AST4220: Cosmology I

Øystein Elgarøy

Contents

Chapter 1

Cosmological models

Cosmology is the study of the universe as a whole. We want to learn about its size, its shape and its age. Also, we want to understand the distribution of matter in the form of galaxies, clusters of galaxies and so on, and how this distribution arose. Even more ambitiously, we want to know how the universe started and how it will end. These are all bold questions to ask, and the fact that we are now getting closer to answering many of them is a testimony to the tremendous theoretical and, perhaps most important, observational effort invested over the past century.

It is not totally inaccurate to say that modern cosmology started with Einstein's theory of general relativity (from now on called GR for short). GR is the overarching framework for modern cosmology, and we cannot avoid starting this course with at least a brief account of some of the most important features of this theory.

1.1 Special relativity: space and time as a unity

Special relativity, as you may recall, deals with inertial frames and how physical quantities measured by observers moving with constant velocity relative to each other are related. The two basic principles are:

1. The speed of light in empty space, c, is the same for all observers.
2. The laws of physics are the same in all inertial frames.

From these principles the strange, but by now familiar, results of special relativity can be derived: the Lorentz transformations, length contraction, time dilation etc. The most common textbook approach is to start from the Lorentz transformations relating the position and time for an event as observed in two different inertial frames. However, all the familiar results can be obtained by focusing instead on the invariance of the spacetime interval (here given in Cartesian coordinates)

$$ds^2 = c^2dt^2 - dx^2 - dy^2 - dz^2 \tag{1.1}$$

for two events separated by the time interval dt and by coordinate distances dx, dy, and dz. The invariance of this quantity for all inertial observers follows directly from the principles of relativity.

To see how familiar results can be derived from this viewpoint, consider the phenomenon of length contraction: Imagine a long rod of length L as measured by an observer at rest in the frame S. Another observer is travelling at speed v relative to the frame S, at rest in the origin of his frame S'. When the observer in S' passes the first end point of the rod, both observers start their clocks, and they both stop them when they see the observer in S' pass the second end point of the rod. To the observer in S, this happens after a time $dt = L/v$. Since the observer in S' is at rest in the origin of his frame, he measures no spatial coordinate difference between the two events, but a time difference $dt' = \tau$. Thus, from the invariance of the interval we have

$$ds^2 = c^2 \left(\frac{L}{v}\right)^2 - L^2 = c^2\tau^2 - 0^2$$

from which we find

$$\tau = \frac{L}{v}\sqrt{1 - \frac{v^2}{c^2}}.$$

Since the observer in S' sees the first end point of the rod receding at a speed v, he therefore calculates that the length of the rod is

$$L' = v\tau = L\sqrt{1 - \frac{v^2}{c^2}} \equiv \gamma L < L. \tag{1.2}$$

Similarly, we can derive the usual time dilation result: moving clocks run at a slower rate (i.e. record a shorter time interval between two given events) than clocks at rest. Consider once again our two observers in S and S' whose clocks are synchronized as the origin of S' passes the origin of S at $t = t' = 0$. This is the first event. A second event, happening at the origin of S' is recorded by both observers after a time Δt in S, $\Delta t'$ in S'. From the invariance of the interval, we then have

$$c^2\Delta t^2 - v^2\Delta t^2 = c^2\Delta t'^2$$

which gives

$$\Delta t = \frac{\Delta t'}{\sqrt{1 - v^2/c^2}} = \frac{\Delta t'}{\gamma} > \Delta t'.$$

This approach to special relativity emphasizes the unity of space and time: in relating events as seen by observers in relative motion, both the time and the coordinate separation of the events enter. Also, the geometrical aspect of special relativity is emphasized: spacetime 'distances' (intervals) play the fundamental role in that they are the same for all observers. These

features carry over into general relativity. General relativity is essential for describing physics in accelerated reference frames and gravitation. A novel feature is that acceleration and gravitation lead us to introduce the concept of curved spacetime. In the following section we will explore why this is so.

1.2 Curved spacetime

In introductory mechanics we learned that in the Earth's gravitational field all bodies fall with the same acceleration, which near the surface of the Earth is the familiar $g = 9.81\ \mathrm{m}/s^2$. This result rests on the fact that the mass which appears in Newton's law of gravitation is the same as that appearing in Newton's second law $\mathbf{F} = m\mathbf{a}$. This equality of gravitational and inertial mass is called the equivalence principle of Newtonian physics. We will use this as a starting point for motivating the notion of curved spacetime and the equivalence of uniform acceleration and uniform gravitational fields.

Consider a situation where you are situated on the floor of an elevator, resting on the Earth's surface. The elevator has no windows and is in every way imaginable sealed off from its surroundings. Near the roof of the elevator there is a mechanism which can drop objects of various masses towards the floor. You carry out experiments and notice the usual things like, e.g. that two objects dropped at the same time also reach the floor at the same time, and that they all accelerate with the same acceleration g. Next we move the elevator into space, far away from the gravitational influence of the Earth and other massive objects, and provide it with an engine which keeps it moving with constant acceleration g. You carry out the same experiments. There is now no gravitational force on the objects, but since the floor of the elevator is accelerating towards the objects, you will see exactly the same things happen as you did when situated on the surface of the Earth: all objects accelerate towards the floor with constant acceleration g. There is no way you can distinguish between the two situations based on these experiments, and so they are completely equivalent: you cannot distinguish uniform acceleration from a uniform gravitational field!

Einstein took this result one step further and formulated his version of the equivalence principle: You cannot make *any* experiment which will distinguish between a uniform gravitational field and being in a uniformly accelerated reference frame!

This has the further effect that a light ray will be bent in a gravitational field. To understand this, consider the situation with the elevator accelerating in outer space. A light ray travels in a direction perpendicular to the direction of motion of the elevator, and eventually enters through a small hole in one of the sides. For an outside observer the light ray travels in a straight line, but to an observer inside the elevator it is clear that the light ray will hit the opposite side at a point which is lower than the point of

entry because the elevator is all the time accelerating upwards. Thus, the light ray will by the observer in the elevator be seen to travel in a curved path. But if we are to take the equivalence principle seriously, this must also mean that a stationary observer in a uniform gravitational field must see the same thing: light will follow a curved path. Since the trajectory of light rays are what we use to define what is meant by a 'straight line', this must mean that space itself is curved. We can interpret the effect of the gravitational field as spacetime curvature.

1.3 Curved spaces: the surface of a sphere

You already have some experience with curved spaces, since we actually live on one! The Earth's surface is spherical, and the surface of a sphere is a two-dimensional curved space. But how can we tell that it is curved? One way is by looking at straight lines. If we define a straight line as the shortest path (lying completely within the surface) between two points in the surface, then in a plane this will be what we normally think of as a straight line. However, it is easy to see that on the surface of a sphere, a straight line defined in this manner will actually be an arc of a circle.

Another, more quantitative way of detecting curvature is to consider the ratio of the circumference and the radius of a circle on the surface. By a circle we mean the set of points on the surface which all lie at a given distance s (measured on the surface!) from a given point P (the center of the circle). In a plane the relationship between the radius and the circumference is the usual $c = 2\pi s$ we all know and love. However, consider a circle on a spherical surface (see fig. 1.1). The circumference of this circle is clearly $c = 2\pi r$. However, the radius, as measured on the surface, is not r but s, and these two quantities are related by

$$r = a \sin\theta \tag{1.3}$$

$$\theta = \frac{s}{a}, \tag{1.4}$$

where a is the radius of the sphere. We therefore find

$$\begin{aligned} c &= 2\pi a \sin\theta = 2\pi \sin\left(\frac{s}{a}\right) \\ &= 2\pi a \left(\frac{s}{a} - \frac{s^3}{6a^3} + \cdots\right) \\ &= 2\pi s \left(1 - \frac{s^2}{6a^2} + \ldots\right), \end{aligned} \tag{1.5}$$

which is smaller than $2\pi s$. This is characteristic of curved spaces: the circumference of a circle does not obey the usual '2π times the radius'-relationship.

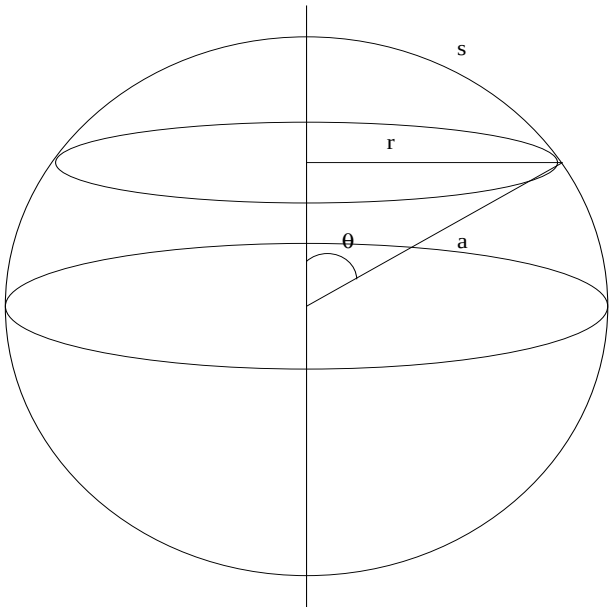

Figure 1.1: Symbols used in the discussion of the curvature of a spherical surface. Note that the circumference of the circle is $2\pi r$, but the radius (the distance from the center to the perimeter) as measured by a creature confined to walk along the surface of the sphere is s.

We can go a bit further and define a quantitative measure of curvature (for two-dimensional spaces), the so-called *Gaussian curvature*, K:

$$K \equiv \frac{3}{\pi} \lim_{s \to 0} \left(\frac{2\pi s - C}{s^3} \right). \tag{1.6}$$

For the spherical, two-dimensional space we find

$$\begin{aligned} K &= \frac{3}{\pi} \lim_{s \to 0} \frac{1}{s^3} \left(2\pi s - 2\pi s + \frac{2\pi s^3}{6a^2} - \ldots \right) \\ &= \frac{1}{a^2}. \end{aligned} \tag{1.7}$$

The Gaussian curvature of the surface of a sphere is thus positive. It is a general feature of positively curved spaces that the circumference of a circle of radius s is less than $2\pi s$. One can also show that there exists negatively curved spaces in two dimensions, one example being the surface of a hyperboloid. For negatively curved surfaces, the circumference of a circle is greater than $2\pi s$.

1.4 The Robertson-Walker line element

In this section we will try to make plausible the form of the line element for a homogeneous and isotropic space. Homogeneous means that, from a given observation point, the density is independent of the distance from the observer. Isotropic means that the observer sees the same density in all directions. Such a space is an excellent approximation to our Universe, so the result in this section is one of the most important in these lectures. It forms the foundation for almost everything we will do later on.

We start by, once again, looking at the two-dimensional surface of a sphere in three dimensions. Let us introduce coordinates (r', ϕ) on this surface in such a way that the circumference of a circle centered at one of the poles is given by $2\pi r'$. We see that $r' = a\sin\theta$, $\theta = s/a$, so

$$s = a\sin^{-1}\left(\frac{r'}{a}\right).$$

If we keep r' fixed ($dr' = 0$) and vary ϕ, we have $ds = r'd\phi$. Keeping constant ϕ and changing r' by dr', we get

$$\begin{aligned} ds &= \frac{ds}{dr'}dr' = a\frac{1}{\sqrt{1-\left(\frac{r'}{a}\right)^2}}\frac{1}{a}dr' \\ &= \frac{dr'}{\sqrt{1-\left(\frac{r'}{a}\right)^2}}. \end{aligned}$$

Since the two coordinate directions are orthogonal and independent, we can then write the line element for this surface as

$$ds^2 = \frac{dr'^2}{1-\left(\frac{r'}{a}\right)^2} + r'^2 d\phi^2.$$

We saw that the Gaussian curvature K for this surface is $K = 1/a^2$, so we can write

$$ds^2 = \frac{dr'^2}{1-Kr'^2} + r'^2 d\phi^2,$$

and introducing a dimensionless coordinate $r = r'/a$, we find

$$ds^2 = a^2\left(\frac{dr^2}{1-kr^2} + r^2 d\phi^2\right), \tag{1.8}$$

where $k \equiv Ka^2 = +1$. We now note that we can describe other spaces by allowing k to be a parameter taking on different values for different spaces. For example, taking $k = 0$, we get

$$ds^2 = a^2(dr^2 + r^2 d\phi^2),$$

which is the line element of the two-dimensional Euclidean plane expressed in polar coordinates. Furthermore, one can show that the negatively curved two-dimensional space (e.g. the surface of a hyperboloid) has a line element on the same form with $k = -1$. So flat, as well as both positively and negatively curved two-dimensional surfaces can be described by the line element (1.8) with $k = -1, 0, +1$. Note that the physical size a enters just as an overall scale factor in the expression.

Let us calculate the path length s in going from $r = 0$ to a finite value of r along a meridian with $d\phi = 0$:

$$s = a \int_0^r \frac{dr'}{\sqrt{1 - kr'^2}},$$

which is equal to $a \sin^{-1}(r)$ for $k = +1$, ar for $k = 0$, and $a \sinh^{-1} r$ for $k = -1$.

Note that in the case $k = +1$ the circumference of a circle $c = 2\pi a \sin(s/a)$ increases until $s = \pi a/2$, then decreases and finally reaches zero for $s = \pi a$. By drawing a sequence of circles from the north to the south pole of a sphere you should be able to see why this is so. This feature is typical of a positively curved space. For $k = -1, 0$ the circumference of a circle in the surface will increase without bounds as s increases. The surface of the sphere is also an example of a closed space. Note that it has a finite surface area equal to $4\pi a^2$, but no boundaries.

So far we have looked at two-dimensional surfaces since they have the advantage of being possible to visualize. Three dimensional surfaces (i.e. the surface of a four-dimensional object) are harder once we go beyond the flat, Euclidean case. But in that case we know that we can write the line element in spherical coordinates as

$$ds^2 = a^2(dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\phi^2) = a^2(dr^2 + r^2 d\Omega^2),$$

where $d\Omega^2 \equiv d\theta^2 + \sin^2\theta d\phi^2$. This space is homogeneous and isotropic. It looks the same at any point and in any direction, and the local curvature is the same at all points. These are properties we normally assume our Universe to possess, an assumption which called the 'The Cosmological Principle'. It has passed all observational tests so far, and thus seems to be a very reasonable starting point for building a cosmological model. However, flat Euclidean space is not the only space satisfying this principle. There are both positively and negatively curved homogeneous and isotropic spaces.

For a positively curved space, we can carry out a 3D version of the analysis we went through for the surface of a sphere. We define angular variables θ and ϕ and a dimensionless radial coordinate r so that a surface through the point with coordinate r has area $4\pi(ar)^2$. We then have

$$ds^2 = a^2(g_{rr}dr^2 + r^2 d\Omega^2).$$

For the surface of a three-sphere

$$x^2 + y^2 + z^2 + w^2 = a^2,$$

we can repeat the two-dimensional analysis and obtain

$$g_{rr} = \frac{a^2}{1 - r^2}.$$

More generally, it can be shown that any isotropic and homogeneous three dimensional space can be described by coordinates of this type and with a line element

$$ds^2 = a^2 \left(\frac{dr^2}{1 - kr^2} + r^2 d\Omega^2 \right) \tag{1.9}$$

where the curvature parameter k again can take on the values -1, 0 and $+1$. This line element describes the spatial structure of our Universe, so at a given time t the spatial part of the line element will be of this form. The factor a will in general be a function of the time (cosmic time) t, so we write $a = a(t)$. It is this feature which will allow us to describe an expanding universe. The time part of the line element is just $c^2 dt^2$, so we can finally write

$$ds^2 = c^2 dt^2 - a^2(t) \left(\frac{dr^2}{1 - kr^2} + r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2 \right). \tag{1.10}$$

This is the Robertson-Walker (RW) line element, and it is the only line element we will ever use. The coordinates r, θ, ϕ are such that the circumference of a circle corresponding to t, r, θ all being constant is given by $2\pi a(t) r$, the area of a sphere corresponding to t and r constant is given by $4\pi a^2(t) r^2$, but the physical radius of the circle and sphere is given by

$$R_{\text{phys}} = a(t) \int_0^r \frac{dr'}{\sqrt{1 - kr'^2}}.$$

I emphasize that the coordinates (r, θ, ϕ) are comoving coordinates: if an object follows the expansion or contraction of space it has fixed coordinates with respect to the chosen origin. The expansion or contraction of space is described entirely by the scale factor $a(t)$. For $k = +1$ the Universe is closed (but without boundaries), and $a(t)$ may be interpreted as the 'radius' of the Universe at time t. If $k = 0, -1$, the Universe is flat/open and infinite in extent.

The time coordinate t appearing in the RW line element is the so-called *cosmic time*. It is the time measured on the clock of an observer moving along with the expansion of the universe. The isotropy of the universe makes it possible to introduce such a global time coordinate. We can imagine that observers at different points exchange light signals and agree to set their

clocks to a common time t when, e.g., their local matter density reaches a certain value. Because of the isotropy of the universe, this density will evolve in the same way in the different locations, and thus once the clocks are synchronized they will stay so.

1.5 Redshifts and cosmological distances

The RW metric contains two unknown quantities: the scale factor $a(t)$ and the spatial curvature parameter k. In order to determine them, we need an equation relating the geometry of the universe to its matter-energy content. This is the subject of the next section. In the present section we will use the RW line element to introduce the notions of cosmic redshift and distances. When doing so, we will consider how light rays propagate in a universe described by the RW line element. Light rays in special relativity move along lines of constant proper time, $ds^2 = 0$. This is easily seen by noting that $ds^2 = 0$ implies

$$\frac{\sqrt{dx^2 + dy^2 + dz^2}}{dt} = \pm c$$

and thus describes motion at the speed of light. This carries over to general relativity since it is always possible locally, at a given point, to find a frame where the line element reduces to that of flat space. And since ds^2 is a scalar, which means that it is the same evaluated in any frame, this means that $ds^2 = 0$ is valid in all reference frames for a light ray.

1.5.1 The cosmic redshift

The redshift of a cosmological object has the advantage of being quite easily measurable: it just requires comparing the wavelengths of spectral lines. In mechanics we are used to interpreting redshift as a consequence of the Doppler effect, an effect of the source of the waves moving through space. However, the cosmological redshift is of a different nature: it can in a certain sense be intepreted as a result of space itself stretching! More conservatively, one can say that it is a result of light propagating in curved spacetime.

Let us consider a train of electromagnetic waves emitted from a point P, as shown in fig. 1.2, and moving towards us at the origin O. The first peak of the wave is emitted at a cosmic time t_e, and the second at an infinitesimally later time $t_e + \delta t_e$. We receive them at times t_o and $t_o + \delta t_o$, respectively. The light wave travels along a line of constant θ and ϕ and follows a path defined by $ds^2 = 0$. Inserting this in the RW line element gives

$$ds^2 = 0 = c^2 dt^2 - a^2(t) \frac{dr^2}{1 - kr^2},$$

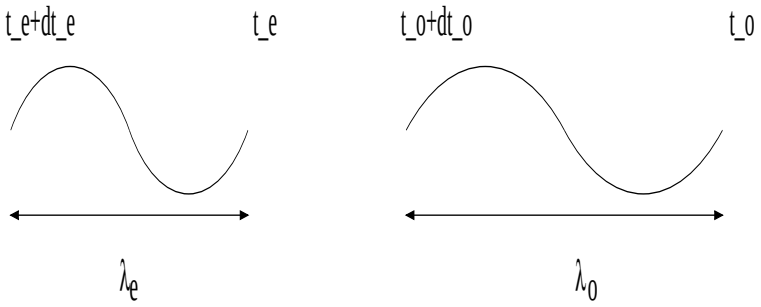

Figure 1.2: An electromagnetic wave travelling through the expanding universe is stretched.

and since $dr < 0$ for $dt > 0$ (the light wave moves towards lower values of r since it is moving towards us at the origin), we have

$$\frac{cdt}{a(t)} = -\frac{dr}{\sqrt{1-kr^2}}.$$

For the first peak we then have

$$\int_{t_e}^{t_o} \frac{cdt}{a(t)} = -\int_r^0 \frac{dr}{\sqrt{1-kr^2}} = \int_0^r \frac{dr}{\sqrt{1-kr^2}},$$

and for the second peak we have similarly

$$\int_{t_e+\delta t_e}^{t_o+\delta t_o} \frac{cdt}{a(t)} = \int_0^r \frac{dr}{\sqrt{1-kr^2}}.$$

We then see that we must have

$$\int_{t_e}^{t_o} \frac{cdt}{a(t)} = \int_{t_e+\delta t_e}^{t_o+\delta t_o} \frac{cdt}{a(t)}.$$

We can split the integrals on each side into two parts:

$$\int_{t_e}^{t_e+\delta t_e} \frac{cdt}{a(t)} + \int_{t_e+\delta t_e}^{t_o} \frac{cdt}{a(t)} = \int_{t_e+\delta t_e}^{t_o} \frac{cdt}{a(t)} + \int_{t_o}^{t_o+\delta t_o} \frac{cdt}{a(t)},$$

and hence

$$\int_{t_e}^{t_e+\delta t_e} \frac{cdt}{a(t)} = \int_{t_o}^{t_o+\delta t_o} \frac{cdt}{a(t)}.$$

Since both integrals now are taken over an infinitesimally short time, we can take the integrand to be constant and get

$$\frac{c\delta t_e}{a(t_e)} = \frac{c\delta t_o}{a(t_o)}.$$

Note that this implies that

$$\delta t_e = \frac{a(t_e)}{a(t_o)}\delta t_o < \delta t_o.$$

This means that pulses recieved with a separation in time δt_o were emitted with a shorter separation in time δt_e by the object.

Since $c\delta t_e = \lambda_e$ and $c\delta t_o = \lambda_o$, we can rewrite the relation above as

$$\frac{\lambda_o}{\lambda_e} = \frac{a(t_o)}{a(t_e)}.$$

This means that in an expanding universe, the wavelength of a light wave upon reception will be longer than at the time of emission by a factor equal to the ratio of the scale factors of the universe at the two times. The cosmic redshift is usually measured by the parameter z defined by

$$1 + z = \frac{\lambda_o}{\lambda_e} = \frac{a(t_o)}{a(t_e)}, \tag{1.11}$$

and measures how much the universe has expanded between the times of emission and reception of the signal.

1.5.2 Proper distance

You may already have thought about one issue that arises when we want to specify distances in cosmology, namely that space is expanding. One way of handling this when calculating distances is to compute them at a given time t. This is the content of the so-called *proper distance*, it is the length of the spatial geodesic (shortest path in space) between two points at a specified time t, so that the scale factor describing the expansion of the universe is held fixed at $a(t)$. Another way of saying this is that the proper distance between two points is the distance as read off on a set of rulers connecting the two points at the time t. It is denoted by $d_{\rm P}(t)$, and can be obtained as follows. Without loss of generality, we can place one point at the origin $(0, 0, 0)$ and let the other point have coordinates (r, θ, ϕ). Along the spatial geodesic (the 'straight line') between the two points, only the coordinate r varies (think of the surface of a sphere!) The time t is fixed, and we are to compute the spatial distance, so the RW line element gives for an infinitesimal displacement along the geodesic

$$|ds| = a(t)\frac{dr'}{\sqrt{1 - kr'^2}}.$$

The proper distance is found by summing up all contributions along the geodesic, hence

$$d_{\rm P}(t) = a(t)\int_0^r \frac{dr'}{\sqrt{1 - kr'^2}} = a(t)\mathcal{S}_k{}^{-1}(r),$$

where $\mathcal{S}_k^{-1}$, is a functional[1] such that $\mathcal{S}_k^{-1}(r) = \sin^{-1}(r)$ for $k = +1$, $\mathcal{S}_k^{-1}(r) = r$ for $k = 0$ and $\mathcal{S}^{-1}(r) = \sinh^{-1}(r)$ for $k = -1$. We see that this results agrees with our intuition for the spatially flat case, $k = 0$: $d_{\rm P}(t) = a(t)r$, which means that the proper distance is then just the comoving coordinate r of the point, which is a constant in time, times the scale factor which describes how much the universe has expanded since a given reference time.

Since $d_{\rm P}$ is a function of t, the relative distance between the two points is increasing as the Universe expands. The relative radial velocity is given by

$$v_{\rm r} = \frac{d}{dt} d_{\rm P}(t) = \dot{a} {\mathcal{S}_k}^{-1}(r) = \frac{\dot{a}}{a} d_{\rm P}(t),$$

where dots denote time derivatives. If we introduce the Hubble parameter $H(t) \equiv \dot{a}/a$, we find that

$$v_{\rm r}(t) = H(t) d_{\rm P}(t), \tag{1.12}$$

which is Hubble's law: at a given time, points in the Universe are moving apart with a speed proportional to their distance. Note that the Hubble parameter is in general a function of time: the Universe does not in general expand at the same rate at all times.

It is worthwhile to note that Hubble's expansion law is a direct consequence of the homogeneity of the universe. Consider, e.g., three galaxies A, B, and C, lying along the same straight line. Let B be at a distance d from A, and let C be at distance d from B, and hence $2d$ from A. Now, let the velocity of B relative to A be v. Assuming homogeneity, then C has to move with speed v relative to B, since it has the same distance from B as B has from A. But then C moves at a velocity $2v$ relative to A, and hence its speed is proportional to its distance from A. We can add more galaxies to the chain, and the result will be the same: the speed of recession of one galaxy with respect to another is proportional to its distance from it. Note that we used the non-relativistic law of addition of velocities in this argument, so for galaxies moving at the speed of light, this kind of reasoning is no good. However, as we probe greater distances, we also probe more distant epochs in the history of the universe. As can be seen from equation (1.12), the Hubble parameter actually varies in time, so we do not expect a strict linear relationship between distance and speed as we probe the universe at great distances.

If we denote the present time by t_0, the best measurements of the current value of the Hubble parameter indicate that $H_0 \equiv H(t_0) = (72 \pm 8)\ \mathrm{km\,s^{-1}\,Mpc^{-1}}$. Note that it is common to introduce the dimensionless Hubble constant by writing

$$H_0 = 100h\ \mathrm{km\,s^{-1}\,Mpc^{-1}}, \tag{1.13}$$

where we have $h \approx 0.72$ today.

[1] That is, a paramatrised familiy of functions.

1.5.3 The luminosity distance

All measured distances to cosmological objects are derived from the properties of the light we receive from them. Since light travels at a finite speed, it is clear that the universe may have expanded by a significant amount during the time the light has travelled towards us. We need to establish relations between distances deduced from the properties of the light we receive and the quantities in the RW metric.

A common measure of distance is the so-called *luminosity distance* $d_{\rm L}$. Consider a source P at a distance d from an observer O. If the source emits an energy per unit time L, and l is the flux (energy per unit time and area) received by the observer, then in a static, Euclidean geometry we would have $l = L/(4\pi d^2)$, and so the distance d would be related to luminosity L and flux l by

$$d = \sqrt{\frac{L}{4\pi l}}.$$

Motivated by this, we define the luminosity distance in general to be given by

$$d_{\rm L} \equiv \sqrt{\frac{L}{4\pi l}}. \tag{1.14}$$

The received flux l is relatively easy to measure, and if we know L, we can then compute $d_{\rm L}$. But how is it related to $a(t)$ and k? Consider a spherical shell centered at P going through O at the time of observation $t_{\rm o}$. Its area is given by definition of the coordinate r as $4\pi a^2(t_{\rm o})r^2$. The photons emitted at P at the time t have had their wavelengths stretched by a factor $a(t_{\rm o})/a(t)$ when they reach O. Furthermore, as illustrated in our discussion of the redshift, wave peaks emitted in a time interval δt at P are received at O in the slightly longer interval $\delta t_{\rm o} = a(t_{\rm o})/a(t)\delta t$, hence reducing further the energy received per unit time at O as compared with the situation at P. The received flux at O therefore becomes

$$l = \frac{L}{4\pi a^2(t_{\rm o})r^2}\left(\frac{a(t)}{a(t_o)}\right)^2, \tag{1.15}$$

and using the definition (1.14) we get

$$d_{\rm L} = \sqrt{\frac{L}{4\pi l}} = a(t_{\rm o})r\frac{a(t_{\rm o})}{a(t)},$$

and using finally the definition of redshift (1.11) we find

$$d_{\rm L} = a(t_{\rm o})r(1+z). \tag{1.16}$$

Not that this definition assumes that we know the intrinsic luminosity L of the source. Sources with this property are called 'standard candles',

and they have been crucial in determining the cosmological distance ladder. Historically, Cepheid variables have been important, and more recently supernovae of type Ia have been used to determine distances out to very large redshifts z and have led to the discovery of accelerated cosmic expansion.

1.5.4 The angular diameter distance

Another common measure of distance is the *angular diameter distance*, $d_{\rm A}$. Recall that a source of known, fixed size D observed at a large distance d ('large' means $d \gg D$) covers an angle $\Delta\theta = D/d$ (in radians) in a static, Euclidean geometry. We define the angular diameter distance so as to preserve this relation in the general case, thus

$$d_{\rm A} \equiv \frac{D}{\Delta\theta}. \tag{1.17}$$

We now have to relate the quantities in this definition to the RW line element. We place the observer at the origin and a source at a radial comoving coordinate r. The proper diameter $D_{\rm P}$ of the source is measured at time t, and we measure that the source has an angular extent $\Delta\theta$ now. Using the RW line element, we find

$$ds^2 = -r^2 a^2(t)(\Delta\theta)^2 = -D_{\rm P}^2,$$

so that

$$D_{\rm P} = a(t) r \Delta\theta.$$

We therefore find

$$d_{\rm A} = \frac{D_{\rm P}}{\Delta\theta} = a(t) r = \frac{a(t)}{a(t_{\rm o})} a(t_{\rm o}) r = \frac{a(t_{\rm o}) r}{1+z}, \tag{1.18}$$

where $t_{\rm o}$ is the time at which the observer O receives the light emitted at time t by the source P. Note that, as with the luminosity distance, an intrinsic property of the source must be known in order to determine the angular diameter distance observationally, in this case the intrinsic size of the source.

Comparing equation (1.18) to equation (1.16) we see that there is a simple relation between $d_{\rm L}$ and $d_{\rm A}$:

$$\frac{d_{\rm L}}{d_{\rm A}} = (1+z)^2, \tag{1.19}$$

and hence this ratio is model-independent.

1.5.5 The comoving coordinate r

The expressions for the luminosity distance and the angular diameter distance of a source P observed at time t_o both involve its comoving radial coordinate r at the time of emission t. We want to relate this to the scale factor $a(t)$ and the spatial curvature parameter k. In order to do this we consider a light ray propagating from the source towards the observer at the origin. The light ray travels at constant θ and ϕ along a null geodesic $ds^2 = 0$, and thus the RW line element gives

$$\begin{aligned} 0 &= c^2dt^2 - \frac{a^2(t)dr^2}{1-kr^2} \\ \Rightarrow \frac{dr}{\sqrt{1-kr^2}} &= -\frac{cdt}{a(t)}, \end{aligned} \tag{1.20}$$

where the $-$ sign is chosen because r decreases ($dr < 0$) as time increases ($dt > 0$) along the path of the light ray. Integrating equation (1.20) we therefore have

$$\mathcal{S}_k^{-1}(r) \equiv \int_0^r \frac{dr'}{\sqrt{1-kr'^2}} = \int_t^{t_o} \frac{cdt'}{a(t')}. \tag{1.21}$$

where $\mathcal{S}_k^{-1}(r)$ is the inverse of the function $\mathcal{S}_k(r)$, the latter being equal to $\sin(r)$ for $k = +1$, r for $k = 0$ and $\sinh(r)$ for $k = -1$. Thus we find that

$$r = \mathcal{S}_k\left[\int_t^{t_o} \frac{cdt'}{a(t')}\right]. \tag{1.22}$$

1.6 The Friedmann equations

We have now seen how we can use the RW metric for an isotropic and homogeneous universe to compute distances and obtain redshifts for astrophysical objects. We have also seen that these expressions depend on the scale factor $a(t)$ and the spatial curvature parameter k. So far we have assumed that these are given, but now we turn to the question of how they can be determined. The key is Einstein's theory of general relativity which is the most fundamental description of gravity we know of. In this theory, gravity is no longer considered a force, but an effect of matter and energy causing spacetime to curve. Thus, free particles are always travelling in straight lines, but what a 'straight line' is, is determined by the geometry of spacetime. And the geometry of spacetime is determined by the matter and energy which is present through the so-called Einstein field equation. To develop the full machinery of GR would take us too far afield here, and we do not really need it. Suffice it to say that the field equation says that the spacetime curvature is proportional to the so-called energy-momentum tensor. Given the RW line element, the field equation is reduced to two differential equations for the scale factor where the spatial curvature enters as a parameter. The

form of these equations can be derived from a Newtonian argument, and you may already have seen how this can be done in earlier courses. In case you haven't, here it is: We assume a homogeneous and isotropic mass distribution of density ρ. Consider a sphere of radius R centered on the origin of our coordinate system. We allow the sphere to expand or contract under its own gravity and write the radius as $R = ra(t)$, where r is a constant, and represents a comoving coordinate. Next, we place a test mass m on the surface of the sphere. From Newtonian theory we know that only the mass M contained within the sphere of radius R will exert a gravitational force on m: if one divides the region outside into spherical shells, one finds that the force from each shell on m vanishes. Thus, the motion of the test mass can be analyzed by considering the mass within R only. The first thing to note is that gravity is a conservative force field so that the mechanical energy is conserved during the motion of the test mass:

$$\frac{1}{2}m\dot{R}^2 - \frac{GMm}{R} = \text{constant} \equiv C',$$

where G is the Newtonian gravitational constant and $\dot{R} = dR/dt$. This we can rewrite as

$$\dot{R}^2 = \frac{2GM}{R} + C,$$

with $C = 2C'/m$. Since $R(t) = ra(t)$, where r is constant, and $M = 4\pi R^3\rho/3$, we find

$$r^2\dot{a}^2 = \frac{2G}{ra(t)}\frac{4\pi}{3}\rho a^3(t)r^3 + C,$$

or,

$$\dot{a}^2 = \frac{8\pi G}{3}\rho a^2 + \frac{C}{r^2}$$

Since both C and r are constants, we can define $C/r^2 \equiv -kc^2$, and get

$$\dot{a}^2 + kc^2 = \frac{8\pi G}{3}\rho a^2, \tag{1.23}$$

and if we, although totally unmotivated, postulate that k is the curvature parameter in the RW line element, then equation (1.23) is of the same form as the result of a full treatment in general relativity.

Instead of using energy conservation, we could have started from Newton's second law applied to the test particle:

$$m\ddot{R} = -\frac{GMm}{R^2},$$

which upon inserting $R = ra(t)$ and the expression for M can be rewritten as

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\rho.$$

Again, this is similar to what a relativistic analysis of the problem gives. However, in the correct treatment it turns out that ρ must include all contributions to the energy density, and in addition there is a contribution from the pressure p of the matter of the form $3p/c^2$. Thus, the correct form of the equation is

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\left(\rho + \frac{3p}{c^2}\right). \tag{1.24}$$

These equations are often called the Friedmann equations.

There are several problems with these 'derivations'. We have assumed that space is Euclidean, and then it is not really consistent to interpret k as spatial curvature. Second, in the correct treatment it turns out that ρ is not simply the mass density but also includes the energy density. These important points are missing in the Newtonian approach. Furthermore, the derivation using conservation of energy assumes that the potential energy can be normalized to zero at infinity, and this is not true if the total mass of the universe diverges as $(ar)^3$, as required by a constant density. If we try to rescue the situation by making the density approach zero at large distances, then the universe is no longer homogeneous, and we can no longer argue that we can center our sphere at any point we wish. The difficulty with the second derivation, based on Newton's gravitational force law, is that we assume that the mass outside the spherical shell we consider does not contribute to the gravitational force. The proof for this assumes that the total mass of the system is finite, and hence breaks down for an infinite universe of constant density. For a careful discussion of Newtonian cosmology the reader is referred to a paper by F. J. Tipler (Americal Journal of Physics **64** (1996) 1311).

Deriving the Friedmann equations using the full apparatus of GR is outside the scope of this course. However, based on a paper by J. C. Baez and E. F. Bunn (American Journal of Physics **73** (2005) 644) I can give you a simple, general relativistic derivation if you are prepared to take on faith that Einstein's field equation implies the following result:

> Given a small ball of freely falling test particles initially at rest with respect to each other, the rate at which the ball begins to shrink is proportional to its volume times the following quantity: the energy density at the center of the ball pluss the pressure in the x direction at that point, plus the pressure in the y direction, plus the pressure in the z direction.

Taking the initial instant to be $t = 0$, the mathematical formulation of the statement above is

$$\left(\frac{\ddot{V}}{V}\right)_{t=0} = -4\pi G\left(\rho + \frac{p_x + p_y + p_z}{c^2}\right), \tag{1.25}$$

where V is the volume, ρc^2 is the energy density (including the rest mass contribution) and p_i is the pressure in the i direction.

Let us now look at an observer in a homogeneous and isotropic universe. Suppose that the observer at some instant $t = 0$ identifies a small ball B of test particles centered on his position. The ball is assumed to expand with the universe, but remains spherical since the universe is isotropic. Let $R(t) = a(t)r$ be the radius of this ball as a function of time. Equation (1.25) cannot be applied to the ball directly, because we assume that the sphere is expanding, and equation (1.25) applies to a situation where the test particles are at rest with respect to each other. But we are of course free to introduce a second ball of test particles, B', centered on the observer, where the test particles are at rest with respect to each other at $t = 0$. We denote its radius by $l(t)$. Since the particles are intially at relative rest, we have $\dot{l}(0) = 0$. Furthermore, we are free to choose B' so that it has the same radius as B at $t = 0$, $l(0) = R(0)$. By construction, equation (1.25) applies to B'. Since the volume of the ball is $V = 4\pi l^3/3$, we find

$$\ddot{V} = 4\pi l^2 \ddot{l} + 8\pi l \dot{l}^2,$$

and because $\dot{l}(0) = 0$ we get

$$\ddot{V}(0) = 4\pi l^2(0)\ddot{l}(0),$$

and

$$\left(\frac{\ddot{V}}{V}\right)_{t=0} = \left(\frac{3\ddot{l}}{l}\right)_{t=0}.$$

Since the universe is isotropic, the pressure is equal in all directions, $p_x = p_y = p_z = p$, and so equation (1.25) gives

$$\left(\frac{3\ddot{l}}{l}\right)_{t=0} = -4\pi G\left(\rho + \frac{3p}{c^2}\right). \tag{1.26}$$

At $t = 0$, $l(0) = R(0)$. Furthermore, the second derivatives are the same: $\ddot{l}(0) = \ddot{R}(0)$. This follows from the equivalence principle, which says that, at any given location, particles in free fall do not accelerate with respect to each other. At the moment $t = 0$, each test particle on the surface of ball B is right next to a test particle on the surface of ball B'. Since they are not accelerating with respect to each other, the observer at the origin must see both particles accelerating in the same way, so $\ddot{l}(0) = \ddot{R}(0)$. We can therefore replace l with R, and since $R(t) = a(t)r$, with r constant, we get

$$\left(\frac{3\ddot{a}}{a}\right)_{t=0} = -4\pi G\left(\rho + \frac{3p}{c^2}\right). \tag{1.27}$$

This result is derived for a very small ball. However, in a homegenous universe, the result applies to balls of all sizes since in such a universe balls

of all radii must expand at the same fractinonal rate. Furthermore, there is nothing special about the time $t = 0$, so equation (1.27) will apply at an arbitrary time t. We therefore have

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\left(\rho + \frac{3p}{c^2}\right), \tag{1.28}$$

which is the same as equation (1.24).

We can also go some way towards deriving the first Friedmann equation (1.23) by first establishing a very useful equation describing the evolution of the energy density with the expansion of the universe. This is done by bringing thermodynamics into the picture. Thermodynamics is a universal theory which also applies in the context of GR. Consider the First Law of thermodynamics:

$$TdS = dE + pdV$$

where T is temperature, S is entropy, E is energy and V is volume. Applying this law to the expansion of the Universe, we have $E = \rho c^2 V \propto \rho c^2 a^3$, because the energy density is ρc^2 and the volume is proportional to a^3 since a measures the linear expansion of the homogeneous and isotropic universe. Homogeneity and isotropy also means that ρ and a are functions of time only, so if we insert these expressions on the right-hand side of the First Law, we get

$$\begin{aligned} dE + pdV &\propto d(\rho c^2 a^3) + pd(a^3) \\ &= 3a^2\dot{a}\rho c^2 + a^3\dot{\rho}c^2 + 3pa^2\dot{a} \\ &= a^3c^2\left[\dot{\rho} + 3\frac{\dot{a}}{a}\left(\rho + \frac{p}{c^2}\right)\right]. \end{aligned}$$

The universe expands adiabatically, $dS = 0$. When you think about it, this is not really surprising, since non-adiabaticity would imply that heat flows into or out of a given infinitesimal volume, which would violate homogeneity and isotropy. But from the equation above we must then have

$$\dot{\rho} = -3\frac{\dot{a}}{a}\left(\rho + \frac{p}{c^2}\right). \tag{1.29}$$

This is a very useful and important equation which will allow us to determine how the energy density of the universe evolves with the expansion. But first of all, let us use it to express the pressure in terms of the energy density and its time derivative:

$$\frac{p}{c^2} = -\frac{a}{3\dot{a}}\dot{\rho} - \rho.$$

Using this relation to eliminate the pressure term from the second Friedmann equation (1.24) we find

$$\ddot{a} = \frac{8\pi G}{3}\rho a + \frac{4\pi G}{3}\frac{a^2}{\dot{a}}\dot{\rho},$$

and multiplying through by $\dot{a}$ we get

$$\dot{a}\ddot{a} = \frac{8\pi G}{3}\rho a\dot{a} + \frac{4\pi G}{3}\dot{\rho}a^2,$$

and we see that both sides of the equation can be expressed as total derivatives:

$$\frac{1}{2}\frac{d}{dt}(\dot{a})^2 = \frac{4\pi G}{3}\frac{d}{dt}(\rho a^2).$$

and so

$$\dot{a}^2 = \frac{8\pi G}{3}\rho a^2 + \text{constant}.$$

This is how far we can go with rigor. We cannot easily relate the constant of integration to the curvature parameter appearing in the RW metric in this approach, but if we postulate that it is equal to $-kc^2$, we see that we get

$$\dot{a}^2 + kc^2 = \frac{8\pi G}{3}\rho a^2, \tag{1.30}$$

which is identical to equation (1.23).

Note that we derived equation (1.23) using equations (1.24) and (1.29). This means that these three equations are not all independent, any two of them taken together will be sufficient to describe the kinematics of the expanding universe.

1.6.1 Time to memorize!

We have now collected some of the most important equations in cosmology. This is therefore a good place for me to summarize them and for you to memorize them. Here they are:

- The Robertson-Walker line element:

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

- Redshift

$$1 + z = \frac{a(t_\text{o})}{a(t_\text{e})}.$$

- The first Friedmann equation:

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

- The second Friedmann equation:

$$\ddot{a} = -\frac{4\pi G}{3}\left(\rho + \frac{3p}{c^2}\right)a.$$

- Adiabatic expansion:

$$\dot{\rho} = -3\frac{\dot{a}}{a}\left(\rho + \frac{p}{c^2}\right).$$

1.7 Equations of state

The Friedmann equations seem to involve four unknowns: the scale factor a, the spatial curvature parameter k, the matter/energy density ρ, and the pressure p. Since only two of the Friedmann equations are independent, we have only two equations for four unknowns. A little thinking shows, however, that the spatial curvature parameter is not a big problem. From equation (1.23) we can write

$$kc^2 = \frac{8\pi G}{3}\rho(t)a^2(t) - \dot{a}^2(t),$$

where I display the time argument explicitly. Now, in solving the differential equations we must always supply some boundary or initial conditions on the solutions. We are free to choose when to impose these boundary conditions, and the most convenient choice is to use the present time, which we will denote by t_0. The present value of the Hubble parameter is given by $H_0 = H(t_0) = \dot{a}(t_0)/a(t_0)$, and if we furthermore define $\rho(t_0) \equiv \rho_0$, we can therefore write

$$\frac{kc^2}{a_0^2} = \frac{8\pi G}{3}\rho_0 - H_0^2.$$

We thus see that if we specify initial conditions by choosing values for H_0 and ρ_0, e.g. by using measurements of them, then the spatial curvature is determined for all times. Thus, this is not a problem.

However, there still remains three unknown functions $a(t)$, $\rho(t)$, and $p(t)$, and we have only two independent equations for them. Clearly, we need one more equation to close the system. The common way of doing this is by specifying an *equation of state*, that is, a relation between pressure p and matter/energy density ρ. The most important cases for cosmology can fortunately be described by the simplest equation of state imaginable:

$$p = w\rho c^2 \tag{1.31}$$

where w is a constant. We will introduce two important cases here and a third case (the cosmological constant) in section 1.9.

1.7.1 Dust: non-relativistic matter

The matter in the universe (e.g. the matter in galaxies) is mostly moving at non-relativistic speeds. Non-relativistic matter in the context of cosmology is often called *dust*, and we will use this term in the following. From thermodynamics we know that the equation of state of an ideal gas of N non-relativistic particles of mass m at temperature T in a volume V at low densities is

$$p = \frac{Nk_{\mathrm{B}}T}{V},$$

where $k_{\rm B}$ is Boltzmann's constant. We rewrite this slightly:

$$p = \frac{Nmc^2}{Vmc^2} k_{\rm B} T = \frac{k_{\rm B} T}{mc^2} \rho c^2,$$

where $\rho = Nm/V$ is the mass density of the gas. Now, we also recall that for an ideal gas the mean-square speed of the particles is related to the temperature as

$$m\langle v^2 \rangle = 3 k_{\rm B} T,$$

and hence

$$p = \frac{\langle v^2 \rangle}{3c^2} \rho c^2.$$

Thus, we see that $w = \langle v^2 \rangle / 3c^2$ for this gas. However, since the particles are non-relativistic we have $v \ll c$, and it is an excellent approximation to take $w \approx 0$ for non-relativistic particles. We will therefore in the following assume that a dust-filled universe has equation of state

$$p = 0, \tag{1.32}$$

that is, dust is pressureless.

1.7.2 Radiation: relativistic matter

For a gas of massless particles, for example photons, the equation of state is also simple:

$$p = \frac{1}{3} \rho c^2, \tag{1.33}$$

and hence $w = 1/3$ in this case. You have probably seen this already in thermodynamics in the discussion of blackbody radiation.

Why do we need to think about radiation? As you may know, the universe is filled with a relic radiation, the cosmic microwave background, with a temperature of around 3 K. Although it gives a negligible contribution to the present energy density of the universe, we will see that it was actually the dominant component in the distant past, and thus we need to take it into consideration when we discuss the early universe. There is also a background radiation of neutrinos. Neutrinos were long considered to be massless, but we now know that at least one of the three types of neutrino has a small mass. However, they are so light that it is an excellent approximation to treat neutrinos as massless in the early universe, and hence they obey the equation of state (1.33).

1.8 The evolution of the energy density

Equipped with the equation of state, we can now proceed to solve equation (1.29) to obtain ρ as a function of the scale factor a. Having done this,

we can then proceed to rewrite equations (1.23) and (1.24) as differential equations for a only.

We start from the general equation of state $p = w\rho c^2$, where w is a constant. Inserting this into equation (1.29) gives

$$\dot{\rho} = -3\frac{\dot{a}}{a}\left(\rho + \frac{w\rho c^2}{c^2}\right) = -3\frac{\dot{a}}{a}(1+w)\rho.$$

Now, recall that $\dot{\rho} = d\rho/dt$ and $\dot{a} = da/dt$, so that we can rewrite this as the differential equation

$$\frac{d\rho}{dt} = -3(1+w)\frac{\rho}{a}\frac{da}{dt},$$

or

$$\frac{d\rho}{\rho} = -3(1+w)\frac{da}{a}.$$

This equation is easily integrated. Since we have agreed to specify boundary conditions at the present time t_0, and chosen $\rho(t_0) = \rho_0$ and $a(t_0) = a_0$, we find

$$\int_{\rho_0}^{\rho} \frac{d\rho'}{\rho'} = -3(1+w)\int_{a_0}^{a} \frac{da'}{a'},$$

which gives

$$\ln\left(\frac{\rho}{\rho_0}\right) = -3(1+w)\ln\left(\frac{a}{a_0}\right),$$

or

$$\rho = \rho_0\left(\frac{a_0}{a}\right)^{3(1+w)}. \tag{1.34}$$

For the case of dust, $w = 0$, this gives

$$\rho = \rho_0\left(\frac{a_0}{a}\right)^{3}, \tag{1.35}$$

which is easy to understand: since the energy density is proportional to the matter density and no matter disappears, the density decreases inversely proportional to the volume, which in turn is proportional to a^3.

For radiation, $w = 1/3$, we find

$$\rho = \rho_0\left(\frac{a_0}{a}\right)^{4}, \tag{1.36}$$

which also has a simple physical interpretation: again there is a factor of $1/a^3$ from the fact that the energy density decreases with the volume, but in addition, since the energy of relativistic particles is inversely proportional to their wavelenghts, which increase in proportion to a, there is an additional factor of $1/a$.

1.9 The cosmological constant

When Einstein had formulated his theory of general relativity, he rapidly recognized the possibility of applying it to the Universe as a whole. He made the simplest assumptions possible consistent with the knowledge at his time: a *static*, homogeneous and isotropic universe, filled with dust. Remember that Einstein did this in 1916, and at that time it was not even clear that galaxies outside our own Milky Way existed, let alone universal expansion! Following in Einstein's footsteps we look for static solutions of equations (1.23,1.24) with $p = 0$. Then:

$$\begin{aligned} \dot{a}^2 + kc^2 &= \frac{8\pi G}{3}\rho a^2 \\ \ddot{a} &= -\frac{4\pi G}{3}\rho a \end{aligned}$$

If the universe is static, then $a(t) = a_0 = \text{constant}$, and $\dot{a} = \ddot{a} = 0$. From the second equation this gives $a = a_0 = 0$ or $\rho = 0$. The first case corresponds to having no universe, and the second possibility is an empty universe. Inserting this in the first equation gives $kc^2 = 0$, hence $k = 0$. So, a static, dust-filled universe must necessarily be empty or of zero size. Both options are in violent disagreement with our existence.

Faced with this dilemma, Einstein could in principle have made the bold step and concluded that since no static solution is possible, the universe must be expanding. However, one should bear in mind that when he made his first cosmological calculations, all observations indicated that the universe is static. Furthermore, there was a strong philosophical bias towards an eternal, static universe since one then did not need to explain how the universe came into existence in the first place. Therefore, Einstein chose to modify his theory so as to allow static solutions. How can this be done? The key lies in the so-called cosmological constant. When Einstein wrote down his field equations, he assumed that they had the simplest form possible. However, it turns out that they can be modified slightly by adding a constant which, in Einstein's way of thinking, corresponds to assigning a curvature to empty space. In fact, there is no a priori reason why this term should be equal to zero. When this so-called cosmological constant term is added, the Friedmann equations turn out to be (for pressureless matter):

$$\dot{a}^2 + kc^2 = \frac{8\pi G}{3}\rho a^2 + \frac{\Lambda}{3}a^2 \tag{1.37}$$

$$\ddot{a} = -\frac{4\pi G}{3}\rho a + \frac{\Lambda}{3}a, \tag{1.38}$$

where Λ is the cosmological constant. Now, a static solution is possible. Take $a = a_0 = \text{constant}$. Then, equation (1.38) gives

$$\Lambda = 4\pi G\rho_0,$$

and inserting this in equation (1.37) we get

$$kc^2 = \frac{8\pi G}{3}\rho_0 a_0^2 + \frac{4\pi G}{3}\rho_0 a_0^2 = 4\pi G\rho_0 a_0^2.$$

Since the right-hand side is positive, we must have $k = +1$. The static universe is therefore closed with the scale factor (which in this case gives the radius of curvature) given by

$$a_0 = \frac{c}{\sqrt{4\pi G\rho_o}} = \frac{c}{\sqrt{\Lambda}}.$$

This model is called the Einstein universe. Einstein himself was never pleased with the fact that he had to introduce the cosmological constant. And it is worth noting that even though the model is static, it is unstable: if perturbed away from the equilibrium radius, the universe will either expand to infinity or collapse. If we increase a from a_0, then the Λ-term will dominate the equations, causing a runaway expansion, whereas if we decrease a from a_0, the dust term will dominate, causing collapse. Therefore, this model is also physically unsound, and this is a far worse problem than the (to Einstein) unattractive presence of Λ.

As I said, Einstein originally introduced the cosmological constant as a contribution to the curvature of spacetime. Throughout the years our understanding of the cosmological constant has led us to consider it instead as a contribution to the energy density and pressure of the universe, since it has turned out to be intimately linked with the energy density of empty space: the vacuum energy. As a consequence of Heisenberg's uncertainty principle, empty space is not empty but has an associated energy density set up by quantum mechanical processes. From this viewpoint we should write the Friedmann equation with dust and cosmological constant as

$$\begin{aligned} \dot{a}^2 + kc^2 &= \frac{8\pi G}{3}(\rho + \rho_\Lambda)a^2 \\ \ddot{a} &= -\frac{4\pi G}{3}\left(\rho + \rho_\Lambda + \frac{3p_\Lambda}{c^2}\right)a, \end{aligned}$$

and if we compare the first equation with (1.37) we see that

$$\rho_\Lambda = \frac{\Lambda}{8\pi G}. \tag{1.39}$$

Inserting this in the second equation and comparing with equation (1.38) we find

$$-\frac{4\pi G}{3}\left(\frac{\Lambda}{8\pi G} + \frac{3p_\Lambda}{c^2}\right) = \frac{\Lambda}{3},$$

which gives

$$p_\Lambda = -\frac{\Lambda}{8\pi G}c^2 = -\rho_\Lambda c^2. \tag{1.40}$$

Notice that $p = -\rho_\Lambda c^2$. This means $w = -1$, and that for $\Lambda > 0$, the pressure is negative! If we consider how the energy density associated with the cosmological constant evolves with time, we can insert this equation of state in equation (1.29). This gives

$$\dot{\rho}_\Lambda = -3\frac{\dot{a}}{a}(\rho_\Lambda - \rho_\Lambda) = 0,$$

so that ρ_Λ = constant = $\Lambda/8\pi G$. The vacuum energy density remains constant as space expands! The concept of negative pressure may seem odd, but such things do occur elsewhere in nature. The pressure in e.g. an ideal gas is positive because we have to do work to compress it. Negative pressure corresponds to the opposite situation when we have to supply work in order to make the system expand. A situation like that occurs with a stretched string: we have to do work in order to stretch if further. It can thus be considered a 'negative pressure' system.

If we insist on a Newtonian interpretation in terms of gravitational forces instead of spacetime geometry, then the cosmological constant is seen to give rise to a repulsive contribution to the gravitational force. This is, of course, necessary in order to have a static universe, since a homogeneous matter distribution starting at rest will collapse. Once Hubble discovered the expansion of the Universe in 1929, the cosmological constant rapidly dropped out of fashion since *expanding* solutions were possible without it. However, it has come back into fashion from time to time, and there is really no compelling theoretical reason to drop it besides simplicity and beauty. Since it can be associated with the vacuum energy, and no one yet knows how to calculate that consistently, the most honest thing to do is to keep Λ in the equations and try to constrain it with observations. In fact, observations made over the last few years have shown that not only is the cosmological constant present, it actually dominates the dynamics of our universe! We will therefore study both models with and without a cosmological constant.

1.10 Some classic cosmological models

We will now make a brief survey of the simplest cosmological models. As a prelude, we consider equation (1.23) rewritten as

$$\left(\frac{\dot{a}}{a}\right)^2 + \frac{kc^2}{a^2} = \frac{8\pi G}{3}\rho.$$

This equation is valid at all times, and so it must also apply at the present time t_0. Since $\dot{a}(t_0)/a(t_0) = H_0$, the present value of the Hubble parameter, we have

$$1 + \frac{kc^2}{a_0^2 H_0^2} = \frac{8\pi G}{3H_0^2}\rho_0.$$

We see that the combination $3H_0^2/8\pi G$ must have the units of a density. It is called the present value of the *critical density*, and denoted by ρ_{c0}. Inserting values for the constants, we have

$$\rho_{c0} = 1.879 \times 10^{-29} h^2 \, \mathrm{g\,cm^{-3}}.$$

Its importance derives from the fact that for a spatially flat universe, $k = 0$, we see from the equation above that

$$1 = \frac{8\pi G}{3H_0^2}\rho_0 = \frac{\rho_0}{\rho_{c0}},$$

so that for a spatially flat universe, the density equals the critical density. It is common to measure densities in units of the critical density and define

$$\Omega_0 \equiv \frac{\rho_0}{\rho_{c0}}. \tag{1.41}$$

Furthermore, one also introduces a 'curvature density parameter',

$$\Omega_{k0} = -\frac{kc^2}{a_0^2 H_0^2}, \tag{1.42}$$

and hence we can write

$$\Omega_0 + \Omega_{k0} = 1. \tag{1.43}$$

1.10.1 Spatially flat, dust- or radiation-only models

Let us consider the simplest case first: a flat universe ($k = 0$) filled with dust ($w = 0$) or radiation ($w = 1/3$), and with a vanishing cosmological constant ($\Lambda = 0$). In this case the Friedmann equations become

$$\begin{aligned} \left(\frac{\dot{a}}{a}\right)^2 &= \frac{8\pi G}{3}\rho_0 \left(\frac{a}{a_0}\right)^{-3(1+w)} \\ \frac{\ddot{a}}{a} &= -\frac{4\pi G}{3}(1+3w)\rho_0 \left(\frac{a}{a_0}\right)^{-3(1+w)}. \end{aligned}$$

Taking the square root of the first equation, we see that it allows both positive and negative $\dot{a}$. However, we know that the universe is expanding now, so we will consider $\dot{a}/a > 0$. The second equation implies that $\ddot{a} < 0$ for $w > -1/3$ which is what we assume in the present discussion. Thus, the second derivative of the scale factor is always negative. Since we know that its first derivative is positive now, this must mean that the scale factor within these models must have been vanishing at some time in the past. This is useful to know when we want to normalize our solution. Let us start with the first equation:

$$\left(\frac{\dot{a}}{a}\right)^2 = H_0^2 \frac{8\pi G}{3H_0^2}\rho_0 \left(\frac{a}{a_0}\right)^{-3(1+w)},$$

where we see that the first factor on the right-hand side is $1/\rho_{c0}$, and since $k = 0$, we have $\rho_0/\rho_{c0} = 1$. Taking the square root of the equation, we therefore have

$$\frac{\dot{a}}{a} = H_0 \left(\frac{a}{a_0}\right)^{-3(1+w)/2},$$

which we rearrange to

$$a_0^{-3(1+w)/2} a^{\frac{1}{2}+\frac{3}{2}w} da = H_0 dt,$$

which means that

$$a_0^{-3(1+w)/2} \int_{a_0}^{a} a'^{\frac{1}{2}+\frac{3}{2}w} da' = \int_{t_0}^{t} H_0 dt',$$

or

$$\frac{2}{3(1+w)} \left(\frac{a}{a_0}\right)^{\frac{3}{2}(1+w)} - \frac{2}{3(1+w)} = H_0(t - t_0).$$

As it stands, this equation is perfectly fine and can be solved for a as a function of t. However, we can simplify it further by using the fact noted earlier that the scale factor must have been equal to zero at some time $t < t_0$. We see that the solution for a will depend on $t - t_0$ only, so we are free to choose the time where the scale factor vanished to be $t = 0$. Imposing $a = 0$ at $t = 0$, we get

$$\frac{2}{3(1+w)} = H_0 t_0,$$

and we can therefore write

$$H_0 t_0 \left(\frac{a}{a_0}\right)^{\frac{3}{2}(1+w)} = H_0 t,$$

which gives

$$a(t) = a_0 \left(\frac{t}{t_0}\right)^{\frac{2}{3(1+w)}}, \tag{1.44}$$

with

$$t_0 = \frac{2}{3(1+w)H_0}. \tag{1.45}$$

We see that the universe expands according to a power law, and that t_0 denotes the current age of the universe (more precisely: the expansion age), since it is the time elapsed from $t = 0$ to the present time t_0. Note that at everything breaks down at $t = 0$: since $a = 0$ there, the density, scaling as a negative power of a, is formally infinite, so we have a zero-size universe with infinite density. Our theory cannot describe such a singular state, so we must regard our extension of our model to $t = 0$ as purely mathematical. As the energy density skyrockets, we must take into account that new physical effects which current theories cannot describe, like for example quantum

gravity, must enter the stage and modify the picture in a way we can only guess at in our present state of knowledge.

Note that the expansion age t_0 is less than $1/H_0$, the value it would have if the universe were expanding at the same rate all the time. Since $\ddot{a} < 0$, the universe is constantly decelerating. We have fixed the scale factor to unity at the present time t_0, and furthermore we have fixed the present expansion rate to be H_0. This explains why the age of the universe in this model is lower than in the case of expansion at a constant rate: since the universe is constantly decelerating, in order to expand at a given rate H_0 now, it must have been decelerating for a shorter time.

We know that the Universe is not radiation-dominated now, but in its early stages it was, and so the radiation-dominated model is of interest. For $w = 1/3$, we get

$$\begin{aligned} a(t) &= a_0 \left(\frac{t}{t_0}\right)^{\frac{1}{2}} \qquad (1.46) \\ t_0 &= \frac{1}{2H_0}. \end{aligned}$$

The case of a dust-filled, flat universe is called the Einstein-de Sitter (EdS) model and was long a favourite among cosmologists. In this case $w = 0$ and we find

$$\begin{aligned} a(t) &= a_0 \left(\frac{t}{t_0}\right)^{\frac{2}{3}} \qquad (1.47) \\ t_0 &= \frac{2}{3H_0}. \qquad (1.48) \end{aligned}$$

If we use $H_0 = 100h\ \mathrm{km\,s^{-1}\,Mpc^{-1}}$, and the current best value $h \approx 0.7$, we find that

$$t_0 = 9.3 \times 10^9\ \mathrm{yrs}.$$

This is a problem for this model, since e.g. the ages of stars in old globular clusters indicate that the universe must be at least 12 billion years old. However, as far as we know the universe was dominated by dust until 'recently', so that this model is still a useful description of a large part of the history of the universe. Also, because of its simplicity, one can calculate a lot of quantities analytically in this model, and this makes it a valuable pedagogical tool.

1.10.2 Spatially flat, empty universe with a cosmological constant

Let us go back to the Friedmann equations and look at the case where there is no matter or radiation, but the universe is made spatially flat by a

cosmological constant Λ. In this case we have

$$\rho = \rho_\Lambda = \frac{\Lambda}{8\pi G} = \text{constant},$$

and the Friedmann equations for $k = 0$ become

$$\begin{aligned} \dot{a}^2 &= \frac{\Lambda}{3}a^2 \\ \ddot{a} &= \frac{\Lambda}{3}a \end{aligned}$$

From the first equation we see that

$$\frac{\dot{a}}{a} = \pm\sqrt{\frac{\Lambda}{3}} = \text{constant},$$

and since $H(t) = \dot{a}/a$ and we seek a solution which is expanding at the rate $H_0 > 0$ at the present time t_0, we have $\sqrt{\Lambda/3} = H_0$. We easily see that the equation

$$\frac{\dot{a}}{a} = H_0$$

has $a(t) = Ae^{H_0 t}$ as general solution, where A is a constant. We also see that this solution satisfies the second Friedmann equation. Furthermore, $a(t_0) = a_0$ gives $A = a_0 e^{-H_0 t_0}$, and hence

$$a(t) = a_0 e^{H_0(t-t_0)}.$$

We notice two peculiar features of this solution. First of all, it describes a universe expanding at an accelerating rate, since $\ddot{a} > 0$, in contrast to the dust- and radiation-filled universes of the previous subsection which were always decelerating. This is because of the negative pressure of the vacuum energy density (recall that $p_\Lambda = -\rho_\Lambda c^2$). Secondly, note that there is no singularity in this case: there is nothing particular happening at $t = 0$, and in fact the scale factor is finite and well-behaved at any finite time in the past and in the future. Since this is a model of a universe with no matter or radiation in it, it obviously does not correspond to the one we live in. However, observations suggest very strongly that at the present epoch in the history of the universe, the cosmological constant gives the largest contribution to the energy density, and makes the universe expand at an accelerating rate. As matter and radiation are diluted away by the expansion, our universe will approach the model considered in this subsection asymptotically.

The model we have found is called the de Sitter model, after the Dutch astronomer Willem de Sitter who first discovered it. He found this solution shortly after Einstein had derived his static universe model in 1917, and interestingly, de Sitter actually thought he had discovered another static

solution of Einstein's equations! By a transformation of the coordinates r and t to new coordinates r' and t' one can actually bring the line element to the static form

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

where $1/R^2 = \Lambda/3$. It attracted some interest after the discovery of the galaxy redshifts, since even from this form of the line element one can show that light will be redshifted when travelling along geodesics in this universe. Even though this model describes a universe completely void of matter, it was thought that the matter density might be low enough for the de Sitter line element to be a good approximation to the present universe. Note, however, that the new time coordinate does not have the same significance as the cosmic time t. It was not until the work of Robertson[2] on the geometry of homogeneous and isotropic universe models that the expanding nature of the de Sitter solution was clarified.

1.10.3 Open and closed dust models with no cosmological constant

We next turn to another class of models where analytic solutions for the scale factor a can be obtained: models with dust (non-relativistic matter, $p = 0$) and curvature. In terms of the density parameter $\Omega_{\rm m0}$ for matter, recalling that $\Omega_{\rm m0} + \Omega_{\rm k0} = 1$, we can write the Friedmann equation for $\dot{a}^2$ as

$$\frac{H^2(t)}{H_0^2} = \Omega_{\rm m0}\left(\frac{a_0}{a}\right)^3 + (1 - \Omega_{\rm m0})\left(\frac{a_0}{a}\right)^2,$$

where $H(t) = \dot{a}/a$. We now have to distinguish between two cases, corresponding to models which expand forever and models which cease to expand at some point and then start to contract. If a model stops expanding, this must mean that $\dot{a} = 0$ for some finite value of a, and hence $H = 0$ at that point. This gives the condition

$$\Omega_{\rm m0}\left(\frac{a_0}{a}\right)^3 + (1 - \Omega_{\rm m0})\left(\frac{a_0}{a}\right)^2 = 0.$$

The first term in this equation is always positive, and so for this equation to be fulfilled the second term must be negative, corresponding to

$$\Omega_{\rm m0} > 1.$$

This again gives $\Omega_{\rm k0} = -kc^2/(a_0 H_0)^2 < 0$, and therefore $k = +1$. It is, of course, possible that the model will continue to expand after this, but if we

[2] H. P. Robertson, 'On the Foundations of Relativistic Cosmology', Proceedings of the National Academy of Science, **15**, 822-829, 1929

consider the Friedmann equation for $\ddot{a}$, we see that in this case $\ddot{a} < 0$ always, which means that the universe will start to contract. Thus we have obtained the interesting result that dust universes with positive curvature (closed dust models) will stop expanding at some point and begin to contract, ultimately ending in a Big Crunch. Models with dust and negative spatial curvature (open dust models), on the other hand, will continue to expand forever since $H \neq 0$ always in that case. This close connection between the matter content and the ultimate fate of the universe is peculiar to dust models. We will later see that the addition of a cosmological constant spoils this nice correspondence completely.

In the closed case the scale factor a has a maximum value $a_{\max}$ given by

$$\Omega_{\mathrm{m}0}\left(\frac{a_0}{a_{\max}}\right)^3 = (\Omega_{\mathrm{m}0} - 1)\left(\frac{a_0}{a_{\max}}\right)^2,$$

and so

$$a_{\max} = a_0 \frac{\Omega_{\mathrm{m}0}}{\Omega_{\mathrm{m}0} - 1}.$$

Recall that we have defined the present value of the scale factor $a(t_0) = a_0$, so this means that, for example, if the density parameter is $\Omega_{\mathrm{m}0} = 2$, the universe will expand to a maximum linear size of twice its present size. Note also that H enters the equations only as H^2, which means that the contraction phase $H < 0$ will proceed exactly as the expansion phase.

Now for the solution of the Friedmann equation. We start with the closed case and note that we can write the equation for H^2 above as

$$\frac{1}{H_0}\frac{da}{dt} = a_0\sqrt{\Omega_{\mathrm{m}0}\frac{a_0}{a} - (\Omega_{\mathrm{m}0} - 1)},$$

or

$$H_0 dt = \frac{da/a_0}{\sqrt{\Omega_{\mathrm{m}0}\frac{a_0}{a} - (\Omega_{\mathrm{m}0} - 1)}}.$$

The simple substitution $x = a/a_0$ simplifies this equation to

$$H_0 dt = \frac{dx}{\sqrt{\frac{\Omega_{\mathrm{m}0}}{x} + (\Omega_{\mathrm{m}0} - 1)}}.$$

Since we start out with $\dot{a} > 0$ and $\ddot{a} < 0$ always, there must have been some point in the past where $a = 0$. We choose this point to be the zero for our cosmic time variable t. Then we can integrate both sides of this equation and find

$$\begin{aligned} H_0 t &= \int_0^{a/a_0} \frac{\sqrt{x}dx}{\sqrt{\Omega_{\mathrm{m}0} - (\Omega_{\mathrm{m}0} - 1)x}} \\ &= \frac{1}{\sqrt{\Omega_{\mathrm{m}0} - 1}} \int_0^{a/a_0} \frac{\sqrt{x}dx}{\sqrt{\alpha - x}}, \end{aligned}$$

where we have defined $\alpha = \frac{\Omega_{\rm m0}}{\Omega_{\rm m0}-1}$. We now introduce a change of variables:

$$x = \alpha \sin^2 \frac{\psi}{2} = \frac{1}{2}\alpha(1-\cos\psi),$$

which gives $dx = \alpha \sin(\psi/2)\cos(\psi/2)d\psi$ and $\sqrt{\alpha - x} = \sqrt{\alpha}\cos(\psi/2)$. Then the integral can be carried out easily:

$$\begin{aligned} H_0 t &= \frac{\alpha}{\sqrt{\Omega_{\rm m0}-1}} \int_0^\psi \sin^2\frac{\psi}{2} d\psi \\ &= \frac{\Omega_{\rm m0}}{(\Omega_{\rm m0}-1)^{3/2}} \frac{1}{2} \int_0^\psi (1-\cos\psi) d\psi \\ &= \frac{1}{2}\frac{\Omega_{\rm m0}}{(\Omega_{\rm m0}-1)^{3/2}}(\psi - \sin\psi). \end{aligned}$$

Thus we have obtained a parametric solution of the Friedmann equation:

$$a(\psi) = \frac{a_0}{2}\frac{\Omega_{\rm m0}}{\Omega_{\rm m0}-1}(1-\cos\psi) \tag{1.49}$$

$$t(\psi) = \frac{1}{2H_0}\frac{\Omega_{\rm m0}}{(\Omega_{\rm m0}-1)^{3/2}}(\psi - \sin\psi), \tag{1.50}$$

where the parameter ψ varies from 0 to 2π, and the scale factor varies from 0 at $\psi = 0$ to the maximum value $a_{\rm max}$ at $\psi = \pi$, and back to zero for $\psi = 2\pi$. It is easy to show that the age of the universe in this model is given by

$$t_0 = \frac{1}{2H_0}\frac{\Omega_{\rm m0}}{(\Omega_{\rm m0}-1)^{3/2}}\left[\cos^{-1}\left(\frac{2}{\Omega_{\rm m0}}-1\right) - \frac{2}{\Omega_{\rm m0}}\sqrt{\Omega_{\rm m0}-1}\right], \tag{1.51}$$

and that the lifetime of the universe is

$$t_{\rm crunch} = t(2\pi) = \frac{\pi\Omega_{\rm m0}}{H_0(\Omega_{\rm m0}-1)^{3/2}}. \tag{1.52}$$

The solution in the open case ($\Omega_{\rm m0} < 1$) proceeds along similar lines. In this case we can manipulate the Friedmann equation for $\dot{a}$ into the form

$$H_0 t = \frac{1}{\sqrt{1-\Omega_{\rm m0}}}\int_0^{a/a_0} \frac{\sqrt{x}dx}{\sqrt{x+\beta}},$$

where $\beta = \Omega_{\rm m0}/(1-\Omega_{\rm m0})$, and then substitute

$$x = \frac{1}{2}\beta(\cosh u - 1) = \beta \sinh^2\frac{u}{2}.$$

Using standard identities for hyperbolic functions the integral can be carried out with the result

$$H_0 t = \frac{\Omega_{\rm m0}}{2(1-\Omega_{\rm m0})^{3/2}}(\sinh u - u),$$

and thus we have the parametric solution

$$a(u) = \frac{a_0}{2}\frac{\Omega_{\rm m0}}{1-\Omega_{\rm m0}}(\cosh u - 1) \quad (1.53)$$

$$t(u) = \frac{\Omega_{\rm m0}}{2H_0(1-\Omega_{\rm m0})^{3/2}}(\sinh u - u), \quad (1.54)$$

where the parameter u varies from 0 to ∞. This model is always expanding, and hence there is no Big Crunch here. The present age of the universe is found to be

$$t_0 = \frac{1}{2H_0}\frac{\Omega_{\rm m0}}{(1-\Omega_{\rm m0})^{3/2}}\left[\frac{2}{\Omega_{\rm m0}}\sqrt{1-\Omega_{\rm m0}} - \cosh^{-1}\left(\frac{2}{\Omega_{\rm m0}} - 1\right)\right]. \quad (1.55)$$

1.10.4 Models with more than one component

We will frequently consider models where more than one component contributes to the energy density of the universe. For example, in the next subsection we will consider a model with matter and radiation. Let us look at the general situation where we have several contributions ρ_i and $p_i = p_i(\rho_i)$ to the energy density and pressure, so that, e.g., the first Friedmann equation becomes

$$H^2 = \frac{8\pi G}{3}\rho = \frac{8\pi G}{3}\sum_i \rho_i.$$

The evolution of ρ is found by solving

$$\dot{\rho} = -3H\left(\rho + \frac{p}{c^2}\right),$$

but this equation can now be written as

$$\sum_i \dot{\rho}_i = -3H\sum_i\left(\rho_i + \frac{p_i}{c^2}\right),$$

or

$$\sum_i [\dot{\rho}_i + 3H\left(\rho_i + \frac{p_i}{c^2}\right)] = 0.$$

As long as $p_i = p_i(\rho_i)$ and does not depend on any of the other contributions to the energy density, the terms in the sum on the left-hand side of the equation are in general independent, and the only way to guarantee that the sum vanishes is for the individual terms to be equal to zero, i.e.,

$$\dot{\rho}_i + 3H\left(\rho_i + \frac{p_i}{c^2}\right) = 0.$$

We have thus shown that when we consider models with more than one component, we can solve for the evolution of the energy density with the scale factor for each component separately, and then plug the results into the Friedmann equations.

1.10.5 Models with matter and radiation

Two components we are quite certain exist in our universe are radiation and matter. To our present best knowledge, the density parameters for these two components are $\Omega_{\rm r0} \approx 8.4 \times 10^{-5}$ and $\Omega_{\rm m0} \approx 0.3$. Since the densities vary as

$$\begin{aligned} \rho_{\rm m} &= \rho_{\rm c0}\Omega_{\rm m0}\left(\frac{a_0}{a}\right)^3 \\ \rho_r &= \rho_{\rm c0}\Omega_{\rm r0}\left(\frac{a_0}{a}\right)^4, \end{aligned}$$

we see that there is a value of a for which the energy densities in the two components are equal. At this value, $a_{\rm eq}$, we have

$$\rho_{\rm c0}\Omega_{\rm m0}\left(\frac{a_0}{a}\right)^3 = \rho_{\rm c0}\Omega_{\rm r0}\left(\frac{a_0}{a}\right)^4,$$

which gives

$$a_{\rm eq} = a_0\frac{\Omega_{\rm r0}}{\Omega_{\rm m0}},$$

or in terms of redshift $1 + z_{\rm eq} = a_0/a_{\rm eq} = \Omega_{\rm m0}/\Omega_{\rm r0} \approx 3570$. We see that $a_{\rm eq} \ll a_0$, so that this corresponds to an early epoch in the history of the universe. For $a < a_{\rm eq}$ radiation dominates the energy density of the universe, whereas for $a > a_{\rm eq}$ the universe is matter dominated. Thus, the early universe was radiation dominated. I will refer to $z_{\rm eq}$ as the redshift of matter-radiation equality.

The Friedmann equation for a universe with matter, radiation, and spatial curvature can be written as

$$\frac{H^2(t)}{H_0^2} = \Omega_{\rm m0}\left(\frac{a_0}{a}\right)^3 + \Omega_{\rm r0}\left(\frac{a_0}{a}\right)^4 + \Omega_{\rm k0}\left(\frac{a_0}{a}\right)^2.$$

How important is the curvature term? Since it drops off with a as a^2 whereas the matter and radiation terms fall as a^3 and a^4 respectively, we would expect the curvature term to be negligible for sufficiently small values of a. Let us see what this means in practice. The curvature term is negligible compared to the matter term if $\Omega_{\rm k0}a_0^2/a^2 \ll \Omega_{\rm m0}a_0^3/a^3$. This gives the condition

$$\frac{a}{a_0} \ll \frac{\Omega_{\rm m0}}{\Omega_{\rm k0}}.$$

To the best of our knowledge, $\Omega_{\rm k0}$ is small, perhaps less than 0.02. In this case, with $\Omega_{\rm m0} = 0.3$, we get

$$\frac{a}{a_0} \ll 15$$

as the condition for neglecting curvature. This result means that the curvature term will only be important in the distant future. But note that this

argument only applies to the expansion rate. Curvature can still be important when we calculate geometrical quantities like distances, even though it plays a negligible role for the expansion rate.

The condition for neglecting curvature term compared to the radiation term is easily shown to be

$$\frac{a}{a_0} \ll \sqrt{\frac{\Omega_{\rm r0}}{\Omega_{\rm k0}}} = \sqrt{\frac{\Omega_{\rm m0}}{\Omega_{\rm k0}}\frac{\Omega_{\rm r0}}{\Omega_{\rm m0}}} \sim 4\sqrt{\frac{a_{\rm eq}}{a_0}} \approx 0.07.$$

In combination, this means that we can ignore the curvature term in the radiation-dominated phase, and well into the matter-dominated phase. This simplifies the Friedmann equation to

$$\frac{H^2(t)}{H_0^2} = \Omega_{\rm m0}\left(\frac{a_0}{a}\right)^3 + \Omega_{\rm r0}\left(\frac{a_0}{a}\right)^4,$$

which can be rewritten as

$$H_0 dt = \frac{ada}{a_0^2\sqrt{\Omega_{\rm r0}}}\left(1+\frac{a}{a_{\rm eq}}\right)^{-1/2}.$$

Carrying out the integration is left as an exercise. The result is

$$H_0 t = \frac{4(a_{\rm eq}/a_0)^2}{3\sqrt{\Omega_{\rm r0}}}\left[1-\left(1-\frac{a}{2a_{\rm eq}}\right)\left(1+\frac{a}{a_{\rm eq}}\right)^{1/2}\right]. \qquad (1.56)$$

From this we can find the age of the universe at matter-radiation equality. Inserting $a = a_{\rm eq}$ in (1.56), we get

$$t_{\rm eq} = \frac{4}{3H_0}\left(1-\frac{1}{\sqrt{2}}\right)\frac{\Omega_{\rm r0}^{3/2}}{\Omega_{\rm m0}^2},$$

which for $h = 0.7$, $\Omega_{\rm m0} = 0.3$, $\Omega_{\rm r0} = 8.4\times10^{-5}$ gives $t_{\rm eq} \approx 47000$ yr. Compared to the total age of the universe, which is more than 10 Gyr, the epoch of radiation domination is thus of negligible duration. We are therefore justified in ignoring it when calculating the total age of the universe.

Equation (1.56) cannot be solved analytically for a in terms of t, but one can at least show that it reduces to the appropriate solutions in the radiation- and matter-dominated phases. For $a \ll a_{\rm eq}$ one finds

$$a(t) \approx a_0(2\sqrt{\Omega_{\rm r0}}H_0 t)^{1/2},$$

which has the same $t^{1/2}$-behaviour as our earlier solution for a flat, radiation-dominated universe. In the opposite limit, $a \gg a_{\rm eq}$ one finds

$$a(t) \approx a_0\left(\frac{3}{2}\sqrt{\Omega_{\rm m0}}H_0 t\right)^{2/3},$$

which corresponds to the behaviour of the flat, matter-dominated Einstein-de Sitter model discussed earlier.

1.10.6 The flat ΛCDM model

Although the models we have considered in the previous subsections are important both historically and as approximations to the actual universe in the radiation dominated era and in the matter dominated era, a combination of cosmological data now seems to point in the direction of a different model: a model where the Universe is dominated by dust (mostly in the form of so-called cold dark matter with the acronym CDM) and a positive cosmological constant. More specifically, the observations seem to prefer a flat model with $\Omega_{\rm m0} \approx 0.3$ and $\Omega_{\Lambda 0} = 1 - \Omega_{\rm m0} \approx 0.7$. Hence we should spend some time on spatially flat models with matter and a cosmological constant. As we will see, the Friedmann equation can be solved analytically in this case.

Let us write the Friedmann equation as

$$\frac{H^2(t)}{H_0^2} = \Omega_{\rm m0}\left(\frac{a_0}{a}\right)^3 + (1 - \Omega_{\rm m0}).$$

As in the case of dust+curvature, we have to distinguish between two cases. For $\Omega_{\rm m0} > 1$, corresponding to $\Omega_\Lambda < 0$, the right hand side of the equation changes sign at a value $a_{\rm max}$, and after that the universe will enter a contracting phase. The value of $a_{\rm max}$ is given by

$$\Omega_{\rm m0}\left(\frac{a_0}{a_{\rm max}}\right)^3 = \Omega_{\rm m0} - 1,$$

i.e.,

$$\frac{a_{\rm max}}{a_0} = \left(\frac{\Omega_{\rm m0}}{\Omega_{\rm m0} - 1}\right)^{1/3}.$$

In this case the Friedmann equation can be rewritten as

$$H_0 dt = \frac{1}{\sqrt{\Omega_{\rm m0} - 1}} \frac{\sqrt{a} da}{\sqrt{\alpha - a^3}},$$

where we have defined $\alpha = \Omega_{\rm m0}/(\Omega_{\rm m0} - 1) = (a_{\rm max}/a_0)^3$. Since $a = 0$ for $t = 0$, we now have to calculate the integral

$$H_0 t = \frac{1}{\sqrt{\Omega_{\rm m0} - 1}} \int_0^a \frac{\sqrt{a} da}{\sqrt{\alpha - a^3}}.$$

The expression in the square root in the denominator suggests that we should try the substitution $a = \alpha^{1/3}(\sin\theta)^{2/3}$. This gives $da = \frac{2}{3}\alpha^{1/3}(\sin\theta)^{-1/3}\cos\theta d\theta$, and $\sqrt{\alpha - a^3} = \alpha^{1/2}\cos\theta$. When we insert all this in the integral, by a miracle everything except the constant factor 2/3 cancels out, and we are left with

$$H_0 t = \frac{2}{3\sqrt{\Omega_{\rm m0} - 1}} \int_0^{\sin^{-1}[(a/a_{\rm max})^{3/2}]} d\theta = \frac{2}{3\sqrt{\Omega_{\rm m0} - 1}} \sin^{-1}\left[\left(\frac{a}{a_{\rm max}}\right)^{3/2}\right].$$

Because of the inverse sine, we see that the universe will collapse in a Big Crunch after at time

$$t_{\rm crunch} = \frac{2\pi}{3H_0}\frac{1}{\sqrt{\Omega_{\rm m0}-1}}.$$

We can also solve for the scale factor a as a function of time and find

$$a(t) = a_0\left(\frac{\Omega_{\rm m0}}{\Omega_{\rm m0}-1}\right)^{1/3}\left[\sin\left(\frac{3}{2}\sqrt{\Omega_{\rm m0}-1}H_0 t\right)\right]^{2/3}.$$

Note that at early times, $a \ll a_{\rm max}$, we have

$$a(t) \approx a_{\rm max}\left(\frac{3}{2}\sqrt{\Omega_{\rm m0}-1}H_0 t\right)^{2/3},$$

and hence $a \propto t^{2/3}$, as expected for a matter-dominated universe.

Although there is no physical reason why the cosmological constant cannot be negative, observations indicate that we live in a universe where it is positive. In this case, corresponding to $\Omega_{\rm m0} < 1$, the right hand side of the Friedmann equation is always positive, and hence the universe is always expanding. In this case there is a value of the scale factor where the contribution to the energy density from matter becomes equal to the contribution from the cosmological constant. This value of the scale factor is given by

$$\Omega_{\rm m0}\left(\frac{a_0}{a_{\rm m\Lambda}}\right)^3 = \Omega_{\Lambda 0} = 1-\Omega_{\rm m0},$$

which gives

$$a_{\rm m\Lambda} = a_0\left(\frac{\Omega_{\rm m0}}{1-\Omega_{\rm m0}}\right)^{1/3}.$$

For $a < a_{\rm m\Lambda}$ matter dominates, and for $a > a_{\rm m\Lambda}$ the cosmological constant dominates. We can write the Friedmann equation as

$$H_0 dt = \frac{1}{\sqrt{1-\Omega_{\rm m0}}}\frac{\sqrt{a}da}{\sqrt{\beta+a^3}},$$

where $\beta = (a_{\rm m\Lambda}/a_0)^3$. Then,

$$H_0 t = \frac{1}{\sqrt{1-\Omega_{\rm m0}}}\int_0^a \frac{\sqrt{a}da}{\sqrt{\beta+a^3}},$$

and by substituting $a = \beta^{1/3}(\sinh u)^{2/3}$ and using the properties of the hyperbolic functions we find that

$$H_0 t = \frac{2}{3\sqrt{1-\Omega_{\rm m0}}}\sinh^{-1}\left[\left(\frac{a}{a_{\rm m\Lambda}}\right)^{3/2}\right]. \tag{1.57}$$

This equation can also be solved for a in terms of t, and this gives

$$a(t) = a_0 \left(\frac{\Omega_{\rm m0}}{1-\Omega_{\rm m0}}\right)^{1/3} \left[\sinh\left(\frac{3}{2}\sqrt{1-\Omega_{\rm m0}}H_0 t\right)\right]^{2/3}. \tag{1.58}$$

The present age of the universe in this model is found by inserting $a = a_0$ in equation (1.57):

$$t_0 = \frac{2}{3H_0\sqrt{1-\Omega_{\rm m0}}} \sinh^{-1}\left(\sqrt{\frac{1-\Omega_{\rm m0}}{\Omega_{\rm m0}}}\right),$$

and for $\Omega_{\rm m0} = 0.3$, $h = 0.7$ this gives $t_0 = 13.5$ Gyr. Thus the ΛCDM model is consistent with the age of the oldest observed objects in the universe. At the value of the scale factor $a_{\rm m\Lambda}$ where the cosmological constant starts to dominate the energy density of the universe, the age of the universe is

$$t_{\rm m\Lambda} = \frac{2}{3H_0\sqrt{1-\Omega_{\rm m0}}} \sinh^{-1}(1),$$

which for $\Omega_{\rm m0} = 0.3$, $h = 0.7$ gives $t_{\rm m\Lambda} = 9.8$ Gyr. Hence, in this model the universe has been dominated by the cosmological constant for the last 3.7 billion years.

The most peculiar feature of the ΛCDM model is that the universe at some point starts expanding at an accelerating rate. To see this, we rewrite the Friedmann equation for $\ddot{a}$ as

$$\begin{aligned}
\frac{\ddot{a}}{a} &= = -\frac{4\pi G}{3}\left(\rho_{\rm m0}\frac{a_0^3}{a^3} + \rho_{\Lambda 0} - 3\frac{p_\Lambda}{c^2}\right) \\
&= -\frac{H_0^2}{2}\frac{8\pi G}{3H_0^2}\left(\rho_{\rm m0}\frac{a_0^3}{a^3} - 2\rho_{\Lambda 0}\right) \\
&= -\frac{H_0^2}{2}\left(\Omega_{\rm m0}\frac{a_0^3}{a^3} - 2\Omega_{\Lambda 0}\right),
\end{aligned}$$

and we see that we get $\ddot{a} > 0$ (which means accelerating expansion) when $\Omega_{\rm m0}a_0^3/a^3 - 2\Omega_{\Lambda 0} < 0$. Intuitively, we would think that the universe should decelerate since we are used to thinking of gravity as an attractive force. However, a positive cosmological constant corresponds to an effective gravitational repulsion, and this then can give rise to an accelerating universe. The crossover from deceleration to acceleration occurs at the value $a_{\rm acc}$ of the scale factor given by

$$a_{\rm acc} = a_0\left(\frac{1}{2}\frac{\Omega_{\rm m0}}{1-\Omega_{\rm m0}}\right)^{1/3} = \left(\frac{1}{2}\right)^{1/3} a_{\rm m\Lambda},$$

and thus it happens slightly before the cosmological constant starts to dominate the energy density of the universe. For our standard values $\Omega_{\rm m0} = 0.3$,

$h = 0.7$, this corresponds to a redshift $z_{\rm acc} = a_0/a_{\rm acc} - 1 \approx 0.67$, and the age of the universe at this point is

$$t_{\rm acc} = \frac{2}{3H_0\sqrt{1-\Omega_{\rm m0}}} \sinh^{-1}\left(\frac{1}{\sqrt{2}}\right) \approx 7.3 \text{ Gyr}.$$

In this model, then, the universe has been accelerating for the last 6.2 billion years.

Finally, let us consider the extreme limits of this model. At early times, when $a \ll a_{\rm m\Lambda}$ we can use that $\sinh^{-1} x \approx x$ for $x \ll 1$ in equation (1.57) to find

$$H_0 t \approx \frac{2}{3\sqrt{1-\Omega_{\rm m0}}} \left(\frac{a}{a_{\rm m\Lambda}}\right)^{3/2},$$

which gives

$$a(t) \approx a_{\rm m\Lambda} \left(\frac{3}{2}\sqrt{1-\Omega_{\rm m0}} H_0 t\right)^{2/3},$$

so $a \propto t^{2/3}$ in the early stages, as expected for a matter-dominated model. In the opposite limit, $a \gg a_{\rm m\Lambda}$, we can use the approximation $\sinh^{-1} x \approx \ln(2x)$ for $x \gg 1$ in (1.57). Solving for a, we find

$$a(t) \approx 2^{-2/3} a_{\rm m\Lambda} \exp(\sqrt{1-\Omega_{\rm m0}} H_0 t),$$

so that $a \propto \exp(\sqrt{1-\Omega_{\rm m0}} H_0 t)$ in the Λ-dominated phase, as we would have expected from our discussion of the de Sitter universe.

1.10.7 Models with matter, curvature and a cosmological constant

Finally, we abandon the restriction to flat models and consider curved universes with dust and a cosmological constant. As we will see, some pretty weird models then emerge as theoretical possibilities. We write the Friedmann equation as

$$\frac{H^2(t)}{H_0^2} = \Omega_{\rm m0}\left(\frac{a_0}{a}\right)^3 + (1 - \Omega_{\rm m0} - \Omega_{\Lambda 0})\left(\frac{a_0}{a}\right)^2 + \Omega_{\Lambda 0}. \tag{1.59}$$

The matter density is always non-negative, and we leave the case $\Omega_{\rm m0} = 0$ as an exercise, and consider $\Omega_{\rm m0} > 0$ here. Then, the first term is always positive. If the cosmological constant is negative, $\Omega_{\Lambda 0} < 0$, the third term will eventually dominate, and the universe will collapse. For positive values of the cosmological constant, both the first and the third term are positive. If $\Omega_{\rm m0} + \Omega_{\Lambda 0} < 1$, the second term is positive, and hence if the universe is expanding at one point in time, it will always be expanding. For $\Omega_{\rm m0} + \Omega_{\Lambda 0} > 1$, however, the second term is negative. In that case, there is a possibility that the right-hand side of (1.59) may become negative for a certain range

of values of a. In that case, we can have 'bouncing' universe models which start out with $a \gg a_0$, contract until the reach the point where the right hand side of (1.59) vanishes, and then enter an expanding phase and expand out to infinity. In these models, then, there is no 'Big Bang', only a 'Big Bounce'.

We wish to derive the lines in the $\Omega_{\rm m0}$-$\Omega_{\Lambda 0}$ plane separating the various classes of models. Models on the border between the 'Big Bang' and 'Big Bounce' classes can be shown to be asymptotic to a static Einstein model in the infinite past. In that case, we know that they must satisfy

$$\begin{aligned} \frac{H^2}{H_0^2} &= 0 = \frac{\Omega_{\rm m0} a_0^3}{a_s^3} + \frac{\Omega_{\rm k0} a_0^2}{a_s^2} + \Omega_{\Lambda s} \\ \frac{\ddot{a}}{a_s} &= 0 = \frac{1}{2} H_0^2 \left(\frac{\Omega_{\rm m0} a_0^3}{a_s^3} - 2\Omega_{\Lambda s} \right), \end{aligned}$$

where a_s is the (quasi-)static value of the scale factor, and $\Omega_{\Lambda s}$ is the corresponding value of $\Omega_{\Lambda s}$ which we want to determine for given $\Omega_{\rm m0}$. Note that for brevity of notation we have reinstated the quantity $\Omega_{\rm k0} = 1 - \Omega_{\rm m0} - \Omega_{\Lambda 0}$. From these two equations it follows immediately that $\Omega_{\Lambda s} = \Omega_{\rm m0} a_0^3/(2a_s^3) = -\Omega_{\rm k0} a_0^2/(3a_s^2)$. Solving for a_s we find $a_s = -3\Omega_{\rm m0} a_0/2\Omega_{\rm k0}$, and inserting this back in the first equality using the constraint on $\Omega_{\rm k0}$, we get the equation

$$\Omega_{\Lambda s} = \frac{4}{27} \frac{(\Omega_{\rm m0} + \Omega_{\Lambda s} - 1)^3}{\Omega_{\rm m0}^2}.$$

Introducing the new variable $x^3 = \Omega_{\Lambda s}/(4\Omega_{\rm m0})$ it is easy to show that this equation can be rewritten as the cubic equation

$$x^3 - \frac{3}{4} x + \frac{\Omega_{\rm m0} - 1}{\Omega_{\rm m0}} = 0.$$

This is on the so-called reduced form (a general cubic equation can always be brought into this form),

$$y^3 + 3py + 2q,$$

with $p = -1/4$, $q = (\Omega_{\rm m0} - 1)/(8\Omega_{\rm m0})$. We are interested in real, positive solutions of the equation. The nature of the solutions is determined by the discriminant, which is given by

$$\Delta = 4(p^3 + q^2) = \frac{1 - 2\Omega_{\rm m0}}{\Omega_{\rm m0}^2}.$$

The theory of equations tells us that there are three real, distinct roots for $\Delta < 0$, corresponding to $\Omega_{\rm m0} > 1/2$. There are three real roots, of which at least two are degenerate for $\Delta = 0$, i.e. $\Omega_{\rm m0} = 1/2$, and for $\Delta > 0$, $\Omega_{\rm m0} < 1/2$. Furthermore, it turns out that for $1/2 < \Omega_{\rm m0} < 1$, only one root is positive, whereas for $\Omega_{\rm m0} > 1$, two roots are positive. One can show that the final result is:

Case 1: For $0 < \Omega_{\rm m0} \leq 1/2$, we have

$$\Omega_{\Lambda s} = 4\Omega_{\rm m0}\left\{\cosh\left[\frac{1}{3}\cosh^{-1}\left(\frac{1-\Omega_{\rm m0}}{\Omega_{\rm m0}}\right)\right]\right\}^3.$$

Case 2: For $1/2 \leq \Omega_{\rm m0} \leq 1$, we have

$$\Omega_{\Lambda s} = 4\Omega_{\rm m0}\left\{\cos\left[\frac{1}{3}\cos^{-1}\left(\frac{1-\Omega_{\rm m0}}{\Omega_{\rm m0}}\right)\right]\right\}^3.$$

Case 3: For $\Omega_{\rm m0} > 1$, we have

$$\Omega_{\Lambda s} = 4\Omega_{\rm m0}\left\{\cos\left[\frac{1}{3}\cos^{-1}\left(\frac{1-\Omega_{\rm m0}}{\Omega_{\rm m0}}\right)\right]\right\}^3.$$

In addition, we get a second solution, corresponding to $a_s > 1$, that is, corresponding to a quasi-static state in the future. This occurs for

$$\Omega_{\Lambda s2} = 4\Omega_{\rm m0}\left\{\cos\left[\frac{1}{3}\cos^{-1}\left(\frac{1-\Omega_{\rm m0}}{\Omega_{\rm m0}}\right)+\frac{4\pi}{3}\right]\right\}^3.$$

For those who are interested in details about the solutions of cubic equations and how to express them in terms of trigonometric and hyperbolic functions, I can recommend the paper by J. P. McKelvey in the American Journal of Physics, volume 52, page 269. Close to the critical line in the top left corner of fig. 1.3 one finds a class of models where the universe spends a considerable amount of time expanding very slowly. These models are called 'loitering' models, and have from time to time been taken off the shelf because they allow, e.g., more time for quasar formation at high redshifts.

1.11 Horizons

Which parts of the universe are visible to us now? And which parts will be visible to us in the future? Given that the speed of light is finite, and that the universe is expanding, these are relevant question to ask, and leads to the introduction of the two concepts *event horizon* and *particle horizon*. The best discussion of these concepts is still Wolfgang Rindler's paper from 1966 (W. Rindler. MNRAS 116, 1966, 662), and I will to a large extent follow his treatment here. The event horizon answers the question: if distant source emits a light ray in our direction now, will it reach us at some point in the future no matter how far away this source is? The particle horizon answers a different question: Is there a limit to how distant a source, which we have received, or are receiving, light from by now, can be? Thus, the event horizon is related to events observable in our future, whereas the particle horizon is related to events observable at present and in our past. The particle horizon

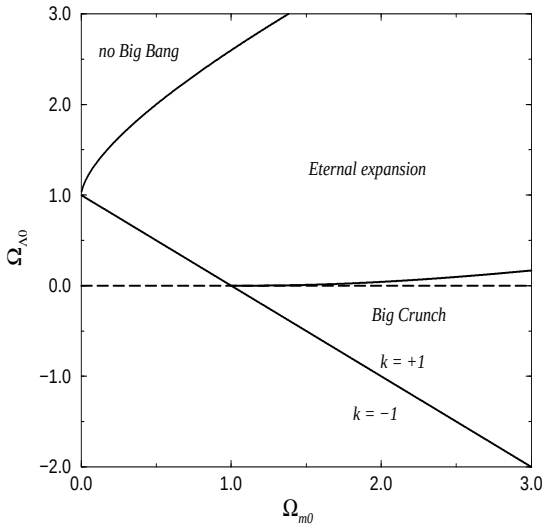

Figure 1.3: A classification of models with matter, cosmological constant, and curvature.

is particularly important because it tells us how large regions of the universe are in causal contact (i.e. have been able to communicate by light signals) at a given time. Since no information, and in particular no physical forces, can be transmitted at superluminal speed, the particle horizon puts a limit on the size of regions where we can reasonably expect physical conditions to be the same.

Let us start by citing Rindler's definitions of the two horizons:

- *Event horizon*: for a given fundamental observer A, this is a hypersurface in spacetime which divides all events into two non-empty classes: those that have been, are, or will be observable by A, and those that are forever outside A's possible powers of observation.
- *Particle horizon*: for a given fundamental observer A and cosmic time t_0, this is a surface in the instantaneous 3-space $t = t_0$ which divides all events into two non-empty classes: those that have already been observable by A at time t_0 and those that have not.

We will place our fundamental observer at the origin at comoving coordinate $r = 0$. Light rays will play an important role in the following, and a light ray going through the origin is described by having $d\theta = 0 = d\phi$, and $ds^2 = 0$, where ds^2 is given by the RW line element. This gives

$$\frac{cdt}{a(t)} = \pm \frac{dr}{\sqrt{1 - kr^2}},$$

where the plus sign is chosen for rays moving away from the origin, the minus sign for rays towards the origin. In what follows it is useful to use the function

$$\mathcal{S}^{-1}(r) = \int_0^r \frac{dr}{\sqrt{1-kr^2}},$$

introduced in our discussion of the proper distance. From that discussion, recall that at a given time t_1, the proper distance from the origin of a source at comoving coordinate r_1 is given by

$$d_{\rm P}(t_1) = a(t_1)\mathcal{S}^{-1}(r_1).$$

Now, r_1 is by definition constant in time, so the equation of motion describing the proper distance of the source from the origin at any given time t is simply

$$d_{\rm P}(t) = a(t)\mathcal{S}^{-1}(r_1).$$

Let us now consider a light ray emitted towards the origin from comoving coordinate r_1 at time t_1. At time t, its comoving radial coordinate is given by

$$\int_{r_1}^{r} \frac{dr}{\sqrt{1-kr^2}} = -\int_{t_1}^{t} \frac{cdt'}{a(t')},$$

from which we find

$$\mathcal{S}^{-1}(r) = \mathcal{S}^{-1}(r_1) - \int_{t_1}^{t} \frac{cdt'}{a(t')} \tag{1.60}$$

and hence the proper distance of this light ray from the origin at a given time t is

$$d_{\rm P}^{l} = a(t)\left[\mathcal{S}^{-1}(r_1) - \int_{t_1}^{t} \frac{cdt'}{a(t')}\right], \tag{1.61}$$

where the superscript l stands for 'light'. The key point to note now is that for the light ray to reach the origin, the expression in the brackets must vanish at some time, otherwise the light ray will always be at a non-zero distance from the origin. We will limit our cases to the situation where $\mathcal{S}^{-1}(r)$ is a strictly increasing function of r, which corresponds to $k = -1, 0$. (The case of a positively curved universe is more subtle, for details see Rindler's original paper.)

1.11.1 The event horizon

Will the light ray emitted by the source at r_1 at time t_1 ever reach the origin? The key question here is whether the integral

$$\int_{t_1}^{\infty} \frac{cdt'}{a(t')},$$

converges to a finite limit. To see this, note that $\mathcal{S}^{-1}(r)$ is a positive, increasing function of r, and that r_1 is constant. If r_1 is so large that

$$\mathcal{S}^{-1}(r_1) > \int_{t_1}^{\infty} \frac{cdt'}{a(t')},$$

then at no finite time t will the expression in brackets in equation (1.61) vanish, and hence the light ray will never reach the origin. It may sound paradoxical that a light ray moving towards the origin at the speed of light (as measured locally) will never reach it, but bear in mind that space is expanding while the light ray is moving (and there is no speed limit on the expansion of space, only on particles moving *through* space). It is a bit like an athlete running towards a moving goal. If the finishing line moves away faster than the athlete can run, he will never reach it. If the integral converges then, there is a maximum value $r_{\rm EH}$ of r_1 such that for $r_1 > r_{\rm EH}$ light emitted from r_1 at t_1 will never reach the origin. We see that this value of r is determined by

$$\mathcal{S}^{-1}(r_{\rm EH}) = \int_{t_1}^{\infty} \frac{cdt'}{a(t')},$$

so that the light ray emitted towards the origin at time t_1 reaches the origin in the infinite future. Light rays emitted at the same time from sources with $\mathcal{S}^{-1}(r) > \mathcal{S}^{-1}(r_{\rm EH})$ will never reach the origin. The time t_1 is arbitrary, so we can replace it by t to make it clear that the event horizon is in general a time-dependent quantity. The proper distance to the event horizon is given by

$$d_{\rm P}^{\rm EH} = a(t) \int_{t}^{\infty} \frac{cdt'}{a(t')}. \tag{1.62}$$

1.11.2 The particle horizon

The event horizon concerns events observable in the future, whereas the particle horizon is related to events which have been, or are being, observed by a given time t (for example now). Again, we consider a source at comoving radial coordinate r_1 which emits a light signal at time t_1, so that the equation of motion of the light signal is again

$$d_{\rm P}^{l} = a(t) \left[\mathcal{S}^{-1}(r_1) - \int_{t_1}^{t} \frac{cdt'}{a(t')} \right]. \tag{1.63}$$

We want to know whether there is a limit to which light rays can have reached the origin by the time t. To maximize the chance of the light reaching the origin, we consider a light ray emitted at the earliest possible moment, which normally means taking $t_1 = 0$ (but in the case of the de Sitter model, where there is no Big Bang, we have to take $t_1 = -\infty$.) Since $a(t) \to 0$ as $t \to 0$, there is a possibility that the integral on the right hand

side diverges. However, in the case where the integral does converge to a finite value, there will be points r_1 so that

$$\mathcal{S}^{-1}(r_1) > \int_0^t \frac{cdt'}{a(t')},$$

and a light ray emitted from r_1 at $t = 0$ will then not yet have reached the origin by time t. We then say that there exist a particle horizon with comoving radial coordinate at time t determined by

$$\mathcal{S}^{-1}(r_{\rm PH}) = \int_0^t \frac{cdt'}{a(t')}, \tag{1.64}$$

and the proper distance of this point from the origin is

$$d_{\rm P}^{\rm PH} = a(t) \int_0^t \frac{cdt'}{a(t')}. \tag{1.65}$$

1.11.3 Examples

First, let us consider the de Sitter model. Recall that in this model we found that the scale factor is given by $a(t) = a_0 \exp[H_0(t - t_0)]$, where t_0 is cosmic time at the current epoch. There is nothing preventing us from defining $t_0 = 0$, so we will do this for simplicity, and hence take $a(t) = a_0 \exp(H_0 t)$. Bear in mind that there is no Big Bang in this model, and the time t can vary from $-\infty$ to $+\infty$. Consider the integral

$$I(t_1, t_2) = \int_{t_1}^{t_2} \frac{cdt}{a(t)} = \frac{c}{a_0} \int_{t_1}^{t_2} e^{-H_0 t} dt = \frac{c}{a_0 H_0}(e^{-H_0 t_1} - e^{-H_0 t_2}). \tag{1.66}$$

First, let $t_1 = t$ be fixed and let t_2 vary. Then we see that I is an increasing function of t_2. Furthermore, we see that I reaches a limiting value as $t_2 \to \infty$:

$$I(t_1 = t, t_2 \to \infty) = \frac{c}{a_0 H_0} e^{-H_0 t}.$$

Thus, there exists an event horizon in this model. Since the de Sitter model we consider here is spatially flat, we have $f(r) = r$, and hence the comoving radial coordinate of the event horizon is

$$r_{\rm EH} = \frac{c}{a_0 H_0} e^{-H_0 t}.$$

At a given time t, there is therefore a maximum radial coordinate, $r_{\rm EH}$, and light signals emitted from sources with $r > r_{\rm EH}$ at this time will never reach the origin. Furthermore, as t increases, $r_{\rm EH}$ decreases, and hence more and more regions will disappear behind the event horizon. This does not mean that they will disappear completely from our sight: we will be receiving light signals emitted before the source disappeared inside the event horizon

all the time to $t = \infty$, but the light will be more and more redshifted. And, of course, no light signal emitted after the source crossed the event horizon will ever be received by us. Note that the proper distance to the event horizon is constant:

$$d_{\rm P}^{\rm EH} = a(t) r_{\rm EH} = \frac{c}{H_0}.$$

Thus, we can look at this in two ways: in comoving coordinates, the observer (at $r = 0$) and the source stay in the same place, whereas the event horizon moves closer to the origin. In terms of proper distances, the origin observer and the event horizon stay in the same place as time goes by, but the source is driven away from us by the expansion and eventually moves past the event horizon.

For the de Sitter model, there is no particle horizon. To see this, fix $t_2 = t$ and let $t_1 \to -\infty$ in the expression for $I(t_1, t_2)$ above. Clearly, the expression diverges. This means that light rays sent out at $t = -\infty$ will have reached the origin by time t, no matter where they are sent from. Hence, in this model, the whole universe is causally connected. This is an important point to note for our discussion of inflation later on.

For our second example, let us consider the flat Einstein-de Sitter model, where $a(t) = a_0 (t/t_0)^{2/3}$, and $H_0 = 2/3t_0$. Once again, we start by calculating the integral

$$I(t_1, t_2) = \int_{t_1}^{t_2} \frac{c dt}{a(t)} = \frac{2c}{a_0 H_0} \left[\left(\frac{t_2}{t_0} \right)^{1/3} - \left(\frac{t_1}{t_0} \right)^{1/3} \right].$$

First, let $t_1 = t$ be fixed and let t_2 vary. We see that I increases without limit as $t_2 \to \infty$, and hence there is no event horizon in this model. Thus, receiving a light signal emitted anywhere in the universe at any time is just a matter of waiting long enough: eventually, the light will reach us. However, for $t_2 = t$ fixed, with t_1 varying, we see that I has a finite limit for $t_1 \to 0$:

$$I(t_1 \to 0, t_2 = t) = \frac{2c}{a_0 H_0} \left(\frac{t}{t_0} \right)^{1/3}.$$

Thus, there is a particle horizon in this model. This means that at time t, there is a limit to how distant a source we can see. The comoving radial coordinate of the particle horizon is given by

$$r_{\rm PH} = \frac{2c}{a_0 H_0} \left(\frac{t}{t_0} \right)^{1/3},$$

and the proper distance to the particle horizon is given by (since $\mathcal{S}^{-1}(r) = r$ in this model)

$$d_{\rm P}^{\rm PH} = a(t) r_{\rm PH} = \frac{2c}{H_0} \left(\frac{t}{t_0} \right).$$

Finally, we note that the ΛCDM model has both a particle horizon (since it behaves as an EdS model at early times) and an event horizon (since it behaves as a dS model at late times). I leave the demonstration of this as an exercise.

1.12 The Steady State model

Almost all of present-day cosmology is carried out within the theoretical framework outlined so far in this chapter, and the general consensus is that the current body of cosmological observations clearly point towards a model where the universe has evolved to its present state from a dense and hot region in the distant past, more than ten billion years ago. However, during the early days of modern cosmology there were quite a few scientists who felt uncomfortable with this picture. If taken to the extreme, the Big Bang model says that the universe with all its matter emerged from a singular point in the finite past. But the model says nothing about how this matter was created. And it is not likely to ever do so, because the physical laws it is built on break down at $t = 0$. During the late 1940's and up to the mid 1960's there was therefore a number of astronomers and physicists who instead preferred a totally different picture of the universe, called the Steady State model. Even though it has been shown beyond all reasonable doubt to be inconsistent with observations, it was a beautiful idea and it is appropriate that we should all know some of the basic properties of the model.

The Steady State model was introduced in two papers in 1948, one by Herman Bondi and Thomas Gold, and one written by Fred Hoyle. Although all three of them were partly working together, Hoyle chose a different starting point than Bondi and Gold. We will here follow the approach of the latter two.

Bondi and Gold agreed that cosmology should be built on the cosmological principle (which says that the universe on large scales is homogeneous and isotropic) because this gives reason to believe the physical laws should be the same everywhere in the universe at a given time. However, they felt that this was a bit too weak. If we are to have any hope of understanding the universe, we also need the physical laws to be the same at all times, and they felt that there was no reason to believe that this was so in a Big Bang-type model of the universe. Thus, they chose as their starting point a stronger version of the cosmological principle, called the Perfect Cosmological Principle (PCP), which states that not only is the universe homogeneous and isotropic, it is also unchanging with time on large scales. Note the qualifier 'on large scales'. Bondi and Gold did not deny that the universe goes through dramatic changes on small scales: stars are born and die, new galaxies are formed, and so on. But averaged over sufficiently large scales, the properties of the universe should not change. This, they felt, ensured that

it was safe to employ physics as we know it to construct a model for the universe. An immediate consequence of the PCP is that the line element, if the universe is described by a metric theory of gravity, should be of the RW form, since space is homogeneous and isotropic. But the temporal aspect of the PCP also allows some further deductions to be made. First of all, the expansion rate $H = \dot{a}/a$ has to be constant, equal to its present value H_0. Then it follows that the scale factor must be given by $a(t) = \exp(H_0 t)$. Furthermore, the spatial curvature at a given time t can be shown to be $k/a^2(t)$. Since this is a, in principle, measurable large-scale property of the universe, it must by the PCP be constant in time. The only way to ensure this, given that $a(t)$ varies with time, is that $k = 0$. Thus, the PCP alone is enough to deduce that the spatial part of the line element is flat, and that line element is of the de Sitter form

$$ds^2 = c^2 dt^2 - e^{2H_0 t}(dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\phi^2).$$

But this is not all. We can also say something about the sign of H_0. Recall that in the de Sitter model the universe is infinitely old. If H_0 is zero, then the universe is static, and since it also is of infinite age, it should be in a state of thermodynamic equilibrium with maximum entropy. In short, the universe should be a dead, quiet place. This obviously does not correspond to the current state of affairs. Also, if $H_0 < 0$, a contracting universe, we would expect to see the light from distant sources being blue-shifted. Since the universe is infinitely old, we should have been cooked in the radiation from distant sources in the universe. Again, this is obviously not the case. So the only possibility is $H_0 > 0$: the universe must expand! We have arrived at this conclusion without using any detailed data on the redshift-distance relationship for galaxies, and this illustrates the power of the PCP.

A question which arises now is how a de Sitter line element can describe a universe filled with matter, since we saw earlier that the de Sitter solution corresponds to an empty universe which expansion is driven by the cosmological constant. The answer is that the Steady State model is not strictly a solution of Einstein's field equations. Bondi and Gold did not deny the validity of general relativity on small scales, but felt that since it had not been tested on cosmological scales, it should take a back seat to the PCP. In Fred Hoyle's version of the model one starts from a modified set of gravitational field equations, and from which the de Sitter line element can arise even in a universe containing matter.

Note, however, that since space is expanding, one would expect matter to become more and more dilute as time goes on. Since a given three-volume V is proportional to $a^3(t) = \exp(3H_0 t)$, if mass is conserved, the mean density should drop as $\exp(-3H_0 t)$. However, by the PCP we should have $\rho = \text{constant} = \rho_0$. The only solution to this dilemma is to postulate

that new matter is constantly being created at the rate

$$\dot{M} = 3H_0 V \rho_0,$$

corresponding to a matter creation rate per unit volume of

$$\dot{Q} \approx 2 \times 10^{-46} \left(\frac{\rho_0}{\rho_{\rm c0}} \right) h^3 \ {\rm g\,cm^{-3}}.$$

Matter appearing spontaneously out of empty space is obviously not a part of standard physics. However, the required rate is seen to be very modest. Hoyle, Bondi and Gold felt that it was much more easier to accommodate a slow, steady creation of matter than a sudden creation of all matter in the universe in a Big Bang. Hoyle also was able to devise GR-like versions of the model which had matter creation as one of its elements.

From a philosophical standpoint, the Steady State model has several virtues. It is simple, and from the PCP alone follow falsifiable predictions about the universe. Furthermore, since in this model the universe is infinitely old, there is no dramatic, unexplained Big Bang event. The universe has always existed, will always exist, and will always be the same. The only problem with the Steady State model is that it is in glaring contradiction with observations. We will see examples of this throughout this course. This is why the overwhelming majority of cosmologists today subscribe to the Big Bang model. However, there are still some people who try to rescue the pieces of the Steady State theory, but this cannot be done without adding 'epicycles' which destroy much of the attractiveness of the original model, and none of the new versions can successfully accommodate all the cosmological data. Thus, I would personally say that there is no empirical or philosophical reason to prefer one of the Steady State model's descendants over the Big Bang model.

1.13 Some observable quantities and how to calculate them

In order to make contact with observations, we need to know how to calculate observables for the Friedmann models. We will limit our attention to models containing a mixture of dust, radiation, a cosmological constant, and curvature. A convenient way of writing the Friedmann equation in this case is

$$\frac{H^2(a)}{H_0^2} = \Omega_{\rm m0} \left(\frac{a_0}{a} \right)^3 + \Omega_{\rm r0} \left(\frac{a_0}{a} \right)^4 + \Omega_{\rm k0} \left(\frac{a_0}{a} \right)^2 + \Omega_{\Lambda 0}, \tag{1.67}$$

or, since $a/a_0 = 1/(1+z)$, we can alternatively write it as

$$\frac{H^2(z)}{H_0^2} = \Omega_{\rm m0}(1+z)^3 + \Omega_{\rm r0}(1+z)^4 + \Omega_{\rm k0}(1+z)^2 + \Omega_{\Lambda 0}. \tag{1.68}$$

By inserting $t = t_0$ in the first equation, or $z = 0$ in the second equation, we see that

$$\Omega_{\rm m0} + \Omega_{\rm r0} + \Omega_{\rm k0} + \Omega_{\Lambda 0} = 1. \tag{1.69}$$

We have already obtained expressions for the age of the universe in some Friedmann models. In general it is not possible to find analytical expressions for the age, so it is useful to have a form which is suited for numerical computations. This is easily done by noting that the definition

$$\frac{\dot{a}}{a} = \frac{1}{a}\frac{da}{dt} = H,$$

can be written as

$$dt = \frac{da}{aH(a)}.$$

If there is a Big Bang in the model so that $a(t = 0) = 0$, then we can find the cosmic time corresponding to the scale factor having the value a as

$$t(a) = \int_0^a \frac{da'}{a'H(a')},$$

and the present age of the universe is

$$t_0 = \int_0^{a_0} \frac{da}{aH(a)}, \tag{1.70}$$

and for given values of the density parameters, this integral can be computed numerically using equation (1.67). We can also write these equations in terms of the redshift z. Note that $1 + z = a_0/a$ implies that

$$dz = -\frac{a_0 da}{a^2} = -(1+z)^2\frac{da}{a_0}$$

so we can write (1.70) as

$$t_0 = -\int_\infty^0 \frac{(1+z)}{(1+z)^2}\frac{dz}{H(z)} = \int_0^\infty \frac{dz}{(1+z)H(z)}. \tag{1.71}$$

However, in numerical computations the form (1.70) is usually more convenient since it only involves integration over the finite interval from 0 to a_0.

The observable distance measures are the luminosity distance given by equation (1.16), and the angular diameter distance (1.18). They both depend on the comoving radial coordinate r, given by equation (1.22), which again is determined by the integral

$$I = \int_t^{t_\circ} \frac{cdt'}{a(t')},$$

where t_o is the time where the light is received by the observer, and t is the time when the light was emitted by the source. Being enormously self-centered, we will mostly take $t_o = t_0$, corresponding to the epoch we are living in, which corresponds to $a = a_0$ and $z = 0$. Using $da = \dot{a}dt$, we can write

$$I = \int_a^1 \frac{cda'}{\dot{a}'a'} = \int_a^1 \frac{cda'}{a'^2 H(a')},$$

and if we want to carry out the integral in terms of redshifts, we can use the same substitution we employed in the age integral to show that

$$I = \frac{c}{a_0 H_0} \int_0^z \frac{dz'}{H(z')/H_0}.$$

As an illustration, the luminosity distance to a source at redshift z can then be written as

$$d_L = (1+z)\mathcal{S}_k\left(\frac{c}{a_0 H_0}\int_0^z \frac{dz'}{H(z')/H_0}\right).$$

By going back to the integral defining $\mathcal{S}_k(r)$, and using that $a_0 = c/H_0\sqrt{|\Omega_{k0}|}$ for $k \neq 0$, one can put this in the more useful form

$$d_L = \frac{(1+z)c}{H_0\sqrt{|\Omega_{k0}|}}\mathcal{S}_k\left(\sqrt{|\Omega_{k0}|}\int_0^z \frac{H_0 dz'}{H(z')}\right), \tag{1.72}$$

where, to remind you, $\mathcal{S}_k(x) = \sinh(x)$ for $k = -1$, $\mathcal{S}_k(x) = x$ for $k = 0$, and $\mathcal{S}_k(x) = \sin(x)$ for $k = +1$.

1.14 Closing comments

We have now gone through the basics of classical cosmology. Given the composition of the universe in terms of matter, radiation, and vacuum energy, and given the isotropy and homogeneity of space, general relativity predicts the evolution of the universe. Except in the case where the universe is dominated by a cosmological constant early on in its history, we see that the prediction is that the universe started expanding from zero size. How this expansion started, and where the matter and other sources of energy density came from, the model says nothing about. Strictly, the time $t = 0$ is not a part of the model, since the density of the universe goes to infinity at this point, and then general relativity breaks down. In this sense, the Big Bang model is really a model for how the universe evolved once the expansion had started. It is somewhat similar to the theory of evolution: it is a (very successful) model for how complex organisms have evolved from simple beginnings, not a theory for how life arose in the first place. However, we are still curious as to how that happened, and, similarly, we are also interested in extending our understanding of the universe all the way back to the beginning.

The field of cosmology is much more than just the building of models of the large-scale structure of spacetime. We also want to understand things like how the elements were formed, and how galaxies and clusters of galaxies were assembled. In the next few chapters we will turn our attention to these questions. To do that, we need to understand the conditions of the early universe. We need to know something about the particle species likely to be present, and how they behave as the temperature changes. Therefore, we will start by reviewing some statistical physics and thermodynamics.

1.15 Exercises

Exercise 1.1

The Copernican principle asserts that no point in the universe is special. That means that an observer would find the universe to have the same large-scale properties regardless of where she or he carries out the observations.. Prove that if we assume that the universe is isotropic and that the Copernican principle is valid, then the universe must be homogeneous.

Exercise 1.2

Consider a one-dimensional creature living on the circumference of a two-dimensional circle of constant radius a. Use Cartesian coordinates (x, y).

a) Show that for a small displacement along the circumference of the circle,

$$dy = -\frac{xdx}{(a^2 - x^2)^{1/2}}.$$

b) Show that the length of a small displacement along the circle measured by the creature is given by

$$dl^2 = \frac{dx^2}{1 - x^2/a^2}.$$

Exercise 1.3

Consider our three-dimensional universe to be embedded in a 4-dimensional Euclidean space. More precisely, assume that we live on the surface of a four-sphere of constant radius a, defined by the equation

$$x^2 + y^2 + z^2 + w^2 = a^2,$$

where w is the coordinate for the extra spatial dimension.

a) Define $r^2 = x^2 + y^2 + z^2$, and show that for a small displacement along the surface of the sphere,

$$dw = -\frac{r dr}{(a^2 - r^2)^{1/2}}.$$

b) Show that the length of a small displacement along the surface of the sphere is given by

$$dl^2 = dx^2 + dy^2 + dz^2 + \frac{r^2 dr^2}{a^2 - r^2}.$$

c) Switch to spherical coordinates (r, θ, ϕ) defined by $x = r\sin\theta\cos\phi$, $y = r\sin\theta\sin\phi$, $z = r\cos\theta$, and show that the spatial line element can be written as

$$dl^2 = \frac{dr^2}{1 - r^2/a^2} + r^2 d\theta^2 + r^2 \sin^2\theta d\phi^2.$$

d) If this universe is expanding, then a is not constant, but a function of time $a = a(t)$. However, if the expansion is slow compared with the time it takes particle or light ray to move the infinitesimal distance dl, then $a(t)$ can be considered constant during this displacement. Taking this assumption to be valid, introduce a new coordinate $u = r/a(t)$ and show that we can rewrite dl^2 as

$$dl^2 = a^2(t)\left(\frac{du^2}{1 - u^2} + u^2 d\theta^2 + u^2 \sin^2\theta d\phi^2\right).$$

Exercise 1.4

Assume that vi can describe the Universe with Newtonian gravity, and consider an expanding ball of matter with constant density ρ. Let the ball's radius be $R = ra(t)$ (r is a constant), and the speed in the radial direction of a small piece of matter at a distance x from the centre of the ball is $v = Hx$, where $H = \dot{a}/a$.

a) Calculate the total gravitational potential energy of the ball.

b) Calculate the total kinetic energy of the ball.

c) Assume that the density is given by $\rho = 3H^2/8\pi G$, and calculate the total energy of the ball.

d) The Heisenberg uncertainty principle for energy and time says that $\Delta E \delta t \geq \hbar/2$. In the light of this principle and the results in this exercise do you think that our Universe can be a quantum fluctuation?

Exercise 1.5

a) The so-called critical energy density is the value of ρ for which $k = 0$ in Friedmann's equations. Show that it is given by

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

where $H(t) = \dot{a}/a$. Using the present value of the Hubble parameter, $H_0 = 72\ {\rm km\,s^{-1}\,Mpc^{-1}}$, calculate the present value of $\rho_{\rm c}$. Give your answer in units of ${\rm g\,cm^{-3}}$ and $M_\odot\,{\rm Mpc^{-3}}$. Calculate the present value of the energy density $\rho_{\rm c}c^2$ in units of ${\rm GeV\,cm^{-3}}$.

b) Starting with Friedmann's equation with a cosmological constant:

$$\dot{a}^2 + kc^2 = \frac{8\pi G}{3}\rho a^2 + \frac{1}{3}\Lambda a^2,$$

and looking at the present epoch, $t = t_0$, show that it can be written as

$$\Omega_0 + \Omega_{\Lambda 0} + \Omega_{k0} = 1,$$

where the so-called density parameters are given by $\Omega_0 = \rho_0/\rho_{c0}$, $\Omega_{\Lambda 0} = \Lambda/3H_0^2$, and $\Omega_{k0} = -kc^2/a_0^2H_0^2$.

c) The so-called deceleration parameter q_0 is defined by

$$q_0 = -\frac{\ddot{a}_0 a_0}{\dot{a}_0^2}.$$

Show that $q_0 > 0$ for $k = 0$, $\Lambda = 0$. Assume that only dust ($p = 0$) contributes to the density. What does this mean ?

d) Consider a model of the universe where the present value of the density parameter for dust $\Omega_0 < 1$, $\Lambda \neq 0$, and $k = 0$. Find an expression for the present deceleration parameter q_0 in this case. What condition must Ω_0 satisfy if the Universe is to expand at an accelerating rate ?

Exercise 1.6

'Phantom energy', a substance with equation of state parameter $w < -1$, has been proposed as an alternative to the cosmological constant for explaining the present accelerated phase of expansion. Assume that we live in a spatially flat universe, dominated by phantom energy with $w = -2$.

a) Determine how the energy density of this component varies with the scale factor a.

b) Integrate the Friedmann equation for $\dot{a}/a$ from our present epoch t_0 ($a(t_0) = a_0$) and into the future to find $a(t)$ for $t > t_0$.

c) What happens as $t - t_0 \to \frac{2}{3H_0}$? Does the expression 'Big Rip' seem appropriate?

Exercise 1.7

Assume a spatially flat universe ($k = 0$) with scale factor given by

$$a(t) = a_0 \left(\frac{t}{t_0}\right)^{2/3}.$$

Here t_0 is the present cosmic time, and a_0 is the present value of the scale factor. We observe an object at cosmic redshift $z = 3$.

a) Calculate the comoving coordinate r of the object and its proper distance from us at $t = t_0$.

b) The radiation we receive from the object contains a message from an advanced civilization. We wish to send a radio signal back to them. If we send it at t_0, at what time (in units of t_0) will our signal reach them?

Exercise 1.8

a) Show that the luminosity distance to an object with a redshift z in a flat ($k = 0$) universe containing non-relativistic matter and vacuum energy can be written as

$$d_{\mathrm{L}} = \frac{c(1+z)}{H_0} \int_0^z \frac{dz'}{\sqrt{\Omega_{\mathrm{m0}}(1+z')^3 + 1 - \Omega_{\mathrm{m0}}}}.$$

, where Ω_{m0} is the density parameter for non-relativistic matter (dust).

b) Evaluate the integral for i) $\Omega_{\mathrm{m0}} = 1$ and ii) $\Omega_{\mathrm{m0}} = 0$.

c) Show that in the limit of very low redshifts d_{L} is approximately given by

$$d_{\mathrm{L}} \approx \frac{cz}{H_0},$$

independent of what the value of Ω_{m0} is .

Fluxes in astronomy are (sadly) usually quoted in terms of *magnitudes*. Magnitudes are related to fluxes via $m = -\frac{5}{2}\log(F) + \text{constant}$, where log denotes the logarithm with base 10. The *apparent magnitude* m is the flux we observe here on Earth, whereas the *absolute magnitude* M is the flux emitted at the source. They are related by

$$m - M = 5\log\left(\frac{d_{\mathrm{L}}}{10\ \mathrm{pc}}\right) + K,$$

where K is a correction for the shifting of spectrum into or out of the wavelength range measured due to the expansion. If we know both m and

M for a source, we can infer its luminosity distance. Objects of known M are called *standard candles*. Supernovae of type Ia are believed to be standard candles, and apparent magnitudes and redshifts have been determined for more than 150 of them.

d) Consider Supernova 1997ap found at redshift $z = 0.83$ with apparent magnitude $m = 24.32$, and Supernova 1992P found at low redshift $z = 0.026$ with apparent magnitude $m = 16.08$. Assuming they both have the same absolute magnitude M, show that the luminosity distance to Supernova 1997ap is given by

$$d_{\rm L}(z = 0.83) = 1.16\frac{c}{H_0}.$$

e) Compare the result in d) with the results in b) for $z = 0.83$. Any comments ?

Exercise 1.9

Use the Friedmann equation for $\ddot{a}$ with a cosmological constant to find the equation for the time evolution of a small, time dependent perturbation η around the Einstein static solution $a = a_0 = c/\sqrt{\Lambda}$, and use this equation to show that the Einstein model is unstable.

Exercise 1.10

The dutch astronomer Willem de Sitter originally published his universe model as an alternative, *static* solution to Einstein's model. In his original solution, the line element is written as

$$ds^2 = \left(1 - \frac{r^2}{R^2}\right) dt^2 - \frac{dr^2}{1 - r^2/R^2} - r^2 d\theta^2 - r^2 \sin^2\theta d\phi^2,$$

where R is a constant. Show that by transforming to a new set of coordinates,

$$\begin{aligned} \overline{r} &= \frac{r}{\sqrt{1 - r^2/R^2}} e^{-t/R}, \\ \overline{t} &= t + \frac{1}{2} R \ln\left(1 - \frac{r^2}{R^2}\right), \end{aligned}$$

the line element can be brought on the form

$$ds^2 = d\overline{t}^2 - e^{2\overline{t}/R}(d\overline{r}^2 + \overline{r}^2 d\theta^2 + \overline{r}^2 \sin^2\theta d\phi^2).$$

Comment on this result.

Exercise 1.11

Derive the expressions (1.51) and (1.55). Discuss the viability of open, flat, and closed dust-only models of the universe given that we know from the oldest stars observed that the age of the universe must be greater than 12 Gyrs.

Exercise 1.12

Consider the solution for a flat universe with negative cosmological constant found in the lecture notes. Find a constraint on Ω_Λ by demanding that the universe must be at least 12 Gyrs.

Exercise 1.13

Discuss the behaviour of universe models situated on the line $\Omega_{\mathrm{m}0} = 0$ in fig. 1.3.

Exercise 1.14

a) Show that $a(t) = ct$ is a solution of the Friedmann equations for a completely empty universe ($\rho = p = 0$) if $k = -1$.

b) Find expressions for the proper distance d_{p} and the angular diameter distance d_{A} as functions redshift z for this model.

Exercise 1.15

For flat EdS and dS models, make i) a t-r and ii) t-d_{P} diagrams showing an observer at the origin, the relevant horizons, and the path of light signals emitted towards the origin by other observers who follow the expansion of the universe.

Exercise 1.16

For the flat ΛCDM model with $\Omega_{\mathrm{m}0} = 0.3$, calculate numerically (using, e.g., MATLAB or MAPLE) the present proper distance to the particle horizon and the event horizon.

Exercise 1.17 (Warning: involves long and tedious calculations!)

Derive Mattig's formula (W. Mattig, Astronomische Nachrichten, 284, 1958, 19) for the luminosity distance in a dust universe, valid for all values of $\Omega_{\mathrm{m}0}$:

$$d_{\mathrm{L}} = \frac{2c}{H_0 \Omega_{\mathrm{m}0}^2} \left[\Omega_{\mathrm{m}0} z + (\Omega_{\mathrm{m}0} - 2)(\sqrt{\Omega_{\mathrm{m}0} z + 1} - 1) \right].$$

Exercise 1.18 (From the exam in AST4220, 2004)

In some model the universe can experience a so-called loitering phase where it spends some time without expanding significantly. We will in this problem define the existence of such a phase by the existence of a redshift $z_{\rm loit} \geq 0$ so that $H'(z_{\rm loit}) = 0$ (where $H'(z) = \frac{dH}{dz}$ and $H = \frac{\dot{a}}{a}$ is the Hubble parameter.)

a) Write down the expression for $H^2(z)$ for a universe containing non-relativistic matter (dust), curvature and a cosmological constant. Express your answer in terms of the present-day density parameters $\Omega_{\rm m0}$, $\Omega_{\rm k0}$ and $\Omega_{\Lambda 0}$.

b) Let $\Omega_{\rm m0} = \frac{1}{2}$ and $\Omega_{\Lambda 0} = 2$. Find $z_{\rm loit}$ and $H(z_{\rm loit})$ for this model. Make a rough, qualitative sketch of $H(z)$ for this case. In the same figure, draw $H(z)$ for a flat universe with $\Omega_{\rm m0} = \frac{1}{2}$.

c) Which one of the two cases in *b*) gives the larger age for the universe? No numerical calculations are required, just try to give a qualitative argument.

d) Show that a loitering phase is impossible in a curved universe which contains only non-relativistic matter. (Hint: remember that we require $z_{\rm loit} \geq 0$.)

e) Show that we can have a loitering phase in a universe with curvature, non-realtivistic matter and a cosmological constant provided that it is closed and $\Omega_{\Lambda 0} \geq \frac{1}{2}\Omega_{\rm m0} + 1$

Exercise 1.19 (From the exam in AST 4220, 2004)

The time evolution of the energy density ρ of a perfect fluid with pressure p is governed by the equation

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

We choose $a_0 = 1$ where a_0 is the scale factor at our present epoch t_0.

a) Find ρ as a function of a for a perfect fluid with equation of state $p = w\rho$, where w is a constant.

b) The deceleration parameter q is defined by

$$q = -\frac{\ddot{a}a}{\dot{a}^2}$$

Determine q for a flat universe which contains a combination of non-relativistic matter and a fluid with equation of state $w = -\frac{1}{3}$.

c) Find q for a universe which contains non-relativistic matter only, but has spatial curvature. Compare with the result in b) and comment.

Exercise 1.20 (Exam preparation)

Do the mid term exams from previous years.

Chapter 2

The early, hot universe

In chapter 1 we learned that in an isotropic and homogeneous universe, there can be no flow of heat through any surface: the univers expands adiabatically. We will start by showing that this immediately leads us to expect the early universe to be a very hot place.

2.1 Radiation temperature in the early universe

We have earlier seen that the universe was dominated by its ultrarelativistic (radiation) component during the first few tens of thousands of years. Then, $a \propto \sqrt{t}$, and $H = \dot{a}/a \propto 1/t$. From your course on statistical physics and thermodynamics you recall that the energy density of a gas of ultrarelativistic particles is proportional to T^4, the temperature to the fourth power. Also, we have found that the variation of the energy density with scale factor for ultrarelativistic particles is $\rho c^2 \propto 1/a^4$. From this, we immediately deduce two important facts:

$$T \propto \frac{1}{a} \propto (1+z), \tag{2.1}$$

$$T \propto \frac{1}{\sqrt{t}}, \tag{2.2}$$

where the last equality follows from the Friedmann equation. This tells us that the temperature of the radiation increases without limit as we go backwards in time towards the Big Bang at $t = 0$. However, there are strong reasons to believe that the physical picture of the universe has to be altered before we reach $t = 0$ and $T = \infty$, so that it is not strictly valid to extrapolate equations (2.1) and (2.2) all the way back to the beginning. The result is based on thermodynamics and classical GR, and we expect *quantum gravity* to be important in the very early universe. We can get an estimate of the energy scale where quantum gravity is important by the following argument: quantum dynamics is important for a particle of mass

m when we probe length scales corresponding to its Compton wave length $2\pi\hbar/(mc)$. General relativistc effects dominate when we probe distances corresponding to the Schwarzschild radius $2Gm/c^2$. Equating the two, we find the characteristic energy scale (neglecting factors of order unity)

$$E_{\rm Pl} = m_{\rm Pl}c^2 \sim \sqrt{\frac{\hbar c^5}{G}} \approx 10^{19}\ {\rm GeV}. \tag{2.3}$$

From Heisenbergs uncertainty principle, the time scale associated with this energy scale is

$$t_{\rm Pl} \sim \frac{\hbar}{E_{\rm Pl}} = \sqrt{\frac{\hbar G}{c^5}} \approx 10^{-43}\ {\rm s}, \tag{2.4}$$

and the corresponding length scale is

$$\ell_{\rm Pl} = \sqrt{\frac{\hbar G}{c^3}} \approx 10^{-35}\ {\rm m}. \tag{2.5}$$

Unfortunately, there is no universally accepted theory of quantum gravity yet. The most common view is that string theory/M-theory holds the key to unlock the secrets of the very early universe, but this framework is still in the making and a lot of work remains before it can be used to make testable predictions for particle physics and the beginning of the universe. Thus, even though the Big Bang model extrapolated back to $t = 0$ says that the universe began at a point of infinite temperature and density, this cannot be looked upon as a prediction of the true state of affairs. New physics is bound to enter the picture before we reach $t = 0$. We really don't know anything about exactly how the universe began, if it began at all. All we can say is that our observable universe has evolved from a very hot and dense phase some 14 billion years ago. One should therefore be very wary of strong philosophical statements based on arguments concerning the Big Bang singularity[1]. It might well not exist.

2.2 Statistical physics: a brief review

If we wish to be more precise about the time-temperature relationship than in the previous section, we need to know how to calculate the statistical properties of gases in thermal equilibrium. The key quantity for doing so for a gas of particles of species i is its distribution function $f_i(\mathbf{p})$. This function tells us what fraction of the particles is in a state with momentum $\mathbf{p}$ at at given temperature T, and it is given by

$$f_i(\mathbf{p}) = \frac{1}{e^{(E_i(p)-\mu_i)/(k_{\rm B}T)} \pm 1}, \tag{2.6}$$

[1] An example of two philosophers battling it out over the implications of the Big Bang model can be found in the book 'Theism, atheism, and Big Bang cosmology' by W. L. Craig and Q. Smith (Clarendon Press, Oxford, 1993)

where $k_{\rm B}$ is Boltzmann's constant, μ_i is the chemical potential of the species, $E_i = \sqrt{\mathbf{p}^2c^2 + m_i^2c^4}$ (m_i is the rest mass of a particle of species i), and the plus sign is for fermions (particles of half-integer spin), whereas the minus sign is chosen if the particles i are bosons (have integer spin). Remember that fermions obey the Pauli principle, which means that any given quantum state can accomodate at most one particle. For bosons, no such restriction applies. Note that $E(p)$ depends only on $p = \sqrt{\mathbf{p}^2}$, and therefore we can write $f_i = f_i(p)$.

Once the distribution function is given, it is easy to calculate equilibrium properties of the gas, like the number density, energy density, and pressure:

$$n_i = \frac{g_i}{(2\pi\hbar)^3}\int f_i(p)d^3p, \tag{2.7}$$

$$\rho_i c^2 = \frac{g_i}{(2\pi\hbar)^3}\int E_i(p)f_i(p)d^3p, \tag{2.8}$$

$$P_i = \frac{g_i}{(2\pi\hbar)^3}\int \frac{p^2}{3E(p)}f_i(p)d^3p, \tag{2.9}$$

where the pressure is denoted by a capital P in this chapter to distinguish it from the momentum p. The quantity g_i is the number of internal degrees of freedom of the particle, and is related to the degeneracy of a momentum state. For a particle of spin S, we normally have $g_i = 2S+1$, corresponding to the number of possible projections of the spin on a given axis. There are, however, important exceptions to this rule. Massless particles, like the photon, are constrained to move at the speed of light. For such particles it turns out that is impossible to find a Lorentz frame where the spin projection vanishes. Thus, for photons, which have $S = 1$, there are only two possible spin projections (polarization states).

It is useful to write the integrals above as integrals over the particle energy E_i instead of the momentum p. Using the relation between energy and momentum,

$$E_i^2 = p^2c^2 + m_i^2c^4,$$

we see that $E_i dE_i = c^2 p dp$, and

$$p = \frac{1}{c}\sqrt{E_i^2 - m_i^2c^4}.$$

Furthermore, since the distribution function depends on p only, the angular part of the integral gives just a factor 4π, and so we get

$$n_i = \frac{g_i}{2\pi^2(\hbar c)^3}\int_{m_ic^2}^{\infty}\frac{(E^2 - m_i^2c^4)^{1/2}EdE}{\exp[(E-\mu_i)/(k_{\rm B}T)]\pm 1}, \tag{2.10}$$

$$\rho_i c^2 = \frac{g_i}{2\pi^2(\hbar c)^3}\int_{m_ic^2}^{\infty}\frac{(E^2 - m_i^2c^4)^{1/2}E^2dE}{\exp[(E-\mu_i)/(k_{\rm B}T)]\pm 1}, \tag{2.11}$$

$$P_i = \frac{g_i}{6\pi^2(\hbar c)^3}\int_{m_ic^2}^{\infty}\frac{(E^2 - m_i^2c^4)^{3/2}dE}{\exp[(E-\mu_i)/(k_{\rm B}T)]\pm 1}. \tag{2.12}$$

We will normally be interested in the limit of non-relativistic particles, corresponding to $m_i c^2/(k_{\rm B}T) \gg 1$, and the ultrarelativistic limit, corresponding to $m_i c^2/(k_{\rm B}T) \ll 1$. We take the latter first, and also assume that $k_{\rm B}T \gg \mu_i$. This assumption, that the chemical potential of the particle species in question is negligible, is valid in most applications in cosmology. This is easiest to see in the case of photons, where one can derive the distribution function in the canonical ensemble (corresponding to fixed particle number, volume, and temperature), with the result that it is of the Bose-Einstein form with $\mu = 0$. With these approximations, let us first calculate the energy density, and consider the case of bosons first. Then

$$\rho_i c^2 \approx \frac{g_i}{2\pi^2(\hbar c)^3} \int_{m_i c^2}^{\infty} \frac{E^3 dE}{\exp(E/k_{\rm B}T) - 1}.$$

We introduce the substitution $x = E/(k_{\rm B}T)$, which gives

$$\rho_i c^2 \approx \frac{g_i (k_{\rm B}T)^4}{2\pi^2(\hbar c)^3} \int_0^{\infty} \frac{x^3 dx}{e^x - 1},$$

where we have taken the lower limit in the integral to be 0, since $m_i c^2 \ll k_{\rm B}T$ by assumption. The integral can be looked up in tables, or you can calculate it yourself in several different ways. For example, we can start by noting that $e^{-x} < 1$ for $x > 0$, so that the expression $1/(1 - e^{-x})$ can be expanded in an infinite geometric series. We can therefore proceed as follows:

$$\begin{aligned}
\int_0^{\infty} \frac{x^3 dx}{e^x - 1} &= \int_0^{\infty} x^3 e^{-x} \frac{1}{1 - e^{-x}} dx \\
&= \int_0^{\infty} x^3 e^{-x} \sum_{n=0}^{\infty} e^{-nx} dx = \sum_{n=0}^{\infty} \int_0^{\infty} x^3 e^{-(n+1)x} dx \\
&= \sum_{n=0}^{\infty} \frac{1}{(n+1)^4} \int_0^{\infty} t^3 e^{-t} dt.
\end{aligned}$$

The last integral is by the definition of the gamma function equal to $\Gamma(4)$. I leave it as an exercise for you to show that $\Gamma(n) = (n-1)!$ when n is a positive integer. Thus the integral is given by

$$\int_0^{\infty} \frac{x^3 dx}{e^x - 1} = 3! \sum_{n=1}^{\infty} \frac{1}{n^4},$$

and the sum is by definition the Riemann zeta function of 4, which can be shown to have the value $\pi^4/90$. Therefore,

$$\int_0^{\infty} \frac{x^3 dx}{e^x - 1} = \Gamma(4)\zeta(4) = 3! \frac{\pi^4}{90} = \frac{\pi^4}{15}.$$

The energy density of a gas of ultrarelativistic bosons is therefore given by

$$\rho_i c^2 = \frac{g_i \pi^2}{30} \frac{(k_{\rm B}T)^4}{(\hbar c)^3}. \tag{2.13}$$

In the case of fermions, we need to evaluate the integral

$$\int_0^\infty \frac{x^3 dx}{e^x + 1}.$$

This can be done by the same method used for the bosonic integral, but the simplest way is to relate the two cases by using the identity

$$\frac{1}{e^x + 1} = \frac{1}{e^x - 1} - \frac{2}{e^{2x} - 1}.$$

Then,

$$\begin{aligned}
\int_0^\infty \frac{x^3 dx}{e^x + 1} &= \int_0^\infty \frac{x^3 dx}{e^x - 1} - 2\int_0^\infty \frac{x^3 dx}{e^{2x} - 1} \\
&= \int_0^\infty \frac{x^3 dx}{e^x - 1} - \frac{2}{2^4}\int_0^\infty \frac{t^3 dt}{e^t - 1} \\
&= \int_0^\infty \frac{x^3 dx}{e^x - 1} - \frac{1}{2^3}\int_0^\infty \frac{x^3 dx}{e^x - 1} \\
&= \frac{7}{8}\int_0^\infty \frac{x^3 dx}{e^x - 1},
\end{aligned}$$

where we have used the substitution $t = 2x$, and the fact that the integration variable is just a 'dummy variable' to rename the integration variable from t to x and thus we have

$$\rho_i c^2 = \frac{7}{8}\frac{g_i \pi^2}{30} \frac{(k_{\rm B}T)^4}{(\hbar c)^3}. \tag{2.14}$$

for ultrarelativistic fermions. The calculation of the number density proceeds along the same lines and is left as an exercise. The result is

$$n_i = \frac{g_i \zeta(3)}{\pi^2} \left(\frac{k_{\rm B}T}{\hbar c}\right)^3 \text{ bosons} \tag{2.15}$$

$$= \frac{3}{4}\frac{g_i \zeta(3)}{\pi^2} \left(\frac{k_{\rm B}T}{\hbar c}\right)^3 \text{ fermions}, \tag{2.16}$$

where $\zeta(3) \approx 1.202$. Furthermore, it is easy to show that the pressure is related to the energy density by

$$P_i = \frac{1}{3}\rho_i c^2, \tag{2.17}$$

for both bosons and fermions.

In the non-relativistic limit, $m_i c^2 \gg k_{\rm B}T$, we expand the energy of a particle in powers of its momentum p, which to second order gives $E_i \approx m_i c^2 + p^2/(2m_i)$, and we find that the particle number density is given by

$$\begin{aligned} n_i &\approx \frac{g_i}{2\pi^2(\hbar c)^3} c^3 \int_0^\infty \frac{p^2 dp}{\exp\left(\frac{m_i c^2 + \frac{p^2}{2m_i} - \mu_i}{k_{\rm B}T}\right) \pm 1} \\ &\approx \frac{g_i}{2\pi^2\hbar^3} \int_0^\infty p^2 \exp\left(\frac{\mu_i - m_i c^2}{k_{\rm B}T}\right) \exp\left(-\frac{p^2}{2m_i k_{\rm B}T}\right) dp \\ &= \frac{g_i}{2\pi^2\hbar^3} \exp\left(\frac{\mu_i - m_i c^2}{k_{\rm B}T}\right) (2m_i k_{\rm B}T)^{3/2} \int_0^\infty x^2 e^{-x^2} dx. \end{aligned}$$

The remaining integral can either be looked up in tables, or be obtained from the more familiar integral $\int_0^\infty \exp(-\alpha x^2)dx = \sqrt{\pi/4\alpha}$ by differentiating on both sides with respect to α. It has the value $\sqrt{\pi}/4$. The final result is then

$$n_i = g_i \left(\frac{m_i k_{\rm B}T}{2\pi\hbar^2}\right)^{3/2} \exp\left(\frac{\mu_i - m_i c^2}{k_{\rm B}T}\right). \tag{2.18}$$

Note that this result is independent of whether the particles are fermions or bosons, i.e., whether they obey the Pauli principle or not. The reason for this is that in the non-relativistic limit the occupation probability for any momentum state is low, and hence the probability that any given state is occupied by more than one particle is negligible. In the same manner one can easily find that

$$\rho_i c^2 \approx n_i m_i c^2, \tag{2.19}$$

$$P_i = n_i k_{\rm B}T, \tag{2.20}$$

again, independent of whether the particles obey Bose-Einstein or Fermi-Dirac statistics. From this we see that

$$\frac{P_i}{\rho_i c^2} = \frac{k_{\rm B}T}{m_i c^2} \ll 1,$$

which justifies our use of $P = 0$ as the equation of state for non-relativistic matter.

In the general case, the energy density and pressure of the universe gets contributions from many different species of particles, which can be both ultrarelativistic, non-relativistic, or something in between. The contribution to the energy density of a given particle species of mass m_i, chemical potential μ_i, and temperature T_i can be written as

$$\rho_i c^2 = \frac{g_i}{2\pi^2} \frac{(k_{\rm B}T_i)^4}{(\hbar c)^3} \int_{x_i}^\infty \frac{(u^2 - x_i^2)^{1/2} u^2 du}{\exp(u - y_i) \pm 1},$$

where $u = E/(k_{\rm B}T_i)$, $x_i = m_i c^2/(k_{\rm B}T_i)$, and $y_i = \mu_i/(k_{\rm B}T_i)$. It is convenient to express the total energy density in terms of the photon temperature, which we will call T, since the photons are still with us, whereas other particles, like muons and positrons, have long since annihilated. We therefore write the total energy density as

$$\rho c^2 = \frac{(k_{\rm B}T)^4}{(\hbar c)^3} \sum_i \left(\frac{T_i}{T}\right)^4 \frac{g_i}{2\pi^2} \int_{x_i}^{\infty} \frac{(u^2 - x_i^2)^{1/2} u^2 du}{\exp(u - y_i) \pm 1}.$$

Similarly, the total pressure can be written as

$$P = \frac{(k_{\rm B}T)^4}{(\hbar c)^3} \sum_i \left(\frac{T_i}{T}\right)^4 \frac{g_i}{6\pi^2} \int_{x_i}^{\infty} \frac{(u^2 - x_i^2)^{3/2} du}{\exp(u - y_i) \pm 1}.$$

We note from our earlier results that the energy density and pressure of non-relativistic particles is exponentially suppressed compared to ultrarelativistic particles. In the early universe (up to matter-radiation equality) it is therefore a good approximation to include only the contributions from ultrarelativistic particles in the sums. With this approximation, and using equations (2.13) and (2.14), we get

$$\rho c^2 \approx \frac{\pi^2}{30} \frac{(k_{\rm B}T)^4}{(\hbar c)^3} \left[\sum_{i=\text{bosons}} g_i \left(\frac{T_i}{T}\right)^4 + \frac{7}{8} \sum_{i=\text{fermions}} g_i \left(\frac{T_i}{T}\right)^4 \right],$$

which we can write compactly as

$$\rho c^2 = \frac{\pi^2}{30} g_* \frac{(k_{\rm B}T)^4}{(\hbar c)^3}, \tag{2.21}$$

where we have defined the *effective number of relativistic degrees of freedom*,

$$g_* = \sum_{i=\text{bosons}} g_i \left(\frac{T_i}{T}\right)^4 + \frac{7}{8} \sum_{i=\text{fermions}} g_i \left(\frac{T_i}{T}\right)^4. \tag{2.22}$$

Since $P_i = \rho c^2/3$ for all ultrarelativistic particles, we also have

$$P = \frac{1}{3}\rho c^2 = \frac{\pi^2}{90} g_* \frac{(k_{\rm B}T)^4}{(\hbar c)^3}. \tag{2.23}$$

Note that g_* is a function of the temperature, but usually the dependence on T is weak.

Using (2.21) in the Friedmann equation gives

$$H^2 = \frac{8\pi G}{3c^2}\rho c^2 = \frac{4\pi^3}{45} \frac{G}{\hbar^3 c^5} g_* (k_{\rm B}T)^4,$$

Using the definition of the Planck energy, $E_{\rm Pl} = \sqrt{\hbar c^5/G}$, we can write this as

$$H = \left(\frac{4\pi^3}{45}\right)^{1/2} g_*^{1/2}(T) \frac{(k_{\rm B}T)^2}{\hbar E_{\rm Pl}} \approx 1.66 g_*^{1/2}(T) \frac{(k_{\rm B}T)^2}{\hbar E_{\rm Pl}}. \tag{2.24}$$

Furthermore, since $H = 1/2t$ in the radiation dominated era, and the Planck time is related to the Planck energy by $t_{\rm Pl} = \hbar/E_{\rm Pl}$, we find

$$\frac{t}{t_{\rm Pl}} = \frac{1}{2}\left(\frac{45}{4\pi^3}\right)^{1/2} g_*^{-1/2}(T) \left(\frac{k_{\rm B}T}{E_{\rm Pl}}\right)^{-2} = 0.301 g_*^{-1/2}(T) \left(\frac{k_{\rm B}T}{E_{\rm Pl}}\right)^{-2}, \tag{2.25}$$

and using the values $E_{\rm Pl} = 1.222 \times 10^{19}$ GeV, $t_{\rm Pl} = 5.391 \times 10^{-44}$ s, we can write this result in the useful form

$$t \approx 2.423 g_*^{-1/2}(T) \left(\frac{k_{\rm B}T}{1\ {\rm MeV}}\right)^{-2}\ {\rm s}. \tag{2.26}$$

Thus, we see that the universe was a few seconds old when the photon temperature had dropped to $k_{\rm B}T = 1$ MeV, if $g_*^{1/2}$ was not much larger than one at that time. The quantity g_* depends on how many particles are ultrarelativistic, their internal degrees of freedom, and their temperature relative to the photon temperature. To get a handle on these factors, we need to make a brief detour into the properties of elementary particles.

2.3 An extremely short course on particle physics

We will confine our attention to the so-called Standard Model of elementary particle physics. This highly successful model should be considered one of the highlights of the intellectual history of the 20th century[2]. The Standard Model summarizes our current understanding of the building blocks of matter and the forces between them. The fermions of the model are the constituents of matter, whereas the bosons transmit the forces between them. Some of the properties of the fermions are summarized in table 1.1. Note that there are three generations of fermions, each heavier than the next. The charges are given in units of the elementary charge e. The fractionally charged particles are the quarks, the building blocks of baryons and mesons, like the neutron, the proton and the pions. The baryons are built from three quarks, whereas the mesons are built from quark-antiquark pairs. The particles with integer charges in the table are called *leptons*. In addition to the properties given in the table, the fermions have important *quantum numbers* which correspond to internal degrees of freedom:

[2]For an introduction at the popular level, I can recommend R. Oerter: 'The theory of almost everything' (Pi Press, New York, 2006). The detailed properties of elementary particles, as well as several highly readable review articles, can be found at the web site of the Particle Data Group: http://pdb.lbl.gov/

Electric charge	$Q = 0$	$Q = -1$	$Q = +2/3$	$Q = -1/3$
1. family	$\nu_e (< 3$ eV)	e (511 keV)	u (1.5-4 MeV)	d (4-8 MeV)
2. family	ν_μ (< 0.19 MeV)	μ (106 MeV)	c (1.15-1.35 GeV)	s (80-130 MeV)
3. family	ν_τ (< 18.2 MeV)	τ (1.78 GeV)	t (170-180 GeV)	b (4.1-4.4 GeV)

Table 2.1: The fermions of the Standard Model. The numbers in the parantheses are the particle rest masses mc^2

- Each quark has three internal degrees of freedom, called *colour*.
- All quarks and leptons have spin 1/2, giving two internal degrees of freedom ($2S + 1 = 2$) associated with spin.
- For each fermion, there is a corresponding antifermion with the same mass and spin, but with the opposite charge.
- Note that the neutrinos are normally approximated as being massless (although we know now that this is not strictly correct). They are the only electrically neutral fermions in the Standard Model, but they have a different charge called *weak hypercharge*, which means that neutrinos and antineutrinos are different particles (at least within the Standard Model). Even though neutrinos have spin 1/2, when they are considered massless they have only one internal degree of freedom associated with the spin. For a given neutrino, only one of two possibilities are realized: either the spin is aligned with the direction of the momentum, or it is anti-aligned. In the first case, we say that they are right-handed, in the second case they are left-handed. In the Standard Model, neutrinos are left-handed, antineutrinos right-handed. This property is closely related to the fact that the weak interaction (the only interactions neutrinos participate in) breaks invariance under parity transformations (reflection in the origin).

Many quantum numbers are important because they are conserved in the interactions between different particles:

- The total spin is always conserved.
- The electric charge is always conserved.
- In the Standard Model, baryon number is under normal circumstances conserved. The baryon number is defined so that the baryon number of any quark is 1/3, and that of its corresponding antiquark is $-1/3$. Thus, e.g., the baryon number of the proton is $+1$, whereas all mesons have baryon number 0.

Particle	Interaction	Mass (mc^2	Electric charge
Photon	Electromagnetic	0	0
Z^0	Weak (neutral current)	91 GeV	0
W^+,W^-	Weak (charged current)	80 GeV	± 1
Gluons, g_i, $i = 1, \ldots, 8$	Strong	0	0

Table 2.2: The gauge bosons of the Standard Model.

- The lepton number is conserved. One can actually define three lepton numbers, one associated with each generation: the electron lepton number, the muon lepton number, and the tau lepton number. Each is defined so that the leptons in each generation has lepton number 1, their corresponding antiparticles have lepton number -1. In all interactions observed so far, each of the three lepton numbers is conserved separately.

The fundamental forces in nature are gravity, electromagnetism, the weak interaction, and the strong interaction. Gravity is special in that it affects all particles, and that it is normally negligible compared with the other forces in elementary particle processes. This is fortunate, since gravity is the only force for which we do not have a satisfactory quantum mechanical description. The other three forces are in the Standard Model mediated by the so-called *gauge bosons*. Some of their properties are summarized in table 1.2. All of the gauge bosons have spin 1, which corresponds to $2S + 1 = 3$ internal degrees of freedom for a massive particle. But since the photon and the gluons are massless, one of these degrees of freedom is removed, so they are left with only two. Since the $W^{\pm}$ and Z^0 bosons are massive, they have the full three internal degrees of freedom.

Of the fermions in the Standard Model, all charged particles feel the electromagnetic force. All leptons participate in weak interactions, but not in the strong interaction. Quarks take part in both electromagnetic, weak, and strong interactions. One of the triumphs of the Standard Model is that it has been possible to find a unified description of the electromagnetic and the weak interaction, the so-called electroweak theory. Efforts to include the strong interaction in this scheme to make a so-called Grand Unified Theory (GUT) have so far met with little success [3].

In addition to the fermions and the gauge bosons, the Standard Model also includes an additional boson, the so-called Higgs boson: A spin-zero particle that is a consequence of a trick implemented in the Standard Model,

[3]Note that some popular accounts of particle physics claim that gravity is included in a GUT. This is wrong, the term is reserved for schemes in which the electromagnetic, the weak, and the strong force are unified. A theory that also includes gravity is often given the somewhat grandiose name 'Theory of Everything' (TOE).

the so-called Higgs mechanism, in order to give masses to the other particles without violating the gauge symmetry of the electroweak interaction. The Higgs boson is still undetected, but if it is to exist without making the Standard Model extremely artificial, its mass should be in a range to be probed at the Large Hadron Collider at CERN.

We can now sum up the total number of degrees of freedom in the Standard Model. For one family of fermions, each of the two quarks in the family has two spin degrees of freedom and three colours, making the contribution from quarks equal to 12. A charged lepton contributes two spin degrees of freedom, whereas the neutrino contributes only 1. The total contribution from the fermions of one family is hence $12 + 2 + 1 = 15$. In addition, each particle in the family has its own antiparticle with the same number of internal degrees of freedom, and there are in total three generations of fermions. Hence, the total number of degrees of freedom for the fermions in the Standard Model is $g^{\rm tot}_{\rm fermions} = 2 \times 3 \times 15 = 90$. As for the bosons, the photon contributes 2 degrees of freedom, $W^{\pm}$ and Z^0 each contribute 3, the eight gluons each contribute 2, whereas the spin-0 Higgs boson only has one internal degree of freedom. The total for the bosons of the Standard Model is hence $g^{\rm tot}_{\rm bosons} = 2 + 3 \times 3 + 8 \times 2 + 1 = 28$. The total number of internal degrees of freedom in the Standard Model is therefore $90 + 28 = 118$.

Having tabulated the masses (which tell us which particles can be considered relativistic at a given temperature) and the number of degrees of freedom of each particle in the Standard Model, we are almost ready to go back to our study of thermodynamics in the early universe. However, we need a few more inputs from particle physics first. Recall that when we apply thermodynamics and statistical mechanics to a system, we assume that it is in thermal equilibrium. This is in general a good approximation for the universe as a whole through most of its history. However, some particle species dropped out of equilibrium at early times and so became decoupled from the rest of the universe. The key to finding out when this happens for a given particle is to compare its total interaction rate with the expansion rate of the Universe. If the interaction rate is much lower than the expansion rate, the particles do not have time enough to readjust their temperature to the temperature of the rest of the universe. The rule of thumb is thus that particles are in thermal equlibrium with other components of the universe if their interaction rate Γ with those components is greater than the expansion rate H.

The interaction rate Γ has units of inverse time and is given by $\Gamma = n\sigma\overline{v}$, where n is the number density, σ is the total scattering cross section, and $\overline{v}$ is the average velocity of the particles in question. The scattering cross section has units of area, and is related to the probability for a given reaction to take place. To learn how to calculate such things for elementary particle interactions involves getting to grips with the machinery of relativistic quantum field theory. However, some of the basic features can be under-

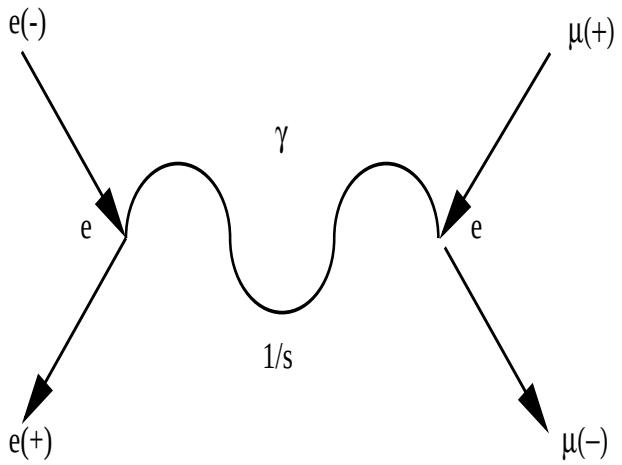

Figure 2.1: Lowest-order Feynman diagram for muon pair production from an electron-positron pair. In the diagram, time flows from left to right.

stood without having to go through all that. What one learns in quantum field theory is that the scattering amplitude for a given process (the cross section is related to the square of the scattering amplitude) can be expanded as a perturbative series in the strength of the interaction governing the process. The perturbation series can be written down pictorially in terms of so-called *Feynman diagrams*, where each diagram can be translated into a mathematical expression giving the contribution of the diagram to the total amplitude. If the interaction strength is weak, it is usually enough to consider the lowest-order diagrams only. An example will make this clearer. Let us consider the (predominantly) electromagnetic process where an electron and a positron annhilate and produce a muon-antimuon pair. The lowest-order diagram for this process is shown in figure 2.1. A point where three lines meet in a diagram is called a vertex. Each vertex gives rise to a factor of the coupling constant describing the strength of the relevant interaction. Since we neglect gravity, there are three different coupling constants which may enter:

- The electromagnetic coupling constant, $g_{\mathrm{EM}} \sim e$, the elementary charge. Often it is replaced with the so-called fine structure constant $\alpha = e^2/(4\pi\epsilon_0\hbar c) \sim 1/137$.
- The weak coupling constant. In the electroweak theory, this is related to the electromagnetic coupling constant, so that $g_{\mathrm{weak}} = e/\sin\theta_W$, where θ_W is the so-called Weinberg angle. From experiments we have $\sin^2\theta_W \approx 0.23$.
- The strong coupling constant, $\alpha_s = g_s^2/4\pi \sim 0.3$.

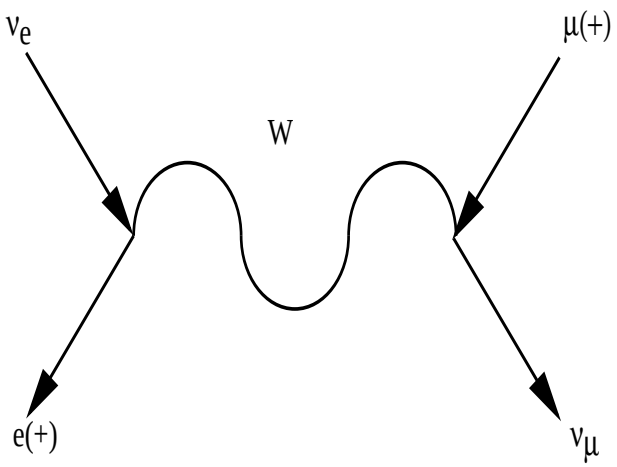

Figure 2.2: Lowest-order Feynman diagram for $\nu_e e^+ \to \nu_\mu \mu^+$.

The line connecting the two vertices, in this case representing a so-called virtual photon, gives rise to a propagator factor $1/(s - m_i^2 c^4)$, where m_i is the mass of the particle in the intermediate state. Here, since the photon is massless, the propagator is simpy $1/s$. The quantity s is the square of the total center-of-mass energy involved in the process. So, apart from some numerical factor, the diagram above, which is the dominating contribution to the cross section, has an amplitude $e \times e \times 1/s$. To find the cross section, we have to square the amplitude, and in addition we have to sum over all initial states of the electrons and all final states of the muons. This gives rise to a so-called phase space factor F. Thus,

$$\sigma \propto F \times \left(\frac{e^2}{s}\right)^2 = F\frac{e^4}{s^2} \propto F\frac{\alpha^2}{s^2}.$$

For center-of-mass energies which are much higher than the rest masses of the particles involved, i.e., for ultrarelativistic particles, the phase space factor can be shown to scale as s, and thus we get

$$\sigma \propto \frac{\alpha^2}{s}.$$

This is often accurate enough for cosmological purposes. A more detailed calculation gives the result $\sigma = 4\pi\alpha^2/(3s)$.

When considering weak interactions, the only important difference from electromagnetic interactions is that the particles mediating this force, the W and Z bosons, are very massive particles. Taking the process $\nu_e e^+ \to \nu_\mu \mu^+$ as an example, the diagram looks as shown in figure 2.2. The two vertices contribute a factor $g_{\rm weak} = e/\sin\theta_W$ each. At the accurace we are working,

we can just take $g_{\rm weak} \sim e$. The propagator gives a factor $1/(s - m_W^2 c^4)$, so that the cross section is

$$\sigma_{\rm weak} \propto F \frac{\alpha^2}{(s - m_W^2 c^4)^2} \sim \frac{\alpha^2 s}{(s - m_W^2 c^4)^2}.$$

For $\sqrt{s} \ll m_W c^2 \approx 80$ GeV, we see that $\sigma_{\rm weak} \sim \alpha^2 s/(m_W c^2)^4$, which is a lot smaller than typical electromagnetic cross sections. Note, however, that at high center-of-mass energies $\sqrt{s} \gg m_W c^2$, the cross section is again of the same order of magnitude as electromagnetic cross sections. This reflects another aspect of electroweak unification: at low energies, the electromagnetic and weak interactions look very different, but at very high energies they are indistinguishable. Note that one often sees weak interaction rates expressed in terms of the so-called Fermi coupling constant, G_F, which is related to the weak coupling constant by

$$\frac{G_F}{\sqrt{2}} = \frac{g_{\rm weak}^2}{8 m_W^2}.$$

We note that if the intermediate state in a Feynman diagram is a fermion, the propagator simply goes as $1/mc^2$ at low energies, where m is the rest mass of the fermion. Thus, for a process like Thomson scattering (photon-electron scattering at low energies), $\gamma e \rightarrow \gamma e$, you can draw the diagram yourself and check that the cross section should scale like α^2/m_e^2.

Finally, to estimate interaction rates, note that for ultrarelativistic particles, $n \propto T^3$, the center-of-mass energy is the typical thermal energy of particles, which is proportional to the temperature, so that $s \propto T^2$, and $\overline{v} \sim c =$ constant. Thus, for a typical weak interaction at energies below the W boson rest mass, where $\sigma \sim \alpha^2 s/(m_W c^2)^4$, we get

$$\Gamma = n \sigma \overline{v} \propto T^3 \times \frac{\alpha^2 T^2}{m_W^4} \propto \frac{\alpha^2 T^5}{m_W^4},$$

which falls fairly rapidly as the universe expands and the temperature drops. For a typical electromagnetic interaction, where $\sigma \propto \alpha^2/s$, we get

$$\Gamma \propto T^3 \times \frac{\alpha^2}{T^2} \propto \alpha^2 T,$$

which drops more slowly with decreasing temperature than the typical weak interaction rate. Note that the proper units for Γ is inverse time. In order to get it expressed in the correct units, we need to insert appropriate factors of $\hbar$, c, and $k_{\rm B}$ in the expressions above. For the weak rate, for example, you can convince yourself that the expression

$$\Gamma = \frac{\alpha^2}{\hbar} \frac{(k_{\rm B} T)^5}{(m_W c^2)^4},$$

has units of inverse seconds.

2.4 Entropy

In situations where we can treat the Universe as being in local thermodynamic equlibrium, the entropy per comoving volume is conserved. To see this, note that the entropy S is a function of volume V and temperature T, and hence its total differential is

$$dS = \frac{\partial S}{\partial V} dV + \frac{\partial S}{\partial T} dT.$$

But from the First Law of thermodynamics, we also have

$$dS = \frac{1}{T}[d(\rho(T)c^2 V) + P(T)dV = \frac{1}{T}\left[(\rho c^2 + P)dV + V\frac{d(\rho c^2)}{dT}dT\right],$$

and comparison of the two expressions for dS gives

$$\begin{aligned} \frac{\partial S}{\partial V} &= \frac{1}{T}(\rho c^2 + P), \\ \frac{\partial S}{\partial T} &= \frac{V}{T}\frac{d(\rho c^2)}{dT}. \end{aligned}$$

From the equality of mixed partial derivatives,

$$\frac{\partial^2 S}{\partial V \partial T} = \frac{\partial^2 S}{\partial T \partial V},$$

we see that

$$\frac{\partial}{\partial T}\left[\frac{1}{T}(\rho c^2 + P)\right] = \frac{\partial}{\partial V}\frac{V}{T}\frac{d(\rho c^2)}{dT},$$

which, after some manipulation, gives

$$dP = \frac{\rho c^2 + P}{T} dT.$$

By using this result and rewriting the First Law as

$$TdS = d[(\rho c^2 + P)V] - VdP,$$

we get

$$TdS = d[(\rho c^2 + P)V] - V\frac{\rho c^2 + P}{T}dT,$$

and hence

$$\begin{aligned} dS &= \frac{1}{T}d[(\rho c^2 + P)V] - (\rho c^2 + P)V\frac{dT}{T^2} \\ &= d\left[\frac{(\rho c^2 + P)V}{T} + \text{const}\right], \end{aligned}$$

so, up to an additive constant,

$$S = \frac{a^3(\rho c^2 + P)}{T}, \tag{2.27}$$

where we have taken $V = a^3$. The equation for energy conservation states that $d[(\rho c^2 + P)V] = V dP$, so that $dS = 0$, which means that the entropy per comoving volume is conserved. It is useful to introduce the *entropy density*, defined as

$$s = \frac{S}{V} = \frac{\rho c^2 + P}{T}. \tag{2.28}$$

Since the energy density and pressure are dominated by the ultrarelativistic particle species at any given time, so is the entropy density. Normalizing everything to the photon temperature T, we have earlier found for bosons that

$$\begin{aligned} \rho_i c^2 &= \frac{\pi^2}{30} g_i \frac{(k_{\rm B}T)^4}{(\hbar c)^3} \left(\frac{T_i}{T}\right)^4, \\ P_i &= \frac{1}{3}\rho_i c^2, \end{aligned}$$

and that the relation between pressure and energy density is the same for fermions, but that there is an additional factor of 7/8 in the expression for the fermion energy density. From equation (2.28) we therefore find that the entropy density can be written as

$$s = \frac{2\pi^2}{45} k_{\rm B} g_{*s} \left(\frac{k_{\rm B}T}{\hbar c}\right)^3, \tag{2.29}$$

where we have introduced a new effective number of degrees of freedom

$$g_{*s} = \sum_{i=\text{bosons}} g_i \left(\frac{T_i}{T}\right)^3 + \frac{7}{8} \sum_{i=\text{fermions}} g_i \left(\frac{T_i}{T}\right)^3. \tag{2.30}$$

In general, $g_{*s} \neq g_*$, but for most of the early history of the universe the difference is small and of little significance.

Since the number density of photons (denoted by n_γ) is

$$n_\gamma = \frac{2\zeta(3)}{\pi^2} \left(\frac{k_{\rm B}T}{\hbar c}\right)^3,$$

we can express the total entropy density in terms of the photon number density as

$$s = \frac{\pi^4}{45\zeta(3)} g_{*s} n_\gamma k_{\rm B} \approx 1.80 g_{*s} n_\gamma k_{\rm B}.$$

The constancy of S implies that $sa^3 = \text{constant}$, which means that

$$g_{*s} T^3 a^3 = \text{constant}, \tag{2.31}$$

As an application of entropy conservation, let us look at what happens with neutrinos as the universe expands. At early times they are in equilibrium with the photons, but as the universe expands, their scattering rate decreases and eventually falls below the expansion rate, and they drop out of equilibrium. A precise treatment of this phenomenon requires the Boltzmann equation from the next section, but a reasonable estimate of the temperature at which this happens can be obtained by equating the scattering rate, given by the typical weak interaction rate discussed earlier,

$$\Gamma = \frac{\alpha^2}{\hbar}\frac{(k_B T)^5}{(m_W c^2)^4},$$

to the Hubble expansion rate

$$H \approx 1.66 g_*^{1/2}(T)\frac{(k_B T)^2}{\hbar E_P l}.$$

This results in

$$k_B T_{\rm dec} = 1.18 g_*^{1/6}(T_{\rm dec})\left[\frac{(m_W c^2)^4}{\alpha^2 E_{\rm Pl}}\right]^{1/3} \approx 4.69 g_*^{1/6}(T_{\rm dec})\ {\rm MeV}.$$

At temperatures of order MeV, the relevant degrees of freedom in the Standard Model are photons, electrons, positrons, and neutrinos. This gives $g_* = 43/4$, and hence

$$k_B T_{\rm dec} \approx 6.97\ {\rm MeV}.$$

What happens to the neutrinos after this? They will continue as free particles and follow the expansion of the universe. Their energies will be redshifted by a factor $a_{\rm dec}/a$, where $a_{\rm dec}$ is the value of the scale factor at $T_{\rm dec}$, and they will continue to follow a Fermi-Dirac distribution with temperature $T_\nu = T_{\rm dec} a_{\rm dec}/a \propto a^{-1}$. Now, conservation of entropy tells us that

$$g_{*s}(aT)^3 = \text{constant},$$

so $T \propto g_{*s}^{-1/3} a^{-1}$ for the particles in the universe still in thermal equilibrium. Hence, the Fermi-Dirac distribution for neutrinos will look like it does in the case when they are in thermal equilibrium with the rest of the universe until g_{*s} changes. This happens at the epoch when electrons and positrons become non-relativistic and annihilate through the process $e^+ + e^- \rightarrow \gamma + \gamma$, at a temperature of $k_{\rm B}T = m_e c^2 \approx 0.511$ MeV. At this temperature, the average photon energy, given roughly by $k_{\rm B}T$, is too small for the collision of two photons to result in the production of an electron-positron pair, which requires an energy of at least twice the electron rest mass. So, after this all positrons and electrons will disappear (except for a tiny fraction of electrons, since there is a slight excess of matter over antimatter in the universe) out of

the thermal history. Before this point, the relativistic particles contributing to g_{*s} are electrons, positrons and photons, giving $g_{*s}(\text{before}) = 2 + \frac{7}{8} \times 2 \times 2 = 11/2$, and after this point only the photons contribute, giving $g_{*s} = 2$. Conservation of entropy therefore gives

$$(aT)_{\text{after}} = \left(\frac{11}{4}\right)^{1/3} (aT)_{\text{before}}.$$

So entropy is transferred from the e^+e^--component to the photon gas, and leads to a temperature increase (or, rather, a less rapid temperature decrease) of the photons. The neutrinos are thermally decoupled from the photon gas, and their temperature follows

$$(aT_\nu)_{\text{before}} = (aT_\nu)_{\text{after}},$$

and thus take no part in the entropy/temperature increase. Therefore, cosmological neutrinos have a lower temperature today than the cosmic photons, and the relation between the two temperatures is given by

$$T_\nu = \left(\frac{4}{11}\right)^{1/3} T. \tag{2.32}$$

2.5 The Boltzmann equation

If we want to study processes involving particle creation, freeze-out of thermal equilibrium etc., it is important to be able to consider systems which are not necessarily in thermal equilbrium. The key equation in this context is the Boltzmann equation.

The Boltzmann equation is trivial when stated in words: the rate of change in the abundance of a given particle is equal to the rate at which it is produced minus the rate at which it is annihilated. Let's say we are interested in calculating how the abundance of particle 1 changes with time in the expanding universe. Furthermore, assume that the only annihilation process it takes part in is by combining with another particle, 2, to form two particles, 3 and 4: $1 + 2 \to 3 + 4$. Also, the reverse process is taking place, $3 + 4 \to 1 + 2$, and in equilbrium the two processes are in balance. The Boltzmann equation which formalizes the statement above looks like this:

$$a^{-3} \frac{d(n_1 a^3)}{dt} = n_1^{(0)} n_2^{(0)} \langle \sigma v \rangle \left[\frac{n_3 n_4}{n_3^{(0)} n_4^{(0)}} - \frac{n_1 n_2}{n_1^{(0)} n_2^{(0)}} \right]. \tag{2.33}$$

In this equation, $n_i^{(0)}$ denotes the number density of species i in thermal equilibrium at temperature T, which we derived in section 2.2, equations (2.15), (2.16), and (2.18), and $\langle \sigma v \rangle$ is the so-called thermally averaged cross section, which basically measures the reaction rate in the medium. Note

that the left-hand side of this equation is of order n_1/t or, since the typical cosmological timescale is $1/H$, $n_1 H$. The right-hand side is of order $n_1^{(0)} n_2^{(0)} \langle \sigma v \rangle$, so we see that if the reaction rate $n_2^{(0)} \langle \sigma v \rangle \gg H$, then the only way for this equation to be fulfilled, is for the quantity inside the square brackets to vanish:

$$\frac{n_3 n_4}{n_3^{(0)} n_4^{(0)}} = \frac{n_1 n_2}{n_1^{(0)} n_2^{(0)}}, \tag{2.34}$$

which therefore can be used when the reaction rate is large compared to the expansion rate of the universe.

Armed with the Boltzmann equation, we can now investigate some of the interesting events in the thermal history of the universe.

2.6 Freeze-out of dark matter

In earlier courses you have (hopefully) seen some of the evidence for the existence of dark matter in the universe: the rotation curves of spiral galaxies, the velocities of galaxies in galaxy clusters, etc. But what is the dark matter? The common-sense option is some form of still-born or dead star: brown dwarfs or black holes. Strange though these objects may be, their origin is in the kind of matter we are familiar with: protons, neutrons and electrons, what we in cosmology call baryonic matter. However, we are confident that the dark matter contributes more than 10 % of the total matter-energy density of the universe, most probably the contribution is around 30 % . And from the cosmic microwave background, we can get a fairly accurate estimate of the total amount of baryonic matter in the universe: it is around 4 % . Thus, there just cannot be enough baryonic dark matter to make up all of the dark matter, and most of it therefore has to be non-baryonic. The neutrinos represent a possible solution, since they are abundant in the universe, are non-baryonic, and are now known to be massive. However, both experimental and cosmological limits on the neutrino masses tell us that they make up at most a couple of percent of the total amount of dark matter. We are forced to the conclusion that the overwhelming amount of dark matter is some substance not yet known to mankind! Particle physicists are, fortunately, willing to provide us with many candidates. In their quest to look beyond the Standard Model and deepen our understanding of it, they have found that a new kind of symmetry, called supersymmetry (or SUSY for short) can provide a very elegant solution of many of the puzzles posed by the Standard Model. However, SUSY dictates that there to each particle of the Standard Model should correspond a SUSY partner with the same mass, but with a spin which differs by 1/2. Thus, the photon should have a massless SUSY partner called the photino with spin 1/2, the electron should have a spin-0 partner of the same mass called the selectron, and so on. But if this were so, we should have seen these partners in the lab a long time ago.

So clearly supersymmetry must be wrong? Well, not exactly. It is possible to keep all the attractive features of SUSY, while still being consistent with what we already know of the particle world by postulating that the supersymmetry is hidden at the energies we have explored so far. We say that SUSY is 'broken', and this can be achieved by assigning higher masses to the SUSY partners. This doesn't sound very convincing, but remember that as a physicist you are allowed to make all kinds of crazy claims if a) your idea has great explanatory power and b) it can be tested experimentally. Fortunately, for supersymmetry to work in the way we want it to, the SUSY partners of the known particles are within reach of present experiments. In fact, if it is not seen at the Large Hadron Collider at CERN, which started operations in 2008, SUSY will probably be dead. Thus, at least to my mind, there is nothing scientifically unsound in taking SUSY seriously and seeing where that leads us.

How can SUSY help us with the dark matter? Well, among all the different new particles it predicts, it also predicts the existence of a lightest supersymmetric particle (LSP). If a quantum number call R-parity is conserved, then the LSP is stable, and hence if it is produced in the early universe, it will stay around forever. This LSP, which in the most studied models is a neutral particle called the neutralino, does not interact directly with electromagnetic radiation, and hence it is a viable dark matter candidate. In addition, it has a cross section typically given by the weak cross section, and as we will see shortly, this very naturally leads to its giving rise to the right amount of dark matter. As a point of terminology, a weakly interacting massive particle like the neutralino is given the acronym 'WIMP'. It is then perhaps predictable that those in favour of black holes, brown dwarfs etc. as the dark matter used the acronym MACHO (Massive Compact Halo Object) for this type of dark matter. However, it has become clear that MACHOs can provide only a small fraction of the required amount of dark matter.

Enough talk, let's get down to some calculations (in the words of the late Indian nuclear physicist Vijay Pandharipande 'Why speculate when you can calculate?'). In the generic WIMP scenario, the WIMPs X can annihilate to light leptons l by the process $X + X \leftrightarrow l + l$. The leptons l will be tightly coupled to the cosmic plasma, and can be taken to be in both kinetic (that is, scattering processes are so rapid that the particle distributions take on their equilibrium forms) and chemical equilibrium (the chemical potentials on each side of the reaction balance). Hence, their number distribution is given by the equilibrium thermal distribution, $n_l = n_l^{(0)}$. Inserting this in equation (2.33), we find

$$a^{-3}\frac{d(n_X a^3)}{dt} = \langle\sigma v\rangle[(n_X^{(0)})^2 - n_X^2],$$

where σ denotes the total annihilation cross section, where we have summed

over all possible leptons l. Now, we know that the temperature scales as $T \propto 1/a$, and that $(aT)^3 \sim S = \text{constant}$, as long as there is no substantial change in the effective number of relativistic degrees of freedom. We can therefore write the left-hand side of the equation as

$$a^{-3}\frac{d}{dt}\left(\frac{n_X}{T^3}a^3T^3\right) = T^3\frac{d}{dt}\left(\frac{n_X}{T^3}\right).$$

Introducing the new variable $Y \equiv n_X/T^3$, we can then rewrite the Boltzmann equation as

$$\frac{dY}{dt} = T^3\langle\sigma v\rangle(Y_{\rm EQ}^2 - Y^2),$$

where $Y_{\rm EQ} = n_X^{(0)}/T^3$. Qualitatively, we expect that at high temperatures the reactions proceed rapidly and the Boltzmann equation can only be satisfied by having $Y = Y_{\rm EQ}$. At low temperatures, when the temperature drops below the rest mass m of the X particle, the abundance becomes suppressed by the exponential factor $\exp(-mc^2/k_{\rm B}T)$, and it then becomes more and more difficult for an X particle to find a partner to annihilate with, and hence they eventually drop out of thermal equilibrium.

It is convenient to rewrite the differential equation using $x \equiv mc^2/k_{\rm B}T$, using the fact that T is proportional to $1/a$, so that $T = C/a$, where C is a constant:

$$\begin{aligned}\frac{d}{dt} &= \frac{dx}{dt}\frac{d}{dx} = \frac{d}{dt}\left(\frac{mc^2}{k_{\rm B}T}\right)\frac{d}{dx} = \frac{da}{dt}\frac{d}{da}(mc^2Ca)\frac{d}{dx} \\ &= \frac{\dot{a}}{a}mc^2Ca\frac{d}{dx} = Hx\frac{d}{dx}.\end{aligned}$$

Furthermore, since we can expect dark matter freeze-out to take place in the radiation-dominated era where $\rho \propto T^4$, we have from the Friedmann equation $H^2 \propto \rho \propto T^4$, and you can convince yourself that we therefore can write

$$H = \frac{H(k_{\rm B}T = mc^2)}{x^2} \equiv \frac{H(m)}{x^2}.$$

It is then straightforward to rewrite the differential equation as

$$\frac{dY}{dx} = \frac{\lambda}{x^2}(Y_{\rm EQ}^2 - Y^2),$$

where

$$\lambda = \left(\frac{mc^2}{k_{\rm B}}\right)^3\frac{\langle\sigma v\rangle}{H(m)}.$$

The thermally averaged annihilation cross section $\langle\sigma v\rangle$ may be temperature dependent, but if we assume it is a constant, we can glean some features of the solution of this equation. For $x \sim 1$, the left-hand-side is $\sim Y$,

whereas the right-hand-side is $\sim \lambda Y^2$. Since λ is typically $\gg 1$, we then have $Y = Y_{\rm EQ}$ for the equation to be fulfilled. The abundance at very late times, Y_∞ can be calculated by noting that at late times, $Y_{\rm EQ}$ drops dramatically, so that $Y \gg Y_{\rm EQ}$, and the equation can be written approximately as

$$\frac{dY}{dx} \approx -\frac{\lambda Y^2}{x^2},$$

valid for $x \gg 1$. We can integrate this equation from freeze-out at x_f up to very late times, $x \to \infty$:

$$\int_{Y_f}^{Y_\infty} \frac{dY}{Y^2} = -\lambda \int_{x_f}^{\infty} \frac{dx}{x^2},$$

and since we can expect $Y_f \gg Y_\infty$, we find the simple relation

$$Y_\infty = \frac{x_f}{\lambda}.$$

After freeze-out, the density of heavy particles decays as a^{-3}, and hence the dark matter energy density today is mc^2 times a_1^3/a_0^3 times the number density of dark matter particles, where a_1 is the scale factor when Y reaches Y_∞, and a_0 is its present value. The number density when $a = a_1$ is given by $Y_\infty T_1^3$. Therefore,

$$\rho_X c^2 = mc^2 \left(\frac{a_1}{a_0}\right)^3 Y_\infty T_1^3 = mc^2 Y_\infty T_0^3 \left(\frac{a_1 T_1}{a_0 T_0}\right)^3.$$

Now, for a similar physical reason to why neutrinos are at a different temperature from the photons today, $a_1 T_1 \neq a_0 T_0$. The photons are heated by the annihilation of heavy particles. Let us assume that $k_{\rm B} T_1 \sim 10$ GeV. At this temperature, the top quark with mass $m_t \approx 178$ GeV has gone non-relativistic and annihilated. The contribution from quarks and antiquarks to the number of degrees of freedom is thus $2 \times 5 \times 3 \times 2 = 60$. From leptons the contribution is $2 \times 6 \times 2 = 24$, from photons 2, and from gluons 8×2. The latter two species are bosons, and so the effective number of relativistic degrees of freedom at $k_{\rm B} T_1 = 10$ GeV is

$$g_{*s}(10\ {\rm GeV}) = 2 + 16 + \frac{7}{8}(30 + 30 + 12 + 12) = 91.5.$$

Today, the only contributions to the number of relativistic degrees of freedom come from photons and neutrinos (here assumed massless). Recalling that the temperature of the neutrinos is lower than the photon temperature, we find

$$g_{*s}(T_0) = 2 + \frac{7}{8} \times 3 \times 2 \times \left(\frac{4}{11}\right)^{4/3} \approx 3.36.$$

Entropy conservation therefore gives

$$\left(\frac{a_1 T_1}{a_0 T_0}\right)^3 = \frac{3.36}{91.5} \approx \frac{1}{30}.$$

So,

$$\rho_X c^2 \approx \frac{mc^2 Y \propto T_0^3}{30},$$

and by dividing by the critical energy density, on can show in the end that

$$\Omega_X = 0.3 h^{-2} \left(\frac{x_f}{10}\right) \left(\frac{g_*(k_{\rm B}T = mc^2)}{100}\right)^{1/2} \frac{10^{-39}\ {\rm cm}^2}{\langle \sigma v \rangle}.$$

This result tells us that a heavy particle with a typical weak cross section will naturaly freeze out with the right density to account for the dark matter in the universe. This is one of the motivations for taking the WIMP scenario seriously.

2.7 Big Bang Nucleosynthesis

One of the biggest successes of the Big Bang model is that it can correctly account for the abundance of the lightest elements (mainly deuterium and helium) in the universe. While the heavy elements we depend on for our existence are cooked in stars, it is hard to account for the abundances of the lightest elements from stellar nucleosynthesis. The early universe turns out to be the natural place for forming these elements, as we will see in this section.

First, a few facts from nuclear physics. A general nucleus consists of Z protons and N neutrons, and is said to have *mass number* $A = Z + N$. The standard notation is to denote a general nucleus X by ${}^A_Z{\rm X}_N$. The number of protons determines the chemical properties of the corresponding neutral atom. Nuclei with the same Z, but with different N are called *isotopes* of the same element. When it is clear from the context what nucleus we are talking about, we sometimes denote the nucleus just by giving its mass number A: ${}^A{\rm X}$. The simplest nucleus is hydrogen, ${}^1_1{\rm H}_0$ (or simply ${}^1{\rm H}$), which is just a proton, p. A proton and a neutron may combine to form the isotope ${}^2{\rm H}$, which is also called the deuteron and denoted by D. One proton and two neutrons form ${}^3{\rm H}$, triton, also denoted by T. The next element is helium, which in its simplest form consists of two protons and one neutron (the neutron is needed for this nucleus to be bound), ${}^3{\rm He}$. By adding a neutron, we get the isotope ${}^4{\rm He}$.

A nucleus X has rest mass $m({}^A_Z{\rm X}_N)$, and its binding energy is defined as the difference between its rest mass energy and the rest mass energy of Z protons and N neutrons:

$$B = [Zm_p + Nm_n]c^2 - m({}^A_Z{\rm X}_N)c^2.$$

Here, $m_p c^2 = 938.272$ MeV is the proton rest mass, and $m_n c^2 = 939.565$ MeV is the neutron rest mass. In many circumstances one can neglect the difference between these two masses and use a common nucleon mass m_N. For the nucleus to exist, B must be positive, i.e., the neutrons and protons must have lower energy when they sit in the nucleus than when they are infinitely separated. Deuterium has a binding energy of 2.22 MeV. The binding energy increases with A up to ^{56}Fe, and after that it decreases. This means that for nuclei lighter than iron, it is energetically favourable to fuse and form heavier elements, and this is the basis for energy production in stars.

The constituents of nuclei, protons and neutrons, are baryons. Since the laws of nature are symmetric with respect to particles and antiparticles, one would naturally expect that there exists an equal amount of antibaryons. As baryons and antibaryons became non-relativistic, they would have annihilated to photons and left us with a universe without baryons and without us. Clearly this is not the case. The laws of nature do actually allow baryons to be overproduced with respect to antibaryons in the early universe, but the detailed mechanism for this so-called baryon asymmetry is not yet fully understood. Since the number of baryons determines the number of nuclei we can form, the baryon number is an important quantity in Big Bang Nucleosynthesis (BBN). It is usually given in terms of the baryon-to-photon ratio, which has the value

$$\begin{aligned} \eta_{\rm b} &= \frac{n_{\rm b}}{n_\gamma} = \frac{n_{\rm b0}(1+z)^3}{n_{\gamma 0}(1+z)^3} = \frac{\rho_{\rm b}/m_N}{n_{\gamma 0}} \\ &= \frac{\rho_{\rm c0}\Omega_{\rm b0}}{n_{\gamma 0} m_N} \approx 2.7 \times 10^{-8} \Omega_{\rm b0} h^2. \end{aligned} \tag{2.35}$$

Since $\Omega_{\rm b0}$ is at most of order 1 (actually it is a few hundreths), we see that photons outnumber baryons by a huge factor.

Given the range of nuclei that exist in nature, one could imagine that following the neutrons and protons and tracing where they end up would be a huge task. However, the problem is simplified by the fact that essentially no elements heavier than ^{4}He are formed. This is because there is no stable nucleus with $A = 5$ from which the building of heavier elements can proceed in steps. Two helium nuclei cannot combine to form the ^{8}Be beryllium nucleus, and proceed from there on to heavier elements, because also this nucleus is unstable. In stars, *three* helium nuclei can combine to form an excited state of ^{12}C, but in the early universe the conditions for this process to proceed are not fulfilled. Also, since ^{4}He has a higher binding energy than D and T, the nucleons will prefer to end up in helium, and thus we need in practice only consider production of helium, at least as a first approximation. However, the formation of the deuteron is an intermediate step on the way to helium. In more detail, the chain of reactions leading to ^{3}He and ^{4}He are

$$d + n \rightarrow {}^3{\rm H} + \gamma$$

$$d + p \rightarrow {}^3\mathrm{He} + \gamma,$$

or

$$\begin{aligned} d + d &\rightarrow {}^3\mathrm{H} + p \\ d + d &\rightarrow {}^3\mathrm{He} + n, \end{aligned}$$

from which one can form ^{4}He as

$$\begin{aligned} {}^3\mathrm{H} + p &\rightarrow {}^4\mathrm{He} + \gamma, \text{ or} \\ {}^3\mathrm{He} + n &\rightarrow {}^4\mathrm{He} + \gamma. \end{aligned}$$

So, the onset of nucleosynthesis is when deuteron production begins. Deuterons are formed all the time in the early universe, but at high temperatures they are immediately broken up by photons with energies equal to the deuteron binding energy 2.22 MeV or higher. Since the mean photon energy is roughly $k_\mathrm{B}T$, one would naively expect that the process of photons breaking up deuterons would become inefficient as soon as $k_\mathrm{B}T \sim 2.22$ MeV. However, this process persists until much lower temperatures are reached. This is because $k_\mathrm{B}T$ is only the *mean* photon energy: there are always photons around with much higher (or lower) energies than this, even though they only make up a small fraction of the total number of photons. But since there are so many more photons than baryons, roughly 10^9 times as many as we saw above, even at much lower temperatures than 2.22 MeV there may be enough high-energy photons around to break up all deuterons which are formed. Let us look at this in more detail. The number density of photons with energy E greater than a given energy E_0 is given by

$$n_\gamma(E \geq E_0) = \frac{1}{\pi^2(\hbar c)^3} \int_{E_0}^{\infty} \frac{E^2 dE}{e^{E/k_\mathrm{B}T} - 1}.$$

We are interested in the situation when $E_0 \gg k_\mathrm{BT}$, and then $e^{E/k_\mathrm{BT}} \gg 1$ in the integrand, so we can write

$$\begin{aligned} n_\gamma(E \geq E_0) &= \frac{1}{\pi^2(\hbar c)^3} \int_{E_0}^{\infty} dE E^2 e^{-E/k_\mathrm{B}T} \\ &= \frac{1}{\pi^2} \left(\frac{k_\mathrm{B}T}{\hbar c}\right)^3 \int_{x_0}^{\infty} x^2 e^{-x} dx \\ &= \frac{1}{\pi^2} \left(\frac{k_\mathrm{B}T}{\hbar c}\right)^3 (x_0^2 + 2x_0 + 2) e^{-x_0}, \end{aligned}$$

where I have introduced the variable $x = E/k_\mathrm{B}T$. Since we have found earlier that the total number density of photons is given by

$$n_\gamma = \frac{2.404}{\pi^2} \left(\frac{k_\mathrm{B}T}{\hbar c}\right)^3,$$

the fraction of photons with energies greater than E_0 is

$$f(E \geq E_0) = 0.416e^{-x_0}(x_0^2 + 2x_0 + 2),$$

where $x_0 = E_0/k_{\rm B}T$. If this fraction is greater than or equal to the baryon-to-photon rato, there are enough photons around to break up all deuterons which can be formed. To be definite, let us take $\eta_{\rm b} = 10^{-9}$. Then deuteron break-up will cease when the temperature drops below the value determined by

$$f(E \geq E_0) = \eta_{\rm b},$$

which gives the equation

$$0.416e^{-x_0}(x_0^2 + 2x_0 + 2) = 10^{-9}.$$

This equation must be solved numerically, and doing so gives $x_0 \approx 26.5$, which means that deuteron break-up by photons is efficient down to temperatures given by

$$k_{\rm B}T = \frac{2.22 \text{ MeV}}{26.5} \approx 0.08 \text{ MeV}.$$

Because of these two facts: essentially no elements heavier than helium, and no production until temperatures below 0.1 MeV, we can split the problem into two parts. First, calculate the neutron abundance at the onset of deuteron synthesis, and then from this calculate the helium abundance.

To calculate the neutron abundance, we must again go by way of the Boltzmann equation. Weak reactions like $p + e^- \leftrightarrow n + \nu_e$ keep the protons and neutrons in equilibrium until temperatures of the order of 1 MeV, but after that one must solve the Boltzmann equation. At these temperatures, neutrons and protons are non-relativistic, and the ratio between their equilibrium number densities is

$$\frac{n_n^{(0)}}{n_p^{(0)}} = \left(\frac{m_p}{m_n}\right)^{3/2} \exp\left[-\frac{(m_n - m_p)c^2}{k_{\rm B}T}\right] \approx e^{-Q/k_{\rm B}T},$$

where $Q = (m_n - m_p)c^2 \approx 1.293$ MeV. For temperatures $k_{\rm B}T \gg Q$, we see that $n_p \approx n_n$, whereas for $k_{\rm B}T \leq Q$, the neutron fraction drops, and would fall to zero if the neutrons and protons were always in equilibrium.

Let us define the neutron abundance as

$$X_n = \frac{n_n}{n_n + n_p}.$$

The Boltzmann equation applied to the generic process $n + \ell_1 \leftrightarrow p + \ell_2$, where ℓ_1 and ℓ_2 are leptons assumed to be in equilibrium (i.e., $n_\ell = n_\ell^{(0)}$), gives

$$a^{-3}\frac{d(n_n a^3)}{dt} = \lambda_{np}(n_p e^{-Q/k_{\rm B}T} - n_n),$$

where $\lambda_{np} = n_\ell^{(0)}\langle\sigma v\rangle$ is the neutron decay rate. We can write the number density of neutrons as $n_n = (n_n + n_p)X_n$, and since the total number of baryons is conserved, $(n_n + n_p)a^3$ is constant, and we can rewrite the Boltzmann equation as

$$\frac{dX_n}{dt} = \lambda_{np}[(1 - X_n)e^{-Q/k_\mathrm{B}T} - X_n].$$

Now, we switch variables from t to $x = Q/k_\mathrm{B}T$, and since $T \propto 1/a$, we get

$$\frac{d}{dt} = \frac{dx}{dt}\frac{d}{dx} = Hx\frac{d}{dx},$$

where

$$H = \sqrt{\frac{8\pi G}{3}\rho},$$

and

$$\rho c^2 = \frac{\pi^2}{30}g_*\frac{(k_\mathrm{B}T)^4}{(\hbar c)^3}.$$

Assuming that $e^\pm$ are still present, we have $g_* = 10.75$. Inserting the expression for the energy density, we can write the Hubble parameter as

$$H(x) = \sqrt{\frac{4\pi^3 G}{45c^2}g_*\frac{Q^4}{(\hbar c)^3}}\frac{1}{x^2} = H(x=1)\frac{1}{x^2},$$

where $H(x = 1) \approx 1.13\ \mathrm{s}^{-1}$. The differential equation for X_n now becomes

$$\frac{dX_n}{dx} = \frac{x\lambda_{np}}{H(x=1)}[e^{-x} - X_n(1 + e^{-x})].$$

To proceed, we need to know λ_{np}. It turns out that there are two processes contributing equally to λ_{np}: $n + \nu_e \leftrightarrow p + e^-$, and $n + e^+ \leftrightarrow p + \overline{\nu}_e$. It can be shown that

$$\lambda_{np} = \frac{255}{\tau_n x^5}(12 + 6x + x^2),$$

where $\tau_n = 885.7$ s is the free neutron decay time. The differential equation can now be solved numerically, with the result shown in figure 2.3. We see that the neutrons drop out of equilibrium at $k_\mathrm{B}T \sim 1$ MeV, and that X_n freezes out at a value ≈ 0.15 at $k_\mathrm{B}T \sim 0.5$ MeV.

On the way from freeze-out to the onset of deuterium production, neutrons decay through the standard beta-decay process $n \to p + e^- + \overline{\nu}_e$. These decays reduce the neutron abundance by a factor e^{-t/τ_n}. The relation between time and temperature found earlier was,

$$t \approx 2.423 g_*^{-1/2}(T)\left(\frac{k_\mathrm{B}T}{1\ \mathrm{MeV}}\right)^{-2}\ \mathrm{s},$$

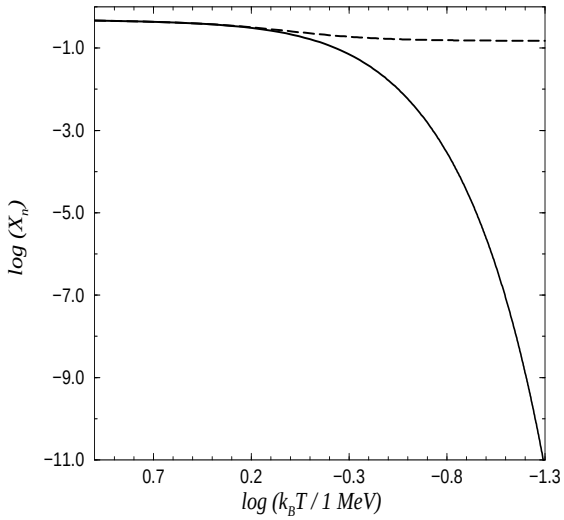

Figure 2.3: Solution of the Boltzmann equation for the neutron abundance (dashed line), along with the equilbrium abundance (full line).

and taking into account that electrons and positrons have now annihilated, we have

$$g_* = 2 + \frac{7}{8} \times 6 \times \left(\frac{4}{11}\right)^{4/3} \approx 3.36.$$

This gives

$$t \approx 132 \left(\frac{0.1 \text{ MeV}}{k_\text{B}T}\right)^2 \text{ s},$$

and by the onset of deuteron production at $k_\text{B}T \approx 0.08$ MeV, this means that the neutron abundance has been reduced by a factor

$$\exp\left[-\frac{132 \text{ s}}{885.7 \text{ s}} \left(\frac{0.1}{0.08}\right)^2\right] \approx 0.79,$$

and hence that at the onset of deuteron production we have $X_n = 0.79 \times 0.15 \approx 0.12$.

We now make the approximation that the light element production occurs instantaneously at the time where deuteron production begins. Since the binding energy of ^{4}He is greater than that of the other light nuclei, production of this nucleus is favoured, and we will assume that all the neutrons go directly to ^{4}He. Since there are two neutrons for each such nucleus, the abundance will be $X_n/2$. However, it is common to define the helium

abundance as

$$X_4 = \frac{4n_{^4\mathrm{He}}}{n_b} = 4 \times \frac{1}{2} X_n = 2X_n,$$

which gives the fraction of mass in ^{4}He. Using our derived value for X_n, we therefore get $X_4 \approx 0.24$. Bearing in mind the simplicity of our calculation, the agreement with more detailed treatments, which give $X_4 \approx 0.22$, is remarkable.

I close this section with a few comments on this result. First of all, we see that the helium abundance depends on the baryon density, mainly through the temperature for the onset of deuteron production, which we found dependend on η_b. A more exact treatment of the problem gives a result which can be fit by the expression

$$X_4 = 0.2262 + 0.0135 \ln(\eta_\mathrm{b}/10^{-10}),$$

so we see that the dependence on the baryon density is weak. By measuring the primordial helium abundance, we can in principle deduce the baryon density of the universe, but since the dependence on η_b is weak, helium is not the ideal probe. Observations of the primordial helium abundance come from the most unprocessed systems in the universe, typically identified by low metallicities. The agreement between theory and observations is excellent.

A more accurate treatment reveals that traces of other elements are produced. Some deuterons survive, because the process $\mathrm{D} + p \rightarrow^3 \mathrm{He} + \gamma$ is not completely efficient. The abundance is typically of order 10^{-4}-10^{-5}. If the baryon density is low, then the reactions proceed more slowly, and the depletion is not as effective. Therefore, low baryon density leads to more deuterium, and the deuterium abundance is quite sensitive to the baryon density. Observations of the deuterium abundance is therefore a better probe of the baryon density than the helium abundance. Measuring the primordial helium abundance typically involves observing absorbtion lines in the spectra of high-redshift quasars. Although this is a field of research bogged by systematic uncertainties, the results indicate a value $\Omega_{\mathrm{b}0} h^2 \approx 0.02$.

There will also be produced a small amount of nuclei with $A = 7$,

$$^4\mathrm{He} +^3 \mathrm{H} \rightarrow^7 \mathrm{Li} + \gamma,$$

and

$$^4\mathrm{He} +^3 \mathrm{He} \rightarrow^7 \mathrm{Be} + \gamma,$$

but these reactions have Coulomb barriers of order 1 MeV, and since the mean nuclear energies at the time of element production are ~ 0.1 MeV and less, these abundances will be small.

The abundance of light elements can also by used to put constraints on the properties and behaviour of elementary particles valid at this epoch in

the history of the universe. An important effect for the helium abundance was the decay of neutrons which reduces the value of the neutron abundance at the onset of deuteron production. This factor depends on the expansion rate of the universe, and if the expansion rate were higher, fewer neutrons would have had time to decay before ending up in helium nuclei, thus increasing the helium abundance. The Hubble parameter is at this epoch proportional to the energy density of relativistic species, and so the helium abundance can be used to constrain the number of relativistic species at the time of Big Bang Nucleosynthesis. Actually, the first constraints on the number of neutrino species N_ν came from this kind of reasoning, and showed that $N_\nu \leq 4$.

2.8 Recombination

The formation of the first neutral atoms is an important event in the history of the universe. Among other things, this signals the end of the age where matter and radiation were tightly coupled, and thus the formation of the cosmic microwave background. For some strange reason, this era is called *recombination*, even though this is the first time electrons and nuclei combine to produce neutral atoms.

We will in this section look exclusively on the formation of neutral hydrogen. A full treatment must of course include the significant amount of helium present, but since one gets the basic picture by focusing on hydrogen only, we will simplify as much as we can. Since the binding energy of the hydrogen atom is $B_{\rm H} = 13.6$ eV, one would guess that recombination should take place at a temperature $k_{\rm B}T = B_{\rm H}$. However, exactly the same effect as in the case of deuteron formation is at work here: since the number of neutral atoms is given by the number of baryons, and the photons outnumber the baryons by a factor of about a billion, even at $k_{\rm B}T$ significantly less than $B_{\rm H}$ there are still enough energetic photons around to keep the matter ionized. Following exactly the same reasoning as in the previous section, one finds that the recombination temperature is given roughly by

$$k_{\rm B}T_{\rm rec} = \frac{B_H}{26.5} \sim 0.5 \text{ eV}.$$

The process responsible for formation of hydrogen is

$$e^- + p \leftrightarrow {\rm H} + \gamma,$$

and as long as this process is in equlibrium, the Boltzmann equation is reduced to

$$\frac{n_e n_p}{n_{\rm H}} = \frac{n_e^{(0)} n_p^{(0)}}{n_{\rm H}^{(0)}}.$$

We note that because of overall charge neutrality, we must have $n_e = n_p$. The number density of free electrons is given by n_e, whereas the total number density of electrons is $n_e + n_{\rm H} = n_p + n_{\rm H}$. The fraction of free electrons is defined as

$$X_e = \frac{n_e}{n_e + n_{\rm H}} = \frac{n_p}{n_p + n_{\rm H}}.$$

The equilibrium number densities are given by

$$\begin{aligned} n_e^{(0)} &= 2\left(\frac{m_e k_{\rm B} T}{2\pi\hbar^2}\right)^{3/2} \exp\left(-\frac{m_e c^2}{k_{\rm B} T}\right), \\ n_p^{(0)} &= 2\left(\frac{m_p k_{\rm B} T}{2\pi\hbar^2}\right)^{3/2} \exp\left(-\frac{m_p c^2}{k_{\rm B} T}\right), \\ n_{\rm H}^{(0)} &= 4\left(\frac{m_{\rm H} k_{\rm B} T}{2\pi\hbar^2}\right)^{3/2} \exp\left(-\frac{m_{\rm H} c^2}{k_{\rm B} T}\right), \end{aligned}$$

where the first factor on the right hand side of these expressions is the number of degrees of freedom. For the ground state of hydrogen, this factor is 4: the proton has spin $\frac{1}{2}$, and the electron with spin $\frac{1}{2}$ has zero angular momentum when hydrogen is in its ground state. The proton and the elctron can then couple to a spin 0 state (which has only one possible value for the total spin projection) or a spin 1 state (which has three), and neglecting the small hyperfine splitting between these two states, this gives a spin degeneracy factor of 4. Substituting these expressions in the equilibrium condition above gives

$$\frac{n_e n_p}{n_{\rm H}} = \left(\frac{m_e k_{\rm B} T}{2\pi\hbar^2}\right)^{3/2} \exp\left(-\frac{B_{\rm H}}{k_{\rm B} T}\right),$$

where in the prefactor the small difference between the mass of the proton and the mass of the hydrogen atom has been neglected, and $B_{\rm H} = m_e c^2 + m_p c^2 - m_{\rm H} c^2$. Using the definition of the free electron fraction, we can now write

$$n_e n_p = (n_e + n_{\rm H})^2 X_e^2,$$

and,

$$n_{\rm H} = (n_e + n_{\rm H})(1 - X_e),$$

and we get the equation

$$\frac{X_e^2}{1 - X_e} = \frac{1}{n_e + n_H}\left(\frac{m_e k_{\rm B} T}{2\pi\hbar^2}\right)^{3/2} \exp\left(-\frac{B_{\rm H}}{k_{\rm B} T}\right).$$

But $n_e + n_{\rm H} = n_p + n_{\rm H} = n_b$, the number density of baryons, which by definition is equal to $\eta_b n_\gamma$, and since the number density of photons is still given by the equilibrium result, we have

$$n_b = \frac{2\zeta(3)}{\pi^2}\left(\frac{k_{\rm B} T}{\hbar c}\right)^3 \eta_b.$$

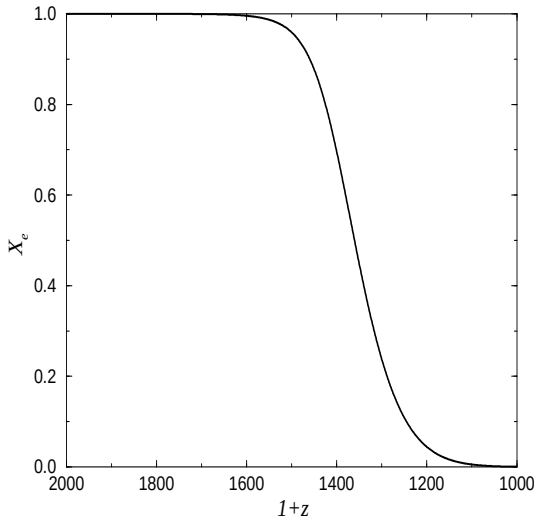

Figure 2.4: The solution of the equation for the free electron fraction in the case $\Omega_{\rm b0}h^2 = 0.02$.

The equation for X_e therefore becomes

$$\begin{aligned}\frac{X_e^2}{1-X_e} &= \frac{1}{4}\sqrt{\frac{\pi}{2}}\frac{1}{\zeta(3)\eta_b}\left(\frac{m_e c^2}{k_{\rm B}T}\right)^{3/2}\exp\left(-\frac{B_{\rm H}}{k_{\rm B}T}\right)\\ &= \frac{0.261}{\eta_b}\left(\frac{m_e c^2}{k_{\rm B}T}\right)^{3/2}\exp\left(-\frac{B_{\rm H}}{k_{\rm B}T}\right).\end{aligned}$$

Since $\eta_b \sim 10^{-9}$, we see that when $k_{\rm B}T \sim B_{\rm H}$, the right hand side of the equation is of order $10^9(m_e c^2/B_{\rm H})^{3/2} \sim 10^{15}$, and since X_e is at most unity, the only way for the equation to be fulfilled is by having $X_e \sim 1$. This reflects what I said in the introduction, namely that recombination takes place at temperatures significantly less than the binding energy of neutral hydrogen. The equation can be solved for various values of $\eta_b = 2.7 \times 10^{-8}\Omega_{\rm b0}h^2$. In figure 2.4 the solution for the free electron fraction is shown as a function of redshift for the canonical value $\Omega_{\rm b0}h^2 = 0.02$. Note that this solution is not accurate once significant recombination starts taking place: as the free electron fraction falls, the rate for recombination also falls, so that eventually the electrons drop out of equilibrium, and the free electron fraction will freeze out at a non-zero value. A full treatment requires the solution of the full Boltzmann equation, but we will not go into that here. The approach above gives a good indication of when the free electron fraction drops significantly, and we see that this takes place at redshifts around $z \sim 1000$. The solution

of the full Boltzmann equation shows that X_e freezes out at a value of a few times 10^{-4}.

During recombination, the scattering rate of photons off electrons drops dramatically. Up to this time, photons could not move freely over very long distances, but after this so-called decoupling of the photons, their mean-free-path became essentially equal to the size of the observable universe. This is therefore the epoch where the universe became transparent to radiation, and the photons present at this stage are observable today as the cosmic microwave background radiation, with a temperature today of about 2.73K.

2.9 Concluding remarks

We have now gone through some of the important epochs in the thermal history of the universe. There are still some 'holes' which need filling out, however. So far, we have only looked at homogeneous models of the universe. But there is clearly some 'clumpiness' in the matter distribution, and the question is how we can account for this. How do structures like clusters of galaxies form? This will be the subject of the final chapter. Another puzzle, as we will see, is why the average density of the universe is so close to the critical one. Most cosmologists believe that the answer to this and some other puzzles in the Big Bang model lies in an epoch of extremely rapid expansion which took place when the universe was around 10^{-35} s old. This epoch is called inflation, and it is the topic of the next chapter.

2.10 Exercises

Exercise 2.1

The Planck mass is $m_{\rm Pl} = E_{\rm Pl}/c^2$, with $E_{\rm Pl}$ given by equation (2.3). Calculate the Planck mass density $\rho_{\rm Pl} = m_{\rm Pl}/\ell_{\rm Pl}^3$. Observations favour a value for $\Omega_{\Lambda 0}$ of 0.7. Calculate $\rho_{\Lambda 0}$, and compare the result with $\rho_{\rm Pl}$.

Exercise 2.2

Prove equations (2.15) and (2.16).

Exercise 2.3

Use conservations laws to determine whether the following reactions are allowed or forbidden.

a) $\nu_\mu + p \rightarrow \mu^+ + n$

b) $\nu_e + p \rightarrow e^- + \pi^+ + p$

c) $\Lambda \rightarrow \pi^+ + e^- + \overline{\nu}_e$

d) $K^+ \rightarrow \pi^0 + \mu^+ + \nu_\mu$.

Recall $\overline{x}$ means the antiparticle of x. Here, π^+ and π^0 are pions with quark content $u\overline{d}$ and $u\overline{u} + d\overline{d}$, respectively. The Λ particle has quark content uds, whereas the kaon K^+ has quark content $u\overline{s}$.

Exercise 2.4

Assume inflation, a period of accelerated expansion in the very early universe which we will discuss in the next chapter, takes place when the temperature of the universe is $k_{\rm B}T = 10^{16}$ GeV. Take $g_* \sim 100$, and estimate the age of the universe when inflation took place.

Exercise 2.5

Show that the number density of cosmic neutrinos today is $n_{\nu,0} = 112 N_\nu \ {\rm cm}^{-3}$, where N_ν is the number of neutrino species. Assume all neutrino species have a mass small enough for the expressions for a relativistic gas to be applicable, and show that the neutrino mass contribution to the present density of the universe is

$$\Omega_{\nu 0} h^2 = \frac{\sum_{i=1}^{N_\nu} m_i c^2}{94 \ {\rm eV}},$$

where m_i is the mass of neutrino species i.

Exercise 2.6

Start from the expressions in (2.11) and (2.12). Assume $\mu_i = 0$, and consider fermions only. Show that by defining variables in an appropriate way, the expressions for the energy density and the pressure can be written as

$$\begin{aligned} \rho_i c^2 &= \frac{g_i (k_{\rm B}T)^4}{2\pi^2 (\hbar c)^3} \int_{x_i}^{\infty} \frac{(u^2 - x_i^2)^{1/2} u^2 du}{e^u + 1}, \\ P_i &= \frac{g_i (k_{\rm B}T)^4}{6\pi^2 (\hbar c)^3} \int_{x_i}^{\infty} \frac{(u^2 - x_i^2)^{3/2} du}{e^u + 1}. \end{aligned}$$

Using MATLAB or whatever tool you deem appropriate to evaluate the integrals numerically, make a plot of the ratio $P_i/\rho_i c^2$. Do the limits $x_i \rightarrow 0$ and $x_i \rightarrow \infty$ agree with your expectations?

Exercise 2.7

Generate plots like the one in figure 2.4 for $\Omega_{\rm b0} h^2 = 0.01$, 0.02, 0.03 and 0.04.

Exercise 2.8

There is a strong link between recombination and *decoupling*: the time when photons stopped interacting significantly with electrons and could move unhindered across the universe. In this exercise you will use the 'scattering rate=expansion rate' critierion to show that this is so. The electrons and photons interact mainly through Thomson scattering with cross section $\sigma_{\rm T} = 0.665 \times 10^{-24}$ cm^2,and their interaction rate is given by $n_e\langle\sigma v\rangle = n_e\sigma_{\rm T}c = X_e n_b \sigma_{\rm T} c$.

a) Explain why the baryon number density can be written as

$$n_b = \frac{\Omega_{\rm b0}a^{-3}}{m_p}\rho_{\rm c0},$$

where m_p is the proton mass.

b) Show that the ratio of the interaction rate to the expansion rate is given by

$$\frac{n_e\sigma_{\rm T}c}{H} = 0.0692 a^{-3} X_e \Omega_{\rm b0} h \frac{H_0}{H}.$$

c) Assuming a universe with matter, radiation, and negligible curvature, show that the ratio in b) can be written as

$$\frac{n_e\sigma_{\rm T}c}{H} = 0.0692 X_e \frac{\Omega_{\rm b0}h^2}{\sqrt{\Omega_{\rm m0}h^2}}(1+z)^{3/2}\left(1+\frac{1+z}{1+z_{\rm eq}}\right)^{-1/2},$$

where $z_{\rm eq}$ is the redshift of matter-radiation equality. Assuming reasonable values for the cosmological parameters, and using $1+z \approx 1000$, show that the photons decouple from the electrons once X_e falls below $\sim 10^{-2}$.

Exercise 2.9 (From the final exam 2003)

a) Write down the equation that determines how the energy density ρ_i of a perfect fluid i varies as the universe expands. Determine how ρ_i depends on the scale factor a and redshift z in the following cases:

1. Dust (with equation of state $p_m = 0$).
2. Radiation (with equation of state $p_r = \frac{\rho_r}{3}$).
3. Cosmological constant (with equation of state $p_\Lambda = -\rho_\Lambda$).

b) Assume that we have three types of neutrinos which are all massless. The energy density of a gas of relativistic bosons is given by

$$\rho = \frac{\pi^2}{30} g \frac{(k_B T)^4}{(\hbar c)^3}$$

where g is the number of internal degrees of freedom for the boson in question. The corresponding expression for fermions is

$$\rho = \frac{7}{8}\frac{\pi^2}{30} g \frac{(k_B T)^4}{(\hbar c)^3}$$

Make use of the fact that the present temperature of the cosmic neutrino background is

$$T_{\nu 0} = \left(\frac{4}{11}\right)^{1/3} T_0$$

where T_0 is the cosmic microwave temperature of 2.73K, and show that the present parameter for radiation (photons plus neutrinos), $\Omega_{r0} = \frac{\rho_{r0}}{\rho_{c0}c^2}$ (where $\rho_{c0} = \frac{3H_0^2}{8\pi G} = 1.879 \times 10^{-26} h^2 \mathrm{kgm}^{-3}$) is given by

$$\Omega_{r0} h^2 = 4.2 \times 10^{-5}$$

where h such that $H_0 = 100 h \mathrm{kms}^{-1}\mathrm{Mpc}^{-1}$

Assuming a value of $\Omega_{m0} = 0.3$ at what redshift z_{eq} was the energy density in the form of realtivistic particles (radiation) equal to the energy density in dust?

c) Show that Friedmann's first equation in the case of a spatially flat universe with dust and radiation can be written as

$$\frac{H^2}{H_0^2} = \frac{\Omega_{r0}}{a^4}\left(1 + \frac{a}{a_{\mathrm{eq}}}\right)$$

where a_{eq} is the scale factor at z_{eq}.

d) Show that the equation in c) can be rewritten as

$$H_0 dt = \frac{a da}{\sqrt{\Omega_{r0}}}\left(1 + \frac{a}{a_{\mathrm{eq}}}\right)^{-1/2}.$$

Integrate this equation and show that

$$H_0 t = \frac{4a_{\mathrm{eq}}^2}{3\sqrt{\Omega_{r0}}}\left[1 - \left(1 - \frac{a}{2a_{\mathrm{eq}}}\right)\left(1 + \frac{a}{a_{\mathrm{eq}}}\right)^{1/2}\right].$$

Useful integral:

$$\int \frac{x dx}{\sqrt{1+x}} = \frac{2}{3}(1+x)^{3/2} - 2(1+x)^{1/2} + C$$

where C is a constant of integration. How old was the universe at $a = a_{\mathrm{eq}}$? How does this compare to the time of decoupling?

e) At what redshift z_Λ is the energy density of dust equal to the vacuum energy density if we assume the universe to be spatially flat with $\Omega_\Lambda = 0.7$, $\Omega_{m0} = 0.3$ today? How old was the universe at z_Λ? How does this compare with the time of decoupling?

f) Which universe components are most important in determining the universe expansion from the last scattering surface to today?

Exercise 2.10 (From the exam in AST4220, 2005)

The temperature of the cosmic microwave background is to lowest order the same in all directions on the sky value $T_0 = 2.73$K. The photons propagated freely through the universe since it became electrically neutral. We will in this problem assume that this happened when the temperature of the photons was 3000 K.

a) Show that $T \propto (1+z)$ and calculate the redshift $z_{\rm dec}$ when the universe became neutral.

b) Write down the Friedmann equation for a spatially flat, matter-dominated universe and use it to show that the present age of the Universe is

$$t_0 = \frac{2}{3H_0},$$

where H_0 is the Hubble parameter. Calculate t_0 for $H_0 = 70\text{kms}^{-1}\text{Mpc}^{-1}$. Use this value of H_0 in the remaining parts of this exercise.

c) Calculate the age $t_{\rm dec}$ of the universe at $z_{\rm dec}$.

d) Calculate $d_P^{\rm PH}(z_{\rm dec})$, the proper distance to the particle horizon at $t_{\rm dec}$.

e) Calculate the proper distance from us out to the redshift $z_{\rm dec}$.

f) Find the angle $\theta_{\rm PH}$ subtended by the particle horizon at $z_{\rm dec}$ on the sky today.

Chapter 3

Inflation

The Big Bang model is extremely successful in accounting for many of the basic features of our universe: the origin of light elements, the formation of the cosmic microwave background, the magnitude-redshift relationship of cosmological objects etc. However, we always want to deepen our understanding and ask further questions. As we will see in the first section, there are several questions we can ask that cannot be answered within the Hot Big Bang model of the universe. These questions indicate that the universe must have started in a very special initial state in order to have the properties that it has today. This does not mean a crisis for the model in the sense that the model is inconsistent, but having the universe start off with fine-tuned initial conditions is not something we like. The idea of inflation, an epoch of accelerated expansion in the very early universe, goes some way towards resolving this issue in that it shows that having an early epoch of accelerated expansion can do away with some of the fine-tuning problems.

3.1 Puzzles in the Big Bang model

Observations tell us that the present universe has a total energy density which is close to the critical one. Why is that so? To see that this is a legitimate question to ask, and indeed a real puzzle, let us consider the first Friedmann equation:

$$\left(\frac{\dot{a}}{a}\right)^2 + \frac{kc^2}{a^2} = \frac{8\pi G}{3}\rho,$$

where ρ is the total energy density. Defining the time-dependent critical density $\rho_{\rm c}(t) = 3H^2/8\pi G$ and the corresponding density parameter $\Omega(t) = \rho(t)/\rho_{\rm c}(t)$, we have after dividing the equation above by H^2:

$$\Omega(t) - 1 = \frac{kc^2}{a^2H^2}.$$

Let us assume that the universe is matter dominated so that $a \propto t^{2/3}$, $H = 2/3t$, and $aH \propto t^{-1/3}$, giving

$$\Omega(t) - 1 \propto t^{2/3}.$$

What are the implications of this equation? It tells us that the deviation of the density from the critical density increases with time. If we have, say $\Omega(t_0) = 1.02$ now, at matter-radiation equality, when $t_{\rm eq} = 47000$ yrs $\sim 1.5 \times 10^{12}$ s, we have

$$\Omega(t_{\rm eq}) - 1 = \left(\frac{1.5 \times 10^{12}}{4.4 \times 10^{17}}\right)^{2/3} (\Omega(t_0) - 1) \sim 4.5 \times 10^{-6}.$$

So the density must have been even closer to the critical one back then. In the radiation-dominated era, $a \propto t^{1/2}$, $H \propto 1/t$, so $aH \propto t^{-1/2}$. At the epoch of Big Bang Nucleosynthesis, $t_{\rm nuc} \sim 60$ s, it then follows that

$$\Omega(t_{\rm nuc}) - 1 = \frac{60}{1.5 \times 10^{12}} \times 4.5 \times 10^{-6} \sim 1.8 \times 10^{-16}.$$

Pushing the evolution back to the Planck time $t_{\rm Pl} \sim 10^{-43}$ s, we find

$$\Omega(t_{\rm Pl}) - 1 \sim 3 \times 10^{-61}.$$

The point of all this numerology is the following: since $1/aH$ is an increasing function of time, the deviation of the density from the critical one also increases with time. This means that in order to have a density close to the critical one today, the density must have been extremely fine-tuned at the beginning of the cosmic evolution. Considering all the possible values the density could have started out with, it seems extremely unlikely that the universe should begin with a value of Ω equal to one to a precision of better than one part in 10^{60}!

The isotropy of the CMB poses another puzzle: we observe that the temperature of the CMB is around 2.7 degrees Kelvin to a precision of about one part in 10^5 across the whole sky. The natural thing to assume is that the physical processes have served to smooth out any large temperature variation that may have existed in the early universe. However, we also know that the size of regions where causal physics can operate is set by the particle horizon. The particle horizon at last scattering, $z_{\rm LSS} \sim 1100$, assuming a matter-dominated universe with negligible spatial curvature, is given by

$$r_{\rm PH}(z_{\rm LSS}) = \int_{z_{\rm LSS}}^{\infty} \frac{cdz}{a_0 H_0 \sqrt{\Omega_{\rm m0}}(1+z)^{3/2}} = \frac{2c}{a_0 H_0 \sqrt{\Omega_{\rm m0}}}(1+z_{\rm LSS})^{-1/2}.$$

The meaning of this number becomes clear if we consider the angular size of this region on the sky today. The radial comoving coordinate of the last scattering surface is given by

$$r(z_{\rm LSS}) = \int_0^{z_{\rm LSS}} \frac{cdz}{a_0 H(z)}.$$

For a spatially flat universe with dust and a cosmological constant, one finds by numerical experiments that to a very good approximation,

$$r(z_{\rm LSS}) \approx \frac{1.94c}{a_0 H_0 \Omega_{\rm m0}^{0.4}}.$$

This gives us the angular size of the particle horizon at last scattering on the sky today as

$$\theta_{\rm PH} = \frac{r_{\rm PH}(z_{\rm LSS})}{r(z_{\rm LSS})} \sim 1.8\Omega_{\rm m0}^{-0.1} \text{ degrees}.$$

How, then, is it possible for regions on the sky today separated by as much as 180 degrees to have almost exactly the same temperature? As in the case of the matter density, nothing prevents us from saying that the uniform temperature was part of the initial conditions of the Big Bang model. However, we might with good reason feel a bit uneasy about having the universe start off in such a special state.

The CMB poses another question for the Big Bang model. Tiny temperature fluctuation have actually been observed, of the order of $\Delta T/T \sim 10^{-5}$. Moreover, they seem to be correlated over scales much larger than the particle horizon at last scattering. How is it possible to set up temperature fluctuations which are correlated on scales which are seemingly causally disconnected? Again, there is nothing to prevent us from making the temperature fluctuations part of the initial conditions of the Big Bang, but most of us would like to have an explanation for why the universe started in such a special state.

Inflation is an attempt at providing a dynamical answer to these question by postulating a mechanism which makes a more general initial state evolve rapidly into a universe like the one we observe. The basic idea can be illustrated by looking at a model we have already considered: that of a universe dominated by vacuum energy.

3.2 The idea of inflation: de Sitter-space to the rescue!

We recall that the de Sitter universe expands at an exponential rate, $a(t) \propto e^{H_0 t}$, where $H_0 = \sqrt{\Lambda/3}$. This gives immediately that $H = H_0$, a constant, and hence $aH \propto e^{H_0 t}$. In contrast to the matter-dominated and radiation-dominated models, we see that $1/aH$ is a decreasing function of time, and

$$\Omega(t) - 1 \propto e^{-2H_0 t}.$$

Thus, if the universe starts off in a de Sitter-like state, any deviations of the density from the critical one will rapidly be wiped out by the expansion.

To put it in geometric terms, if a region of the universe was not spatially flat to begin with, the enormous expansion rate would blow it up and make its radius of curvature infinitesimally small. The horizon problem can also be solved by postulating the existence of de Sitter-expansion in the early universe, because we recall that there is no particle horizon in de Sitter space, and hence no limit on the size of regions which can be causally connected at a given time. The simplest way to think of this is perhaps that the enormous expansion can make a region which is initially small enough for physical conditions to be the same everywhere, but which may possible have a significant spatial curvature, blow up to be an almost flat region of the size of the observable universe.

A numerical example, borrowed from Barbara Ryden's textbook 'Introduction to cosmology' (Addison Wesley, 2003), may serve to make these ideas more precise. Suppose that the universe started out as radiation-dominated, went through a brief period of inflation, after which it returned to radiation-dominated expansion. More specifically, assume that the scale factor is given by

$$\begin{aligned} a(t) &= a_i \left(\frac{t}{t_i}\right)^{1/2}, \ t < t_i \\ &= a_i e^{H_i(t-t_i)}, \ t_i < t < t_f \\ &= a_i e^{H_i(t_f-t_i)} \left(\frac{t}{t_f}\right)^{1/2}, \ t > t_f, \end{aligned}$$

where t_i is the time where inflation starts, t_f is the time inflation ends, and H_i is the Hubble parameter during inflation. We see that in the course of the inflationary epoch, the scale factor grows by a factor

$$\frac{a(t_f)}{a(t_i)} = e^N,$$

where N, the so-called number of e-foldings, is given by

$$N = H_i(t_f - t_i).$$

If the characteristic timescale during inflation, $1/H_i$, is small compared with the duration of inflation, $(t_f - t_i)$, we see that N will be large, and a will increase by a huge factor. To be specific, let us assume that inflation starts at $t_i \sim 10^{-36}$ s, and that $H_i \sim 1/t_i \sim 10^{36}$ s^{-1}, and furthermore that $t_f - t_i \sim 100/H_i \sim 10^{-34}$ s. Then

$$\frac{a(t_f)}{a(t_i)} \sim e^{100} \sim 10^{43}.$$

During the inflationary epoch we will have

$$\Omega(t) - 1 \propto e^{-2H_i(t-t_i)},$$

and so we see that the flatness problem is easily solved: suppose the universe had $\Omega(t_i) - 1 \sim 1$ at the beginning of inflation. The exponential expansion would then drive Ω to be extremely close to 1 at the end of inflation:

$$\Omega(t_f) - 1 = e^{-2N}(\Omega(t_i) - 1) \sim e^{-200} \sim 10^{-87}.$$

The horizon problem is also solved. The proper distance to the particle horizon is at any time given by

$$d_{\rm PH}(t) = a(t) \int_0^t \frac{cdt'}{a(t')},$$

and so it had the size

$$d_{\rm PH}(t_i) = a_i \int_0^{t_i} \frac{cdt}{a_i(t/t_i)^{1/2}} = 2ct_i$$

at the beginning of inflation. At the end of inflation, we find that the proper distance to the particle horizon is given by

$$\begin{aligned} d_{\rm PH}(t_f) &= a_i e^N \left(\int_0^{t_i} \frac{cdt}{a_i(t/t_i)^{1/2}} + \int_{t_i}^{t_f} \frac{cdt}{a_i \exp[H_i(t-t_i)]} \right) \\ &\sim e^N \times c \left(2t_i + \frac{1}{H_i} \right). \end{aligned}$$

Inserting numbers, we find that $d_{\rm PH}(t_i) = 2ct_i \sim 6 \times 10^{-28}$ m. To put this number into perspective, recall that the typical size of an atomic nucleus is 10^{-15} m. The size of the particle horizon immediately after inflation is on the other hand

$$d_{\rm PH}(t_f) \sim e^N \times 3ct_i \sim 2 \times 10^{16} \text{ m} \sim 0.8 \text{ pc!}$$

So, in the course of 10^{-34} s, the size of the particle horizon is increased from a subnuclear to an astronomical scale. The net result is that the horizon size is increased by a factor $\sim e^N$ compared to what it would have been without inflation. After inflation, the horizon size evolves in the usual way, but since it started out enormously larger than in the calculation which lead us to the horizon problem, we see that this problem is now solved. From another point of view, the size of the visible universe today is set by the proper distance to the last scattering surface, and this is given by

$$d_{\rm P}(t_0) \sim 1.4 \times 10^4 \text{ Mpc}.$$

If inflation ended at $t_f \sim 10^{-34}$; s, that corresponds to $a_f \sim 2 \times 10^{-27}$. Thus, at the time inflation ended, the part of the universe currently observable would fit into a sphere of proper size

$$d_{\rm P}(t_f) = a_f d_{\rm P}(t_0) \sim 0.9 \text{ m}.$$

So, immediately after inflation, the observable universe was less than a meter in radius! And even more amazingly, prior to inflation, this region was a factor e^{-N} smaller, which means that its size was

$$d_{\rm P}(t_i) = e^{-N} d_{\rm P}(t_f) \sim 3 \times 10^{-44} \text{ m!}$$

The vast regions of space visible to us thus could have started out as a Planck-length sized nugget! Note also that the size of this region is much smaller than the particle horizon at the beginning of inflation, and thus there is no problem with understanding the isotropy of the CMB.

How many e-foldings of inflation do we need to be consistent with present constraints on the curvature of the universe? Observations of the temperature fluctuations in the CMB provide the most sensitive probe of the spatial geometry, and the best constraint we have at the time of writing comes from the Wilkinson Microwave Anisotropy Probe (WMAP) satellite: $|\Omega(t_0) - 1| \leq 0.02$. Assuming the universe was matter dominated back to $t_{\rm eq}$, we find that

$$|\Omega(t_{\rm eq}) - 1| \leq |\Omega(t_0) - 1| \left(\frac{t_{\rm eq}}{t_0}\right)^{2/3} \sim 0.02 \times \left(\frac{1.5 \times 10^{12} \text{ s}}{4.5 \times 10^{17} \text{ s}}\right)^{2/3} \sim 4.5 \times 10^{-6}.$$

From there and back to the end of inflation, we take the universe to be radiation dominated, and hence

$$|\Omega(t = 10^{-34} \text{ s}) - 1| \leq 4.5 \times 10^{-6} \left(\frac{10^{-34} \text{ s}}{1.5 \times 10^{12} \text{ s}}\right) \sim 3 \times 10^{-52}.$$

Since inflation reduces $|\Omega - 1|$ by a factor $\sim \exp(-2N)$, we find, assuming $|\Omega - 1| \sim 1$ at the beginning of inflation, we need

$$e^{-2N} \sim 3 \times 10^{-52},$$

and hence $N \sim 60$.

So, we see that the idea of an inflationary epoch neatly solves the conundrums of the standard Big Bang model. However, the model we considered here is too simplistic in that it provided no mechanism for inflation to end. If inflation were driven by constant vacuum energy, it would never end, and the Universe would continue to inflate forever. For this reason, one must come up with more detailed models which preserve the nice features of the simple picture painted in this section. The way this is usually done is by introducing one or several so-called scalar fields in the very early universe.

3.3 Scalar fields and inflation

In earlier physics courses you have come across the concept of a field in the form of e.g. the electric and magnetic fields. These are *vector fields*: prescriptions for associating a vector with a given point in space at a given time.

By analogy, a *scalar field* is a rule for associating a real (or complex) number with a point in space at a given time. As a concrete example from everyday life, the temperature of the Earth's atmosphere can be considered a scalar field. Scalar fields also appear in theoretical particle physics. The most famous example is the Higgs field which is introduced in the electroweak theory to provide the elementary particles with rest masses. Sadly, none of the fundamental scalar fields which have been introduced in particle physics and cosmology have been observed. However, a detection of the Higgs boson (the particle associated with the Higgs field) may be just around the corner. If it exists, it will probably be found when the Large Hadron Collider starts operating at CERN in 2008.

The main thing we need to know about a scalar field is that it has a kinetic and a potential energy associated with it, and hence an energy density and a pressure. We will in the following consider a homogeneous scalar field ϕ. Homogeneity means that ϕ is a function of time only, not of the spatial coordinates. Then, measuring ϕ in units of energy, the energy density of the field is given by

$$\rho_\phi c^2 = \frac{1}{2\hbar c^3}\dot{\phi}^2 + V(\phi), \tag{3.1}$$

and the pressure is given by

$$p_\phi = \frac{1}{2\hbar c^3}\dot{\phi}^2 - V(\phi), \tag{3.2}$$

where $V(\phi)$ is the potential energy of the field. One important thing you should note is that if the field varies slowly in time, in the sense that

$$\frac{\dot{\phi}^2}{2\hbar c^3} \ll V(\phi),$$

then the scalar field will have an equation of state given approximately by $p_\phi = -\rho_\phi c^2$, and it will behave like a cosmological constant. This is the key idea behind using a scalar field to drive inflation.

We will assume that the scalar field dominates the energy density and pressure of the universe, and that we can neglect the curvature (which will be driven rapidly to zero anyway if inflation works the way it is supposed to). The first of the Friedmann equations then reads

$$H^2 = \frac{8\pi G}{3c^2}\rho_\phi c^2 = \frac{8\pi G}{3c^2}\left(\frac{1}{2\hbar c^3}\dot{\phi}^2 + V(\phi)\right). \tag{3.3}$$

As the second equation to use, we will choose the adiabatic expansion equation

$$\dot{\rho}c^2 = -3H(\rho c^2 + p).$$

From equation (3.1) we get

$$\dot{\rho_\phi} c^2 = \frac{\dot{\phi}\ddot{\phi}}{\hbar c^3} + \frac{dV}{d\phi}\dot{\phi},$$

and from (3.1) and (3.2) we see that $\rho_\phi c^2 + p_\phi = \dot{\phi}^2/(\hbar c^3)$. Hence, the equation for the scalar field becomes

$$\ddot{\phi} + 3H\dot{\phi} + \hbar c^3 V'(\phi) = 0, \tag{3.4}$$

where $V'(\phi) = dV/d\phi$. This equation is very interesting, because it is an exact analog to the equation of motion of a particle of unit mass moving along the x-axis in a potential well $V(x)$, and subject to a frictional force proportional to its velocity $\dot{x}$. Newton's second law applied to the motion of this particle gives

$$\ddot{x} = -b\dot{x} - V'(x),$$

that is $\ddot{x} + b\dot{x} + V'(x) = 0$. So we can think of ϕ as the coordinate of a particle rolling down the potential $V(\phi)$ and with a frictional force $3H\dot{\phi}$ supplied by the expansion of the universe. In the more familiar classical mechanics example, you may recall that the particle will reach a terminal velocity when $\ddot{x} = 0$, given by $\dot{x} = -V'(x)/b$. After this point, the particle will move with constant velocity. Similarly, at some point the scalar field will settle down to motion down the potential at constant 'velocity' given by $3H\dot{\phi} = -\hbar c^3 V'(\phi)$, that is,

$$\dot{\phi} = -\frac{\hbar c^3}{3H}\frac{dV}{d\phi}.$$

Let us assume that the field has reached this terminal velocity. We will have inflation if the energy of the field behaves like a cosmological constant, and we have seen that the criterion for this is $\dot{\phi}^2 \ll \hbar c^3 V$. Inserting the terminal velocity for the scalar field in this criterion gives

$$\left(\frac{dV}{d\phi}\right)^2 \ll \frac{9H^2 V}{\hbar c^3}.$$

Since the potential energy of the scalar field dominates if this condition is fulfilled, the Hubble parameter is given by

$$H^2 = \frac{8\pi G}{3c^2}V,$$

and inserting this in the condition above gives

$$\left(\frac{dV}{d\phi}\right)^2 \ll \frac{24\pi G}{\hbar c^5}V^2 = \frac{24\pi}{E_{\mathrm{Pl}}^2}V^2,$$

or

$$\frac{2}{3}\frac{E_{\rm Pl}^2}{16\pi}\left(\frac{V'}{V}\right)^2 \ll 1.$$

It is usual to define the so-called *slow-roll parameter* ϵ by

$$\epsilon = \frac{E_{\rm Pl}^2}{16\pi}\left(\frac{V'}{V}\right)^2, \tag{3.5}$$

and we see that the condition above becomes $\epsilon \ll 1$. It is also possible to derive a further condition, this time on the curvature of the potential V'', related to the fact that inflation must last for a sufficiently long time. We will not have $\ddot{\phi} = 0$ all the time, but as long as $\ddot{\phi} \ll \hbar c^3 V'(\phi)$, we can ignore it in the equation of motion for the scalar field. From $3H\dot{\phi} = -\hbar c^3 V'(\phi)$ we get

$$3H\ddot{\phi} = -\hbar c^3 V''(\phi)\dot{\phi},$$

where we have used that H is approximately constant during inflation. This relation then gives

$$\ddot{\phi} = -\hbar c^3 \frac{\dot{\phi}}{3H} V''(\phi),$$

and using

$$\dot{\phi} = -\frac{\hbar c^3}{3H} V'(\phi),$$

we find

$$\ddot{\phi} = \frac{(\hbar c^3)^2}{9H^2} V'V'',$$

so the condition on $\ddot{\phi}$ becomes

$$\frac{\hbar c^3}{9H^2} V'V'' \ll V',$$

i.e.,

$$\frac{\hbar c^3}{9H^2} V'' \ll 1.$$

But, since $H^2 = 8\pi G V/3c^2$, this can be rewritten as

$$\frac{\hbar c^3}{9}\frac{3c^2}{8\pi G}\frac{V''}{V} \ll 1,$$

or,

$$\frac{1}{3}\frac{E_{\rm Pl}^2}{8\pi}\frac{V''}{V} \ll 1.$$

Defining

$$\eta = \frac{E_{\rm Pl}^2}{8\pi}\frac{V''}{V}, \tag{3.6}$$

the condition can be written (since V'' in principle can be negative)

$$|\eta| \ll 1.$$

When $\epsilon \ll 1$ and $|\eta| \ll 1$ the equations (3.3) and (3.4) reduce to

$$H^2 \approx \frac{8\pi G}{3c^2} V(\phi) \tag{3.7}$$

$$3H\dot{\phi} \approx -\hbar c^3 V'(\phi). \tag{3.8}$$

These two equations are called the slow-roll approximation (SRA). The conditions $\epsilon \ll 1$ and $|\eta| \ll 1$ are necessary for this approximation to be applicable (in most normal cases they are also sufficient). One of the nice features is that if the condition on ϵ is fulfilled, then inflation is guaranteed to take place. To see this, note that inflation takes place if $\ddot{a} > 0$, and hence $\ddot{a}/a > 0$ (since a is positive). Since

$$\dot{H} = \frac{d}{dt}\left(\frac{\dot{a}}{a}\right) = \frac{\ddot{a}}{a} - H^2,$$

this condition can be reformulated as

$$-\frac{\dot{H}}{H^2} < 1.$$

By taking the time derivative of equation (3.7) we get $2H\dot{H} = 8\pi G V'\dot{\phi}/3c^2$, so

$$\dot{H} = \frac{4\pi G}{3c^2} V' \frac{\dot{\phi}}{H}.$$

We can find $\dot{\phi}/H$ by dividing (3.8) by (3.7):

$$\frac{3H\dot{\phi}}{H^2} = -\hbar c^3 \frac{3c^2}{8\pi G}\frac{V'}{V},$$

which gives

$$\frac{\dot{\phi}}{H} = -\frac{\hbar c^5}{8\pi G}\frac{V'}{V} = -\frac{E_{\rm Pl}^2}{8\pi}\frac{V'}{V}.$$

By inserting this in the expression for $\dot{H}$ above, we find

$$\dot{H} = -\frac{4\pi G}{3c^2}\frac{E_{\rm Pl}^2}{8\pi}\frac{(V')^2}{V}.$$

If we now use equation (3.7) again, we get

$$-\frac{\dot{H}}{H^2} = \frac{4\pi G}{3c^2}\frac{3c^2}{8\pi G}\frac{1}{V}\frac{E_{\rm Pl}^2}{8\pi}\frac{(V')^2}{V} = \frac{E_{\rm Pl}^2}{16\pi}\left(\frac{V'}{V}\right)^2 = \epsilon,$$

and so we see that $\ddot{a} > 0$ if $\epsilon < 1$. In scalar field models of inflation, $\epsilon = 1$ is usually taken to mark the end of inflation.

Within the SRA we can derive a useful expression for the number of e-foldings that have taken place at a given time t. This number is defined as

$$N = \ln\left[\frac{a(t_{\rm end})}{a(t)}\right], \tag{3.9}$$

where $t_{\rm end}$ is the time when inflation ends. Note that defined this way, N measures how many e-foldings are left until inflation ends, since we see that $N(t_{\rm end}) = 0$, and when $t = t_i$, at the start of inflation, $N(t_i) = N_{\rm tot}$, the total number of e-foldings produced by inflation. Thus, N is a decreasing function of time. Since $\int \dot{a}dt/a = \int da/a = \ln a$, we can write

$$N(t) = \int_t^{t_{\rm end}} H(t)dt,$$

and by dividing (3.7) by (3.8) we get

$$N(t) = -\frac{8\pi}{E_{\rm Pl}^2}\int_t^{t_{\rm end}} \frac{V}{V'}\dot{\phi}dt = \frac{8\pi}{E_{\rm Pl}^2}\int_{\phi_{\rm end}}^{\phi} \frac{V}{V'}d\phi, \tag{3.10}$$

where $\phi_{\rm end} = \phi(t_{\rm end})$ can be found from the criterion $\epsilon(\phi_{\rm end}) = 1$.

3.3.1 Example: inflaction in a ϕ^2 potential

Let us look at an example. We will consider inflation driven by the evolution of a scalar field with potential energy

$$V(\phi) = \frac{1}{2}\frac{m^2c^4}{(\hbar c)^3}\phi^2,$$

and hence an energy density

$$\rho_\phi c^2 = \frac{1}{2}\frac{1}{\hbar c^3}\dot{\phi}^2 + \frac{1}{2}\frac{m^2c^4}{(\hbar c)^3}\phi^2.$$

The ground state for the field is the state of minimum energy, which in this case is given by the field being at rest ($\dot{\phi} = 0$) at the bottom of the potential well at $\phi = 0$ ($V(\phi = 0) = 0$, see figure 3.1.) We imagine that for some reason, the field starts out at a large, non-zero value ϕ_i, and hence with a large potential energy. Similarly to a ball being released from far up the side of a hill, the scalar field will try to 'roll down' to the minimum energy state at $\phi = 0$. If it rolls sufficiently slowly, the potential energy can be treated as essentially constant for a significant portion of the way down to the minimum, and hence the universe will inflate. The slow-roll conditions involve the parameters ϵ and η, so let us start by evaluating them:

$$\epsilon = \frac{E_{\rm Pl}^2}{16\pi}\left(\frac{V'}{V}\right)^2 = \frac{E_{\rm Pl}^2}{4\pi\phi^2},$$

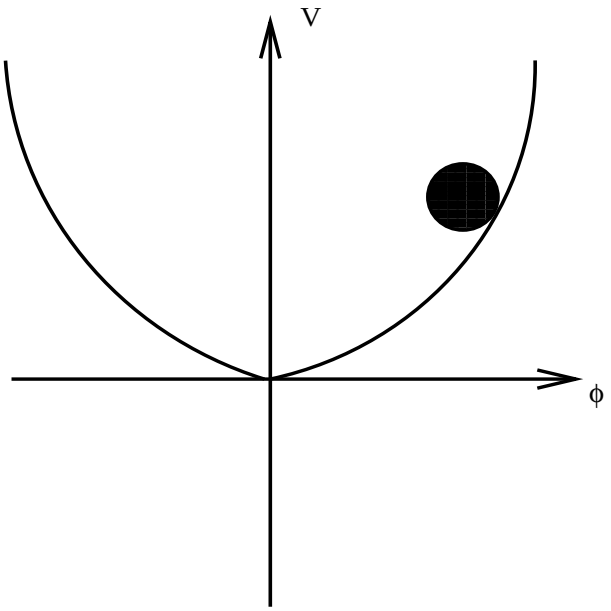

Figure 3.1: The inflaton depicted as a ball rolling down a potential well.

and

$$\eta = \frac{E_{\rm Pl}^2}{8\pi}\frac{V''}{V} = \frac{E_{\rm Pl}^2}{4\pi\phi^2} = \epsilon.$$

The criterion for the SRA to be valid hence becomes

$$\phi \gg \frac{E_{\rm Pl}}{2\sqrt{\pi}} \equiv \phi_{\rm end},$$

and inflation will be over when $\phi \sim \phi_{\rm end}$.

Inserting the potential in the SRA equations (3.7) and (3.8) gives

$$\begin{aligned} H^2 &= \frac{4\pi G}{3}\frac{m^2c^2}{(\hbar c)^3}\phi^2 = \frac{4\pi}{3}\frac{m^2c^4}{\hbar^2}\frac{\phi^2}{E_{\rm Pl}^2} \\ 3H\dot{\phi} &= -\frac{m^2c^4}{\hbar^2}\phi. \end{aligned}$$

Taking the square root of the first equation and inserting it in the second, we get

$$\sqrt{12\pi}\frac{mc^2}{\hbar}\frac{\dot{\phi}\phi}{E_{\rm Pl}} + \frac{m^2c^4}{\hbar^2}\phi = 0,$$

i.e.,

$$\dot{\phi} = -\frac{E_{\rm Pl}}{\sqrt{12\pi}}\frac{mc^2}{\hbar},$$

which can be trivially integrated to give

$$\phi(t) = \phi_i - \frac{mc^2 E_{\rm Pl}}{\hbar\sqrt{12\pi}}t,$$

where for convenience we take inflation to begin at $t_i = 0$. Inserting this result in the equation for H, we get

$$H = \sqrt{\frac{4\pi}{3}}\frac{mc^2}{\hbar E_{\rm Pl}}\left(\phi_i - \frac{mc^2 E_{\rm Pl}}{\hbar\sqrt{12\pi}}t\right),$$

and since $H = \dot{a}/a = da/adt$, we get

$$\int_{a_i}^{a(t)}\frac{da}{a} = \sqrt{\frac{4\pi}{3}}\frac{mc^2}{\hbar E_{\rm Pl}}\int_0^t\left(\phi_i - \frac{mc^2 E_{\rm Pl}}{\hbar\sqrt{12\pi}}t\right)dt,$$

and finally,

$$a(t) = a_i \exp\left[\sqrt{\frac{4\pi}{3}}\frac{mc^2}{\hbar E_{\rm Pl}}\left(\phi_i t - \frac{mc^2 E_{\rm Pl}}{2\hbar\sqrt{12\pi}}t^2\right)\right].$$

We can find the total number of e-foldings produced for a given initial field value ϕ_i by using (3.10):

$$N = \frac{8\pi}{E_{\rm Pl}^2}\int_{\phi_{\rm end}}^{\phi_i}\frac{V d\phi}{V'} = \frac{8\pi}{E_{\rm Pl}^2}\int_{E_{\rm Pl}/\sqrt{4\pi}}^{\phi_i}\frac{1}{2}\phi d\phi = \left(\frac{\phi_i\sqrt{2\pi}}{E_{\rm Pl}}\right)^2 - \frac{1}{2}.$$

As we have seen earlier, we need about 60 e-foldings for inflation to be useful. This gives a condition on the initial value of ϕ in this model: $N = 60$ requires

$$\phi_i = \frac{11}{2\sqrt{\pi}}E_{\rm Pl} \approx 3.10 E_{\rm Pl}.$$

Now, I have said earlier that we don't know the correct laws of physics when the energy of the system reaches the Planck energy and beyond. It seems we may be in trouble then, since the field has to start out at a value greater than $E_{\rm Pl}$ in this model. However, the value of the field is in itself of little consequence, it is not directly observable. As long as the energy density, given by $V(\phi_i)$, is less than the Planck energy density, $E_{\rm Pl}/l_{\rm Pl}^3$, we should be in business. This can be achieved by choosing the mass of the field, m, low enough. How low? The value of the potential is

$$V(\phi_i) = \frac{1}{2}\frac{m^2c^4}{(\hbar c)^3}\phi_i^2 = \frac{121}{8\pi}\frac{E_{\rm Pl}^2 m^2 c^4}{(\hbar c)^3}.$$

This should be compared to the Planck energy density

$$\rho_{\rm Pl}c^2 = \frac{c^7}{\hbar G^2},$$

and $V(\phi_i)$ will therefore be much less than $\rho_{\rm Pl}c^2$ if m satisfies

$$mc^2 \ll \left[\frac{(\hbar c)^3}{E_{\rm Pl} l_{\rm Pl}^3}\right]^{1/2} = E_{\rm Pl}.$$

Therefore, as long as the mass of the scalar field is much smaller than the Planck mass, we should be safe.

3.3.2 Reheating

Once the slow-roll conditons have broken down, the scalar field will start oscillating about the minimum of the potential. In the example with $V(\phi) \propto \phi^2$ above, the field will speed up as it approaches the minimum, and then go into a phase where it oscillates around $\phi = 0$. Since energy is conserved, you might think that the field would bounce back up to the value from which it started, but the friction term $3H\dot{\phi}$ in its equation of motion (3.4) means that the field will lose energy and the oscillations will be damped.

So far we have assumed that the scalar field is free. However, realistically it will be coupled to other fields and particles. These couplings can be modelled as an additional friction term $\Gamma\dot{\phi}$ in the equation of motion of the scalar field. Thus, the energy originally stored in the inflaton field will go into creating the particles that we know and love. This process, where the scalar field undergoes damped oscillations and transfers its energy back into 'normal' particles is called *reheating*. After the reheating phase, the universe will enter a radiation-dominated era and will evolve as in the standard Big Bang model.

3.4 Fluctuations

So far we have assumed that the scalar field responsible for inflation is homogeneous. But quantum mechanics limits how homogeneous the field can be. The Heisenberg uncertainty principle for energy and time limits how precisely we can know the value of the field in a given time interval, and as a consequence of this inflation will begin and end at different times in different regions of space. We will soon show that this leads to perturbations in the energy density. This is an important result, because these perturbations may have been the seeds of the density perturbations that later became the large-scale structures in our Universe.

The Heisenberg uncertainty principle for energy and time states that in the time interval Δt the precision ΔE with which the energy of a system can be measured is limited by

$$\Delta t \Delta E \sim \hbar.$$

Inflation takes place at an energy scale which I will denote by mc^2. For a quadratic inflaton potential, m is the mass of the field. There is, unfortunately, at the moment no theory that predicts the value of mc^2, but it is widely believed that the GUT scale 10^{15} GeV is where the action is. I will first consider the time just before inflation starts. The typical energy per particle is then $k_{\rm B}T \sim mc^2$, and from the relationship between temperature and time in the early universe (derived in chapter 2) I find

$$k_{\rm B}T = mc^2 \sim E_{\rm Pl}\sqrt{\frac{t_{\rm Pl}}{t}},$$

so that

$$t \sim \frac{\hbar E_{\rm Pl}}{m^2c^4}.$$

The order of magnitude of the fluctuations in the energy per particle is therefore

$$\Delta E \sim \frac{\hbar}{t} \sim \frac{m^2c^4}{E_{\rm Pl}},$$

and the relative fluctuations have amplitude

$$\frac{\Delta E}{E} \sim \frac{1}{mc^2}\frac{m^2c^4}{E_{\rm Pl}} \sim \frac{mc^2}{E_{\rm Pl}}.$$

The energy density is given by $\rho \propto T^4 \propto E^4$, and so I find

$$\frac{\Delta\rho}{\rho} \sim \frac{d\rho}{\rho} \sim \frac{1}{E^4}4E^3dE \sim \frac{dE}{E} \sim \frac{\Delta E}{E},$$

so that the fluctuations in the energy density are of the same order of magnitude as the fluctuations in the energy per particle. Notice that the amplitude of the fluctuations depends on the energy scale m of inflation. If this was the whole truth, we could have determined this energy scale by measuring the amplitude of the fluctuations. In reality things are unfortunately not that simple. As I will show next, a more detailed estimate of the amplitude shows that it depends on both the inflaton potential V and its derivative.

Fluctuations in the scalar field ϕ arise because inflation ends at different times in different patches of the universe. If I consider two patches where inflation ends within a time interval Δt, I can write

$$|\Delta\phi| = |\dot{\phi}|\Delta t,$$

so that

$$\Delta t = \left|\frac{\Delta\phi}{\dot{\phi}}\right|.$$

A more careful treatment of the time development of the density perturbations shows that the most important quantity is their amplitude as they

cross the horizon during inflation. This amplitude is determined by the difference in the amount by which the two patches have expanded,

$$\frac{\Delta\rho}{\rho} \sim H\Delta t \sim H\left|\frac{\Delta\phi}{\dot{\phi}}\right|.$$

The first equality above may not seem obvious, so I will try to justify it. I compare to volume elements containing the same total energy U. In the course of the inflationary epoch one element is stretched by a factor a, the other by $a + \Delta a$. This leads to a difference in energy density after inflation given by

$$\begin{aligned}
\Delta\rho &= \frac{U}{a^3} - \frac{U}{(a+\Delta a)^3} \\
&= E\left(\frac{1}{a^3} - \frac{1}{(a+\dot{a}\Delta t)^3}\right) \\
&= \frac{U}{a^3}\left(1 - \frac{1}{\left(1+\frac{\dot{a}}{a}\Delta t\right)^3}\right) \\
&\approx \frac{U}{a^3}\left[1 - \left(1 - 3\frac{\dot{a}}{a}\Delta t\right)\right] \\
&= 3H\Delta t\rho,
\end{aligned}$$

so that

$$\frac{\Delta\rho}{\rho} = 3H\Delta t \sim H\Delta t.$$

The natural time scale during inflation is the Hubble time $1/H$, and applying the uncertainty principle to the field ϕ

$$\frac{1}{H}|\Delta\phi| \sim \hbar,$$

that is

$$|\Delta\phi| \sim \hbar H,$$

so that

$$\frac{\Delta\rho}{\rho} \sim \frac{\hbar H^2}{|\dot{\phi}|}.$$

Next I want to apply the equations of the slow-roll approximation (SRA),

$$\begin{aligned}
H^2 &= \frac{8\pi\hbar c^3}{3E_{\rm Pl}^2}V(\phi) \\
\dot{\phi} &= -\frac{\hbar c^3}{3H}V'(\phi).
\end{aligned}$$

If I insert these equations in the expression for $\Delta\rho/\rho$, I find

$$\begin{aligned}\frac{\Delta\rho}{\rho} &\sim \hbar\frac{\hbar c^3}{E_{\rm Pl}^2}V\frac{H}{\hbar c^3 V'} \\ &\sim \frac{\hbar}{E_{\rm Pl}^3}\frac{V}{V'}H \\ &\sim \frac{(\hbar c)^{3/2}}{E_{\rm Pl}^3}\frac{V^{3/2}}{V'}.\end{aligned}$$

The ratio $\Delta\rho/\rho$ can be determined from observations. To take one example, the amplitude of the temperature fluctuations in the cosmic microwave background over angular scales of a few degrees on the sky are proportional to $\Delta\rho/\rho$. The NASA satellites COBE and WMAP have carried out such observations, and their results show that $\Delta\rho/\rho \sim 10^{-5}$. Unfortunately we cannot come up with a theoretical prediction to compare this number with as long as we don't know what the correct model of inflation is. Neither can we go backwards from the observations to, e.g., the energy scale of inflation, because the amplitude of the density perturbations also depend on the value of ϕ when the perturbations crossed the horizon.

But there is still hope. Another prediction of inflation is that there will also be produced gravitational waves, and that their amplitude is determined directly by the energy scale of inflation. This is the topic of the next subsection.

3.4.1 Inflation and gravitational waves

General relativity predicts the existence of waves in the gravitational field, in the same way as there are waves in the electromagnetic field. This kind of wave does not exist in Newtonian gravitation, it is a unique prediction of general relativity. At the time of writing these waves have still not been detected directly, but we have strong indirect evidence for their existence from the Hulse-Taylor binary pulsar. The rate of energy loss in this system matches very precisely the prediction from general relativity of the amount of energy radiated in the form of gravitational waves. Combined with the fact that general relativity has been proven to be correct whenever and wherever it has been tested, this gives us good reason to take gravitational waves seriously.

Why are there no gravitational waves in Newtonian theory? It is easy to see why this is the case if we reformulate the theory in terms of the gravitational potential Φ. Outside a spherical mass distribution of total mass M we have the familiar result

$$\Phi(r) = -\frac{GM}{r},$$

where r is the distance from the centre of the mass distribution. More generally the gravitational potential in a point $\vec{x}$ outside a mass distribution with density distribution $\rho(\vec{x},t)$ can be shown to be given by

$$\Phi(\vec{x},t) = -G\int \frac{\rho(\vec{y},t)}{|\vec{x}-\vec{y}|}d^3y. \tag{3.11}$$

This equation shows why gravitational waves do not exist in Newtonian theory. The same time t appears on both sides of the equation, and this means that a change in ρ will be transfered immediately to the gravitational potential at any point outside the mass distribution. Waves have to propagate at a finite speed, so it does not make sense to talk of gravitational waves in this situation.

The local version of equation (3.11) is found by using the relation

$$\nabla^2\frac{1}{|\vec{x}-\vec{y}|} = -\delta(\vec{x}-\vec{y}).$$

This gives

$$\nabla^2\Phi(\vec{x},t) = 4\pi G\rho(\vec{x},t).$$

Again we see that changes in ρ are instantly communicated to Φ. This flies in the face of what we have learned in special relativity. Without introducing general relativity (which, of course, is what one really has to do) we can try to make a minimal modification to the equation that will leave it consistent with special relativity:

$$\Phi(\vec{x},t) = -G\int \frac{\rho\left(\vec{y},t-\frac{|\vec{x}-\vec{y}|}{c}\right)}{|\vec{x}-\vec{y}|}d^3y. \tag{3.12}$$

We now see that Φ at time t depends on the source at an earlier time $t-|\vec{x}-\vec{y}|/c$, consistent with the time a light signal needs to travel from the point $\vec{y}$ in the source to the point $\vec{x}$ outside it. I have here taken it for granted that the information travels at the speed of light. More generally it can trave at a speed $v<c$, and to prove that $v=c$, one has to use general relativity. The local version of (3.12) is

$$\nabla^2\Phi - \frac{1}{c^2}\frac{\partial^2\Phi}{\partial t^2} = 4\pi G\rho,$$

which should remind you of wave equations you have come across before. Gravitational waves travelling in vacuum where $\rho=0$ follow the equation

$$\nabla^2\Phi - \frac{1}{c^2}\frac{\partial^2\Phi}{\partial t^2} = 0,$$

which has plane wave solutions

$$\Phi(\vec{x},t) = Ae^{i(\vec{k}\cdot\vec{x}-\omega t)},$$

where $\omega = c|\vec{k}|$.

For those who like Lagrangians and actions, I note in passing that the action for this modified version of Newtonian gravity is

$$S = \int d^3x dt \left[-\rho\Phi - \frac{1}{8\pi G}(\nabla\Phi)^2 + \frac{1}{8\pi G}\left(\frac{1}{c}\frac{\partial\Phi}{\partial t}\right)^2 \right].$$

What kind of sources can give rise to gravitational waves? First of all, the mass density of the source must vary in time. Next, the mass distribution must have a certain amount of structure. A radially oscillating spherical source does not generate gravitational waves. In electromagnetism it is common to decompose the spatial structure of a charge distribution in multipoles: dipole, quadrupole, octupole, etc. We can do the same thing with a mass distribution. If the source oscillates at a characteristic frequency ω, one can show that the radiated power (energy per time) in a multipole mode of order ℓ ($\ell = 1$ is dipole, $\ell = 2$ quadrupole, etc.) is given by

$$P(\ell) \propto \left(\frac{\omega}{c}\right)^{2\ell+2} |Q_{\ell m}|^2,$$

where

$$Q_{\ell m} = \int d^3x r^\ell Y^*_{\ell m}(\theta, \phi)\rho,$$

is the multipole moment. The spherical harmonics $Y_{\ell m}$ appear in this expression. You may recall from quantum mechanics that they carry angular momentum given by ℓ. The electromagnetic field has angular mometum equal to 1, and can therefore be sourced by a dipole distribution. In general relativity one finds that gravitational waves have angular momentum 2, and they therefore need a mass distribution with at least a quadrupole moment as their source. If we return to inflation for a moment, the scalar field has angular momentum equal to 0, and can therefore source gravitational waves directly. However, the gravitational field will have quantum fluctuations, and some of these fluctuations will have a quadrupole moment. So quantum fluctuations in the inflationary epoch can give rise to gravitational waves.

We can determine the amplitude of the gravitational waves generated by quantum fluctuations by combining Heisenberg's uncertainty principle with a little dimensional analysis. We define a dimensionless fluctuation in the gravitational field Φ by $\Delta\Phi/\Phi$, where Φ is the smooth value the field would have had in the absence of waves. The natural time scale in the inflationary epoch is the Hubble time $1/H$. The right hand side of the uncertainty principle is Planck's constant $\hbar$ which has dimensions energy times seconds. We therefore need an energy scale on the left hand side, and the most natural choice is the Planck energy $E_{\rm Pl}$, since this is believed to be the energy scale of quantized gravity. Thus,

$$\frac{1}{H}\frac{\Delta\Phi}{\Phi}E_{\rm Pl} \sim \hbar,$$

which gives

$$\frac{\Delta\Phi}{\Phi} \sim \frac{\hbar H}{E_{\rm Pl}} \propto (\hbar c)^{3/2}\frac{V^{1/2}}{E_{\rm Pl}},$$

where I have used the SRA equation $H^2 \propto V^{1/2}$. This equation shows us something extremely interesting: the amplitude of the gravitational waves produced in the inflationary epoch gives us direct information about the potential V and hence about the energy scale of inflation. This is an important motivation to look for them.

3.4.2 The connection to observations

Once inflation gets going, most of the perturbations in the inflaton field will be swept outside the horizon. Think of the perturbations produced as a Fourier series where each term has a definite wavelength. The wavelength is strethced by the expansion and rapidly becomes greater than the Hubble length $1/H$, which varies slowly in the inflationary epoch. Once outside the horizon, there is no longer any communication between peaks and troughs in the term corresponding to this wavelength, and it will therefore be 'frozen in' as a classical perturbation outside the horizon. The same applies to the gravitational field: they too will be stretched outside the horizon and become classical perturbations. Later in the history of the universe the modes will re-enter the horizon, and we will follow their fate after this point in chapter 4. An important point to bear in mind is that inflation generates sensible initial conditions for the formation of structure in the Universe.

An important question is when perturbations on length scales observable today crossed outside the horizon in the inflationary epoch. A useful rule of thumb turns out to be that this happened about 50 e-foldings before the end of inflation. We can determine the value of the inflaton, ϕ_*, at that time by solving the equation

$$50 = \frac{8\pi}{E_{\rm Pl}^2}\int_{\phi_{\rm end}}^{\phi_*}\frac{V}{V'}d\phi.$$

We have seen that

$$\begin{aligned}\frac{\Delta\rho}{\rho} &\sim \frac{(\hbar c)^{3/2}}{E_{\rm Pl}^3}\frac{V^{3/2}}{V'}\\ \frac{\Delta\Phi}{\Phi} &\sim \frac{(\hbar c)^{3/2}}{E_{\rm Pl}^2}V^{1/2}.\end{aligned}$$

If I form the ration of these two amplitudes, I find that

$$r \equiv (\Delta\Phi/\Phi)/(\Delta\rho/\rho) \sim E_{\rm Pl}\frac{V'}{V} \propto \sqrt{\epsilon}.$$

A more detailed calculation gives

$$r = 3\sqrt{\epsilon}.$$

This is a clear and unambigous prediction of inflation: the ratio of the amplitudes of the gravitational waves and the density perturbations have to satisfy this relation if inflation is driven by a single scalar field. This an important reason for looking for gravitational waves from inflation: they will give a crucial test of the whole concept of inflation. The most promising method for looking for these waves is probably precise measurements of the polarization of the cosmic microwave background. In more advanced treatments one shows that gravitational waves give rise to a characteristic polarization pattern if they are present.

Let us look at an example. Assume that inflation is driven by a scalar field with a quadratic potential, $V(\phi) \propto \phi^2$. In an earlier example we found that the slow-roll parameter ϵ for this potential was given by

$$\epsilon = \frac{E_{\rm Pl}^2}{4\pi\phi^2},$$

and that inflation ends when the field has dercreased to the value

$$\phi_{\rm end} = \frac{E_{\rm Pl}}{2\sqrt{\pi}}.$$

Note that we can also write

$$\epsilon = \frac{\phi_{\rm end}^2}{\phi^2}.$$

I wish to calculate the ratio r defined above, and to do this I need to find the value of ϵ when the field has the value ϕ_* corresponding to the epoch where scales observable in the Universe today disappeared outside the horizon. As I stated earlier I find thsi value by solving the equation

$$50 = \frac{8\pi}{E_{\rm Pl}^2}\int_{\phi_{\rm end}}^{\phi_*}\frac{V'}{V}d\phi = \frac{8\pi}{E_{\rm Pl}^2}\int_{\phi_{\rm end}}^{\phi_*}\frac{1}{2}\phi d\phi.$$

The integral is easily evaluated, and the resulting equation just as easliy solved with the result

$$\left(\frac{\phi_*}{\phi_{\rm end}}\right)^2 = 101,$$

so that

$$\epsilon(\phi_*) = \frac{1}{101}.$$

This model therefore predicts that

$$r = 3\sqrt{\frac{1}{101}} \approx 0.3.$$

Gravitational waves from the inflationary epoch have sadly not been detected at the time of writing. So far we only have upper limits on their amplitude. The WMAP satellite has found an upper limit of $r < 0.65$. A quadratic inflaton potential is thus well within the limits of what observations allow, but it is not far from what one might be able to rule out in the near future.

3.4.3 Optional material: the spectrum of density perturbations

Inflationary models give noe clear prediction of the amplitude of the density perturbations as long as we don't know the energy scale of inflation. But one thing they can predict is how the amplitude varies with length scale. From the expressions

$$\frac{\Delta E}{E} \sim \frac{mc^2}{E_{\rm Pl}},$$

and

$$\frac{\Delta\rho}{\rho} \sim \frac{(\hbar c)^{3/2}}{E_{\rm Pl}^3}\frac{V^{3/2}}{V'}$$

we see that no specific length scale is picked out by the fluctuations. That does not exclude that the amplitude varies with lengt scale, but what it does tell us is that the variations will follow a power-law (in contrast to, e.g., an exponential variation, which has a characteristic damping length). We can determine this power-law if we approximate spacetime during inflation by a flat de Sitter-space. We have seen earlier that a de Sitter-universe is invariant under time translations and will look the same at all epochs. This is understandable since it is empty. Furthermore, the vacuum energy ρ_Λ is constant, the Hubble parameter H is constant, and the latter fact means that the Hubble length $1/H$ also is constant. The Universe is effectively in a stationary state. No place and no time is preferred.

Einstein's field equations connect the line element and the mass-energy density of the Universe. Perturbations in the energy density will therefore give rise to perturbations in the line element. But in de Sitter space the perturbations in the line element must be the same on all length scales while they are inside the horizon, otherwise we could use a change in the amplitude to separate one epoch from another. The line element is determined by the gravitational potential Φ, and when the situation is time-independent we can determine Φ from the equation

$$\nabla^2\Phi = 4\pi G\rho,$$

where ρ is a constant. In spherical coordinates I can write this equation as

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

and this gives

$$r^2 \frac{\partial \Phi}{\partial r} = \frac{4\pi G}{3} \rho r^3,$$

and after yet another integration I find

$$\Phi = \frac{2\pi G}{3} \rho r^3,$$

where I have chosen $\Phi(r = 0) = 0$. On an arbitrary length scale $\lambda < 1/H$ the fluctuation in Φ caused by the fluctuation in ρ will be

$$\Delta\Phi = \frac{2\pi G}{3} \Delta\rho \lambda^2.$$

At the horizon $1/H$ I have

$$\Phi = \frac{2\pi G}{3H^2} \rho,$$

so that

$$\frac{\Delta\Phi}{\Phi} = H^2 \Lambda^2 \frac{\Delta\rho}{\rho}.$$

But in this stationary state $\Delta\Phi/\Phi$ must be independent of λ, and since H is constant I must have

$$\frac{\Delta\rho}{\rho} \sim \frac{1}{\lambda^2}.$$

This is known as a scale-invariant spectrum of density perturbations, of the Harrison-Zeldovich spectrum. It is scale-invariant in the sence that the fluctuations in the gravitational potential are independent of the length scale. This result is valid in a de Sitter universe. In more realistic models for inflation the density perturbations will still to a good approximation follow a power-law, but with a different exponent. The main cause of this deviation from scale-invariance is the fact that the Hubble parameter varies as the scalar field slowly rolls towards the minimum of its potential, and the density perturbations on a given length scale will therefore depend on when the mode crossed outside the horizon.

3.5 Exercises

Exercise 3.1

Consider inflation driven by a scalar field ϕ with the potential $V(\phi) = \lambda\phi^4/(\hbar c)^3$, where λ is a positive constant. Assume that the field is 'rolling' towards $\phi = 0$ from the positive side, so that $\phi > 0$ always.

a) Find the value(s) of ϕ where the slow-roll conditions break down.

b) Assume that inflation ends when $\epsilon = 1$. Calculate the number of e-foldings at the end of inflation when the field had the value $\phi_{\rm i}$ initially.

c) Show that the solutions of the slow-roll equations with initial conditions $\phi = \phi_{\rm i}$, $a = a_{\rm i}$ at $t = t_{\rm i}$ are given by

$$\begin{aligned} \phi &= \phi_i \exp\left[-\sqrt{\frac{2\lambda E_{\rm Pl}^2}{3\pi\hbar^2}}(t-t_i)\right], \\ a &= a_i \exp\left(\frac{\pi\phi_i^2}{E_{\rm Pl}^2}\left\{1-\exp\left[-\sqrt{\frac{8\lambda E_{\rm Pl}^2}{3\pi\hbar^2}}(t-t_i)\right]\right\}\right). \end{aligned}$$

d) Use the solution for ϕ to calculate the time at which inflation ends.

e) Show that the number of e-foldings calculated using the solution for a agrees with what you found in b).

f) Expand the solution for a in powers of $(t - t_{\rm i})$ and show that inflation is approximately exponential in the beginning. Calculate the time constant κ in $a \approx \exp(\kappa t)$ and show that it is equal to the Hubble parameter in the slow-roll approximation.

Exercise 3.2

Some scalar field models of inflation can be solved exactly without using the slow-roll approximation. An example is so-called power-law inflation, defined by the inflaton potential

$$V(\phi) = V_0 \exp\left(-\sqrt{\frac{16\pi}{p}}\frac{\phi}{E_{\rm Pl}}\right),$$

where V_0 and p are positive constants

a) Write down the equations which govern the time evolution of the scale factor a and the scalar field ϕ

b) Show by substitution in the equations from a) that

$$\begin{aligned} a(t) &= Ct^p \\ \phi(t) &= E_{\rm Pl}\sqrt{\frac{p}{4\pi}}\ln\left(\sqrt{\frac{8\pi V_0 l_{\rm Pl}^3}{E_{\rm Pl}p(3p-1)}}\frac{t}{t_{\rm Pl}}\right) \end{aligned}$$

are solutions of these equations, where C is a constant.

c) What condition must p satisfy to have inflation?

d) Find the slow-roll parameters ϵ og η. Will inflation end in this model?

Exercise 3.3 (From the exam in AST4220, 2005)

In this problem we will use a set of units where $\hbar = c = 1$. Observations of the present state of he Universe reveal that it is currently in an accelerated phase of expansion. This can be explained by introducing a positive cosmological constant, $\Lambda > 0$, but there are alternatives. We will consider one alternative in this problem: a homogeous scalar field $\phi(t)$ (NOTE: Not the field that drove inflation.) Assume that the field follows the equation of state $p_\phi = w\rho_\phi$, that is the only contribution to the mass-energy density of the Universe and that the Universe is spatially flat and completely dominated by the scalar field. In the lectures we have shown that for such a universe model,

$$\rho_\phi(a) = \frac{\rho_\phi^0}{a^{3(1+w)}}, \qquad a(t) = \left(\frac{t}{t_0}\right)^{\frac{2}{3(1+w)}}$$

a) Find expressions for $H(t)$ and $\rho_\phi(t)$. What condition must w satisfy if we want accelerated expansion?

b) The energy density and pressure of the scalar field are given by equations (3.1) and (3.2) respectively. If we assume the equation of state given in the introduction to be correct, show that ϕ can be written in terms of a as

$$\phi(a) = \phi_0 + \sqrt{\frac{3\,(1+w)}{8\pi G}}\ln a$$

and that

$$V(\phi) = \frac{1}{2}\,(1-w)\,\rho_\phi^0 \exp\left[-\sqrt{24\pi G\,(1+w)}\,(\phi-\phi_0)\right]$$

where ϕ_0 is the value of the scalar field for $a = a(t_0) = 1$.

c) The potential energy for this particular scalar field model is often written as

$$V(\phi) = V_0 e^{-\lambda\phi\sqrt{8\pi G}}$$

where λ is a positive constant. What is λ and what is V_0? Find the condintion λ must satisfy if we want both accelerated expansion and $p_\phi + \rho_\phi > 0$.

Exercise 3.4 (From the continuation exam in AST4220, 2006)

In this problem we will be using units where $\hbar = 1 = c$ and also make use of the socalled reduced Planck mass $M_{\rm PL} = (8\pi G)^{-1/2}$.

Models for the inflationary epoch in the very early universe make use of a homogenous scalar field $\phi = \phi(t)$ with energy density and pressure given by equations (3.1) and (3.2) respectively. We will assume throughout this problem that $\dot{\phi} > 0$.

a) Show that in a universe dominated by this scalar field the dynamics of the universe and the scalar field are determined by the equations

$$H^2 = \frac{1}{3M_{\rm PL}^2}\left[V(\phi) + \frac{1}{2}\dot{\phi}^2\right], \qquad \ddot{\phi} + 3H\dot{\phi} = -\frac{dV}{d\phi}.$$

b) Use the equations in a) to show that

$$\dot{\phi} = -2M_{\rm PL}^2 H'(\phi), \qquad \text{where} \qquad H'(\phi) = \frac{dH}{d\phi}$$

c) Use the result in b) to show that the first Friedmann equation can be written as

$$[H'(\phi)]^2 - \frac{3}{2M_{\rm PL}^2}H^2(\phi) = -\frac{1}{2M_{\rm PL}^4}V(\phi) \tag{3.13}$$

An important property of the solutions of the equations governing the inflationary phase is that they have a so-called attractor. In practice, this means that regardless of the initial value of ϕ, the Hubble parameter $H(\phi)$ will quickly end up on the same curve in the ϕ-H-plane. You will now demonstrate that this is the case by considering linear, homogeneous perturbations around solutions of equation (3.13): Take $H(\phi) = H_0(\phi) + \delta H(\phi)$ where $H_0(\phi)$ is a solution to (3.13).

d) Show that we to first order in the perturbation δH have

$$H_0'\delta H' = \frac{3}{2M_{\rm PL}^2}H_0\delta H,$$

where again $'$ denotes differentiation with respect to ϕ.

e) Show that the equation in d) has the general solution

$$\delta H(\phi) = \delta H(\phi_i)\exp\left(\frac{3}{2M_{\rm PL}^2}\int_{\phi_i}^{\phi}\frac{H_0(\phi)}{H_0'(\phi)}d\phi\right) \tag{3.14}$$

where ϕ_i is the initial value of the scalar field ϕ. Explain why this result shows that the perturbation δH quickly dies out.

Chapter 4

Structure formation

Except for briefly mentioning that quantum fluctuations of the inflaton field will produce density perturbations, we have so far assumed that the universe is homogeneous. While this is a valid and useful approximation for understanding the large-scale properties of the universe, it clearly cannot be the whole story. We all know that the matter in the universe is not smoothly distributed. It is clumpy, and the clumps come in a range of sizes: from planets via stars and clusters of stars, to galaxies, clusters of galaxies and superclusters. If the universe were completely homogeneous to begin with, it would have stayed so forever, so there must have been initial perturbations in the density. One of the great achievements of inflationary models is to provide a concrete mechanism for producing inhomogeneities in the very early universe. A major challenge in cosmology is to understand how these inhomogeneities grow and become the structures we see in the universe today. The inhomogeneities produced in inflation also lead to small fluctuations in the temperature of the cosmic microwave background (CMB), and the study of these fluctuations is currently one of the most active fields in cosmology.

Perturbations in the density are commonly characterized by the so-called density contrast

$$\Delta(\mathbf{x}, t) = \frac{\rho(\mathbf{x}, t) - \rho_0(t)}{\rho_0(t)}, \tag{4.1}$$

where $\rho_0(t)$ is the spatially averaged density field at time t, and $\rho(\mathbf{x}, t)$ is the local density at the point $\mathbf{x}$ at the same time. We distinguish between two cases:

- $\Delta < 1$: the inhomogeneities are in the linear regime, and we can use linear perturbation theory.
- $\Delta > 1$: the inhomogeneities are starting to collapse and form gravitationally bound structures. Non-linear theory must be used in this case.

We will only consider the first case.

Since the universe on large scales is described by general relativity, one would think that we have to study the Einstein equations to understand the growth of density perturbations. Formally this is correct, but it turns out that a lot of the physics can be understood, both quantitatively and qualitatively, by Newtonian theory if we restrict ourselves to scales smaller than the particle horizon and speeds less than the speed of light. Furthermore, the evolution of perturbations on scales larger than the particle horizon turns out to be fairly simple, and we can use the size of a perturbation at the time it enters the horizon as an initial condition in our Newtonian treatment.

4.1 Non-relativistic fluids

We will start with the simplest situation, a universe with just one component, and find the equations describing small density perturbations. The fundamental equations are:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0 \tag{4.2}$$

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla)\mathbf{v} = -\frac{1}{\rho}\nabla p - \nabla \phi \tag{4.3}$$

$$\nabla^2 \phi = 4\pi G \rho, \tag{4.4}$$

where ρ is the density, $\mathbf{v}$ is the velocity field, p is the pressure, and ϕ is the gravitational potential (do not confuse it with the scalar field of inflation in the previous chapter). The equations are called, respectively, the continuity equation, the Euler equation, and Poisson's equation. The partial derivatives describe time variations in the quantities at a fixed point in space. This description is often called Eulerian coordinates. The equations can also be written in a different form where one follows the motion of a particular fluid element. This is called the Lagrangian description of the fluid. Derivatives describing the time evolution of a particular fluid element are written as total derivatives d/dt, and one can show that

$$\frac{d}{dt} = \frac{\partial}{\partial t} + (\mathbf{v} \cdot \nabla). \tag{4.5}$$

Note that the effect of the operator $(\mathbf{v} \cdot \nabla)$ on a scalar function f is given by

$$(\mathbf{v} \cdot \nabla) f = v_x \frac{\partial f}{\partial x} + v_y \frac{\partial f}{\partial y} + v_z \frac{\partial f}{\partial z}, \tag{4.6}$$

in Cartesian coordinates, whereas the effect on a vector field $\mathbf{A}$ is given by

$$(\mathbf{v} \cdot \nabla)\mathbf{A} = \left(v_x \frac{\partial A_x}{\partial x} + v_y \frac{\partial A_x}{\partial y} + v_z \frac{\partial A_x}{\partial z} \right) \mathbf{e}_x$$

$$+ \quad \left(v_x \frac{\partial A_y}{\partial x} + v_y \frac{\partial A_y}{\partial y} + v_z \frac{\partial A_y}{\partial z}\right) \mathbf{e}_y$$
$$+ \quad \left(v_x \frac{\partial A_z}{\partial x} + v_y \frac{\partial A_z}{\partial y} + v_z \frac{\partial A_z}{\partial z}\right) \mathbf{e}_z. \tag{4.7}$$

In Lagrangian form the equations (4.2)-(4.4) can be written as

$$\frac{d\rho}{dt} = -\rho(\nabla \cdot \mathbf{v}) \tag{4.8}$$
$$\frac{d\mathbf{v}}{dt} = -\frac{1}{\rho}\nabla p - \nabla\phi \tag{4.9}$$
$$\nabla^2\phi = 4\pi G\rho. \tag{4.10}$$

The transition from (4.3) to (4.9) is easily seen; the transition from (4.2) to (4.8) can be shown by writing out (4.2):

$$\begin{aligned} \frac{\partial\rho}{\partial t} + \nabla\cdot(\rho\mathbf{v}) &= \frac{\partial\rho}{\partial t} + \rho(\nabla\cdot\mathbf{v}) + \mathbf{v}\cdot\nabla\rho \\ &= \frac{\partial\rho}{\partial t} + (\mathbf{v}\cdot\nabla)\rho + \rho(\nabla\cdot\mathbf{v}) = 0, \end{aligned}$$

and the desired result follows.

We could imagine starting by studying perturbations around a uniform state where ρ and p are constant in space and $\mathbf{v} = 0$. Unfortunately, such a solution does not exist. The reason for this is that we would then have

$$\begin{aligned} \frac{\partial\rho}{\partial t} &= 0 \\ \frac{\partial\mathbf{v}}{\partial t} &= 0 = -\frac{1}{\rho}\nabla p - \nabla\phi = -\nabla\phi \\ \nabla^2\phi &= 4\pi G\rho \end{aligned}$$

From the second equation follows $\nabla^2\phi = 0$, and from the last equation we then see that $\rho = 0$, which means that the universe is empty, and therefore not very exciting. Clearly, we cannot start from the solution for a static medium. But this is not a disaster, since we are at any rate interested in perturbations around an expanding background. In this case the unperturbed problem has a non-trivial solution, namely the matter-dominated expanding solution we found in chapter 1. Let us call the solution $\mathbf{v}_0$, ρ_0, p_0 and ϕ_0. These quantities obey, by definition, equations (4.8)-(4.10). We now add small perturbations to these solutions, and write the full quantities as

$$\mathbf{v} = \mathbf{v}_0 + \delta\mathbf{v} \tag{4.11}$$
$$\rho = \rho_0 + \delta\rho \tag{4.12}$$
$$p = p_0 + \delta p \tag{4.13}$$
$$\phi = \phi_0 + \delta\phi. \tag{4.14}$$

We assume the perturbations are so small that it is sufficient to expand the equations to first order in them. Furthermore, we assume that the unperturbed pressure p_0 is homogeneous, $\nabla p_0 = 0$. With these assumptions, we can derive the equations for the perturbed quantities. From (4.8):

$$\frac{d}{dt}(\rho_0 + \delta\rho) = -(\rho_0 + \delta\rho)\nabla \cdot (\mathbf{v}_0 + \delta\mathbf{v}),$$

and written out in full detail, this becomes

$$\begin{aligned} \frac{d\rho_0}{dt} + \frac{d}{dt}\delta\rho &= -\rho_0\nabla\cdot\mathbf{v}_0 - \rho_0\nabla\cdot\delta v \\ &- \delta\rho\nabla\cdot\mathbf{v}_0 - \delta\rho\nabla\cdot\delta\mathbf{v}, \end{aligned}$$

where we see that the last term is of second order in the perturbations and therefore should be neglected in first-order perturbation theory. If we use the fact that ρ_0 obeys equation (4.8), several terms cancel and we are left with

$$\frac{d}{dt}\delta\rho = -\rho_0\nabla\cdot\delta\mathbf{v} - \delta\rho\nabla\cdot\mathbf{v}_0.$$

We divide this equation by ρ_0:

$$\frac{1}{\rho_0}\frac{d}{dt}\delta\rho = -\nabla\cdot\delta\mathbf{v} - \frac{\delta\rho}{\rho_0}\nabla\cdot\mathbf{v}_0,$$

and use (4.8) in the last term on the right-hand side so that

$$\frac{1}{\rho_0}\frac{d}{dt}\delta\rho = -\nabla\cdot\delta\mathbf{v} + \frac{\delta\rho}{\rho_0^2}\frac{d\rho_0}{dt}.$$

If we move the last term on the right-hand side over to the left side, we see that the equation can be written as

$$\frac{d}{dt}\left(\frac{\delta\rho}{\rho_0}\right) \equiv \frac{d\Delta}{dt} = -\nabla\cdot\delta\mathbf{v}. \tag{4.15}$$

Next we look at the left-hand side of (4.9):

$$\begin{aligned} \frac{d}{dt}(\mathbf{v}_0 + \delta\mathbf{v}) &= \left[\frac{\partial}{\partial t} + (\mathbf{v}_0 + \delta\mathbf{v})\cdot\nabla\right](\mathbf{v}_0 + \delta\mathbf{v}) \\ &= \frac{\partial\mathbf{v}_0}{\partial t} + [(\mathbf{v}_0 + \delta\mathbf{v})\cdot\nabla]\mathbf{v}_0 + \frac{\partial}{\partial t}\delta\mathbf{v} + [(\mathbf{v}_0 + \delta\mathbf{v})\cdot\nabla]\delta\mathbf{v} \\ &= \frac{\partial\mathbf{v}_0}{\partial t} + (\mathbf{v}_0\cdot\nabla)\mathbf{v}_0 + (\delta\mathbf{v}\cdot\nabla)\mathbf{v}_0 + \frac{d}{dt}\delta\mathbf{v}. \end{aligned}$$

The right-hand side becomes

$$\begin{aligned} -\frac{1}{\rho_0 + \delta\rho}\nabla(p_0 + \delta p) - \nabla(\phi_0 + \delta\phi) &= -\frac{1}{\rho_0}\frac{1}{1 + \frac{\delta\rho}{\rho_0}}\nabla\delta p - \nabla\phi_0 - \nabla\delta\phi \\ &= -\frac{1}{\rho_0}\nabla\delta p - \nabla\phi_0 - \nabla\delta\phi. \end{aligned}$$

We now equate the left-hand side and the right-hand side and use that $\mathbf{v}_0$, p_0 and ϕ_0 are solutions of (4.3) (with $\nabla p_0 = 0$). This leaves us with

$$\frac{d}{dt}\delta\mathbf{v} + (\delta\mathbf{v}\cdot\nabla)\mathbf{v}_0 = -\frac{1}{\rho_0}\nabla\delta p - \nabla\delta\phi. \tag{4.16}$$

The perturbed version of (4.9) is easily found, since Poisson's equation is linear and ϕ_0 and ρ_0 are solutions of the unperturbed version:

$$\nabla^2\delta\phi = 4\pi G\delta\rho. \tag{4.17}$$

Equations (4.15,4.16,4.17) are the linearized equations describing how the perturbations evolve with time.

Since we consider a uniformly expanding background it will be convenient to change from physical coordinates $\mathbf{x}$ to comoving coordinates $\mathbf{r}$,

$$\mathbf{x} = a(t)\mathbf{r}, \tag{4.18}$$

where $a(t)$ is the scale factor. We then have

$$\delta\mathbf{x} = \delta[a(t)\mathbf{r}] = \mathbf{r}\delta a(t) + a(t)\delta\mathbf{r},$$

and the velocity can be written as

$$\begin{aligned}\mathbf{v} = \mathbf{v}_0 + \delta\mathbf{v} &= \frac{\delta\mathbf{x}}{\delta t}\\ &= \mathbf{r}\frac{\delta a(t)}{\delta t} + a(t)\frac{\delta\mathbf{r}}{\delta t}\\ &= \dot{a}\mathbf{r} + a(t)\mathbf{u}\\ &= H\mathbf{x} + a(t)\mathbf{u}\end{aligned}$$

The first term $\mathbf{v}_0$ is given by the Hubble expansion, whereas the velocity perturbation is

$$\delta\mathbf{v} = a(t)\frac{\delta\mathbf{r}}{\delta t} \equiv a(t)\mathbf{u}. \tag{4.19}$$

The velocity $\mathbf{u}$, describes deviations from the smooth Hubble flow, and is often called the peculiar velocity. Equation (4.16) can hence be rewritten as

$$\frac{d}{dt}(a\mathbf{u}) + (a\mathbf{u}\cdot\nabla)(\dot{a}\mathbf{r}) = -\frac{1}{\rho_0}\nabla\delta p - \nabla\delta\phi.$$

We replace the ∇ operator in physical coordinates with ∇ in co-moving coordinates. They are related by

$$\nabla = \frac{1}{a}\nabla_c,$$

where the index c denotes 'co-moving'. We then get

$$\frac{d}{dt}(a\mathbf{u}) + \left(a\mathbf{u}\cdot\frac{1}{a}\nabla_c\right)(\dot{a}\mathbf{r}) = -\frac{1}{\rho_0}\frac{1}{a}\nabla_c\delta p - \frac{1}{a}\nabla_c\delta\phi.$$

The second term on the left-hand side can be rewritten as

$$\begin{aligned}(\mathbf{u}\cdot\nabla_c)(\dot{a}\mathbf{r}) &= \dot{a}(\mathbf{u}\cdot\nabla_c)\mathbf{r} \\ &= \dot{a}\sum_{i,j=x,y,z} u_i\frac{\partial}{\partial r_i}r_j\mathbf{e}_j \\ &= \dot{a}\sum_{i,j=x,y,z} u_i\mathbf{e}_j\delta_{ij} \\ &= \dot{a}\sum_{i=x,y,z} u_i\mathbf{e}_i = \dot{a}\mathbf{u},\end{aligned}$$

so that we have

$$\dot{a}\mathbf{u} + a\dot{\mathbf{u}} + \dot{a}\mathbf{u} = -\frac{1}{\rho_0 a}\nabla_c\delta p - \frac{1}{a}\nabla_c\delta\phi,$$

and finally

$$\dot{\mathbf{u}} + 2\frac{\dot{a}}{a}\mathbf{u} = -\frac{1}{\rho_0 a^2}\nabla_c\delta p - \frac{1}{a^2}\nabla_c\delta\phi. \tag{4.20}$$

Note that we have three equations for four unknowns: $\delta\rho$, $\mathbf{u}$, $\delta\phi$, and δp. We therefore need one more equation to close the system, and we get this by specializing to an adiabatic system where the pressure perturbations are related to the density perturbations by

$$\delta p = c_s^2\delta\rho, \tag{4.21}$$

where c_s is the sound speed in the system. With this extra condition, (4.20) can be rewritten as

$$\dot{\mathbf{u}} + 2\frac{\dot{a}}{a}\mathbf{u} = -\frac{c_s^2}{\rho_0 a^2}\nabla_c\delta\rho - \frac{1}{a^2}\nabla_c\delta\phi. \tag{4.22}$$

We are primarily interested in the time development of the density perturbation $\delta\rho$, and we will therefore find an equation where only this quantity appears as an unknown. We can achieve this by first taking the divergence of equation (4.22):

$$\nabla_c\cdot\dot{\mathbf{u}} + 2\frac{\dot{a}}{a}\nabla_c\mathbf{u} = -\frac{c_s^2}{\rho_0 a^2}\nabla_c^2\delta\rho - \frac{1}{a^2}\nabla_c^2\delta\phi.$$

From (4.17) in co-moving coordinates we have

$$\frac{1}{a^2}\nabla_c^2\delta\phi = 4\pi G\delta\rho,$$

and therefore

$$\nabla_c\cdot\dot{\mathbf{u}} + 2\frac{\dot{a}}{a}\nabla_c\mathbf{u} = -\frac{c_s^2}{\rho_0 a^2}\nabla_c^2\delta\rho - 4\pi G\delta\rho. \tag{4.23}$$

From equation (4.15) we get

$$\frac{d\Delta}{dt} = -\nabla \cdot \delta\mathbf{v} = -\frac{1}{a}\nabla_c \cdot (a\mathbf{u}) = -\nabla_c \cdot \mathbf{u},$$

and

$$\frac{d^2\Delta}{dt^2} = -\nabla_c \cdot \dot{\mathbf{u}},$$

which inserted in (4.23) results in

$$\frac{d^2\Delta}{dt^2} + 2\frac{\dot{a}}{a}\frac{d\Delta}{dt} = \frac{c_s^2}{\rho_0 a^2}\nabla_c^2\delta\rho + 4\pi G\delta\rho, \tag{4.24}$$

where $\Delta = \delta\rho/\rho_0$. This is the desired equation for $\delta\rho$.

We write the density perturbation as a Fourier series

$$\Delta(\mathbf{r}, t) = \sum_{\mathbf{k}} \Delta_k(t) e^{i\mathbf{k}_c\cdot\mathbf{r}}, \tag{4.25}$$

where $\mathbf{k}_c = a\mathbf{k}$ is the co-moving wave number vector, so that

$$\mathbf{k}_c \cdot \mathbf{r} = a\mathbf{k} \cdot \mathbf{r} = \mathbf{k} \cdot (a\mathbf{r}) = \mathbf{k} \cdot \mathbf{x},$$

where $\mathbf{k}$ is the physical wave number vector. Since equation (4.24) is linear, there will be no coupling between different Fourier modes, and the result will be a set of independent equations for each mode on the same form as the equation we will now find. In other words, there is no severe restriction involved in the assumption (4.25). We see that

$$\nabla_c^2\delta\rho = \nabla_c^2(\rho_0\Delta) = -k_c^2\rho_0\Delta = -a^2k^2\rho_0\Delta,$$

so that equation (4.24) can be written

$$\frac{d^2\Delta_k}{dt^2} + 2\frac{\dot{a}}{a}\frac{d\Delta_k}{dt} = \Delta_k(4\pi G\rho_0 - k^2c_s^2). \tag{4.26}$$

We will in the following analyze this equation. It describes the time evolution of a perturbation on a physical length scale $d \sim 1/k$, where $k = |\mathbf{k}|$.

4.2 The Jeans length

Even though we are interested in perturbations around an expanding background, it is useful to first look at the case $\dot{a} = 0$. We look for solutions with time dependence $\Delta_k(t) = \Delta_k \exp(-i\omega t)$, so that $\ddot{\Delta}_k(t) = -\omega^2\Delta_k(t)$. If we insert this in equation (4.26), we see that ω must obey the dispersion relation

$$\omega^2 = c_s^2k^2 - 4\pi G\rho_0. \tag{4.27}$$

This dispersion relation describes either acoustic oscillations (sound waves) or instabilities, depending on the sign of the right-hand side. An important quantity is therefore the value of the wave number k for which the right-hand side is equal to zero. This value is often called the Jeans wave number $k_{\rm J}$, and is given by

$$k_{\rm J} = \frac{\sqrt{4\pi G\rho_0}}{c_s}. \tag{4.28}$$

and the corresponding wave length is called the Jeans length,

$$\lambda_{\rm J} = \frac{2\pi}{k_{\rm J}} = c_s\sqrt{\frac{\pi}{G\rho_0}}. \tag{4.29}$$

For $k > k_{\rm J}$ ($\lambda < \lambda_{\rm J}$) the right-hand side of equation (4.27) is positive, so that ω is real, and we then have solutions of the perturbation equation of the form

$$\Delta(\mathbf{x}, t) = \Delta_k e^{i(\mathbf{k}\cdot\mathbf{x}-\omega t)},$$

where $\omega = \pm\sqrt{c_s^2k^2 - 4\pi G\rho_0}$. These represent periodic variations in the local density, i.e., acoustic oscillations. In this case, the pressure gradient is strong enough to stabilize the perturbations against collapse.

For $k < k_{\rm J}$ ($\lambda > \lambda_{\rm J}$) the right-hand side of (4.27) is negative, so that ω is imaginary. The solutions are then of the form

$$\Delta(\mathbf{x}, t) = \Delta_k e^{\pm\Gamma t},$$

where

$$\Gamma = \sqrt{4\pi G\rho_0 - c_s^2k^2} = \left[4\pi G\rho_0\left(1 - \frac{\lambda_{\rm J}^2}{\lambda^2}\right)\right]^{1/2}. \tag{4.30}$$

We see that we get one exponentially decaying and one exponentially growing solution. The latter represents a perturbation which collapses and finally forms a gravitationally bound subsystem. The growth rate for this mode is Γ, which for perturbations on scales $\lambda \gg \lambda_{\rm J}$ is approximately given by $\Gamma \approx \sqrt{4\pi G\rho_0}$, and the typical collapse time is then $\tau \sim 1/\Gamma \sim (G\rho_0)^{-1/2}$. The physics of this result can be understood from the stability condition for a spherical region of uniform density ρ: for the region to be in equilibrium, the pressure gradient must balance the gravitational forces. For a spherical shell at a distance r from the centre of the sphere, the condition is

$$\frac{dp}{dr} = -\frac{G\rho M(<r)}{r^2},$$

where $M(<r) \sim \rho r^3$ is the mass contained within the distance r from the centre. For this equation to be fulfilled, the pressure must increase towards the centre of the sphere, and we approximate $dp/dr \sim p/r$. We therefore get

$$p = G\rho^2r^2$$

at equilibrium, and if we take $c_s^2 = p/\rho$, we get stability when

$$r = \frac{c_s}{\sqrt{G\rho}} \sim \lambda_{\rm J}.$$

For $r > \lambda_{\rm J}$ the pressure gradient is too weak to stabilize the region, and hence it will collapse. We also note that $\lambda_{\rm J} \sim c_s\tau$, so the Jeans length can be interpreted as the distance a sound wave covers in a collapse time.

4.3 The Jeans instability in an expanding medium

The analysis in the previous subsection was valid for density perturbations in a static background, $\dot{a} = 0$. However, in cosmology we are interested in expanding backgrounds. Let us write equation (4.26) as

$$\frac{d^2\Delta_k}{dt^2} + 2\frac{\dot{a}}{a}\frac{d\Delta_k}{dt} = 4\pi G\rho_0\left(1 - \frac{\lambda_{\rm J}^2}{\lambda^2}\right)\Delta_k, \tag{4.31}$$

where the term with $\dot{a}/a$ will modify the analysis in the previous subsection. This term can be compared to a friction term: in addition to the pressure gradient, the expansion of the universe will work against gravity and try to prevent the collapse of a density perturbation. Let us consider the case $\lambda \gg \lambda_{\rm J}$, so that the equation simplifies to

$$\frac{d^2\Delta_k}{dt^2} + 2\frac{\dot{a}}{a}\frac{d\Delta_k}{dt} = 4\pi G\rho_0\Delta_k.$$

In the case where the background universe is the Einstein-de Sitter universe with $\Omega_{\rm m0} = 1$, $a = a_0(t/t_0)^{2/3}$, this equation has simple solutions. We then have $\dot{a}/a = 2/(3t)$ and $4\pi G\rho_0 = 2/(3t^2)$, so that the equation becomes

$$\frac{d^2\Delta_k}{dt^2} + \frac{4}{3t}\frac{d\Delta_k}{dt} - \frac{2}{3t^2}\Delta_k = 0. \tag{4.32}$$

We look for a solution of the form $\Delta_k = Kt^n$, where K is a constant. Inserted in equation (4.32), we find that n must satisfy

$$n^2 + \frac{1}{3}n - \frac{2}{3} = 0,$$

which has $n = -1$ and $n = 2/3$ as solutions. We see that we have damped, decaying mode $\Delta_k \propto 1/t$ and a growing mode $\Delta_k \propto t^{2/3} \propto a \propto 1/(1+z)$. The expansion of the universe has hence damped the growth of the perturbations and turned exponential growth into power-law growth.

A comment on the $\dot{a}/a$ term: we have taken this from solutions of the Friedmann equations for a homogeneous universe, that is we have neglected the perturbations. This is the correct approach in first-order perturbation theory, because equation (4.26) is already of first order in the perturbation Δ_k. Had we included corrections of first order in Δ_k in the equations for $\dot{a}/a$, these would have given corrections of second order to equation (4.26), and they can therefore be neglected in first-order perturbation theory.

4.4 Perturbations in a relativistic gas

The formalism in the preceding sections describe perturbations in a non-relativistic fluid. If the fluid is relativistic, we need a more general formalism. The professional way of doing this is to use the formalism of general relativity and in addition take into account that a fluid description is not really appropriate for e.g. photons, since they should be described by the Boltzmann equation for their distribution function. We will here content ourselves with formulating and solving the relativistic fluid equations in the radiation dominated epoch of the universe for redshifts $1+z > 4\times 10^4 \Omega_{\rm m0} h^2$.

In the relativistic case one can show that one gets two equations expressing conservation of energy and momentum:

$$\frac{\partial \rho}{\partial t} = -\nabla \cdot \left[\left(\rho + \frac{p}{c^2} \right) \mathbf{v} \right] \tag{4.33}$$

$$\frac{\partial}{\partial t} \left(\rho + \frac{p}{c^2} \right) = \frac{\dot{p}}{c^2} - \left(\rho + \frac{p}{c^2} \right) (\nabla \cdot \mathbf{v}). \tag{4.34}$$

In the special case $p = \rho c^2/3$ both equations reduce to

$$\frac{d\rho}{dt} = -\frac{4}{3} \rho (\nabla \cdot \mathbf{v}). \tag{4.35}$$

The analogue of the Euler equation turns out to be

$$\left(\rho + \frac{p}{c^2} \right) \left[\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla) \mathbf{v} \right] = -\nabla p - \left(\rho + \frac{p}{c^2} \right) \nabla \phi, \tag{4.36}$$

while the analogue of the Poisson equation is

$$\nabla^2 \phi = 4\pi G \left(\rho + \frac{3p}{c^2} \right), \tag{4.37}$$

which for $p = \rho c^2/3$ gives

$$\nabla^2 \phi = 8\pi G \rho. \tag{4.38}$$

We see that for the special case of a relativistic gas, $p = \rho c^2/3$ the equations reduce to the same form as in the non-relativistic case, except that the numerical coefficients which enter are slightly different. It should therefore come as no surprise that after a similar analysis of linear perturbations as in the non-relativistic case, we end up with an equation very similar to equation (4.26):

$$\frac{d^2 \Delta_k}{dt^2} + 2 \frac{\dot{a}}{a} \frac{d\Delta_k}{dt} = \left(\frac{32\pi G \rho_0}{3} - k^2 c_s^2 \right) \Delta_k, \tag{4.39}$$

and the Jeans length in the relativistic case is therefore

$$\lambda_{\rm J} = c_s \left(\frac{3\pi}{8 G \rho_0} \right)^{1/2}, \tag{4.40}$$

where $c_s = c/\sqrt{3}$.

For modes with $\lambda \gg \lambda_{\rm J}$ equation (4.39) becomes

$$\frac{d^2\Delta_k}{dt^2} + 2\frac{\dot{a}}{a}\frac{d\Delta_k}{dt} - \frac{32\pi G\rho_0}{3}\Delta_k = 0, \tag{4.41}$$

and since we in the radiation dominated phase have $\dot{a}/a = 1/2t$, $\rho_0 = 3/(32\pi G t^2)$, we get the equation

$$\frac{d^2\Delta_k}{dt^2} + \frac{1}{t}\frac{d\Delta_k}{dt} - \frac{1}{t^2}\Delta_k = 0. \tag{4.42}$$

We seek solutions of the form $\Delta_k \propto t^n$, and find that n must satisfy

$$n^2 - 1 = 0,$$

i.e., $n = \pm 1$. The growing mode is in this case $\Delta_k \propto t \propto a^2 \propto (1+z)^{-2}$.

4.5 A comment on the perturbations in the gravitational potential

The equation for the growing mode in the gravitational potential ϕ was

$$\nabla^2\delta\phi \propto \delta\rho = \rho_0\Delta.$$

If we seek solutions $\delta\phi = \delta\phi_k \exp(i\mathbf{k}_c \cdot \mathbf{r})$, we find that

$$\frac{1}{a^2}\nabla_c^2\delta\phi = -\frac{k_c^2}{a^2}\delta\phi_k e^{i\mathbf{k}_c\cdot\mathbf{r}} \propto \rho_0\Delta_k e^{i\mathbf{k}_c\cdot\mathbf{r}},$$

which gives

$$\delta\phi_k \propto \rho_0 a^2 \Delta_k, \tag{4.43}$$

Since $\rho_0 \propto a^{-3}$, $\Delta_k \propto a$ for dust, and $\rho_0 \propto a^{-4}$, $\Delta_k \propto a^2$ for radiation, we find that $\delta\phi_k$ is constant in both cases. Therefore the perturbations in ϕ_k, and therefore also in ϕ, are independent of time in both the matter-dominated and radiation-dominated phases if the universe is geometrically flat. In particular, we have that ϕ is constant in an Einstein-de Sitter universe to first order in perturbation theory.

4.6 The Meszaros effect

So far we have only considered the case where the universe contains one component. The real situation is of course more complicated than this. We know that the universe contains both radiation, neutrinos, baryons, dark matter, possibly a cosmological constant etc. In realistic calculations of structure formation, we must take all these components into account.

We will now consider a simple case where an analytic solution can be found: the growth of perturbations in the matter density ρ_m in the radiation dominated phase. In this phase we can consider the radiation to be unperturbed on scales inside the particle horizon. We can then use equation (4.26) for non-relativistic matter, but take $\dot{a}/a$ from the Friedmann equations for a universe with matter and radiation. We also assume that we can neglect non-gravitational interactions between radiation and matter, which should be a good approximation since most of the matter is dark. We will also limit ourselves to consider perturbations on scales $\lambda \gg \lambda_{\rm J}$, so that the equation we have to solve is

$$\ddot{\Delta}_k + 2\frac{\dot{a}}{a}\dot{\Delta}_k - 4\pi G\rho_m\Delta_k = 0. \tag{4.44}$$

To solve this equation, it is convenient to change variable from t to a, so that

$$\frac{d}{dt} = \frac{da}{dt}\frac{d}{da} = \dot{a}\frac{d}{da} \tag{4.45}$$

$$\frac{d^2}{dt^2} = \frac{d}{dt}\left(\dot{a}\frac{d}{da}\right) = \dot{a}^2\frac{d^2}{da^2} + \ddot{a}\frac{d}{da}. \tag{4.46}$$

Furthermore, we introduce

$$y = \frac{a}{a_{\rm eq}}, \tag{4.47}$$

where $a_{\rm eq}$ is the scale factor at matter-radiation equality, determined by

$$\rho_m(a_{\rm eq}) = \rho_r(a_{\rm eq}),$$

where $\rho_m = \rho_{m0}a^{-3}$, $\rho_r = \rho_{r0}a^{-4}$, so that

$$a_{\rm eq} = \frac{\rho_{r0}}{\rho_{m0}}. \tag{4.48}$$

We also see that

$$\frac{\rho_m}{\rho_r} = \frac{\rho_{m0}a^{-3}}{\rho_{r0}a^{-4}} = \frac{a}{\rho_{r0}/\rho_{m0}} = \frac{a}{a_{\rm eq}} = y. \tag{4.49}$$

The Friedmann equations can then be written as

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}(\rho_m + \rho_r) = \frac{8\pi G}{3}\rho_r\left(1 + \frac{\rho_m}{\rho_r}\right)$$
$$= \frac{8\pi G}{3}\rho_r(1+y), \tag{4.50}$$

and

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}(\rho_m + \rho_r + 3p_r) = -\frac{4\pi G}{3}(\rho_m + 2\rho_r)$$
$$= -\frac{4\pi G}{3}\rho_r(2+y). \tag{4.51}$$

We express d/da by d/dy:

$$\frac{d}{da} = \frac{dy}{da}\frac{d}{dy} = \frac{1}{a_{\rm eq}}\frac{d}{dy} \tag{4.52}$$

$$\frac{d^2}{da^2} = \frac{1}{a_{\rm eq}^2}\frac{d^2}{dy^2}. \tag{4.53}$$

Inserting all of this into equation (4.44), we get

$$\frac{2}{3}\rho_r y^2(1+y)\frac{d^2\Delta_k}{dy^2} - \frac{1}{3}\rho_r y(2+y)\frac{d\Delta_k}{dy} + \frac{4}{3}\rho_r y(1+y)\frac{d\Delta_k}{dy} - \rho_m\Delta_k = 0,$$

which after some manipulations gives

$$\frac{d^2\Delta_k}{dy^2} + \frac{2+3y}{2y(1+y)}\frac{d\Delta_k}{dy} - \frac{3}{2y(1+y)}\Delta_k = 0. \tag{4.54}$$

By substitution one easily sees that this equation has the growing solution

$$\Delta_k \propto 1 + \frac{3}{2}y, \tag{4.55}$$

which means that in the course of the entire radiation-dominated phase from $y = 0$ to $y = 1$ the perturbations grow by the modest factor

$$\frac{\Delta_k(y=1)}{\Delta_k(y=0)} = \frac{1+\frac{3}{2}}{1} = \frac{5}{2}.$$

That perturbations in the matter density cannot grow significantly in the radiation-dominated phase is known as the Meszaros effect. It can be understood qualitatively by comparing the collapse time for a density perturbation with the expansion time scale for the universe. We have seen that the collapse time is $\tau_c \sim 1/\sqrt{G\rho_m}$, whereas the expansion time scale is

$$\tau_H = \frac{1}{H} = \frac{a}{\dot{a}} \approx \left(\frac{3}{8\pi G\rho_r}\right)^{1/2} \sim \frac{1}{\sqrt{G\rho_r}}.$$

Since $\rho_r > \rho_m$ in this epoch, we have $\tau_H < \tau_c$. In other words: in the radiation-dominated phase the universe expands faster than a density perturbation can collapse.

4.7 The statistical properties of density perturbations

We have so far considered solutions of equation (4.26) of the form

$$\Delta(\mathbf{x}, t) = \Delta_k(t)e^{i\mathbf{k}\cdot\mathbf{x}}.$$

This is not so restrictive as it may seem, since we can write the general solution as a Fourier series

$$\Delta(\mathbf{x}, t) = \sum_{\mathbf{k}} \Delta_k(t) e^{-i\mathbf{k}\cdot\mathbf{x}}, \tag{4.56}$$

where

$$\Delta_k(t) = \frac{1}{V} \int \Delta(\mathbf{x}, t) e^{i\mathbf{k}\cdot\mathbf{x}} d^3x. \tag{4.57}$$

Here V is some large normalization volume and

$$\frac{1}{V} \int e^{i(\mathbf{k}-\mathbf{k}')\cdot\mathbf{x}} d^3x = \delta_{\mathbf{k},\mathbf{k}'}. \tag{4.58}$$

In the limit $V \to \infty$, we can write $\Delta(\mathbf{x}, t)$ as a Fourier integral

$$\Delta(\mathbf{x}, t) = \frac{1}{(2\pi)^3} \int \Delta_k e^{-i\mathbf{k}\cdot\mathbf{x}} d^3k, \tag{4.59}$$

where

$$\Delta_k(t) = \int \Delta(\mathbf{x}, t) e^{i\mathbf{k}\cdot\mathbf{x}} d^3x. \tag{4.60}$$

We will take the liberty of using both these descriptions, according to which is most convenient. Since the differential equation for $\Delta(\mathbf{x}, t)$ is linear, (4.56) or (4.59) will by insertion in the perturbation equation give a set of independent equation for each Δ_k mode, all of the same form as equation (4.26). In other words: there is no loss of generality in the way we treated the problem in earlier subsections. Since the equations are linear, there will be no coupling between modes with different $\mathbf{k}$, and perturbations on different length scales therefore evolve independently. Note that this applies in linear perturbation theory only. In the non-linear regime, perturbations on different length scales can and will couple, and this is one of the reasons why non-linear perturbation theory is more complicated.

Observationally we are mostly interested in the statistical properties of Δ. Earlier in the course we have seen that the likely origin of the density perturbations are quantum fluctuations in the inflationary epoch of the universe. We can therefore consider $\Delta(\mathbf{x}, t)$ as a stochastic field. The simplest inflationary models predict that the initial perturbations $\Delta_{\rm in}(\mathbf{x}, t)$ had a Gaussian distribution

$$p(\Delta_{\rm in}) \propto \exp\left(-\frac{\Delta_{\rm in}^2}{2\sigma^2}\right).$$

As we have seen, perturbations will evolve in the time after inflation, but as long as the evolution is linear, a Gaussian field will remain a Gaussian field. When the perturbations reach the non-linear regime, different modes will be coupled, and we can in general get non-Gaussian fluctuations. But scales

within the linear regime can be expected to follow a Gaussian distribution. This means that they are fully characterized by their mean and standard deviation, and their mean (i.e., average over all space) is by definition equal to zero, since Δ is the local deviation from the mean density. The other quantity we need to characterize the distribution is then $\langle\Delta^2\rangle$, where

$$\langle\ldots\rangle = \frac{1}{V}\int \ldots d^3x,$$

is the spatial average. By using (4.56) we get

$$\Delta^2(\mathbf{x},t) = \sum_{\mathbf{k},\mathbf{k}'} \Delta_k(t)\Delta_{k'}(t)e^{-i(\mathbf{k}+\mathbf{k}')\cdot\mathbf{x}}. \tag{4.61}$$

We therefore find that

$$\begin{aligned}\langle\Delta^2(\mathbf{x},t)\rangle &= \frac{1}{V}\int \Delta^2(\mathbf{x},t)d^3x = \frac{1}{V}\sum_{\mathbf{k},\mathbf{k}'}\Delta_k\Delta_{k'}\int e^{-i(\mathbf{k}+\mathbf{k}')\cdot\mathbf{x}}d^3x \\ &= \sum_{\mathbf{k},\mathbf{k}'}\Delta_k\Delta_{k'}\delta_{\mathbf{k},-\mathbf{k}'} = \sum_{\mathbf{k}}\Delta_k\Delta_{-k}.\end{aligned}$$

Since $\Delta(\mathbf{x},t)$ is a real function, and it does not matter whether we sum over all $\mathbf{k}$ or all $-\mathbf{k}$, we must have

$$\begin{aligned}\Delta^*(\mathbf{x},t) &= \sum_{\mathbf{k}}\Delta_k^* e^{i\mathbf{k}\cdot\mathbf{x}} \\ &= \sum_{\mathbf{k}}\Delta_k e^{-i\mathbf{k}\cdot\mathbf{x}} = \sum_{\mathbf{k}}\Delta_{-k}e^{i\mathbf{k}\cdot\mathbf{x}}\end{aligned}$$

which gives

$$\Delta_{-k}(t) = \Delta_k^*(t). \tag{4.62}$$

Therefore,

$$\begin{aligned}\langle\Delta^2(\mathbf{x},t)\rangle &= \sum_{\mathbf{k}}|\Delta_k(t)|^2 \\ &= \frac{1}{(2\pi)^3}\int|\Delta_k(t)|^2d^3k \equiv \frac{1}{(2\pi)^3}\int P(\mathbf{k},t)d^3k, \qquad (4.63)\end{aligned}$$

where we have defined the *power spectrum* of the density fluctuations as

$$P(\mathbf{k},t) \equiv |\Delta_k(t)|^2. \tag{4.64}$$

This quantity then gives the standard deviation of the fluctuations on the length scale associated with the wave number k and therefore the strength of the fluctuations on this scale. In normal circumstances, P will be independent of the direction of $\mathbf{k}$ (this is because Δ obeys a differential equation

which is invariant under spatial rotations, and if the initial conditions are rotationally invariant, the solutions will also be so. Inflationary models usually give rise to rotationally invariant initial conditions), and we get

$$\langle \Delta^2(\mathbf{x},t)\rangle = \frac{1}{2\pi^2}\int_0^\infty k^2 P(k) dk. \tag{4.65}$$

An important observational quantity is the two-point correlation function (hereafter called just the correlation function) $\xi(\mathbf{r},t)$ for the distribution of galaxies. It is defined by counting the number of galaxies with a given separation r. If we consider the contribution to this from two small volumes dV_1 around position $\mathbf{x}$ and dV_2 around position $\mathbf{x}+\mathbf{r}$, for a completely uniform distribution of galaxies this will be given by $dN_{12} = \bar{n}^2 dV_1 dV_2$. If there are deviations from a uniform distribution, we can write the contribution as

$$dN_{12} = \bar{n}^2[1+\xi(\mathbf{r},t)]dV_1 dV_2, \tag{4.66}$$

where we have defined the correlation function ξ so that it gives the deviation from a completely uniform, random distribution of galaxies. We next assume that the distribution of galaxies is directly proportional to the distribution of matter. This is a dubious assumption on small scales, but has been tested and seems to hold on large scales. We can then write

$$\begin{aligned} dN_{12} &= \langle \rho(\mathbf{x},t)\rho(\mathbf{x}+\mathbf{r},t)\rangle dV_1 dV_2 \\ &= \rho_0^2 \langle [1+\Delta(\mathbf{x},t)][1+\Delta(\mathbf{x}+\mathbf{r},t)]\rangle dV_1 dV_2 \\ &= \rho_0^2[1+\langle \Delta(\mathbf{x},t)\Delta(\mathbf{x}+\mathbf{r},t)\rangle] dV_1 dV_2, \end{aligned} \tag{4.67}$$

where we have used $\langle \Delta \rangle = 0$. We therefore see that

$$\xi(\mathbf{r},t) = \langle \Delta(\mathbf{x},t)\Delta(\mathbf{x}+\mathbf{r},t)\rangle. \tag{4.68}$$

We can now derive a relation between the correlation function and the power spectrum:

$$\begin{aligned} \xi(\mathbf{r},t) &= \langle \Delta(\mathbf{x},t)\Delta(\mathbf{x}+\mathbf{r},t)\rangle = \langle \Delta(\mathbf{x},t)\Delta^*(\mathbf{x}+\mathbf{r},t)\rangle \\ &= \langle \sum_{\mathbf{k},\mathbf{k}'} \Delta_k(t)\Delta^*_{k'}(t) e^{-i\mathbf{k}\cdot\mathbf{x}} e^{i\mathbf{k}'\cdot(\mathbf{x}+\mathbf{r})}\rangle \\ &= \sum_{\mathbf{k},\mathbf{k}'} \Delta_k(t)\Delta^*_{k'}(t) e^{-i\mathbf{k}'\cdot\mathbf{r}} \frac{1}{V}\int e^{i(\mathbf{k}'-\mathbf{k})\cdot\mathbf{x}} d^3x \\ &= \sum_{\mathbf{k}} |\Delta_k(t)|^2 e^{-i\mathbf{k}\cdot\mathbf{r}} \\ &= \frac{1}{(2\pi)^3}\int |\Delta_k(t)|^2 e^{-i\mathbf{k}\cdot\mathbf{r}} d^3k \\ &= \frac{1}{(2\pi)^3}\int P(\mathbf{k},t) e^{-i\mathbf{k}\cdot\mathbf{r}} d^3k. \end{aligned} \tag{4.69}$$

We have now shown that the correlation function ξ is the Fourier transform of the power spectrum P. If P is independent of the direction of $\mathbf{k}$, so that $P(\mathbf{k},t) = P(k,t)$, we can simplify the expression further:

$$\begin{aligned}\xi(\mathbf{r},t) = \xi(r,t) &= \frac{1}{(2\pi)^3}\int_0^{2\pi} d\phi \int_{-1}^{+1} d(\cos\theta) \int_0^\infty dk k^2 P(k,t) e^{-ikr\cos\theta} \\ &= \frac{1}{4\pi^2}\int_0^\infty dk k^2 P(k,t) \frac{1}{ikr}(e^{ikr} - e^{-ikr}) \\ &= \frac{1}{2\pi^2}\int_0^\infty dk k^2 P(k,t) \frac{\sin(kr)}{kr}, \end{aligned} \tag{4.70}$$

where we have chosen the direction of the k_z axis along $\mathbf{r}$. We see that in this case $\xi(r,t)$ is also isotropic, and is given by the integral of $P(k,t)$ weighted by a filter function which damps contributions from values of k where $k > 1/r$.

4.8 Fluctuations in the cosmic microwave background

The temperature fluctuations in the cosmic microwave background (CMB) are an important source of information about the universe. We will in the following section look at the physics behind the fluctuations on angular scales of a few degrees or less, the so-called acoustic peaks.

The mean temperature of the CMB is $T_0 \approx 2.73$ K. However, there are small deviations from the mean temperature depending on the direction of observation. The relative deviation from the mean is written

$$\frac{\Delta T}{T_0}(\theta,\phi) = \frac{T(\theta,\phi) - T_0}{T_0}, \tag{4.71}$$

and it is practical to decompose $\Delta T/T_0$ in spherical harmonics:

$$\frac{\Delta T}{T_0}(\theta,\phi) = \sum_{\ell=0}^{\infty}\sum_{m=-\ell}^{\ell} a_{\ell m} Y_{\ell m}(\theta,\phi), \tag{4.72}$$

where the spherical harmonics obey the orthogonality relation

$$\int Y^*_{\ell m} Y_{\ell' m'} d\Omega = \delta_{\ell\ell'}\delta_{mm'}. \tag{4.73}$$

The coefficients $a_{\ell m}$ are given by

$$a_{\ell m} = \int \frac{\Delta T}{T_0}(\theta,\phi) Y_{\ell m}(\theta,\phi) d\Omega. \tag{4.74}$$

The standard prediction from inflationary models is that the coefficients $a_{\ell m}$ have a Gaussian distribution with uniformly distributed phases between 0

and 2π. Then each of the $2\ell+1$ coefficients $a_{\ell m}$ associated with multipole ℓ will give an independent estimate of the amplitude of the temperature fluctuation on this angular scale. The power spectrum of the fluctuations is assumed to be circular symmetric around each point (that is, independent of ϕ), so that $a^*_{\ell m}a_{\ell m}$ averaged over the whole sky gives an estimate of the power associated with multipole ℓ:

$$C_\ell = \frac{1}{2\ell+1}\sum_{m=-\ell}^{\ell} a^*_{\ell m}a_{\ell m} = \langle |a_{\ell m}|^2\rangle. \tag{4.75}$$

If the fluctuations are Gaussian, the power spectrum C_ℓ gives a complete statistical description of the temperature fluctuations. It is related to the two-point correlation function of the fluctuations by

$$C(\theta) = \langle \frac{\Delta T(\mathbf{n}_1)}{T_0}\frac{\Delta T(\mathbf{n}_2)}{T_0}\rangle = \frac{1}{4\pi}\sum_{\ell=0}^{\infty}(2\ell+1)C_\ell P_\ell(\cos\theta), \tag{4.76}$$

where $\mathbf{n}_1$ and $\mathbf{n}_2$ are unit vectors in the two directions of observation, $\cos\theta = \mathbf{n}_1\cdot\mathbf{n}_2$, and P_ℓ is the Legendre polynomial of degree ℓ.

We will in the following look at the so-called acoustic oscillations in the power spectrum of the CMB. These have their origin in the physics in the baryon-photon plasma present around the epoch of recombination. In the description of these oscillations, we must then take into account that we are dealing with a system with (at least) three components: photons, baryons, and dark matter. The dark matter dominates the energy density and the gravitational fields present, but does not interact in other ways with the photons and the baryons. The latter two are coupled two each other by Thomson scattering, and as a first approximation we can assume that they are so strongly coupled to each other that we can treat the photons and the baryons as a single fluid. In this fluid we have

$$n_\gamma \propto n_b \propto \rho_b \tag{4.77}$$

$$n_\gamma \propto T^3, \tag{4.78}$$

which gives $T \propto \rho_b^{1/3}$ and

$$\frac{\Delta T}{T} \equiv \Theta_0 = \frac{1}{3}\frac{\Delta\rho_b}{\rho_b} = \frac{1}{3}\Delta_b. \tag{4.79}$$

In other words, the fluctuations in the temperature are determined by the density perturbations in the baryonic matter. The equation describing the time evolution of these is of the form

$$\frac{d^2\Delta_b}{dt^2} + 2\frac{\dot a}{a}\frac{d\Delta_b}{dt} = \text{gravitational term} - \text{pressure term}. \tag{4.80}$$

If we make the approximation that gravity is dominated by the dark matter with density ρ_D, and that the pressure term is dominated by the baryon-photon plasma with speed of sound c_s, we get

$$\frac{d^2\Delta_b}{dt^2} + 2\frac{\dot{a}}{a}\frac{d\Delta_b}{dt} = 4\pi G\rho_D\Delta_D - \Delta_b k^2 c_s^2. \tag{4.81}$$

In addition, we will assume that we can neglect the Hubble friction term and take $\dot{a} \approx 0$. Inserting $\Theta_0 = \Delta_b/3$ we get

$$\frac{d^2\Theta_0}{dt^2} = \frac{4\pi G\Delta_D\rho_D}{3} - k^2 c_s^2\Theta_0. \tag{4.82}$$

We can relate the first term on the right-hand side to the fluctuations in the gravitational potential via Poisson's equation

$$\nabla^2\delta\phi = 4\pi G\rho_D\Delta_D.$$

For a single Fourier mode $\delta\phi = \phi_k \exp(i\mathbf{k}\cdot\mathbf{x})$ we find by substitution

$$\phi_k = -\frac{4\pi G\rho_D\Delta_D}{k^2}, \tag{4.83}$$

so that

$$\frac{d^2\Theta_0}{dt^2} = -\frac{1}{3}k^2\phi_k - k^2 c_s^2\Theta_0. \tag{4.84}$$

We look at adiabatic perturbations, and the entropy is dominated by the photons,

$$S \propto T^3 V \propto \frac{T^3}{m_b/V} \propto \frac{T^3}{\rho_b} \propto \frac{\rho_r^{3/4}}{\rho_b}, \tag{4.85}$$

where m_b is the baryon mass, and we recall that $\rho_r \propto T^4$, so we have

$$\frac{\delta S}{S} = \frac{3}{4}\frac{\delta\rho_r}{\rho_r} - \frac{\delta\rho_b}{\rho_b} = 3\frac{\delta T}{T} - \frac{\delta\rho_b}{\rho_b} = 0, \tag{4.86}$$

so that

$$\Delta_b = \frac{\delta\rho_b}{\rho_b} = 3\frac{\delta T}{T} = \frac{3}{4}\frac{\delta\rho_r}{\rho_r}. \tag{4.87}$$

The speed of sound is given by

$$c_s = \left(\frac{\partial p}{\partial\rho}\right)_S^{1/2}. \tag{4.88}$$

In the photon-baryon plasma we have $\rho = \rho_b + \rho_r$ and $p = p_b + p_r \approx p_r = \rho_r c^2/3$. Therefore we get

$$\begin{aligned} c_s^2 &= \frac{\delta p}{\delta\rho} = \frac{\delta\rho_r c^2/3}{\delta\rho_b + \delta\rho_r} \\ &= \frac{c^2}{3}\frac{1}{1+\frac{\delta\rho_b}{\delta\rho_r}}, \end{aligned} \tag{4.89}$$

so that

$$\begin{aligned} c_s &= \frac{c}{\sqrt{3}}\left[1+\left(\frac{\delta\rho_b}{\delta\rho_r}\right)_S\right]^{-1/2} \\ &= \frac{c}{\sqrt{3}}\left(1+\frac{3}{4}\frac{\rho_b}{\rho_r}\right)^{-1/2} \\ &= \frac{c}{\sqrt{3(1+\mathcal{R})}}, \end{aligned} \tag{4.90}$$

where $\mathcal{R} \equiv 3\rho_b/4\rho_r$.

We will simplify the problem further by assuming that ϕ_k and c_s are independent of time. Then equation (4.84) is a simple oscillator equation, and by substitution one can show that

$$\begin{aligned} \Theta_0(t) &= \left[\Theta_0(0)+\frac{(1+\mathcal{R})}{c^2}\phi_k\right]\cos(kc_st) \\ &+ \frac{1}{kc_s}\dot{\Theta}_0(0)\sin(kc_st)-\frac{(1+\mathcal{R})}{c^2}\phi_k \end{aligned} \tag{4.91}$$

is a solution. After recombination, the photons will propagate freely towards us, so we see the fluctuations today more or less as they were at the time $t = t_{\rm rec}$ of recombination. Then,

$$kc_st_{\rm rec} = k\lambda_S, \tag{4.92}$$

where λ_S is the so-called sound horizon: the distance a sound wave with speed c_s has covered by the time $t_{\rm rec}$. The temperature fluctuations can therefore be written as

$$\begin{aligned} \Theta_0(t_{\rm rec}) &= \left[\Theta_0(0)+\frac{(1+\mathcal{R})}{c^2}\phi_k\right]\cos(k\lambda_S) \\ &+ \frac{1}{kc_S}\dot{\Theta}_0(0)\sin(k\lambda_S)-\frac{(1+\mathcal{R})}{c^2}\phi_k. \end{aligned} \tag{4.93}$$

We therefore get oscillations in k space, which become oscillations in ℓ space after projection on the sky. We see that the initial conditions enter via the terms containing $\Theta_0(0)$ and $\dot{\Theta}_0(0)$. The case $\Theta(0) \neq 0$, $\dot{\Theta}(0) = 0$ are called adiabatic initial conditions, while the case $\Theta(0) = 0$, $\dot{\Theta}(0) \neq 0$ is called isocurvature initial conditions. The simplest inflationary modes give rise to adiabatic initial conditions.

Another thing we have not yet taken into account is the fact that the oscillations take place within gravitational potential wells with amplitude ϕ_k. The *observed* oscillation is therefore, for adiabatic initial conditions,

$$\Theta_0(t_{\rm rec})+\frac{\phi_k}{c^2} = \left[\Theta_0(0)+\frac{(1+\mathcal{R})}{c^2}\phi_k\right]\cos(k\lambda_S)-\frac{\mathcal{R}}{c^2}\phi_k. \tag{4.94}$$

The term in the angular brackets correspond to horizon-scale fluctuations, the so-called Sachs-Wolfe effect, and one can show that

$$\Theta_0(0) + \frac{\phi_k}{c^2} = \frac{\phi_k}{3c^2}, \tag{4.95}$$

and that the observed temperature fluctuations therefore can be written as

$$\left(\frac{\Delta T}{T_0}\right)_{\rm eff} = \frac{\phi_k}{3c^2}(1+3\mathcal{R})\cos(k\lambda_S) - \frac{\mathcal{R}}{c^2}\phi_k. \tag{4.96}$$

The first extremal value occurs for $k\lambda_S = \pi$, which gives

$$\left(\frac{\Delta T}{T_0}\right)_{\rm eff} = -\frac{\phi_k}{3c^2}(1+6\mathcal{R}), \tag{4.97}$$

and the next one occurs for $k\lambda_S = 2\pi$, giving

$$\left(\frac{\Delta T}{T_0}\right)_{\rm eff} = \frac{\phi_k}{3c^2} \tag{4.98}$$

so we see that the ratio of the first and the second extremal value (which corresponds roughly to the ratio of the first and second peak in the power spectrum) can be used to determine the $\mathcal{R}$, which again gives the baryon density $\Omega_{b0}h^2$.

4.9 Exercises

Exercise 4.1 (From the exam in AST4220, 2003)

In this problem you will determine the time evolution of perturbations in the matter density ρ_m in different epochs of the history of the universe on scales much larger than the Jeans length $\lambda_{\rm J}$. The equation describing the time evolution of a Fourier mode of the density perturbations, $\Delta_{\mathbf{k}}(t)$, is

$$\frac{d^2\Delta_{\mathbf{k}}}{dt^2} + 2\frac{\dot{a}}{a}\frac{d\Delta_{\mathbf{k}}}{dt} = 4\pi G\rho_m\Delta_{\mathbf{k}}(t),$$

where ρ_m is the average matter density

a) Describe briefly, without any algebra, how this equation is derived.

b) Show that the equation gives the following results:

1. In a radiation dominated universe ($\rho_m \approx 0$):

$$\Delta_{\mathbf{k}}(t) = B_1 + B_2 \ln t$$

2. In a matter dominated universe ($\rho_m = \rho_{\rm c}$):

$$\Delta_{\mathbf{k}}(t) = C_1 t^{-1} + C_2 t^{2/3}$$

3. In a universe dominated by a cosmological constant (de Sitter universe, $\rho_m \approx 0$):

$$\Delta_{\mathbf{k}}(t) = D_1 + D_2 e^{-2H_\Lambda t},$$

where $H_\Lambda = \sqrt{8\pi G\rho_\Lambda/3}$.

In the expressions above, B_1, B_2, C_1, C_2 and D_1, D_2 are constants of integration.

c) Explain physically why the perturbations grow slowly or not at all in the radiation dominated epoch and in the de Sitter universe.

Exercise 4.2 (From the exam in AST4220, 2004)

In this problem you will study the growth of density perturbations in an Einstein-de Sitter universe (a flat universe with matter only, and no cosmological constant.)

a) Write down expressions for $a(t)$, $H(t) = \dot{a}/a$ and the matter density $\rho_{\rm m}(t)$ in this model for the unperturbed case.

Assume that the matter consists of two components: some form of cold, non-relativistic dark matter, and massive neutrinos. Also, assume that the neutrinos have so high thermal velocities that they do not clump, so that the only density perturbations are those in the cold dark matter. Denote the density parameter of the neutrinos by Ω_ν, and that of the cold dark matter by $\Omega_{\rm cdm}$, so that $\Omega_{\rm m} = \Omega_{\rm cdm} + \Omega_\nu$, and define $f_\nu = \Omega_\nu/\Omega_{\rm m}$.

b) Start from the equation for the time evolution of a Fourier mode Δ_k of the density perturbations, derived in the lectures, and justify that it can be written as

$$\ddot{\Delta}_k + \frac{4}{3t}\dot{\Delta}_k = \frac{2}{3}(1 - f_\nu)\frac{\Delta_k}{t^2},$$

in the situation considered in this problem.

c) Assume that this equation has a power-law solution $\Delta_k \propto t^\alpha$, and show that the growing mode solution is

$$\alpha = \frac{1}{6}\left[5\sqrt{1 - \frac{24}{25}f_\nu} - 1\right],$$

and that

$$\alpha \approx \frac{2}{3}\left(1 - \frac{3}{5}f_\nu\right),$$

for $f_\nu \ll 1$.

d) Density perturbations only start to grow after matter-radiation equality at a redshift $1+z_{\rm eq} \approx 23900\Omega_{\rm m}h^2$. Show that the perturbations in this model by the current epoch ($a = a_0 \equiv 1$, $z = 0$) have grown by a factor

$$\frac{\Delta_k(z=0)}{\Delta_k(z=z_{\rm eq})} = (1+z_{\rm eq})e^{-\frac{3}{5}f_\nu \ln(1+z_{\rm eq})}.$$

(Hint: write the solution for Δ_k in terms of the scale factor a.)

e) For $\Omega_{\rm m} = 1$, $h = 0.5$, compare the growth of density perturbations from $z_{\rm eq}$ to $z = 0$ in the cases $f_\nu = 0.1$ and $f_\nu = 0$.

Exercise 4.3

Assume that we are in the matter dominated epoch (i.e., pretend that there is no cosmic acceleration), and that the matter density in the universe consists of two components: a component of dark matter X which only interacts with the other components through gravity, and ordinary baryonic matter B. The total background density (mean density) in the universe is then $\rho_b = \rho_b^X + \rho_b^B$. We assume that the density of baryonic matter is much smaller than the density of dark matter, $\rho_b^B \ll \rho_b^X$. Density fluctuations in the two components are given by

$$\Delta_X(\mathbf{x},t) = \frac{\rho^X(\mathbf{x},t) - \rho_b^X(t)}{\rho_b^X(t)},$$

and

$$\Delta_B(\mathbf{x},t) = \frac{\rho^B(\mathbf{x},t) - \rho_b^B(t)}{\rho_b^B(t)},$$

a) Explain briefly, without calculations, that for scales much larger than the Jeans length, but much smaller than the particle horizon of the universe, the time evolution of linear perturbations in the two components are given by

$$\begin{aligned} \frac{\partial^2\Delta_X}{\partial t^2} + 2\frac{\dot{a}}{a}\frac{\partial\Delta_X}{\partial t} &= 4\pi G\rho_b^X\Delta_X, \\ \frac{\partial^2\Delta_B}{\partial t^2} + 2\frac{\dot{a}}{a}\frac{\partial\Delta_B}{\partial t} &= 4\pi G\rho_b^X\Delta_X. \end{aligned}$$

b) Show that the growing solution for density fluctuations in the dark matter component is

$$\Delta_X(\mathbf{x},t) = C(\mathbf{x})t^{2/3}.$$

c) While the fluctuations in the baryonic component only can grow after recombination, fluctuations in the dark matter component can start

growing earlier (when the energy density in radiation equals the energy density in matter). Show that if the fluctuations in the dark matter are given by the growing solution found in b), the fluctuations in the baryonic component are given by ($t_{\rm rec}$ and $\Delta_X^{\rm rec}$ denote time and density fluctuation at recombination):

$$\Delta_B(\mathbf{x},t) = \Delta_X^{\rm rec}(\mathbf{x})\left[\left(\frac{t}{t_{\rm rec}}\right)^{2/3} + 2\left(\frac{t}{t_{\rm rec}}\right)^{-1/3} - 3\right].$$

d) Assume that recombination happens instantaneously at redshift $z_{\rm rec} = 999$ and that fluctuations in the dark matter then have amplitude 10^{-3}. Compute the ratio between fluctuations in the baryonic component and in the dark matter component at redshifts given by $1+z = 1000$, 500, 100, 10, and today. Explain why there might be a problem for a purely baryonic model that we today have bound structures in the universe (galaxies etc.) while the fluctuations on galaxy scales in the cosmic microwave background radiation are much smaller than 10^{-3}. Why is the dark, weakly interacting matter hypothesis one possible way of solving this potential problem?

Exercise 4.4 (From the exam in AST4220, 2004)

Consider a matter-dominated universe with $\Omega_{\rm m0} < 1$.

a) Explain why we can neglect the curvature term in the first Friedmann equation early in the matter-dominated period. Show that the Hubble parameter can be written

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{H_0^2\Omega_{\rm m0}}{a^3}$$

b) Determine $a(t)$.

c) We will now consider how density perturbations grow on scales much larger than the Jeans length in this model. Show that we get the same growing mode, $\Delta_{\mathbf{k}}(t) \propto t^{2/3}$, as in the Einstein-de Sitter model.

d) If we want to distinguish between an open universe and the Einstein-de Sitter model by observing the growth of perturbations, should we use observations made at high rdshifts? Justify your conclusion.