figure 6.2: The results published by the "Supernovae Cosmology Project" in 1998 (see Perlmutter et al., Astrophys.J. 517 (1999) 565-586). The various curves represent the effective magnitude–redshift relation, computed for various choices of Ω_m and Ω_Λ. This plot is equivalent to a luminosity distance – redshift relation (effective magnitude and luminosity distance can be related in a straightforward way: $m_B \propto (\log[d_L] + \text{cst})$). The solid black curves account for three examples of a universe with positive/null/negative curvature and no cosmological constant. The dashed blue curves correspond to three spatially flat universes with different values of Ω_Λ. For a given value of H_0, all the curves are asymptotically equal at short distance, probing only the Hubble law. The yellow points are short–distance SNIa's: we can check that they are approximately aligned. The red points, at redshifts between 0.2 and 0.9, show that distant supernovae are too faint to be compatible with a flat matter–dominated universe $(\Omega_m, \Omega_\Lambda)$ =(1,0).

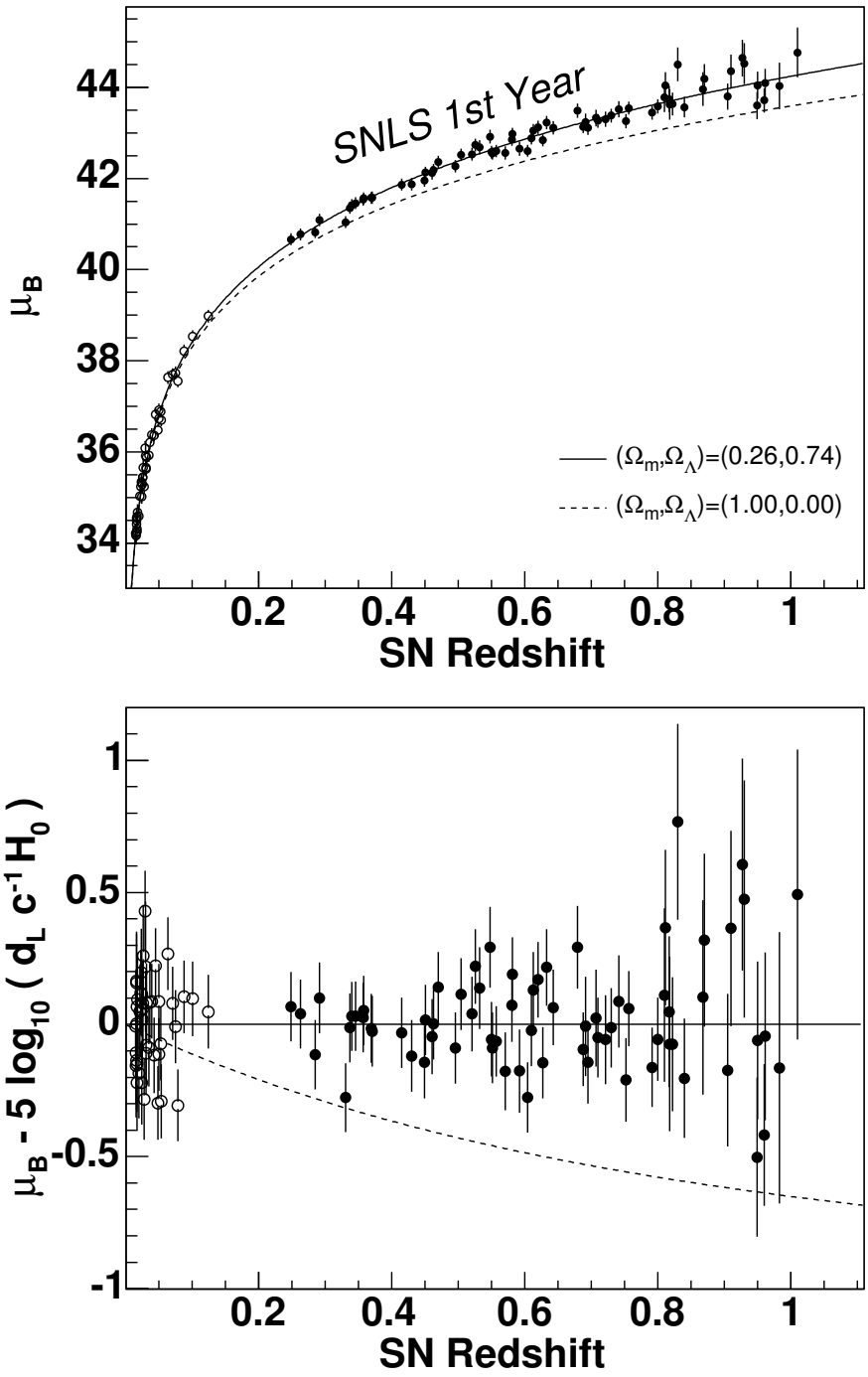


Figure 6.3: (Top panel) Same kind of luminosity distance – redshift diagram as in the previous figure, but for more recent data published by the SNLS collaboration in 2005. (Lower panel) Same data points and errors, divided by the theoretical prediction for the best fit ΛCDM model. *Plot taken from Astronomy and Astrophysics 447: 31-48, 2006 [e-Print: astro-ph/0510447] by Pierre Astier et al.*

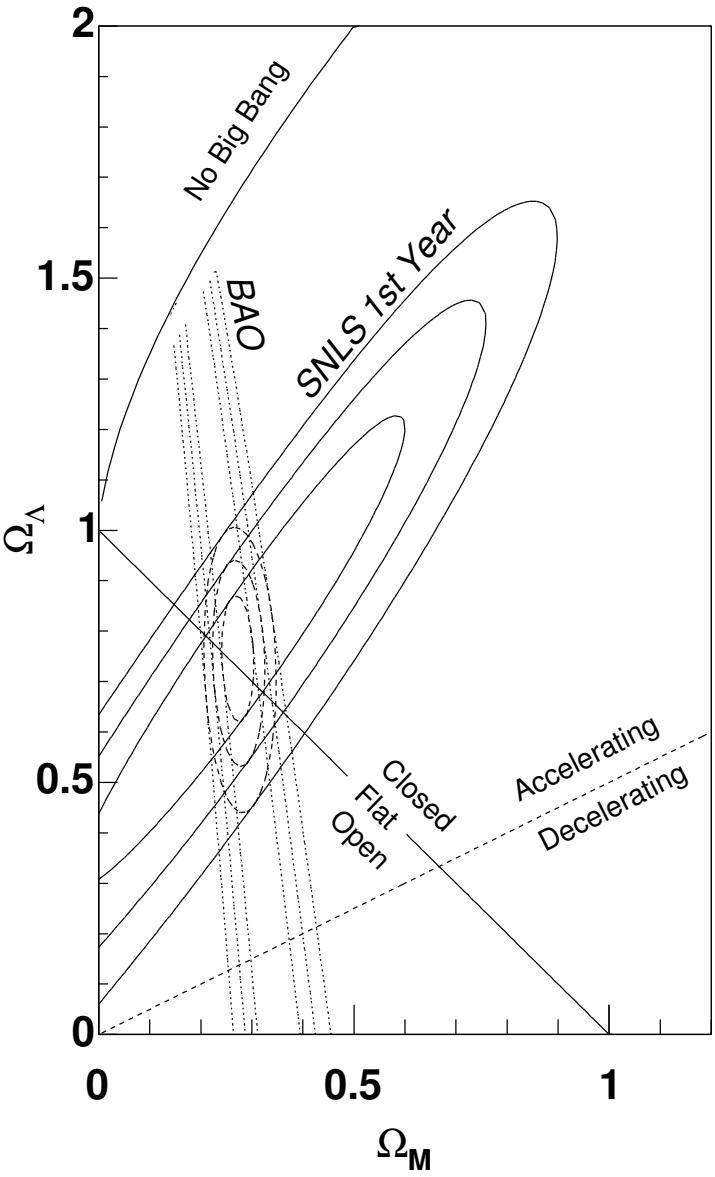


Figure 6.4: Contours at 68.3%, 95.5% and 99.7% confidence levels in the (Ω_m, Ω_Λ) plane from the SNLS supernovae data (solid contours), the SDSS baryon acoustic oscillations (see section **??**, dotted lines), and the joint confidence contours (dashed lines). These plots are all assuming a ΛCDM cosmology, as we are doing in this chapter. *Plot taken from Astronomy and Astrophysics 447: 31-48, 2006 [e-Print: astro-ph/0510447] by Pierre Astier et al.*

Hence, supernovae data strongly suggest the existence of a cosmological constant today ($\Omega_\Lambda > 0$). In fact, the small luminosity of high-redshift supernovae suggests that the universe is currently in accelerated expansion. The supernovae data does not say whether the parameter Ω_k is negligible, positive or negative.

6.6 CMB temperature anisotropies

The order of magnitude of CMB anisotropies was predicted many years before being measured. By extrapolating from the present inhomogeneous structure back to the time of decoupling, many cosmologists in the 80's expected $\delta T/\bar{T}$ to be at least of order 10^{-6} – otherwise, clusters of galaxies could not have formed today.

Many experiments were devoted to the detection of these anisotropies. The first successful one was COBE, a NASA satellite carrying an interferometer of exquisite sensitivity. In 1992, COBE mapped the anisotropies all over the sky, and found an average amplitude $\delta T/\bar{T} \sim 10^{-5}$ (see figure 6.5). This was in perfect agreement with the theoretical predictions – another big success for cosmology. The COBE experiment had an angular resolution of a few degrees:

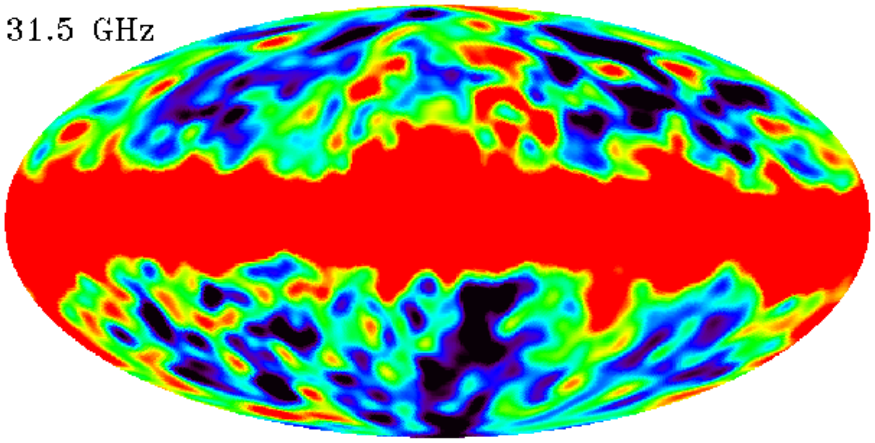


Figure 6.5: The first genuine "picture of the baby universe" at the time of decoupling, 380 000 years after the initial singularity, and 13.8 billion years before the present epoch. Each blue (resp. red) spot corresponds to apparently colder (resp. warmer) photons, and hence, given what we have learnt about the Sachs-Wolfe effect, to a warmer (resp. colder) region of the universe at that time. This map, obtained by the NASA satellite COBE in 1994 (see C. L. Bennett et al., Astrophys.J. 464 (1996) L1-L4), covers the entire sky: so, it pictures a huge sphere centered on us (on the picture, the sphere has been projected onto an ellipse, where the upper and lower points represent the direction of the poles of the Milky way). Away from the central red stripe, which corresponds to photons emitted from our own galaxy, the fluctuations are only of order 10^{-5} with respect to the average value $T_0 = 2.7255$ K. They are the "seeds" for the present structure of the universe: each blue spot corresponds to a small over-density of photons and baryons at the time of decoupling, that has been enhanced later, leading to galaxies and clusters of galaxies today.

so, anisotropies seen under one degree or less were smoothed by the detector. In a Fourier decomposition, it means that COBE could only measure the spectrum of wavelengths larger than the sound horizon at decoupling. So, it was not probing the acoustic oscillations, but only the flat plateau. Hence, after 1992, considerable efforts were devoted to the design of new experiments with better angular resolution, in order to probe smaller wavelengths, check the existence of the acoustic peaks, compare them with theoretical predictions and measure the related cosmological parameters.

For instance, some decisive progresses were made with Boomerang, a US–

Italian–Canadian balloon, carrying some detectors called bolometers. In 2001, Boomerang published the map of figure 6.6. It was focused on a small patch of the sky, but with much better resolution than COBE (a few arc–minutes).

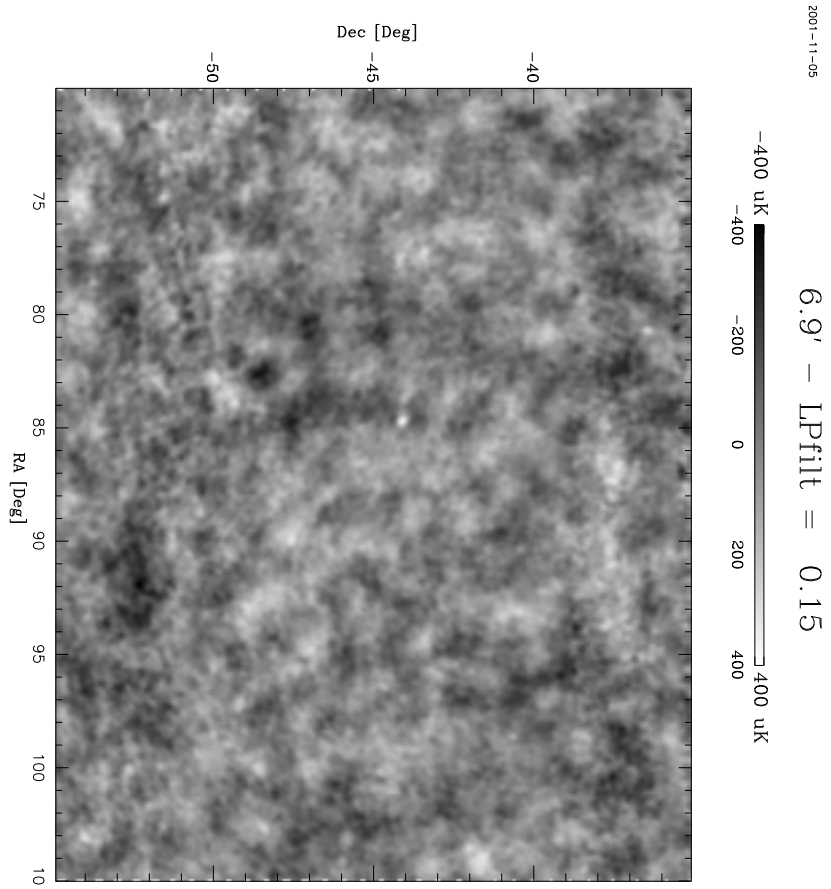


Figure 6.6: The map of CMB anisotropies obtained by the balloon experiment Boomerang in 2001 (see S. Masi et al., Prog.Part.Nucl.Phys. 48 (2002) 243-261). Unlike COBE, Boomerang only analyzed a small patch of the sky, but with a much better angular resolution of a few arc–minutes. The dark spots correspond to colder CMB photons.

The Fourier decomposition of the Boomerang map clearly showed the first three acoustic peaks (see figure 6.7). Let us recall that the angular size of the first peak probes the angular diameter distance at the redshift of photon decoupling, and depends heavily on the spatial curvature parameter Ω_k. The position of the first peak measured by Boomerang was perfectly consistent with $\Omega_k = 0$. Boomerang brought the first convincing arguments in favor of an exactly flat or at least nearly flat universe. The combination of Boomerang data with supernovae observations started to show that the preferred values of Ω_m and Ω_Λ were around 0.3 and 0.7 respectively.

After COBE, there have been two more CMB satellite producing full-sky CMB maps: the NASA satellitle WMAP (results in 2003-2013), and the ESA satellite Planck (results in 2013-2016). These satellite experiments are complementary to several ground-based and ballon-borne instruments looking at small patches of the sky with very high resolution.

The beautiful CMB anisotropy map of Planck and the inferred power spec-

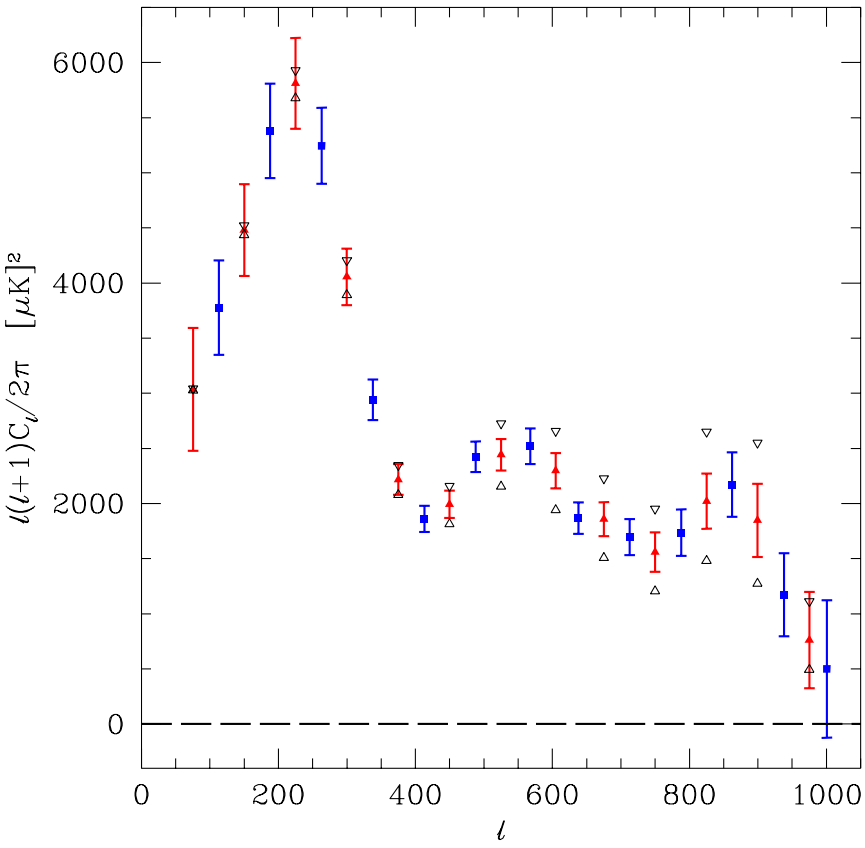


Figure 6.7: The temperature spectrum C_l measured by the Boomerang map revealed the structure of the first three acoustic oscillations (see C. B. Netterfield et al., Astrophys.J. 571 (2002) 604-614).

trum are shown in Figures 6.8 and 6.9. Amazingly, the temperature spectrum is still very well fitted by the ΛCDM model. The error bars of Planck data are very small, so even a tiny deviation from ΛCDM could in principle be observed, if it existed. Hence, the Planck data provides very strong constraints on all the parameters that may account for additional physical ingredients in the Universe. Planck alone measures the following values of ΛCDM parameters:

$$\omega_{\rm b} = 0.02226 \pm 0.00023 \tag{6.13}$$

$$\omega_{\rm cdm} = 0.1186 \pm 0.0020 \quad \text{or} \quad \omega_{\rm m} = 0.1415 \pm 0.0019 \tag{6.14}$$

$$\Omega_\Lambda = 0.692 \pm 0.012 \quad \text{or} \quad h = 0.678 \pm 0.009 \tag{6.15}$$

$$\ln(10^{10} A_s) = 3.062 \pm 0.029 \tag{6.16}$$

$$n_s = 0.9677 \pm 0.0060 \tag{6.17}$$

$$\tau_{\rm reio} = 0.066 \pm 0.016 \tag{6.18}$$

implying an age of $t_0 = 13.799 \pm 0.038$ Gyr. All these error bars are at the 68% Confidence Level (CL), i.e. at "1 σ".

We won't review here the limits on extensions of the ΛCDM model, excepted for the two cases that we have already discussed: is the universe really flat ($\Omega_k = 0$)? And is the radiation density (and neutrino density) really well

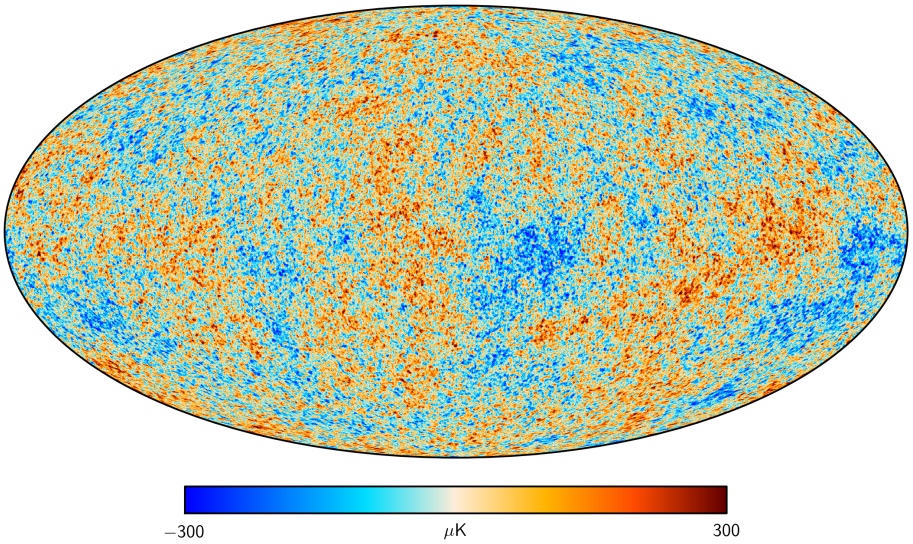


Figure 6.8: The full-sky map of CMB anisotropies obtained by the ESA satellite Planck in 2015 (see arXiv:1502.01582).

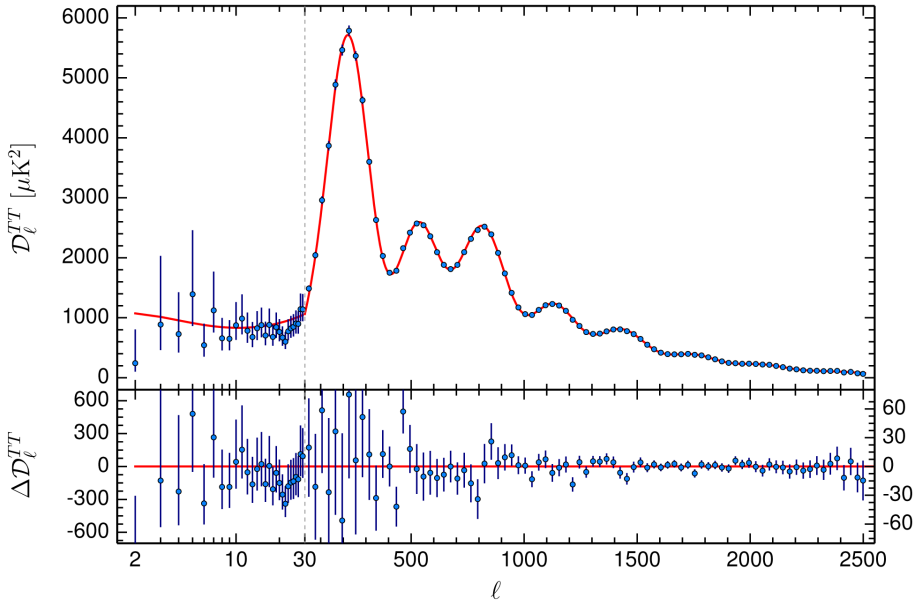


Figure 6.9: *Top.* The blue dot show the temperature power spectrum measured by Planck (the y-axis shows $D_l^{TT} \equiv l(l+1)C_l/[2\pi] \times T_0^2$ in units of $[\mu\mathrm{K}]^2$). For $l < 30$, the data is show for each mutipole l and the l-scale is logarithmic. For $l \geq 30$, the data is presented in bins (i.e., is averaged over a group of consecutive l values), in linear scale. The red curve is the best 6-parameter ΛCDM fit. *Bottom.* Plot of the residuals, i.e., of the data points after subtraction of the best fit. For details see arXiv:1502.01589.

accounted by eq. (6.2) with $N_{\rm eff} = 3.046$? Currently, the Planck data alone give $\Omega_k = -0.005 \pm 0.017$ (95%CL) and $N_{\rm eff} = 2.99 \pm 0.20$ (68%CL), again well compatible with the standard assumptions.

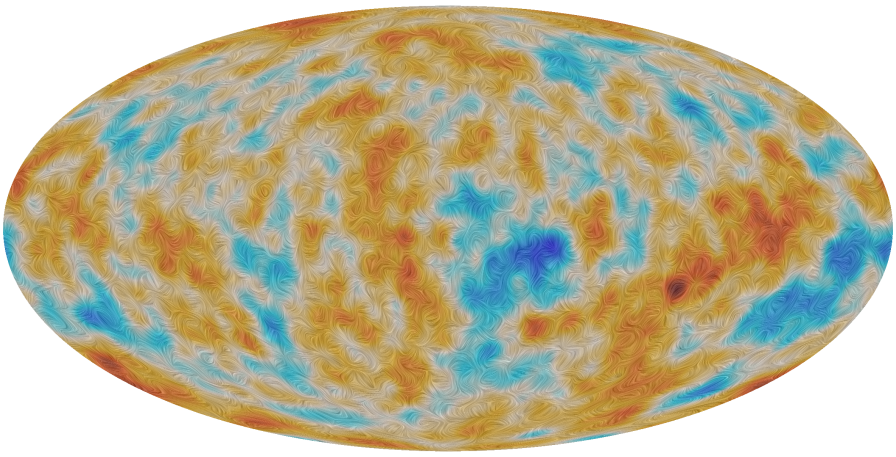


Figure 6.10: The full-sky map of CMB polarisation obtained by the ESA satellite Planck in 2015 (see arXiv:1502.01582). The colours indicate the degree of polarisation (blue = unpolarised, read = very polarised). The patterns show in each point the orientation of the polarisation plane. In order to get a clear plot, the polarisation patterns have been smoothed over a scale of 5 deg. This means that the actual data contains even more information on small scales than shown in this figure.

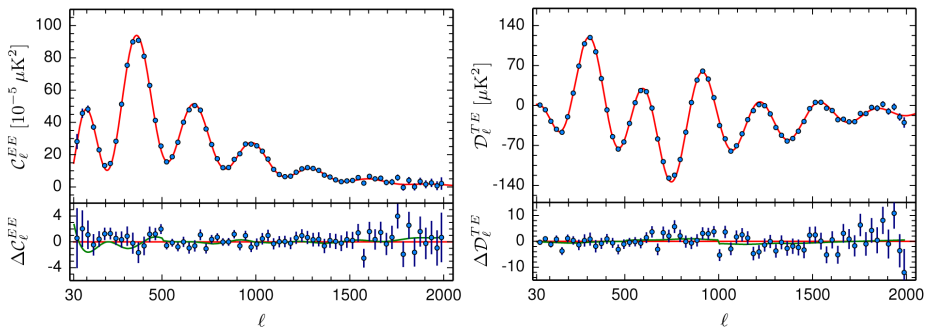


Figure 6.11: *Left.* Same kind of plots as in figure 6.9, but for the polarisation spectrum C_l^{EE} instead of the temperature spectrum $C_l \equiv C_l^{TT}$. *Right.* Same for the cross-correlation spectrum between temperature and polarisation, C_l^{TE}. In both plots, the red curves are obtained by measuring the best-fitting cosmological parameters of the ΛCDM model using the temperature spectrum only, and then, computing theoretical predictions for C_l^{EE} and C_l^{TE}. Hence the very good agreement between the red curves and the data points is a beautiful consistency check of the ΛCDM model, and of our understanding of CMB physics. For details see arXiv:1502.01589.

In this course we did not describe in details the physics of CMB polarisation, just briefly introduced in section 5.2.7. Let us just mention that we can do theoretical predictions for the CMB polarisation spectrum, just like for the temperature spectrum. The temperature spectrum, noted C_l in this course, is often called more precisely C_l^{TT} where TT means "temperature × temperature". One can also build a "temperature × polarisation" spectrum C_l^{TE} and a "polarisation × polarisation" spectrum C_l^{EE}. Figures 6.10 and 6.11 show the Planck polarisation map, and the comparison of predicted and observed polari-

sation spectra. The fits are very good and provide another consistency check of the ΛCDM model.

The polarisation signal mentioned here is what experts call "E-polarisation", i.e. the curl-free component of polarisation patterns. The other component, called "B-polarisation", is much smaller and more difficult to observe. It is however very interesting, because it could reveal the existence of primordial tensor perturbations in the universe (see section 5.2.8). Future CMB experiments are designed primarily for measuring these B modes.

6.7 Other observations not discussed here

Due to time limits, we do not address in this course other techniques which might become particularly important in the future: the study of the galaxy and cluster correlation function, allowing to measure the matter power spectrum $P(k, z)$ reviewed in section 5.3; the study of galaxy cluster abundances as a function of redshift; surveys of peculiar velocities; analyses of Lyman-α forests in the spectrum of quasars; galaxy weak lensing and cosmic shear; CMB weak lensing; the study of the 21cm absorption line in gas clouds; etc. The future of observational and theoretical cosmology is mainly related to these observables, but in order to discuss them, we would need another semester...

Chapter 7

Inflation

7.1 Motivations for inflation

7.1.1 Flatness problem

Today, Ω_k is measured to be at most of order 10^{-2}, possibly much smaller, while $\Omega_r \equiv \rho_r/\rho_{crit} \simeq \rho_r/(\rho_\Lambda + \rho_m)$ is of order 10^{-4}. Since ρ_k^{eff} scales like a^{-2}, while radiation scales like a^{-4}, the hierarchy between ρ_r and ρ_k^{eff} increases as we go back in time. If t_i is some initial time, t_0 is the time today, and we assume for simplicity that the ratio ρ_k^{eff}/ρ_r is at most equal to one today, we obtain

$$\frac{\rho_k^{eff}(t_i)}{\rho_r(t_i)} \leq \left(\frac{a(t_i)}{a(t_0)}\right)^2 = \left(\frac{\rho_r(t_0)}{\rho_r(t_i)}\right)^{1/2} . \tag{7.1}$$

Today, the radiation energy density $\rho_r(t_0)$ is of the order of $(10^{-4}\mathrm{eV})^4$. If the early universe reached the order of the Planck density $(10^{18}\mathrm{GeV})^4$ at the Planck time t_P, then at that time the ratio was

$$\frac{\rho_k^{eff}(t_P)}{\rho_r(t_P)} = \frac{(10^{-4}\mathrm{eV})^2}{(10^{18}\mathrm{GeV})^2} \sim 10^{-62} . \tag{7.2}$$

Even if the universe never reached such an energy, the hierarchy was already huge when ρ_r was of order, for instance, of $(1\,\mathrm{TeV})^4$.

If we try to build a mechanism for the birth of the classical universe (when it emerges from a quantum gravity phase), we will be confronted to the problem of predicting an initial order of magnitude for the various terms in the Friedmann equation: matter, spatial curvature and expansion rate. The Friedmann equation gives a relation between the three, but the question of the relative amplitude of the spatial curvature with respect to the total matter energy density, i.e. of the hierarchy between ρ_k^{eff} and ρ_r, is an open question. We could argue that the most natural assumption is to start from contributions sharing the same order of magnitude; this is actually what one would expect from random initial conditions at the end of a quantum gravity stage. The flatness problem can therefore be formulated as: why should we start from initial conditions in the very early universe such that ρ_k^{eff} should be fine-tuned to a fraction 10^{-62} of the total energy density in the universe?

The whole problem comes from the fact that the ratio ρ_k^{eff}/ρ_r (or more generally $\Omega_k \equiv \rho_k^{eff}/\rho_{crit}$) increases with time: i.e., a flat universe is an unstable solution of the Friedmann equation. Is this a fatality, or can we choose a framework in which the flat universe would become an attractor solution? The answer to this question is yes, even in the context of ordinary general relativity.

We noticed earlier that $|\Omega_k|$ is proportional to $(aH)^{-2}$, i.e. to $\dot{a}^{-2}$. So, as long as the expansion is decelerated, $\dot{a}$ decreases and $|\Omega_k|$ increases. If instead the expansion is accelerated, $\dot{a}$ increases and $|\Omega_k|$ decreases: the curvature is diluted and the universe becomes asymptotically flat.

Inflation is precisely defined as an initial stage during which the expansion is accelerated. One of the motivations for inflation is simply that if this stage is long enough, $|\Omega_k|$ will be driven extremely close to zero, in such way that the evolution between the end of inflation and today does not allow to reach again $|\Omega_k| \sim 1$.

We can search for the *minimal quantity of inflation* needed for solving the flatness problem. For addressing this issue, we should study a cosmological scenario where inflation takes place between times t_i and t_f such that $|\Omega_k| \sim 1$ at t_i, and $|\Omega_k| \sim 1$ again today at t_0. Let us compute the duration of inflation in this model. This will give us an absolute *lower bound* on the needed amount of inflation in the general case. Indeed, we could assume $|\Omega_k| \gg 1$ at t_i (since there could be a long stage of decelerated expansion before inflation); this would just require more inflation. Similarly, we could assume $|\Omega_k| \ll 1$ today at t_0, requiring again more inflation.

So, we assume that between t_i and t_f the scale factor grows from a_i to a_f, and for simplicity we will assume that the expansion is exactly De Sitter (i.e., exponential) with a constant Hubble rate H_i, so that the total density $\rho_{\rm inf}$ is constant between t_i and t_f. We assume that at the end of inflation all the energy $\rho_{\rm inf}$ is converted into a radiation energy ρ_r, which decreases like a^{-4} between t_f and t_0. Finally, we assume that ρ_k^{eff} (which scales like a^{-2}) is equal to $\rho_{\rm inf}$ at t_i and to ρ_r at t_0. With such assumptions, we can write

$$\frac{\rho_k^{\rm eff}(a_0)}{\rho_k^{\rm eff}(a_i)} = \left(\frac{a_i}{a_0}\right)^2 = \frac{\rho_r(a_0)}{\rho_{\rm inf}(a_i)} = \frac{\rho_r(a_0)}{\rho_{\rm inf}(a_f)} = \frac{\rho_r(a_0)}{\rho_r(a_f)} = \left(\frac{a_f}{a_0}\right)^4 \tag{7.3}$$

and we finally obtain the relation

$$\frac{a_f}{a_i} = \frac{a_0}{a_f} \ . \tag{7.4}$$

So, the condition for the minimal duration of inflation reads

$$\frac{a_f}{a_i} \geq \frac{a_0}{a_f} \ , \tag{7.5}$$

which can be summarized in one sentence: there should be as much expansion during inflation as after inflation. A convenient measure of expansion is the so-called *e-fold number* defined as

$$N \equiv \ln\, a \ . \tag{7.6}$$

The scale factor is physically meaningful up to a normalization constant, so the e-fold number is defined modulo a choice of origin. The amount of expansion between two times t_1 and t_2 is specified by the number of e-folds $\Delta N = N_2 - N_1 = \ln(a_2/a_1)$. So, the condition on the absolute minimal duration of inflation reads

$$(N_f - N_i) \geq (N_0 - N_f) \tag{7.7}$$

i.e., the number of inflationary e-folds should be greater or equal to the number of post-inflationary e-folds $\Delta N \equiv N_0 - N_f$. There is no upper bound on $(N_f - N_i)$: for solving the flatness problem, inflation could be arbitrarily long.

It is easy to compute ΔN as a function of the energy density at the end of inflation, $\rho_r(a_f)$. We know that today $\rho_r(a_0)$ is of the order of $(10^{-4}{\rm eV})^4$, and

we will see in section 7.3.2 that the inflationary energy scale is at most of the order of $(10^{16}\mathrm{G}eV)^4$, otherwise current observations of CMB anisotropies would have detected primordial gravitational waves. This gives

$$\Delta N = \ln \frac{a_0}{a_f} = \ln \left(\frac{\rho_r(a_f)}{\rho_r(a_0)} \right)^{1/4} \leq \ln 10^{29} \sim 67 \ . \tag{7.8}$$

We conclude that if inflation takes place around the $10^{16}\mathrm{G}eV$ scale, it should last for a minimum of 67 e-folds. If it takes place at lower energy, the condition is weaker. The lowest scale for inflation considered in the literature (in order not to disturb too much the predictions of the standard inflationary scenario) is of the order of 1 TeV. In this extreme case, the number of post-inflationary e-folds would be reduced to

$$\Delta N \sim \ln 10^{16} \sim 37 \tag{7.9}$$

and the flatness problem can be solved with only 37 e-folds of inflation.

7.1.2 Horizon problem

We recall that the causal horizon $d_H(t_1, t_2)$ is defined as the physical distance at time t_2 covered by a particle emitted at time t_1 and travelling at the speed of light. If the origin of spherical comobile coordinates is chosen to coincide with the point of emission, the physical distance at time t_2 can be computed by integrating over small distance elements dl between the origin and the position r_2 of one particle,

$$d_H(t_1, t_2) = \int_0^{r_2} dl = \int_0^{r_2} a(t_2) \frac{dr}{\sqrt{1 - k\, r^2}} \ . \tag{7.10}$$

In addition, the geodesic equation for ultra-relativistic particles gives $ds = 0$, i.e., $dt = a(t) dr / \sqrt{1 - k\, r^2}$, which can be integrated along the trajectory of the particles,

$$\int_{t_1}^{t_2} \frac{dt}{a(t)} = \int_0^{r_2} \frac{dr}{\sqrt{1 - k\, r^2}} \ . \tag{7.11}$$

We can now replace in the expression of d_H and get

$$d_H(t_1, t_2) = a(t_2) \int_{t_1}^{t_2} \frac{dt}{a(t)} \ . \tag{7.12}$$

Usually, the result is presented in this form. However, for the following discussion, it is particularly useful to eliminate the time from the integral by noticing that $dt = da/(aH)$,

$$d_H(a_1, a_2) = a_2 \int_{a_1}^{a_2} \frac{da}{a^2 H(a)} \ , \tag{7.13}$$

where the Hubble parameter is seen now as a function of a. Let us assume that t_1 and t_2 are two times during Radiation Domination (RD). We know from the Friedmann equation that during RD on has $H \propto a^{-2}$, so we can parametrize the Hubble rate as $H(a) = H_2 \ (a_2/a)^2$. We obtain

$$d_H(a_1, a_2) = a_2 \int_{a_1}^{a_2} \frac{da}{a_2^2 H_2} = \frac{1}{H_2} \frac{(a_2 - a_1)}{a_2} \ . \tag{7.14}$$

If the time t_2 is much after t_1 so that $a_2 \gg a_1$, the expression for the horizon does not depend on a_1,

$$d_H(a_1, a_2) \simeq \frac{1}{H_2} \ . \tag{7.15}$$

So, during RD, the horizon equals the Hubble radius at time t_2 (in agreement with the result of eq. (**??**) with $n = 1/2$). During matter domination, the horizon is still close to the Hubble radius, modulo a factor of order one.

The horizon represents the causal distance in the universe. Suppose that a physical mechanism is turned on at time t_1. Since no information can travel faster than light, the physical mechanism cannot affect distances larger than $d_H(t_1, t_2)$ at time t_2. So, the horizon provides the *coherence scale* of a given mechanism. For instance, if a phase transition creates bubbles or patches containing a given vacuum phase, the scale of homogeneity (i.e., the maximum size of the bubble, or the scale on which a patch is nearly homogeneous) is given by $d_H(t_1, t_2)$ where t_1 is the time at the beginning of the transition.

Before photon decoupling, the Planck temperature of photons at a given point depends on their local density. A priori, we can expect that the universe will emerge from a quantum gravity stage with random values of the local density. The coherence length, or characteristic scale on which the density is nearly homogeneous, is given by $d_H(t_1, t_2)$. We have seen that if t_1 and t_2 are two times during radiation domination, this quantity cannot exceed $R_H(t_2)$, even in the most favorable limit in which t_1 is chosen to be infinitely close to the initial singularity. We conclude that at time t_2, the photon temperature should not be homogeneous on scales larger than $R_H(t_2)$.

CMB experiments map the photon temperature on our last-scattering-surface at the time of photon decoupling. So, we expect CMB maps to be nearly homogeneous on a characteristic scale $R_H(t_{dec})$. This scale is very easy to compute: knowing that $H(t_0)$ is of the order of $(h/3000)\,\mathrm{M}pc^{-1}$ with $h \simeq 0.7$, we can extrapolate $H(t)$ back to the time of equality, and find that the distance $R_H(t_{dec})$ subtends an angle of order of a few degrees in the sky - instead of encompassing the diameter of the last scattering surface. So, it seems that the last scattering surface is composed of several thousands causally disconnected patches. However, the CMB temperature anisotropies are only of the order of 10^{-5}: in other words, the full last scattering surface is extremely homogeneous. This appears as completely paradoxical in the framework of the Hot Big Bang scenario.

What is the origin of this problem? When we computed the horizon, we integrated $(a^2 H)^{-1}$ over da and found that the integral was converging with respect to the boundary a_1: so, even by choosing the initial time to be infinitely early, the horizon is bounded by a function of a_2. If the integral was instead divergent, we could obtain an infinitely large horizon at time t_2 simply by choosing a_1 to be small enough. The convergence of the integral

$$\int_{a_1}^{a_2} \frac{da}{a^2 H(a)} = \int_{a_1}^{a_2} \frac{da}{a\dot{a}} \tag{7.16}$$

with respect to $a_1 \to 0$ depends precisely on the fact that the expansion is accelerated or decelerated. For linear expansion, the integrand is $1/a$, the limiting case between convergence and divergence. If it is decelerated, $\dot{a}$ decreases and the integral converges. If it is accelerated, $\dot{a}$ increases and the integral diverges in the limit $a_1 \to 0$.

So, if the radiation dominated phase is preceded by an infinite stage of accelerated expansion, one can reach an arbitrarily large value for the horizon at the time of decoupling. In fact, in order to explain the homogeneity of the last scattering surface, we only need to boost the horizon by a factor of $\sim 10^3$ with respect to the Hubble radius at that time. This can be fulfilled with a rather small amount of accelerated expansion.

Let us take an exemple and assume that between a_i and a_f, the acceleration is exponential, $a = e^{\alpha t}$. In this case, the Hubble parameter $\dot{a}/a$ is constant over

this period: let's call it $H_{\rm inf}$. The horizon computed between a_i and a_f reads:

$$d_H(a_i, a_f) = a_f \int_{a_i}^{a_f} \frac{da}{a^2 H_{\rm inf}} = \frac{1}{H_{\rm inf}} \left(\frac{a_f}{a_i} - 1 \right) \simeq \frac{1}{H_{\rm inf}} \ \frac{a_f}{a_i} \ . \tag{7.17}$$

So, at the end of inflation, the horizon is larger than the Hubble radius $R_H = 1/H_{\rm inf}$ by a factor a_f/a_i, i.e, by the exponential of the number of inflationary e-folds. After, the horizon will keep growing in the usual way,

$$\begin{aligned} d_H(a_i, a_2) &= a_2 \int_{a_i}^{a_2} \frac{da}{a^2 H(a)} \\ &= \frac{a_2}{H_{\rm inf}} \left(\frac{1}{a_i} - \frac{1}{a_f} \right) + a_2 \int_{a_f}^{a_2} \frac{da}{a^2 H(a)} \\ &\simeq \frac{1}{H_{\rm inf}} \ \frac{a_2}{a_i} + \frac{1}{H_2} \ , \end{aligned} \tag{7.18}$$

and remains much larger than the Hubble radius $\frac{1}{H_2}$.

The condition for solving the horizon problem can be shown to be exactly the same as for solving the flatness problem: the number of inflationary e-fold should be at least equal to that of post-inflationary e-folds. If it is larger, then the size of the observable universe is even smaller with respect to the causal horizon.

7.1.3 Origin of perturbations

Since our universe is inhomogeneous, one should find a physical mechanism explaining the origin of cosmological perturbations. Inhomogeneities can be expanded in comoving Fourier space. Their physical wavelength

$$\lambda(t) = \frac{2\pi a(t)}{k} \tag{7.19}$$

is stretched with the expansion of the universe. During radiation domination, $a(t) \propto t^{1/2}$ and $R_H(t) \propto t$. So, the Hubble radius grows with time faster than the perturbation wavelengths. We conclude that observable perturbations were originally super-Hubble fluctuations (i.e., $\lambda > R_H \ \Leftrightarrow \ k < 2\pi a H$). Actually, the discussion of the horizon problem already showed that at decoupling the largest observable fluctuations are super-Hubble fluctuations. Even if we take a smaller scale, e.g. the typical size of a galaxy cluster $\lambda(t_0) \sim 1$ Mpc, we find that the corresponding fluctuations were clearly super-Hubble fluctuations for instance at the time of Nucleosynthesis. We have seen that in the Hot Big Bang scenario (without inflation) the Hubble radius $R_H(t_2)$ gives an upper bound on the causal horizon $d_H(t_1, t_2)$ for whatever value of t_1. So, super-Hubble fluctuations are expected to be out of causal contact. The problem is that it is impossible to find a mechanism for generating coherent fluctuations on acausal scales. There are two possible solutions to this issue:

- we can remain in the framework of the Hot Big Bang scenario and assume that perturbations are produced causally when a given wavelength enters into the horizon. In this case, there should be not coherent fluctuations on super-Hubble scales, i.e. the power spectrum of any kind of perturbation should fall like white noise in the limit $k \ll aH$. This possibility is now ruled out for at least two reasons. First, the observation of CMB anisotropies on angular scales greater than one degree (i.e., super-Hubble scales at that time) is consistent with coherent fluctuations rather than white noise. Second, the observations of acoustic peaks in the power spectrum of CMB anisotropies is a clear proof that cosmological perturbations

are generated much before Hubble crossing, in such way that all modes with a given wavelength entering inside the Hubble radius before photon decoupling experience coherent acoustic oscillations (i.e. oscillate with the same phase).

- we can modify the cosmological scenario in such way that all cosmological perturbations observable today were inside the causal horizon when they were generated at some early time (we will study a concrete generation mechanism in section 7.3).

So, our goal is to find a paradigm such that the largest wavelength observable today, which is $\lambda_{max}(t_0) \sim R_H(t_0)$ (see section **??**), was already inside the causal horizon at some early time t_i. If before t_i the universe was in decelerated expansion, then the causal horizon at that time was of order $R_H(t_i)$. How can we have $\lambda_{max} \leq R_H$ at t_i and $\lambda_{max} \sim R_H$ today? If between t_i and t_0 the universe is dominated by radiation or matter, it is impossible since the Hubble radius grows faster than the physical wavelengths. However, in general,

$$\frac{\lambda(t)}{R_H(t)} = \frac{2\pi a(t)}{k}\frac{\dot{a}(t)}{a(t)} = \frac{2\pi\dot{a}(t)}{k} \ , \tag{7.20}$$

so that during accelerated expansion the physical wavelengths grow faster than the Hubble radius. So, if between some time t_i and t_f the universe experiences some inflationary stage, it is possible to have $\lambda_{max} < R_H$ at t_i: the scale λ_{max} can then exit the Hubble radius during inflation and re-enter approximately today (see Figure 7.1).

It is easy to show that once again, the minimal number of inflationary e-folds requested for solving this problem should be at least equal to that of post-inflationary e-folds.

One could argue that the argument on the origin of fluctuations is equivalent to that of the horizon problem, reformulated in a different way. Anyway, for understanding inflation it is good to be aware of the two arguments, even if they are not really independent from each other.

7.1.4 Monopoles

We will not enter here into the details of the monopole problem. Just in a few words, some phase transitions in the early universe are expected to create "dangerous relics" like magnetic monopoles, with a very large density which would dominate the total density of the universe. These relics are typically non-relativistic, with an energy density decaying like a^{-3}: so, they are not diluted, and the domination of radiation and ordinary matter can never take place.

Inflation can solve the problem provided that it takes place after the creation of dangerous relics. During inflation, monopoles and other relics will decay like a^{-3} (a^{-4} in the case of relativistic relics) while the leading vacuum energy is nearly constant: so, the energy density of the relics is considerably diluted, typically by a factor $(a_f/a_i)^3$, and today they are irrelevant. The condition on the needed amount of inflation is much weaker than the condition obtained for solving the flatness problem, since dangerous relics decay faster than the effective curvature density ($\rho_k^{eff} \propto a^{-2}$).

7.2 Slow-roll scalar field inflation

So, the first three problems of section 7.1 can be solved under the assumption of a long enough stage of accelerated expansion in the early universe. How can

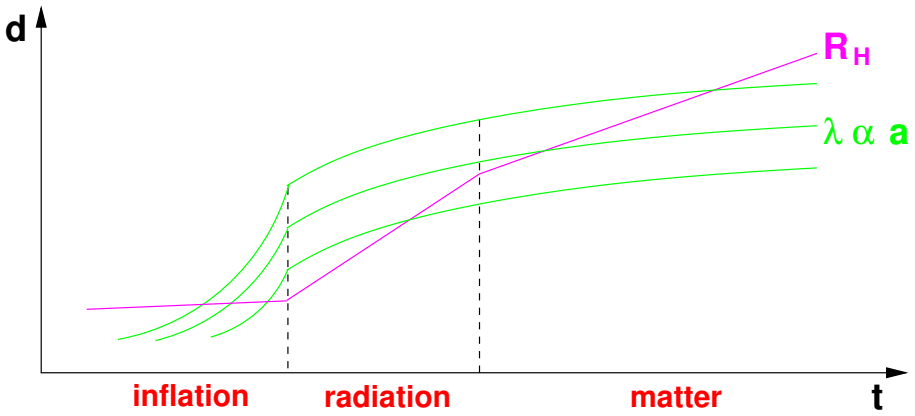


Figure 7.1: Comparison of the Hubble radius with the physical wavelength of a few cosmological perturbations. During the initial stage of accelerated expansion (called inflation), the Hubble radius grows more slowly than each wavelength. So, cosmological perturbations originate from inside R_H. Then, the wavelength of each mode grows larger than the Hubble radius during inflation and re–enters during radiation or matter domination.

this be implemented in practice?

First, by combining the Friedman equation (2.47) in a flat universe with the conservation equation (2.48), it is easy to find that

$$\ddot{a} > 0 \qquad \Rightarrow \qquad \rho + 3p < 0. \tag{7.21}$$

What type of matter corresponds to such an unusual relation between density and pressure? A positive cosmological constant can do the job:

$$p_\Lambda = -\rho_\Lambda \qquad \Rightarrow \qquad \rho_\Lambda + 3p_\Lambda = -2\rho_\Lambda < 0. \tag{7.22}$$

But since a cosmological constant is... constant, it cannot be responsible for an initial stage of inflation: otherwise this stage would go on forever, and there would be no transition to radiation domination.

Let us consider instead the case of a scalar field (i.e., a field of spin zero, represented by a simple function of time and space, and invariant under Lorentz transformations). The general action for a scalar field in curved space-time

$$S = -\int d^4x\sqrt{|g|}\,(\mathcal{L}_g + \mathcal{L}_\varphi) \tag{7.23}$$

involves the Lagrangian of gravitation

$$\mathcal{L}_g = \frac{R}{16\pi\mathcal{G}} \tag{7.24}$$

and that of the scalar field

$$\mathcal{L}_\varphi = \frac{1}{2}\partial_\mu\varphi\partial^\mu\varphi - V(\varphi) = \frac{1}{2}g^{\mu\nu}\partial_\mu\varphi\partial_\nu\varphi - V(\varphi) \tag{7.25}$$

where $V(\phi)$ is the scalar potential. The variation of the action with respect to $g_{\mu\nu}$ enables to define the energy-momentum tensor

$$T_{\mu\nu} = \partial_\mu\varphi\partial_\nu\varphi - \mathcal{L}_\varphi g_{\mu\nu} \tag{7.26}$$

and the Einstein tensor $G_{\mu\nu}$, which are related through the Einstein equations

$$G_{\mu\nu} = 8\pi\mathcal{G}\, T_{\mu\nu} \ . \tag{7.27}$$

Instead, the variation of the action with respect to φ gives Klein-Gordon equation

$$\frac{1}{\sqrt{|g|}}\partial_\mu \left[\sqrt{|g|}\partial^\mu\varphi\right] + \frac{\partial V}{\partial\varphi} = 0 \ . \tag{7.28}$$

The same equation could have been obtained using a particular combination of the components of $T_{\mu\nu}$ and their derivatives, which vanish by virtue of the Bianchi identities (in other word, the Klein-Gordon equation is contained in the Einstein equations).

Let us now assume that the homogeneous Friedmann universe with flat metric

$$g_{\mu\nu} = \mathrm{d}iag\left(1, -a(t)^2, -a(t)^2, -a(t)^2\right) \tag{7.29}$$

is filled by a homogeneous classical scalar field $\bar{\varphi}(t)$. One can show that the corresponding energy-momentum tensor is diagonal, $T^\nu_\mu = \mathrm{d}iag(\rho, -p, -p, -p)$, with

$$\rho = \frac{1}{2}\dot{\bar{\varphi}}^2 + V(\varphi) \ , \tag{7.30}$$

$$p = \frac{1}{2}\dot{\bar{\varphi}}^2 - V(\varphi) \ . \tag{7.31}$$

The Friedmann equation reads

$$G^0_0 = 3H^2 = 8\pi\mathcal{G}\,\rho \tag{7.32}$$

and the Klein-Gordon equation

$$\ddot{\bar{\varphi}} + 3H\dot{\bar{\varphi}} + \frac{\partial V}{\partial\varphi}(\bar{\varphi}) = 0 \ . \tag{7.33}$$

These two independent equations specify completely the evolution of the system. However it is worth mentioning that the full Einstein equations provide another relation

$$G^i_i = \left(2\frac{\ddot{a}}{a} + \left(\frac{\dot{a}}{a}\right)^2\right) = -8\pi\mathcal{G}\, p \ . \tag{7.34}$$

The combination $\dot{G}^0_0 + 3H(\dot{G}^0_0 - G^i_i)$ vanishes (it is one of the Bianchi identities), and gives a conservation equation $\dot{\rho} + 3H(\rho + p) = 0$, which is nothing but the Klein-Gordon equation. Finally, the combination $G^i_i - G^0_0$ provides a very useful relation

$$\dot{H} = -4\pi\mathcal{G}\,\dot{\bar{\varphi}}^2 \tag{7.35}$$

which is consistent with the fact that the Hubble parameter can only decrease.

The condition $p < -\rho/3$ reads $\dot{\bar{\varphi}}^2 < V$: when the potential energy dominates over the kinetic energy, the universe expansion is accelerated. In the limit of zero kinetic energy, the energy-momentum tensor would be that of a cosmological constant, and the expansion would be exponential (this is called "De Sitter expansion") and everlasting. For a long, finite stage of acceleration we must require that the *first slow-roll condition*

$$\frac{1}{2}\dot{\bar{\varphi}}^2 \ll V(\bar{\varphi}) \tag{7.36}$$

holds over an extended period. Since the evolution of the scalar field is given by a second-order equation, the above condition could apply instantaneously but

not for an extended stage, in particular in the case of oscillatory solutions. If we want the first slow-roll condition to hold over an extended period, we must impose that the time-derivative of this condition also holds (in absolute value). This gives the *second slow-roll condition*

$$|\ddot{\bar{\varphi}}| \ll \left|\frac{\partial V}{\partial \varphi}(\bar{\varphi})\right| \tag{7.37}$$

which can be rewritten, by virtue of the Klein-Gordon equation, as

$$|\ddot{\bar{\varphi}}| \ll 3H\,|\dot{\bar{\varphi}}| \ . \tag{7.38}$$

When these two conditions hold, the Friedmann and Klein-Gordon equations become

$$3H^2 \simeq 8\pi\mathcal{G}\,V(\bar{\varphi}) \ , \tag{7.39}$$

$$\dot{\bar{\varphi}} \simeq -\frac{1}{3H}\frac{\partial V}{\partial \varphi}(\bar{\varphi}) \ . \tag{7.40}$$

The two slow-roll conditions can be rewritten as conditions either on the slowness of the variation of $H(t)$, or on the flatness of the potential $V(\varphi)$.

So, a particular way to obtain a stage of accelerated expansion in the early universe is to introduce a scalar field, with a flat enough potential. Scalar field inflation has been proposed in 1979 by Guth. Starting from 1979 and during the 80's, most important aspects of inflation were studied in details by Starobinsky, Guth, Hawking, Linde, Mukhanov and other people. Finally, during the 90's, many ideas and models were proposed in order to make contact between inflation and particle physics. The purpose of scalar field inflation is not only to provide a stage of accelerated expansion in the early universe, but also, a mechanism for the generation of matter and radiation particles, and another mechanism for the generation of primordial cosmological perturbations. Let us summarize how it works in a very sketchy way.

Slow-roll. First, let us assume that just after the initial singularity, the energy density is dominated by a scalar field, with a potential flat enough for slow–roll. In any small region where the field is approximately homogeneous and slowly–rolling, accelerated expansion takes place: this small region becomes exponentially large, encompassing the totality of the present observable universe. Inside this region, the causal horizon becomes much larger than the Hubble radius, and any initial spatial curvature is driven almost to zero – so, some of the main problems of the standard cosmological model are solved. After some time, when the field approaches the minimum its potential, one of the two slow-roll conditions breaks down, and inflation ends: the expansion becomes decelerated again.

Reheating. At the end of inflation, the kinetic energy of the field is bigger than the potential energy; in general, the field is quickly oscillating around the minimum of the potential. According to the laws of quantum field theory, the oscillating scalar field will decay into fermions and bosons. This could explain the origin of all the particles filling our universe. The particles probably reach quickly a thermal equilibrium: this is why this stage is called "reheating".

Generation of primordial perturbations. Finally, the theory of scalar field inflation also explains the origin of cosmological perturbations – the ones leading to CMB anisotropies and large scale structure formation. Using again quantum field theory in curved space–time, it is possible to compute the amplitude of

the small quantum fluctuations of the scalar field φ (as well as the quantum fluctuations of the metric $h_{\mu\nu}$). The physical wavelengths of these fluctuations grow quickly, like in figure 7.1. So, they are initially inside the Hubble radius, where we can apply the laws of quantum mechanics in flat space-time (as long as $k \ll aH$, the modes do not see the curvature of space-time). In the opposite limit, when a wavelength is stretched to scales larger than the Hubble length, it is possible to show that the modes experience a kind of quantum–to–classical transition, in the sense that they become indistinguishable from classical stochastic fluctuations: hence, the primordial fluctuations have a random distribution (as expected), but we don't need to employ the formalism of quantum mechanics (wave functions, etc.) in order to describe their statistics. In addition, the initial quantum fluctuations $\delta\varphi$ are assumed to be vacuum fluctuations (corresponding to the fundamental state of the field $\delta\varphi$). As a consequence, the probability distribution of each mode $\delta\varphi(\mathbf{k})$ after the transition can be showed to be a Gaussian, depending only on k. Hence, at a given time, all information about the statistics of the field is contained in the power spectrum $\langle|\varphi(\mathbf{k})|^2\rangle$, which is a function of k.

7.3 Inflationary perturbations

7.3.1 Scalar perturbations

The perturbations of the scalar field $\delta\varphi$ are coupled with those of the scalar metric fluctuations: for instance, ϕ and ψ in the longitudinal gauge. At first order in perturbation theory, it is easy to show that $\phi = \psi$, so the problem of scalar perturbations during inflation reduces to the evolution of two quantities only, $\delta\varphi$ and ϕ. In addition, the linearized Einstein equations provides a relation between $\delta\varphi$ and ϕ: they are not independent, and their evolution is dictated by a single equation of motion.

As explained above, quantum field theory allows to exactly follow the evolution and the quantum–to–classical transition of the fields $\delta\varphi$ and ϕ during inflation. So, it is possible to compute exactly the power spectrum of $\delta\varphi$ and ϕ for observable modes, i.e. on wavelengths much larger than the Hubble radius at the end of inflation. But how can we related these spectra to the initial conditions at the beginning of radiation domination, after the end of inflation?

We could fear that such a relation could be very difficult to compute, and could depend on the mechanism through which the scalar field decays into radiation and matter... Fortunately, this is not the case: the relation between the spectrum of fluctuations at the end of inflation and at the beginning of radiation domination is trivial. The reason is that when the wavelength of a mode $\phi(\mathbf{k})$ becomes much larger than the Hubble radius, the perturbation freezes out. Hence it is not affected by the decay of the scalar field during reheating. But when the radiation and matter particles are formed during reheating, they are sensitive to the gravitational potential, and more particles accumulate in the potential wells. So, the gravitational potential behaves like a mediator between the scalar field perturbations during inflation and the radiation/matter perturbations in the radiation/matter–dominated universe. If we can compute the power spectrum $\langle|\phi(\mathbf{k})|^2\rangle$ at the end of inflation, we are done, because this power spectrum remains the same at the beginning of radiation domination on super-Hubble scale; then the initial condition described in eq. (**??**) apply, and the evolution of all radiation/matter perturbations is entirely determined.

It is far beyond the level of these notes to compute the evolution of primordial perturbations during inflation. However, we should stress that it can be studied

in a very precise way using quantum field theory. The result for the primordial spectrum of scalar metric perturbations during inflation/radiation domination and on super-Hubble scales $k \ll aH$ reads:

$$\langle |\phi(\mathbf{k})|^2 \rangle = 2 \left(\frac{8\pi G}{3k} \right)^3 \frac{V^3}{V'^2} \ , \tag{7.41}$$

where V and V', which are both functions of φ, should be evaluated *with the value of the field corresponding to the time of Hubble crossing during inflation for each mode* $\mathbf{k}$, i.e., with the value $\bar{\varphi}(t)$ at the time t when $k = aH$. Hence, the primordial spectrum depends on k not only through the above k^{-3} factor, but also through the V^3/V'^2 factor. However, since the field is in slow-roll, V^3/V'^2 does not vary a lot between the time at which the largest and the smallest observable wavelengths cross the Hubble radius during inflation. Hence, the dependence of V^3/V'^2 on k is small, and the above spectrum is close to a scale-invariant spectrum, $\langle |\phi(\mathbf{k})|^2 \rangle \propto k^{-3}$. However, the deviation from exact scale-invariance (i.e. the value of the spectral index n minus one) depends crucially on the evolution of this ratio V^3/V'^2 with time and scale. By taking the derivative of the above equation, one could show that $n-1$ is indeed related to the ratios V''/V and V'/V evaluated when observable scales cross the horizon.

In the previous chapters, we saw that CMB and large scale structure observations allow to reconstruct the cosmological evolution during radiation/matter/Λ domination, as well as the primordial spectrum $\langle |\phi(\mathbf{k})|^2 \rangle$. In particular, the amplitude A and spectral index n (defined in eq. (**??**)) of the primordial spectrum $\langle |\phi(\mathbf{k})|^2 \rangle$ can be measured. According to the above results, these observations provide a measurement of V^3/V'^2 and of its evolution with φ within a small interval. Hence the potential $V(\phi)$ can be reconstructed to some extent from observations. It is quite remarkable that current observations provide a way to constrain the physical mechanism governing the evolution of the universe at extremely high energy (considerably higher than during Nucleosynthesis) and extremely early times (a tiny fraction of second after the initial singularity).

7.3.2 Tensor perturbations (gravitational waves)

The same mechanism which produces stochastic fluctuations of $\delta\varphi$ and ϕ (more precisely, of the scalar metric perturbations) on cosmological scales produces also stochastic fluctuations of the tensor metric perturbations, i.e., of the tensor h_{ij} defined in eq. (**??**). These perturbations are called gravitational waves, since inside the Hubble radius they have oscillatory solutions: there are deformation of our space-time manifold, propagating like waves with the velocity of light. Unlike scalar perturbations, they do not couple with matter fields or scalar fields within linear perturbation theory. Like electromagnetic waves, gravitational waves can propagate in the vacuum without being damped.

It is possible to compute the primordial spectrum of gravitational waves (i.e., the primordial spectrum of the components of h_{ij}) using the same formalism as for the scalar metric fluctuation ϕ. The result reads

$$\langle |h(\mathbf{k})|^2 \rangle = \frac{2}{3} (8\pi G)^2 \, k^{-3} V \ , \tag{7.42}$$

where V is evaluated like for scalar perturbations, i.e. with the value $\bar{\varphi}(t)$ at the time t when $k = aH$. Here h stands for the components of h_{ij} (we don't give the exact definition of h here for concision).

Hence, inflation is also expected to fill the universe with a random background of gravitational waves which could be detected today, at least in principle. Unfortunately, this background of gravitational waves is so low that its

detection is unlikely with the current generation (and even the next generation) of gravitational wave detectors (VIRGO, LIGO, etc.) However, there is a chance to detect it in the CMB: gravitational waves of primordial origin are expected to contribute to the CMB spectrum on the largest angular scales, as shown in figure 7.2. The shape of the tensor contribution to the CMB spectrum can be

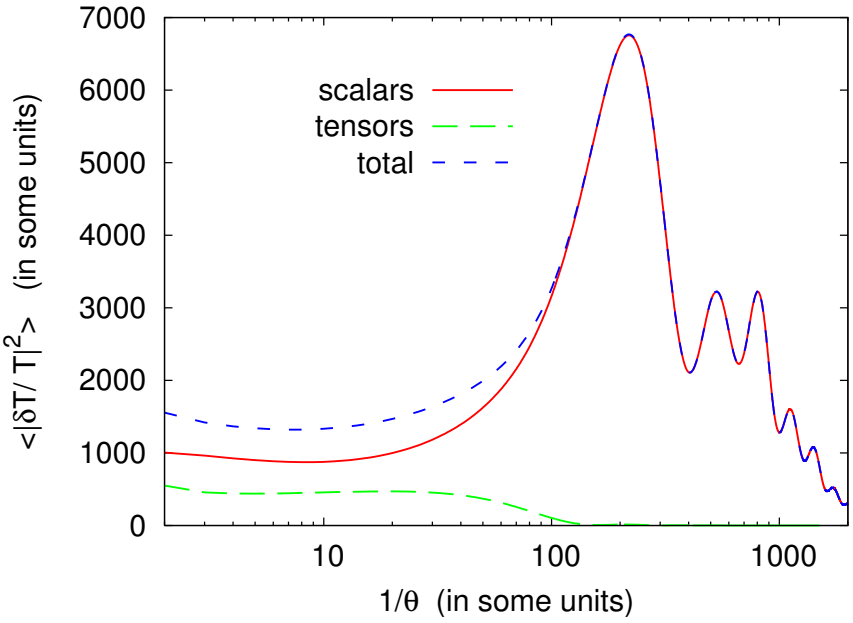


Figure 7.2: The red solid line shows the same reference CMB temperature spectrum as in the previous figures. The dashed green line shows the additional contribution from tensor perturbations for the same cosmological model, assuming a primordial tensor amplitude such that roughly one third of the observed CMB power spectrum on large angular scales would come from tensors. The dashed blue curve shows the total spectrum which would be observed in this model.

computed with the same kind of numerical code as for scalar perturbations. The main uncertainty is not on the shape, but on the amplitude of this contribution. Equation (7.42) shows that the amplitude depends on V during inflation, i.e. on the energy scale of inflation. The condition for the tensor contribution to be roughly of the same order as the scalar one in the large scale CMB spectrum is roughly that $V \sim (10^{16}\text{GeV})^4$ during inflation, i.e. that the energy scale of inflation is of the order of 10^{16}GeV (coincidentally, this turns out to be the order of magnitude of GUT symmetry breaking).

So far, the observation of CMB anisotropies is consistent with a spectrum arising only from scalar perturbations. A large tensor contribution cannot be present, because it would lead to an increase in the ratio between the amplitude of the large-scale plateau region and that of the small-scale peak region, at odds with observation. Hence, CMB temperature maps allow to put an upper limit on the energy scale of inflation: roughly, it has to be smaller than 10^{16}GeV. Future observations of the CMB will be able to test the possible contribution of tensors with better precision: hence, in the next years, cosmologists hope either to push this bound further down, or to detect the background of primordial gravitational waves produced by inflation.

7.4 Success of the theory of inflation

Let us summarize the positive outcomes of the theory of inflation:

1. it provides a simple solution to the flatness, horizon and monopole problems.

2. it includes in some unavoidable way a mechanism for producing primordial fluctuations starting from simple initial conditions, i.e. from a perfectly homogeneous scalar field with vacuum fluctuations dictated by quantum mechanics.

3. these perturbations have almost automatically the properties which are necessary in order to explain observations: they are generated very early and on super-Hubble scales; they have a Gaussian statistics and obey to adiabatic initial conditions; they have a nearly scale-invariant primordial spectrum.

4. inflation provides a mechanism for the generation of a thermal bath of particles in the early universe (the so-called reheating phase occurring after or during the scalar field oscillations and decay). Unfortunately, this mechanism is very difficult to probe experimentally: reheating does not have clear observable signatures, unlike the mechanism for the generation of primordial fluctuations.

5. thanks to the theory of inflation, it is possible to provide a self-consitent explanation for the global properties of our universe without making any assumption about quantum gravity (during inflation, one quantizes only metric perturbations, not the metric itself: hence inflation is based on quantum field theory in curved space-time, but NOT on quantum gravity). In fact, in inflationary cosmology, what happens *before* inflation is usually not important: our universe only keeps track of what happened during the last $\sim$60 e-folds of inflation and after inflation.

The third point is the most convincing argument in favor of inflation. Before the first observations of CMB anisotropies, it was impossible to know whether our universe was described by such initial conditions (primordial perturbations on super-Hubble scales, Gaussian, adiabatic and nearly scale-invariant). So, cosmologist were studying various possible scenarios for the generation of perturbations. The main alternative would be to assume that they are generated during a phase transition (e.g. a spontaneous symmetry breaking). In this case, they would appear inside the Hubble radius and would be non-Gaussian, non-adiabatic and far from scale invariance. As we have seen before, the observation of CMB anisotropies has confirmed the four generic predictions of inflation, as far as primordial perturbations are concerned. Alternative theories are discarded (at least as a dominant mechanism for the generation of primordial perturbations) and most people agree that inflation is a very likely scenario. It is a striking example of a predictive and elegant theory (with few assumptions leading to many observable consequences validated by observations).

The negative outcomes of the theory of inflation are the following:

1. inflation is based on a scalar field (usually called the *inflaton*), but we don't know anything about its origin and its relationship with other known fields/particles. However it is possible to assume some connection between inflation and particle physics (the inflaton could be a Higgs field, the size of an extra dimension, etc.) The difficulty is then to find a good reason for which the inflaton potential would be flat enough for fulfilling the slow-roll

conditionc. This argument readily excludes the possibility that the inflaton would be the usual electroweak Higgs field (in the context of standard particle physics and gravity). However, it could still be a component of the Higgs field associated with the breaking of the GUT symmetry (although this question is very subtle and related to supersymmetry, supergravity, etc.) There are many interesting research activities in this direction.

2. inflation predicts a background of gravitational waves which have not been detected. Hopefully, this is only a matter of sensitivity: future CMB experiments might see these primordial gravitational waves. If they do, there would be one more very convincing evidence in favor of inflation (primordial gravitational waves would be the "smoking gun" of inflation), and the previous issue would become much more interesting and promising, since we would finally know the energy scale of inflation, as well as the details of the inflaton potential $V(\varphi)$ within some interval $\Delta\varphi$. As long as we don't see these primordial gravitational waves, we can only constrain the function V^3/V'^2 (see eq. (7.41)). There exist one-parameter families of potentials all giving the same combination V^3/V'^2, and hence the same primordial scalar spectrum. Hence, the scale of inflation will remain unknown until primordial gravitational waves are observed. Of course, this might never occur, in which case the theory of inflation would not be excluded, but its existence would not be proved or disproved as convincingly as one would like to.

Conclusions

Modern cosmology offers a detailed and self-consistent scenario, able to explains most (if not all) observations of the global properties of the universe. The most impressive success of the past years is the fact that cosmological perturbation theory (with initial conditions motivated by inflation) allowed to predict the non-trivial spectrum of CMB temperature fluctuations much before it was actually observed; the good agreement between, on the one hand, the CMB data obtained by WMAP and more recently by Planck, and on the other hand, the predictions of the minimal ΛCDM scenario, is one of the greatest successes of modern science.

However, the observations of the last decade reveal that the universe contains nearly 26% of dark matter and 69% of cosmological constant (or of another fluid leading to accelerated expansion today, generically called dark energy), which are both of completely unknown nature and origin. This two issues are now the main challenges in cosmology (the third challenge being to understand the nature and origin of the inflaton, by first measuring its energy scale through CMB B-polarisation).

Bibliography

[1] *An Introduction to Modern Cosmology*, by Andrew Liddle, John Wiley & Sons, Chichester, 2003.
[2] *Modern cosmology*, by Scott Dodelson, New York, NY: Academic Press, 2003.
[3] *The Early universe*, by Edward W. Kolb and Michael S. Turner, Redwood City, CA: Addison-Wesley, 1990.
[4] *Physical Foundations of Cosmology*, by Viatcheslav Mukhanov, Cambridge Univ. Press, 2005.
[5] *Cosmology*, by Steven Weinberg, Oxford Univ. Press, 2008.
[6] *Neutrino Cosmology*, by Julien Lesgourgues, Gianpiero Mangano, Gennaro Miele, Sergio Pastor, Cambridge Univ. Press, 2013.