AST3220 – Cosmology I

Øystein Elgarøy

Contents

Chapter 1

The Robertson-Walker line element

1.1 What is cosmology and why should you care?

Congratulations on choosing to study cosmology! You probably had some ideas about what cosmology is when you signed up, perhaps you even read the course description. But before we dive into the subject, I want to say a few words about what *I* think cosmology is, and why it is worth studying.

In one way ore another, astronomers have been doing cosmology throughout all ages. For example, Ptolemy's system of circles and epicycles was a model for the whole Universe, because at the time the known universe consisted of the Sun, the planets, Moon, and the fixed stars. When the Copernican system, perfected by Kepler, replaced Ptolemy's, this could be seen as a change in model for the Universe. But after Galileo's observational discoveries with his telescope and Newton's laws of motion and gravity, it was clear that the fixed stars were not so fixed after all. Newton worried about the stability of a system of stars under their own mutually attractive gravitational forces. Would it collapse? He argued (erroneously) that it could be stable if the distribution of stars extended to infinity in all directions.

Eventually astronomers came to understand that our solar system is part of a large structure: our Galaxy, the Milky Way system. One of the first to try to map out its structure was the great observer William Herschel. Since he was trying to work out a picture of what was then believed to be the whole universe, his work was in essence cosmology.

Less than a century has passed since we first began to understand the true enormity of the Universe. In 1924, Edwin Hubble was able to estimate the distance to the mysterious Andromeda nebula. Astronomers had been aware of the existence of several nebulae for some time, and had been discussing whether they were part of the Milky Way or not. The size and structure of the Milky Way had been roughly worked out by Harlow Shapley in the early 1900-s, so if the nebulae were external to the Milky Way, astronomers realized that they would have to be gigantic systems of stars like the Milky Way, galaxies in their own right, to be so clearly visible to us. So when Hubble estimated the distance to the Andromeda nebula and found it to be far outside our Galaxy, he transformed our perceptions of what the Universe is, and therefore what cosmology is about.

With Einstein's new theory of gravity, general relativity, it became clear that space, matter and time are tied together. The general belief was that the Universe was static and eternal, and Einstein constructed a cosmological model consistent with this belief. Alexander Friedmann and Georges Lemaitre were more excited by solutions of Einstein's equation which described dynamic universes. A new discovery by Hubble in 1929 proved them right: The galaxies are moving a way from us in the way you would expect if space is expanding.

Today we know that the Universe started expanding about 14 billion years ago. It is possibly spatially infinite, but it is at any rate much larger than the part we can see, the observable universe. Inside the observable universe we estimate that there are about 1000 billion galaxies. They are not distributed randomly in space, but seem to form a sponge-like structure with walls and voids.

In one sense, cosmology is what cosmologists do. And cosmologists do many different things. But in general, what we want to do is to find out what the universe consists of, and how it came to be the way it looks today. Since it has a history, we know that the Universe was once very different from how we see it now. Some of us are mostly concerned with studying the very earliest phases of the history, and this area has strong connections with particle physics. Others are interested in understanding how the first galaxies formed, and others again try to figure out how the galaxies came to be distributed in the way we see them. And others yet struggle to understand why, contrary to earlier expectations, the Universe now seems to be expanding at an accelerating rate. Since we are all trying to understand the same Universe and the same history, there are, of course, connections between all these areas.

What will we study in this course? We will mostly be dealing with building models for the Universe on scales large enough for the details of the galaxy distribution to be insignificant. We will also be interested in combining these models with particle, nuclear, and atomic physics to understand the evolution of the Universe during its first 400 000 years or so. You will learn some basic theoretical concepts and tools, and how to relate them to observable quantities. In the final parts of the course, we will study the first stages of structure formation in the Universe, and consider the most popular idea for the origin of these structures: the inflationary universe.

For students who want to specialise in astronomy, this course is compulsory. But I guess if you are the type who likes astronomy, cosmology does anyway sound somewhat interesting to you. For the rest of you, you have decided freely to follow a course that you don't really need. So, again, I can assume that you chose it because it sounded interesting. However, I think it is still worthwhile to ask the question of why we should bother. It is too rarely asked about most things, in my opinion.

Let me be clear on one thing: The emphasis in this course is not on giving you skills and knowledge which will be of immediate use in a career in industry. Model building, solving equations, testing theories against observations, these are all transferable skills, so I am not saying that this course won't give you anything that you can use if you decide to leave academia at some point in the future. But this is not why I teach this course, and if this is your main goal, you could spend your time more wisely on other courses. I teach this course because I love cosmology. I think it is remarkable and beautiful how we can use basic principles of physics to deduce the history of the Universe. We will reach some stunning conclusions, for example that we and everything we see around us are probably the result of tiny quantum fluctuations that arose some 10^{-35} seconds after the Universe started to expand. We will see how we need to understand the smallest constituents of matter if we want to understand the biggest structures in the Universe. We will see how cosmology has provided physics with some of its biggest puzzles today, like the nature of dark matter, and the energy of empty space. In my mind, no physicist, in fact no person, can call him/herself truly educated without at least an aquintance with the questions, ideas and methods of cosmology. You should study cosmology because it is fun, exciting and good for your soul. I enjoy teaching it for the same reasons. While my excitement may not always shine through in my lectures and in these notes, I am after all an old, cynical curmudgeon, I am passionate about the subject, and I want you to be so, too. We are about to see some of the highlights of the intellectual achievements of the human species. Let's get started.

1.2 Making a principle out of a necessity

The Universe is by definition everything that (physically) exists. It is necessarily a fairly complicated system to study as a whole. If cosmologists had to care about every tiny detail, the project could not even get off the ground. Fortunately, simplifications can be made. For example, the shenanigans of the human species have so far not had any consequences on a cosmic scale. Therefore, we decide to not care about humans and leave them to the sociologists, psychologists, and historians. The solar system arose late in the history of the Universe, and it seems to have no special significance on a larger scale, so we leave its study to the solar physicists and the planetary scientists. The smallest structures of interest in cosmology, are in fact galaxies. Anything smaller than that, we will in general not care about.

The distribution of galaxies is complicated, but it shows some patterns and regularities, and as cosmologists we want to understand them. But we need an even simpler point to start from. Fortunately, if you look at the distribution of galaxies with a very course filter that blurs structures smaller than a few hundred Mpc (megaparsecs), it looks very uniform. The distribution of matter, averaged over distances of this order, is to a high degree homogeneous. It also looks isotropic, i.e., the same in all directions. This is fortunate, because if it were otherwise, it would be very difficult set up mathematical models. We now have some pretty good evidence that the matter distribution is homogeneous and isotropic on large scales. But this assumption was being made already at the start of modern cosmology, when the evidence for it was much weaker. As is often the case, when the evidence for your claim is weak, you make it sound more convincing by elevating it to a principle. Thus, we have:

- **The Cosmological Principle:** Averaged over sufficiently large scales, the matter distribution in the Universe is homogeneous and isotropic.

Mathematically, this means that if we describe the matter density by a function ρ, it can not be a function of position: It must be the same everywhere. It can at most be a function of time alone. The density may vary in time, but at any given time (we will return to what we mean by 'time' here), it is the same everywhere in the Universe. So the Cosmological Principle can be formulated more simply as $\rho = \rho(t)$ (on sufficiently large scales.)

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 one of the important features of this theory.

1.3 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)

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

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.

1.4 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 $g = 9.81$ 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.

1.5 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 on 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

$$
\begin{aligned}
r &= a\sin\theta \qquad (1.3)\\
\theta &= \frac{s}{a}, \qquad (1.4)
\end{aligned}
$$

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

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

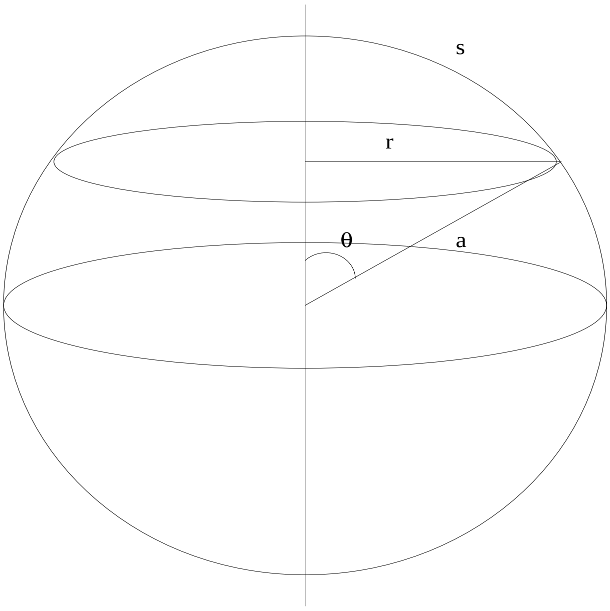

Figure 1.1: Symbols used in the discussion of the curvature of a spherical surface. Note that the circumference of the circle is $2\pi r$, but the radius (the distance from the center to the perimeter) as measured by a creature confined to walk along the surface of the sphere is s.

$$= 2\pi s \left(1 - \frac{s^2}{6a^2} + \ldots\right), \tag{1.5}$$

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.

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 smaller 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 $2\pi s$.

1.6 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 (5.15) 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 the flat 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, i.e., it satisfies the Cosmological Principle. 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^2dt^2, so we can finally write

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

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 finite (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.

1.7 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.

1.7.1 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

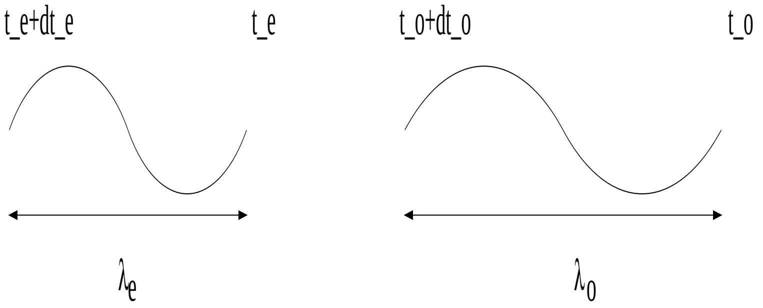

Figure 1.2: An electromagnetic wave travelling through the expanding universe is stretched.

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

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.

1.7.2 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 $d_{\mathrm{P}}(t)$, 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

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

where $\mathcal{S}_k^{-1}(r) = \sin^{-1} r$ for $k = +1$, $\mathcal{S}_k^{-1}(r) = r$ for $k = 0$ and $\mathcal{S}_k^{-1}(r) = \sinh^{-1} r$ for $k = -1$. We see that this results agrees with our intuition for the spatially flat case, $k = 0$: $d_\mathrm{P}(t) = a(t)r$, 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_P is a function of t, the relative distance between the two points is increasing as the Universe expands. The rate of increase of this distance is

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

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

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

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 $H_0 \equiv H(t_0) = (67.8 \pm 0.9)\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$.[1] Note that it is common to introduce the dimensionless

[1] This is the value given in the astrophysical constants table compiled by the Particle Data Group (pdg.lbl.gov), and it is derived from the temperature fluctuations in the cosmoic microwave background. There is a slight tension between this result and more traditional measurements based on observations of the distance-redshift relationship, which tend to give larger values.

Hubble constant by writing

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

where we have $h \approx 0.68$ today.

1.7.3 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_{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

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

The received flux l is relatively easy to measure, and if we know L, we can then compute d_{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_{o}. Its area is given by definition of the coordinate r as $4\pi a^2(t_{\text{o}})r^2$. The photons emitted at P at the time t have had their wavelengths stretched by a factor $a(t_{\text{o}})/a(t)$ 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 $\delta t_{\text{o}} = a(t_{\text{o}})/a(t)\delta t$, 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

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

and using the definition (1.14) we get

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

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

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

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.

1.7.4 The angular diameter distance

Another common measure of distance is the *angular diameter distance*, d_{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

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

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_{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

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

so that

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

We therefore find

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

where t_{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_{L} and d_{A}:

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

and hence this ratio is model-independent.

1.7.5 The comoving coordinate r

The expressions for the luminosity distance and the angular diameter distance of a source P observed at time $t_{\rm o}$ 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

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

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

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

Chapter 2

Newtonian cosmology

2.1 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 spherical region 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

$$\dot{a}^2 + kc^2 = \frac{8\pi G}{3}\rho a^2, \tag{2.1}$$

and if we, although totally unmotivated, postulate that k is the curvature parameter in the RW line element, then equation (7.41) 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

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\left(\rho + \frac{3p}{c^2}\right). \tag{2.2}$$

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).

We can also go some way towards deriving the first Friedmann equation (7.41) 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

$$\dot{\rho} = -3\frac{\dot{a}}{a}\left(\rho + \frac{p}{c^2}\right). \tag{2.3}$$

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 (7.42) 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

$$\dot{a}^2 + kc^2 = \frac{8\pi G}{3}\rho a^2, \tag{2.4}$$

which is identical to equation (7.41).

Note that we derived equation (7.41) using equations (7.42) and (7.43). 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.

2.1.1 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

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

- 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:

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

Chapter 3

Models of the Universe

In this chapter we will build models of the Universe by making assumptions about it contents, plugging them into the Friedmann equations, and see what happens.

3.1 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 (7.41) 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. Therefore,k is not a problem. Since it can only take on the three discrete values -1, 0, or $+1$, we can just solve the equations separately for each of the three values.

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:

$$p = w\rho c^2 \tag{3.1}$$

where w is a constant. We will introduce two important cases here and a third case (the cosmological constant) in section 3.3.

3.1.1 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_{B} is Boltzmann's constant. We rewrite this slightly:

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

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

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

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

$$p = 0, \tag{3.2}$$

that is, dust is pressureless.

3.1.2 Radiation: relativistic matter

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

$$p = \frac{1}{3}\rho c^2, \tag{3.3}$$

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, so 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 (3.3).

3.2 The evolution of the energy density

Equipped with the equation of state, we can now proceed to solve equation (7.43) to obtain ρ as a function of the scale factor a. Having done this, we can then proceed to rewrite equations (7.41) and (7.42) 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 (7.43) 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

$$\rho = \rho_0\left(\frac{a_0}{a}\right)^{3(1+w)}. \qquad (3.4)$$

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

$$\rho = \rho_0 \left(\frac{a_0}{a}\right)^3 , \tag{3.5}$$

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

$$\rho = \rho_0 \left(\frac{a_0}{a}\right)^4 , \tag{3.6}$$

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 $1/a$.

3.3 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. Bear in mind that Einstein did this in 1917, 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 (2.1,2.2) 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 $a(t) = a_0 =$ 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. 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 spacetime. 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):

$$\begin{aligned} \dot{a}^2 + kc^2 &= \frac{8\pi G}{3}\rho a^2 + \frac{\Lambda}{3}a^2 \qquad (3.7) \\ \ddot{a} &= -\frac{4\pi G}{3}\rho a + \frac{\Lambda}{3}a, \qquad (3.8) \end{aligned}$$

where Λ is the cosmological constant. Now, a static solution is possible. Take $a = a_0 =$ constant. Then, equation (3.8) gives

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

and inserting this in equation (3.7) 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. Physicists later realized that empty space, the vacuum, has both energy and pressure, and that this gives rise to a term of exactly the same form as the cosmological constant on the right-hand side of the Friedmann equations. To save labour, I move Einstein's cosmological constant over to this side of the equations and lump them together in an *effective* cosmological constant, which I will still call Λ, and write the Friedmann equations 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 (3.7) we see that

$$\rho_\Lambda = \frac{\Lambda}{8\pi G}. \tag{3.9}$$

Inserting this in the second equation and comparing with equation (3.8) we find

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

which gives

$$p_\Lambda = -\frac{\Lambda}{8\pi G}c^2 = -\rho_\Lambda c^2, \tag{3.10}$$

and his means $w = -1$, so 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 (7.43). This gives

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

so that $\rho_\Lambda = \text{constant} = \Lambda/8\pi G$. 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 a positive 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 now the consensus is that is should be included. 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.

3.4 Some classic cosmological models

We will now make a brief survey of the simplest cosmological models. As a prelude, we consider equation (7.41) 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 $\rho_{\mathrm{c}0}$. Inserting values for the constants, we have

$$\rho_{\mathrm{c}0} = 1.879 \times 10^{-29} h^2 \ \mathrm{g\,cm^{-3}}.$$

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

$$1 = \frac{8\pi G}{3H_0^2}\rho_0 = \frac{\rho_0}{\rho_{\mathrm{c}0}},$$

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

$$\Omega_{i0} \equiv \frac{\rho_0}{\rho_{\mathrm{c}0}}, \tag{3.11}$$

where the subscript 0 denotes that we are considering the density at the present time, and the subscript i is for the component in question, for example dust $i = \mathrm{m}$, radiation $i = \mathrm{r}$, or a cosmological constant $i = \Lambda$. Furthermore, one also introduces a 'curvature density parameter',

$$\Omega_{k0} = -\frac{kc^2}{a_0^2 H_0^2}, \tag{3.12}$$

and hence we can write

$$\Omega_{i0} + \Omega_{k0} = 1. \tag{3.13}$$

3.4.1 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

$$a(t) = a_0 \left(\frac{t}{t_0}\right)^{\frac{2}{3(1+w)}}, \tag{3.14}$$

with

$$t_0 = \frac{2}{3(1+w)H_0}. \tag{3.15}$$

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 new physical effects which current theories cannot describe. 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 expanding 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

$$\begin{aligned} a(t) &= a_0 \left(\frac{t}{t_0}\right)^{\frac{1}{2}} \qquad (3.16) \\ t_0 &= \frac{1}{2H_0}. \end{aligned}$$

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

$$\begin{aligned} a(t) &= a_0 \left(\frac{t}{t_0}\right)^{\frac{2}{3}} && (3.17) \\ t_0 &= \frac{2}{3H_0}. && (3.18) \end{aligned}$$

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

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

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 're-cently', 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.

3.4.2 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

$$\dot{a}^2 = \frac{\Lambda}{3} a^2$$

$$\ddot{a} = \frac{\Lambda}{3} a$$

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 Hubble's law, 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[1] on the geometry of homogeneous and isotropic universe models that the expanding nature of the de Sitter solution was clarified.

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

3.4.3 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 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 energy content and the ultimate fate of the universe only holds in models where all components have $w > -1/3$. We will later see that the addition of a cosmological constant spoils this nice correspondence.

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

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

and so

$$a_{\rm max} = a_0\frac{\Omega_{\rm m0}}{\Omega_{\rm m0} - 1}.$$

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 $\Omega_{\rm m0} = 2$, 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 $\alpha = \frac{\Omega_{\mathrm{m}0}}{\Omega_{\mathrm{m}0}-1}$. 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:

$$\begin{aligned} H_0 t &= \frac{\alpha}{\sqrt{\Omega_{\mathrm{m}0} - 1}}\int_0^{\psi}\sin^2\frac{\psi}{2}d\psi \\ &= \frac{\Omega_{\mathrm{m}0}}{(\Omega_{\mathrm{m}0} - 1)^{3/2}}\frac{1}{2}\int_0^{\psi}(1 - \cos\psi)d\psi \\ &= \frac{1}{2}\frac{\Omega_{\mathrm{m}0}}{(\Omega_{\mathrm{m}0} - 1)^{3/2}}(\psi - \sin\psi). \end{aligned}$$

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

$$\begin{aligned} a(\psi) &= \frac{a_0}{2}\frac{\Omega_{\mathrm{m}0}}{\Omega_{\mathrm{m}0} - 1}(1 - \cos\psi) && (3.19) \\ t(\psi) &= \frac{1}{2H_0}\frac{\Omega_{\mathrm{m}0}}{(\Omega_{\mathrm{m}0} - 1)^{3/2}}(\psi - \sin\psi), && (3.20) \end{aligned}$$

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

$$t_0 = \frac{1}{2H_0}\frac{\Omega_{\rm m0}}{(\Omega_{\rm m0}-1)^{3/2}}\left[\cos^{-1}\left(\frac{2}{\Omega_{\rm m0}}-1\right) - \frac{2}{\Omega_{\rm m0}}\sqrt{\Omega_{\rm m0}-1}\right], \qquad (3.21)$$

and that the lifetime of the universe is

$$t_{\rm crunch} = t(2\pi) = \frac{\pi\Omega_{\rm m0}}{H_0(\Omega_{\rm m0}-1)^{3/2}}. \qquad (3.22)$$

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

$$a(u) = \frac{a_0}{2}\frac{\Omega_{\rm m0}}{1-\Omega_{\rm m0}}(\cosh u - 1) \qquad (3.23)$$

$$t(u) = \frac{\Omega_{\rm m0}}{2H_0(1-\Omega_{\rm m0})^{3/2}}(\sinh u - u), \qquad (3.24)$$

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

$$t_0 = \frac{1}{2H_0}\frac{\Omega_{\rm m0}}{(1-\Omega_{\rm m0})^{3/2}}\left[\frac{2}{\Omega_{\rm m0}}\sqrt{1-\Omega_{\rm m0}} - \cosh^{-1}\left(\frac{2}{\Omega_{\rm m0}}-1\right)\right]. \qquad (3.25)$$

3.4.4 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.

3.4.5 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 $1/a^2$ whereas the matter and radiation terms fall as $1/a^3$ and $1/a^4$ 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 0.005. In this case, with $\Omega_{\rm m0} = 0.3$, we get

$$\frac{a}{a_0} \ll 60$$

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

$$\frac{a}{a_0} \ll \sqrt{\frac{\Omega_{\rm r0}}{\Omega_{\rm k0}}} = \sqrt{\frac{\Omega_{\rm m0}}{\Omega_{\rm k0}}\frac{\Omega_{\rm r0}}{\Omega_{\rm m0}}} \sim 4\sqrt{\frac{a_{\rm eq}}{a_0}} \approx 0.52.$$

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

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

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

$$H_0 t = \frac{4(a_{\rm eq}/a_0)^2}{3\sqrt{\Omega_{\rm r0}}}\left[1-\left(1-\frac{a}{2a_{\rm eq}}\right)\left(1+\frac{a}{a_{\rm eq}}\right)^{1/2}\right]. \qquad (3.26)$$

From this we can find the age of the universe at matter-radiation equality. Inserting $a = a_{\rm eq}$ in (3.26), 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 $h = 0.7$, $\Omega_{\rm m0} = 0.3$, $\Omega_{\rm r0} = 8.4\times 10^{-5}$ 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. But we will see later in the course that many of the important events in the history of the Universe took place in the radiation-dominated era.

Equation (3.26) 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.

3.4.6 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

$$H_0 t = \frac{2}{3\sqrt{1-\Omega_{\rm m0}}}\sinh^{-1}\left[\left(\frac{a}{a_{\rm m\Lambda}}\right)^{3/2}\right]. \tag{3.27}$$

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

$$a(t) = a_0\left(\frac{\Omega_{\rm m0}}{1-\Omega_{\rm m0}}\right)^{1/3}\left[\sinh\left(\frac{3}{2}\sqrt{1-\Omega_{\rm m0}}H_0 t\right)\right]^{2/3}. \tag{3.28}$$

The present age of the universe in this model is found by inserting $a = a_0$ in equation (3.27):

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

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 (3.27) 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

$$a(t) \approx a_{\mathrm{m}\Lambda}\left(\frac{3}{2}\sqrt{1-\Omega_{\mathrm{m}0}}H_0t\right)^{2/3},$$

so $a \propto t^{2/3}$ in the early stages, as expected for a matter-dominated model. In the opposite limit, $a \gg a_{\mathrm{m}\Lambda}$, we can use the approximation $\sinh^{-1}x \approx \ln(2x)$ for $x \gg 1$ in (3.27). Solving for a, we find

$$a(t) \approx 2^{-2/3}a_{\mathrm{m}\Lambda}\exp(\sqrt{1-\Omega_{\mathrm{m}0}}H_0t),$$

so that $a \propto \exp(\sqrt{1-\Omega_{\mathrm{m}0}}H_0t)$ in the Λ-dominated phase, as we would have expected from our discussion of the de Sitter universe.

3.5 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 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

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

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

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

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

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

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

$$\mathcal{S}_k^{-1}(r) = \mathcal{S}_k^{-1}(r_1) - \int_{t_1}^t \frac{cdt'}{a(t')} \tag{3.29}$$

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

$$d_{\rm P}^l = a(t)\left[\mathcal{S}_k^{-1}(r_1) - \int_{t_1}^t \frac{cdt'}{a(t')}\right], \tag{3.30}$$

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 $\mathcal{S}_k^{-1}(r)$ 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.)

3.5.1 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

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

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

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

then at no finite time t will the expression in brackets in equation (3.30) 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

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

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 $\mathcal{S}_k^{-1}(r) > \mathcal{S}_k^{-1}(r_{\rm EH})$ 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

$$d_{\rm P}^{\rm EH} = a(t) \int_{t}^{\infty} \frac{c dt'}{a(t')}. \tag{3.31}$$

3.5.2 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

$$d_{\rm P}^{l} = a(t) \left[\mathcal{S}_k^{-1}(r_1) - \int_{t_1}^{t} \frac{c dt'}{a(t')}\right]. \tag{3.32}$$

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

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

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

$$\mathcal{S}_k^{-1}(r_{\rm PH}) = \int_0^t \frac{cdt'}{a(t')}, \tag{3.33}$$

and the proper distance of this point from the origin is

$$d_{\rm P}^{\rm PH} = a(t) \int_0^t \frac{cdt'}{a(t')}. \tag{3.34}$$

3.5.3 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

$$I(t_1, t_2) = \int_{t_1}^{t_2} \frac{cdt}{a(t)} = \frac{c}{a_0} \int_{t_1}^{t_2} e^{-H_0 t} dt = \frac{c}{a_0 H_0}(e^{-H_0 t_1} - e^{-H_0 t_2}). \tag{3.35}$$

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 $\mathcal{S}_k^{-1}(r) = r$, 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

$$I(t_1, t_2) = \int_{t_1}^{t_2} \frac{cdt}{a(t)} = \frac{2c}{a_0H_0}\left[\left(\frac{t_2}{t_0}\right)^{1/3} - \left(\frac{t_1}{t_0}\right)^{1/3}\right].$$

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 $t_1 \to 0$:

$$I(t_1 \to 0, t_2 = t) = \frac{2c}{a_0H_0}\left(\frac{t}{t_0}\right)^{1/3}.$$

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

$$r_{\rm PH} = \frac{2c}{a_0H_0}\left(\frac{t}{t_0}\right)^{1/3},$$

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

3.6 Ages and distances

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

$$\frac{H^2(a)}{H_0^2} = \Omega_{\rm m0}\left(\frac{a_0}{a}\right)^3 + \Omega_{\rm r0}\left(\frac{a_0}{a}\right)^4 + \Omega_{\rm k0}\left(\frac{a_0}{a}\right)^2 + \Omega_{\Lambda 0}, \tag{3.36}$$

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

$$\frac{H^2(z)}{H_0^2} = \Omega_{\rm m0}(1+z)^3 + \Omega_{\rm r0}(1+z)^4 + \Omega_{\rm k0}(1+z)^2 + \Omega_{\Lambda 0}. \tag{3.37}$$

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

$$\Omega_{\rm m0} + \Omega_{\rm r0} + \Omega_{\rm k0} + \Omega_{\Lambda 0} = 1. \tag{3.38}$$

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

$$t_0 = \int_0^{a_0} \frac{da}{aH(a)}, \tag{3.39}$$

and for given values of the density parameters, this integral can be computed numerically using equation (3.36). When evaluating the intergral numerically, it is useful to introduce a dimensionless variable x through $a = a_0 x$. Substituting this in the integral gives

$$t_0 = \int_0^{1} \frac{dx}{xH(x)}. \tag{3.40}$$

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 (3.39) as

$$t_0 = -\int_\infty^0 \frac{(1+z)}{(1+z)^2}\frac{dz}{H(z)} = \int_0^\infty \frac{dz}{(1+z)H(z)}. \tag{3.41}$$

However, in numerical computations the form (3.40) is usually more convenient since it only involves integration over the finite interval from 0 to a_0.

We have looked at three different measures of distance: The proper distance, the luminosity distance, and the angular diameter distance. At the present epoch, they are given by

$$\begin{aligned} d_{\mathrm{P}} &= a_0 \mathcal{S}_k^{-1}(r) && (3.42)\\ d_{\mathrm{L}} &= a_0(1+z)r && (3.43)\\ d_{\mathrm{A}} &= \frac{a_0}{1+z} r, && (3.44) \end{aligned}$$

so they all depend on the comoving radial coordinate r of the object which distance we are measuring. If we consider first a spatially flat universe ($k = 0$), r is given by

$$r = \int_{t_e}^{t_0} \frac{cdt}{a(t)}, \tag{3.45}$$

where t_e is the time at which the light we observe today at t_0 was emitted. We rewrite the integral by substituting $dt = da/\dot{a}$:

$$\begin{aligned} r &= \int_{a(t_e)}^{a_0} \frac{cda}{a\dot{a}} \\ &= \int_{a_e}^{a_0} \frac{cda}{a^2 \frac{H(a)}{H_0} H_0} \\ &= \frac{c}{H_0} \int_{a_e}^{a_0} \frac{da}{a^2 H(a)/H_0}, && (3.46) \end{aligned}$$

where $a_e \equiv a(t_e)$. We typically want to find the distance to an object at a given redshift z, since the redshift is a directly measurable quantity. The redshift of the object is given by $1+z = a_0/a_e$, and we introduce a new substitution $1+z' = a_0/a$. For $a = a_0$ this gives $z' = 0$, while for $a = a_e$ we have $z' = z$. Furthermore,

$$dz' = -\frac{a_0}{a^2}da = -\frac{1}{a_0}\left(\frac{a_0}{a}\right)^2 da = -\frac{1}{a_0}(1+z')^2 da, \tag{3.47}$$

so

$$da = -\frac{a_0 dz'}{(1+z')^2}. \tag{3.48}$$

The expression for r then becomes

$$\begin{aligned} r &= \frac{c}{H_0}\int_z^0 \frac{1}{\frac{a_0^2}{(1+z')^2}\frac{H(z')}{H_0}}\left(-\frac{a_0}{(1+z')^2}\right)dz' \\ &= -\frac{c}{a_0H_0}\int_z^0 \frac{dz'}{H(z')/H_0} \\ &= \frac{c}{a_0H_0}\int_0^z \frac{dz'}{H(z')/H_0} \end{aligned} \tag{3.49}$$

The different distances follow easily, for example

$$d_{\rm L} = a_0(1+z)\frac{c}{a_0H_0}\int_0^z \frac{dz'}{H(z')/H_0} = \frac{c(1+z)}{H_0}\int_0^z \frac{dz'}{H(z')/H_0}, \tag{3.50}$$

and we see that the present value of the scale factor, a_0, cancels out.

What do we do when $k \neq 0$? In that case, we have seen that

$$r = \mathcal{S}_k\left(\int_{t_e}^{t_0}\frac{cdt}{a(t)}\right), \tag{3.51}$$

where $\mathcal{S}_k(x) = \sin(x)$ for $k = +1$, and $\mathcal{S}_k(x) = \sinh(x)$ for $k = -1$ (for $k = 0$ $\mathcal{S}_k(x) = x$ and we are back in the situation considered in the previous section). The integral we have looked at already, so we can write

$$r = \mathcal{S}_k\left(\frac{c}{a_0H_0}\int_0^z \frac{dz'}{H(z')/H_0}\right), \tag{3.52}$$

but now it is no longer obvious that the present value of the scale factor (which is not so easy to measure) will cancel out in the final results for the distances. However, we note that now

$$\Omega_{\rm k0} = -\frac{kc^2}{a_0^2H_0^2} \neq 0. \tag{3.53}$$

For $k = +1$ we can write

$$-\Omega_{\rm k0} = \frac{c^2}{a_0^2H_0^2}, \tag{3.54}$$

and since $-\Omega_{\mathrm{k0}} > 0$, we get

$$a_0 = \frac{c}{H_0\sqrt{-\Omega_{\mathrm{k0}}}}. \tag{3.55}$$

For $k = -1$ we have

$$\Omega_{\mathrm{k0}} = \frac{c^2}{a_0^2 H_0^2}, \tag{3.56}$$

and

$$a_0 = \frac{c}{H_0\sqrt{\Omega_{\mathrm{k0}}}}. \tag{3.57}$$

We can summarize both cases by writing

$$a_0 = \frac{c}{H_0\sqrt{|\Omega_{\mathrm{k0}}}}, \tag{3.58}$$

and we now have a_0 expressed in terms of quantities we can more easily obtain. Substituting for a_0 in the expression for r gives

$$\begin{aligned} r &= \mathcal{S}_k\left(\frac{c}{H_0}\frac{H_0\sqrt{|\Omega_{\mathrm{k0}}}}{c}\int_0^z \frac{dz'}{H(z')/H_0}\right) \\ &= \mathcal{S}_k\left(\sqrt{|\Omega_{\mathrm{k0}}}\int_0^z \frac{dz'}{H(z')/H_0}\right), \end{aligned} \tag{3.59}$$

and the final expression for, e.g., the luminosity distance becomes

$$d_{\mathrm{L}} = a_0(1+z)r = \frac{c(1+z)}{H_0\sqrt{|\Omega_{\mathrm{k0}}}}\mathcal{S}_k\left(\sqrt{|\Omega_{\mathrm{k0}}}\int_0^z \frac{dz'}{H(z')/H_0}\right). \tag{3.60}$$

Note that this result includes the case $k = 0$. Both $\sin(x)$ and $\sinh(x)$ are approximately equal to x as $x \to 0$, so

$$\begin{aligned} \lim_{|\Omega_{\mathrm{k0}}|\to 0} d_{\mathrm{L}} &= \frac{c}{H_0\sqrt{|\Omega_{\mathrm{k0}}|}}(1+z)\sqrt{|\Omega_{\mathrm{k0}}}\int_0^z \frac{dz'}{H(z')/H_0} \\ &= \frac{c(1+z)}{H_0}\int_0^z \frac{dz'}{H(z')/H_0}, \end{aligned} \tag{3.61}$$

in agreement with equation (3.50).

Chapter 4

The early universe

So far we have only been concerned with the space-time geometry of the Universe. The cosmological principle determined the form of the metric up to the spatial curvature k and the scale factor $a(t)$. The spatial curvature is determined by the total density, and we used the Friedmann equations to find $a(t)$ in some more or less realistic models.

Now it is time to take a look at the history of the contents of the Universe. This history begins in the radiation dominated era.

4.1 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:

$$T \propto \frac{1}{a} \propto (1+z), \tag{4.1}$$

$$T \propto \frac{1}{\sqrt{t}}, \tag{4.2}$$

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 (4.1) and (4.2) 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)

$$E_{\rm P} = m_{\rm P}c^2 \sim \sqrt{\frac{\hbar c^5}{G}} \approx 10^{19}\ \text{GeV}. \tag{4.3}$$

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

$$t_{\rm P} \sim \frac{\hbar}{E_{\rm P}} = \sqrt{\frac{\hbar G}{c^5}} \approx 10^{-43}\ \text{s}, \tag{4.4}$$

and the corresponding length scale is

$$\ell_{\rm P} = \sqrt{\frac{\hbar G}{c^3}} \approx 10^{-35}\ \text{m}. \tag{4.5}$$

Unfortunately, there is no universally accepted theory of quantum gravity yet. The most common view is that string 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.

4.2 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

[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)

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

$$f_i(\mathbf{p}) = \frac{1}{e^{(E_i(p)-\mu_i)/(k_\mathrm{B}T)} \pm 1}, \tag{4.6}$$

where k_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:

$$n_i = \frac{g_i}{(2\pi\hbar)^3}\int f_i(p)d^3p, \tag{4.7}$$

$$\rho_i c^2 = \frac{g_i}{(2\pi\hbar)^3}\int E_i(p)f_i(p)d^3p, \tag{4.8}$$

$$P_i = \frac{g_i}{(2\pi\hbar)^3}\int \frac{(pc)^2}{3E(p)}f_i(p)d^3p, \tag{4.9}$$

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 $E_idE_i = c^2pdp$, 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

$$n_i = \frac{g_i}{2\pi^2(\hbar c)^3}\int_{m_ic^2}^{\infty}\frac{(E^2 - m_i^2c^4)^{1/2}EdE}{\exp[(E-\mu_i)/(k_\mathrm{B}T)] \pm 1}, \tag{4.10}$$

$$\rho_i c^2 = \frac{g_i}{2\pi^2(\hbar c)^3}\int_{m_i c^2}^{\infty}\frac{(E^2 - m_i^2c^4)^{1/2}E^2 dE}{\exp[(E-\mu_i)/(k_{\rm B}T)] \pm 1}, \tag{4.11}$$

$$P_i = \frac{g_i}{6\pi^2(\hbar c)^3}\int_{m_i c^2}^{\infty}\frac{(E^2 - m_i^2c^4)^{3/2} dE}{\exp[(E-\mu_i)/(k_{\rm B}T)] \pm 1}. \tag{4.12}$$

We will normally be interested in the limit of non-relativistic particles, corresponding to $m_ic^2/(k_{\rm B}T) \gg 1$, and the ultrarelativistic limit, corresponding to $m_ic^2/(k_{\rm B}T) \ll 1$. 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 $m_ic^2 \ll k_{\rm B}T$ 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 $\zeta(x)$ evaluated at $x = 4$, 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

$$\rho_i c^2 = \frac{g_i \pi^2}{30} \frac{(k_{\rm B}T)^4}{(\hbar c)^3}. \tag{4.13}$$

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

$$\rho_i c^2 = \frac{7}{8} \frac{g_i \pi^2}{30} \frac{(k_{\rm B}T)^4}{(\hbar c)^3}. \tag{4.14}$$

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

$$\begin{aligned}
n_i &= \frac{g_i \zeta(3)}{\pi^2} \left(\frac{k_{\rm B}T}{\hbar c}\right)^3 \text{ bosons} \qquad (4.15) \\
&= \frac{3}{4} \frac{g_i \zeta(3)}{\pi^2} \left(\frac{k_{\rm B}T}{\hbar c}\right)^3 \text{ fermions,} \qquad (4.16)
\end{aligned}$$

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

$$P_i = \frac{1}{3}\rho_i c^2, \tag{4.17}$$

for both bosons and fermions.

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

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

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

$$n_i = g_i\left(\frac{m_ik_\mathrm{B}T}{2\pi\hbar^2}\right)^{3/2}\exp\left(\frac{\mu_i-m_ic^2}{k_\mathrm{B}T}\right). \tag{4.18}$$

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

$$\rho_ic^2 \approx n_im_ic^2, \tag{4.19}$$

$$P_i = n_ik_\mathrm{B}T, \tag{4.20}$$

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

$$\frac{P_i}{\rho_ic^2} = \frac{k_\mathrm{B}T}{m_ic^2} \ll 1,$$

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

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

where $u = E/(k_{\rm B}T_i)$, $x_i = m_ic^2/(k_{\rm B}T_i)$, 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

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

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 (4.13) and (4.14), 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

$$\rho c^2 = \frac{\pi^2}{30}g_*\frac{(k_{\rm B}T)^4}{(\hbar c)^3}, \tag{4.21}$$

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

$$g_* = \sum_{i=\text{bosons}} g_i\left(\frac{T_i}{T}\right)^4 + \frac{7}{8}\sum_{i=\text{fermions}} g_i\left(\frac{T_i}{T}\right)^4. \tag{4.22}$$

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

$$P = \frac{1}{3}\rho c^2 = \frac{\pi^2}{90}g_*\frac{(k_{\rm B}T)^4}{(\hbar c)^3}. \tag{4.23}$$

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

Using (4.21) in the Friedmann equation gives

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

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

$$H = \left(\frac{4\pi^3}{45}\right)^{1/2} g_*^{1/2}(T)\frac{(k_{\rm B}T)^2}{\hbar E_{\rm P}} \approx 1.66 g_*^{1/2}(T)\frac{(k_{\rm B}T)^2}{\hbar E_{\rm P}}. \tag{4.24}$$

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

$$\frac{t}{t_{\rm P}} = \frac{1}{2}\left(\frac{45}{4\pi^3}\right)^{1/2} g_*^{-1/2}(T)\left(\frac{k_{\rm B}T}{E_{\rm P}}\right)^{-2} = 0.301 g_*^{-1/2}(T)\left(\frac{k_{\rm B}T}{E_{\rm P}}\right)^{-2}, \tag{4.25}$$

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

$$t \approx 2.423 g_*^{-1/2}(T)\left(\frac{k_{\rm B}T}{1\ {\rm MeV}}\right)^{-2}\ {\rm s}. \tag{4.26}$$

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.

4.3 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:

$$a^{-3}\frac{d(n_1a^3)}{dt} = n_1^{(0)}n_2^{(0)}\langle\sigma v\rangle\left[\frac{n_3n_4}{n_3^{(0)}n_4^{(0)}} - \frac{n_1n_2}{n_1^{(0)}n_2^{(0)}}\right]. \tag{4.27}$$

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_1H. 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:

$$\frac{n_3n_4}{n_3^{(0)}n_4^{(0)}} = \frac{n_1n_2}{n_1^{(0)}n_2^{(0)}}, \tag{4.28}$$

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

Conversely, if the reaction rate is much smaller than the expansion rate, the right-hand side of the Boltzmann equation is practically zero, which means that the number of particles of type 1 is constant. This species is then ofen said to have been *frozen out*. As a rule of thumb, we say that this happens when the interaction rate drops below the expansion rate.

4.4 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:

- 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.

[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://pdg.lbl.gov/

Electric charge	$Q = 0$	$Q = -1$	$Q = +2/3$	$Q = -1/3$
1. family	ν_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 4.1: The fermions of the Standard Model. The numbers in the paranthеses are the particle rest masses mc^2

- 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.

- 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.

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, $i = 1, \dots, 8$	Strong	0	0

Table 4.2: The gauge bosons of the Standard Model.

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, the so-called Higgs mechanism, in order to give masses to the other particles without violating the gauge symmetry of the electroweak interaction.

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

[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).

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 understood 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

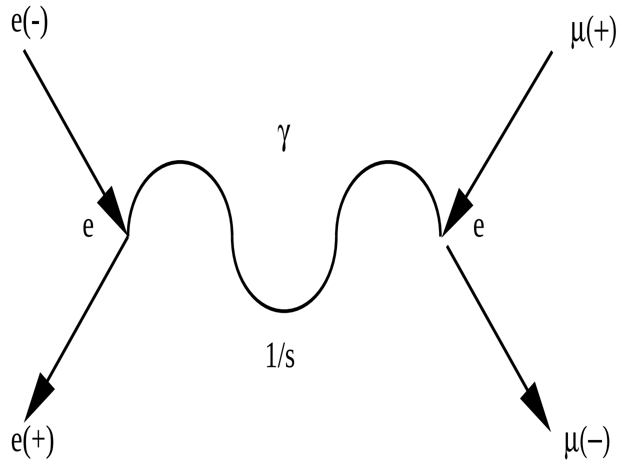

Figure 4.1: Lowest-order Feynman diagram for muon pair production from an electron-positron pair. In the diagram, time flows from left to right.

and a positron annhilate and produce a muon-antimuon pair. The lowest-order diagram for this process is shown in figure 4.1. 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, $\alpha_s = g_s^2/4\pi \sim 0.3$.

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

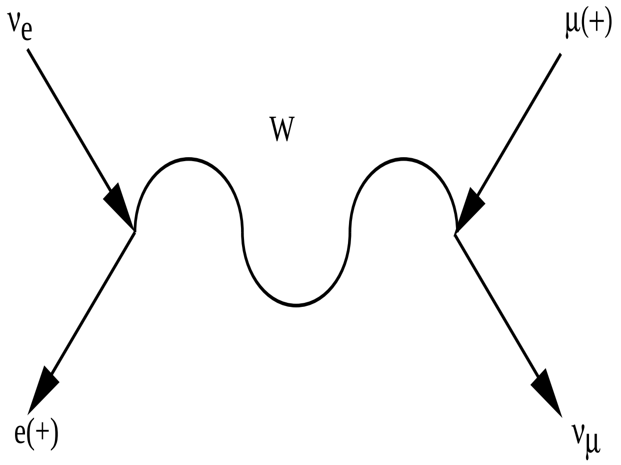

Figure 4.2: Lowest-order Feynman diagram for $\nu_e e^+ \to \nu_\mu \mu^+$.

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 4.2. The two vertices contribute a factor $g_{\text{weak}} = e/\sin\theta_W$ each. At the accuracy we are working, we can just take $g_{\text{weak}} \sim e$. The propagator gives a factor $1/(s - m_W^2 c^4)$, so that the cross section is

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

For $\sqrt{s} \ll m_W c^2 \approx 80$ GeV, we see that $\sigma_{\text{weak}} \sim \alpha^2 s/(m_W c^2)^4$, 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

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

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), $\gamma e \to \gamma e$, 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_{B} in the expressions above. For the weak rate, for example, you can convince yourself that the expression

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

has units of inverse seconds.

4.5 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

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

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,

$$S = \frac{a^3(\rho c^2 + P)}{T}, \tag{4.29}$$

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

$$s = \frac{S}{V} = \frac{\rho c^2 + P}{T}. \tag{4.30}$$

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 (4.30) we therefore find that the entropy density can be written as

$$s = \frac{2\pi^2}{45} k_{\rm B} g_{*s}\left(\frac{k_{\rm B}T}{\hbar c}\right)^3, \tag{4.31}$$

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

$$g_{*s} = \sum_{i=\text{bosons}} g_i \left(\frac{T_i}{T}\right)^3 + \frac{7}{8}\sum_{i=\text{fermions}} g_i \left(\frac{T_i}{T}\right)^3. \tag{4.32}$$

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

$$s = \frac{2\pi^4}{45\zeta(3)} g_{*s} n_\gamma k_{\rm B} \approx 1.80 g_{*s} n_\gamma k_{\rm B}.$$

The constancy of S implies that sa^3 = constant, which means that

$$g_{*s} T^3 a^3 = \text{constant}, \tag{4.33}$$

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,

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

to the Hubble expansion rate

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

This results in

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

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

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

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 $e^+ + e^- \to \gamma + \gamma$, 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

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

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

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

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

$$T_\nu = \left(\frac{4}{11}\right)^{1/3} T. \tag{4.34}$$

4.6 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 ${}^A_Z\mathrm{X}_N$. 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: ${}^A\mathrm{X}$. The simplest nucleus is hydrogen, ${}^1_1\mathrm{H}_0$ (or simply ${}^1\mathrm{H}$), which is just a proton, p. A proton and a neutron may combine to form the isotope ${}^2\mathrm{H}$, which is also called the deuteron and denoted by D. One proton and two neutrons form ${}^3\mathrm{H}$, 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), ${}^3\mathrm{He}$. By adding a neutron, we get the isotope ${}^4\mathrm{He}$.

A nucleus X has rest mass $m({}^A_Z\mathrm{X}_N)$, 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:

$$B = [Zm_p + Nm_n]c^2 - m({}^A_Z\mathrm{X}_N)c^2.$$

Here, $m_pc^2 = 938.272$ MeV is the proton rest mass, and $m_nc^2 = 939.565$ 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 ${}^{56}\mathrm{Fe}$, 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

$$\begin{aligned}\eta_{\rm b} &= \frac{n_{\rm b}}{n_\gamma} = \frac{n_{\rm b0}(1+z)^3}{n_{\gamma 0}(1+z)^3} = \frac{\rho_{\rm b}/m_N}{n_{\gamma 0}} \\ &= \frac{\rho_{\rm c0}\Omega_{\rm b0}}{n_{\gamma 0} m_N} \approx 2.7\times 10^{-8}\Omega_{\rm b0}h^2. \end{aligned} \tag{4.35}$$

Since $\Omega_{\rm b0}$ is at most of order 1 (actually it is a few parts in 100), 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

$$\begin{aligned} d+n &\to {}^3\mathrm{H}+\gamma \\ d+p &\to {}^3\mathrm{He}+\gamma, \end{aligned}$$

or

$$\begin{aligned} d+d &\to {}^3\mathrm{H}+p \\ d+d &\to {}^3\mathrm{He}+n, \end{aligned}$$

from which one can form ^{4}He as

$$\begin{aligned} {}^3\mathrm{H}+p &\to {}^4\mathrm{He}+\gamma, \text{ or} \\ {}^3\mathrm{He}+n &\to {}^4\mathrm{He}+\gamma. \end{aligned}$$

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 $k_{\rm B}T$, one would naively expect that the process of photons breaking up deuterons would become inefficient as soon as $k_{\rm B}T \sim 2.22$ MeV. However, this process persists until much lower temperatures are reached. This is because $k_{\rm B}T$ 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

$$n_\gamma(E \geq E_0) = \frac{1}{\pi^2(\hbar c)^3}\int_{E_0}^{\infty}\frac{E^2 dE}{e^{E/k_{\rm B}T}-1}.$$

We are interested in the situation when $E_0 \gg k_{\rm BT}$, and then $e^{E/k_{\rm BT}} \gg 1$ in the integrand, so we can write

$$\begin{aligned} n_\gamma(E \geq E_0) &= \frac{1}{\pi^2(\hbar c)^3}\int_{E_0}^{\infty} dE E^2 e^{-E/k_{\rm B}T} \\ &= \frac{1}{\pi^2}\left(\frac{k_{\rm B}T}{\hbar c}\right)^3\int_{x_0}^{\infty} x^2 e^{-x}dx \\ &= \frac{1}{\pi^2}\left(\frac{k_{\rm B}T}{\hbar c}\right)^3 (x_0^2+2x_0+2)e^{-x_0}, \end{aligned}$$

where I have introduced the variable $x = E/k_{\rm B}T$. Since we have found earlier that the total number density of photons is given by

$$n_\gamma = \frac{2.404}{\pi^2}\left(\frac{k_{\rm B}T}{\hbar c}\right)^3,$$

the fraction of photons with energies greater than E_0 is

$$f(E \geq E_0) = 0.416 e^{-x_0}(x_0^2+2x_0+2),$$

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

$$a^{-3}\frac{d(n_na^3)}{dt}=\lambda_{np}(n_pe^{-Q/k_{\rm B}T}-n_n),$$

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

$$\frac{dX_n}{dt}=\lambda_{np}[(1-X_n)e^{-Q/k_{\rm B}T}-X_n].$$

Now, we switch variables from t to $x=Q/k_{\rm B}T$, 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

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

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 4.3. We see that the neutrons drop out of equilibrium at $k_{\rm B}T \sim 1$ MeV, and that X_n freezes out at a value ≈ 0.15 at $k_{\rm B}T \sim 0.5$ 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,

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

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

$$t \approx 132\left(\frac{0.1\ \mathrm{MeV}}{k_{\rm B}T}\right)^2\ \mathrm{s},$$

and by the onset of deuteron production at $k_{\rm B}T \approx 0.08$ MeV, this means that the neutron abundance has been reduced by a factor

$$\exp\left[-\frac{132\ \mathrm{s}}{885.7\ \mathrm{s}}\left(\frac{0.1}{0.08}\right)^2\right] \approx 0.79,$$

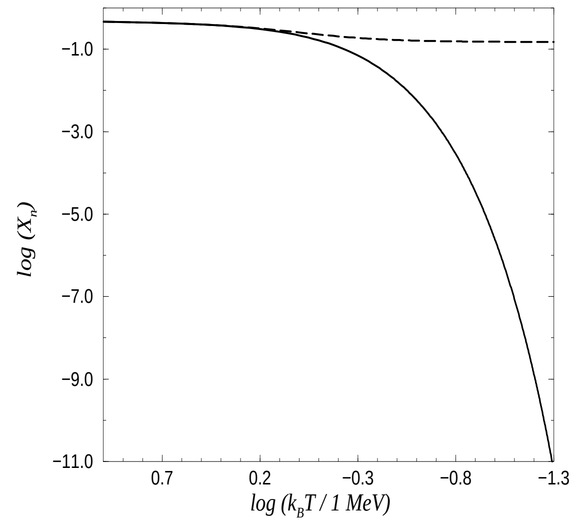

Figure 4.3: Solution of the Boltzmann equation for the neutron abundance (dashed line), along with the equilbrium abundance (full line).

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 $\eta_{\rm b}$ 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 $\Omega_{\rm b0}h^2 \approx 0.02$.

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 $N_\nu \leq 4$.

4.7 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 $B_{\mathrm{H}} = 13.6$ eV, one would guess that recombination should take place at a temperature $k_{\mathrm{B}}T = B_{\mathrm{H}}$. 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 $k_{\mathrm{B}}T$ significantly less than B_{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

$$k_{\mathrm{B}}T_{\mathrm{rec}} = \frac{B_H}{26.5} \sim 0.5 \text{ eV}.$$

The process responsible for formation of hydrogen is

$$e^- + p \leftrightarrow \mathrm{H} + \gamma,$$

and as long as this process is in equlibrium, the Boltzmann equation is reduced to

$$\frac{n_e n_p}{n_{\mathrm{H}}} = \frac{n_e^{(0)} n_p^{(0)}}{n_{\mathrm{H}}^{(0)}}.$$

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 $n_e + n_{\mathrm{H}} = n_p + n_{\mathrm{H}}$. The fraction of free electrons is defined as

$$X_e = \frac{n_e}{n_e + n_{\mathrm{H}}} = \frac{n_p}{n_p + n_{\mathrm{H}}}.$$

The equilibrium number densities are given by

$$\begin{aligned}
n_e^{(0)} &= 2\left(\frac{m_e k_{\mathrm{B}}T}{2\pi\hbar^2}\right)^{3/2} \exp\left(-\frac{m_e c^2}{k_{\mathrm{B}}T}\right), \\
n_p^{(0)} &= 2\left(\frac{m_p k_{\mathrm{B}}T}{2\pi\hbar^2}\right)^{3/2} \exp\left(-\frac{m_p c^2}{k_{\mathrm{B}}T}\right), \\
n_{\mathrm{H}}^{(0)} &= 4\left(\frac{m_{\mathrm{H}} k_{\mathrm{B}}T}{2\pi\hbar^2}\right)^{3/2} \exp\left(-\frac{m_{\mathrm{H}} c^2}{k_{\mathrm{B}}T}\right),
\end{aligned}$$

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

$$n_b = \frac{2\zeta(3)}{\pi^2}\left(\frac{k_{\rm B}T}{\hbar c}\right)^3 \eta_b.$$

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.

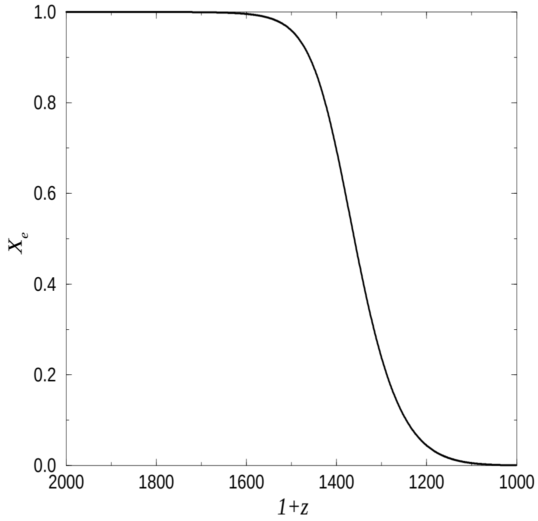

Figure 4.4: The solution of the equation for the free electron fraction in the case $\Omega_{\mathrm{b}0}h^2 = 0.02$.

The equation can be solved for various values of $\eta_b = 2.7 \times 10^{-8}\Omega_{\mathrm{b}0}h^2$. In figure 4.4 the solution for the free electron fraction is shown as a function of redshift for the canonical value $\Omega_{\mathrm{b}0}h^2 = 0.02$. 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.

Chapter 5

Structure formation

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. As we will see in the final chapter, 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). Measuring the statistical properties of the large-scale distribution of matter and the temperature variations in the CMB is one of the most active and important fields in cosmology and are the most important ways of learning about dark matter, dark energy, and inflation.

Perturbations in the density are commonly characterized by the so-called density contrast

$$\Delta(\mathbf{x}, t) = \frac{\rho(\mathbf{x}, t) - \rho_0(t)}{\rho_0(t)}, \tag{5.1}$$

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.

5.1 Non-relativistic fluids

We will start with the simplest situation, a universe with just one component, and find the equations describing small density perturbations. If we treat this component as a fluid, the fundamental equations are:

$$\begin{aligned} \frac{\partial\rho}{\partial t} + \nabla\cdot(\rho\mathbf{v}) &= 0 \qquad (5.2)\\ \frac{\partial\mathbf{v}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{v} &= -\frac{1}{\rho}\nabla p - \nabla\phi \qquad (5.3)\\ \nabla^2\phi &= 4\pi G\rho, \qquad (5.4)\end{aligned}$$

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

$$\frac{d}{dt} = \frac{\partial}{\partial t} + (\mathbf{v}\cdot\nabla). \qquad (5.5)$$

Note that the effect of the operator $(\mathbf{v}\cdot\nabla)$ on a scalar function f is given by

$$(\mathbf{v}\cdot\nabla)f = v_x\frac{\partial f}{\partial x} + v_y\frac{\partial f}{\partial y} + v_z\frac{\partial f}{\partial z}, \qquad (5.6)$$

in Cartesian coordinates, whereas the effect on a vector field $\mathbf{A}$ is given by

$$\begin{aligned}(\mathbf{v}\cdot\nabla)\mathbf{A} &= \left(v_x\frac{\partial A_x}{\partial x} + v_y\frac{\partial A_x}{\partial y} + v_z\frac{\partial A_x}{\partial z}\right)\mathbf{e}_x\\ &+ \left(v_x\frac{\partial A_y}{\partial x} + v_y\frac{\partial A_y}{\partial y} + v_z\frac{\partial A_y}{\partial z}\right)\mathbf{e}_y\\ &+ \left(v_x\frac{\partial A_z}{\partial x} + v_y\frac{\partial A_z}{\partial y} + v_z\frac{\partial A_z}{\partial z}\right)\mathbf{e}_z. \qquad (5.7)\end{aligned}$$

In Lagrangian form the equations (5.2)-(5.4) can be written as

$$\begin{aligned}
\frac{d\rho}{dt} &= -\rho(\nabla \cdot \mathbf{v}) && (5.8)\\
\frac{d\mathbf{v}}{dt} &= -\frac{1}{\rho}\nabla p - \nabla\phi && (5.9)\\
\nabla^2\phi &= 4\pi G\rho. && (5.10)
\end{aligned}$$

The transition from (5.3) to (5.9) is easily seen; the transition from (5.2) to (5.8) can be shown by writing out (5.2):

$$\begin{aligned}
\frac{\partial\rho}{\partial t} + \nabla\cdot(\rho\mathbf{v}) &= \frac{\partial\rho}{\partial t} + \rho(\nabla\cdot\mathbf{v}) + \mathbf{v}\cdot\nabla\rho\\
&= \frac{\partial\rho}{\partial t} + (\mathbf{v}\cdot\nabla)\rho + \rho(\nabla\cdot\mathbf{v}) = 0,
\end{aligned}$$

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. Let us call the solution $\mathbf{v}_0$, ρ_0, p_0 and ϕ_0. These quantities obey, by definition, equations (5.8)-(5.10). We now add small perturbations to these solutions, and write the full quantities as

$$\begin{aligned}
\mathbf{v} &= \mathbf{v}_0 + \delta\mathbf{v} && (5.11)\\
\rho &= \rho_0 + \delta\rho && (5.12)\\
p &= p_0 + \delta p && (5.13)\\
\phi &= \phi_0 + \delta\phi. && (5.14)
\end{aligned}$$

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 (5.8):

$$\frac{d}{dt}(\rho_0 + \delta\rho) = -(\rho_0 + \delta\rho)\nabla\cdot(\mathbf{v}_0 + \delta\mathbf{v}),$$

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 (5.8), 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 (5.8) 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

$$\frac{d}{dt}\left(\frac{\delta\rho}{\rho_0}\right)\equiv\frac{d\Delta}{dt}=-\nabla\cdot\delta\mathbf{v}. \tag{5.15}$$

Next we look at the left-hand side of (5.9):

$$\begin{aligned}\frac{d}{dt}(\mathbf{v}_0+\delta\mathbf{v}) &= \left[\frac{\partial}{\partial t}+(\mathbf{v}_0+\delta\mathbf{v})\cdot\nabla\right](\mathbf{v}_0+\delta\mathbf{v})\\ &= \frac{\partial\mathbf{v}_0}{\partial t}+[(\mathbf{v}_0+\delta\mathbf{v})\cdot\nabla]\mathbf{v}_0+\frac{\partial}{\partial t}\delta\mathbf{v}+[(\mathbf{v}_0+\delta\mathbf{v})\cdot\nabla]\delta\mathbf{v}\\ &= \frac{\partial\mathbf{v}_0}{\partial t}+(\mathbf{v}_0\cdot\nabla)\mathbf{v}_0+(\delta\mathbf{v}\cdot\nabla)\mathbf{v}_0+\frac{d}{dt}\delta\mathbf{v}.\end{aligned}$$

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 (5.3) (with $\nabla p_0=0$). This leaves us with

$$\frac{d}{dt}\delta\mathbf{v}+(\delta\mathbf{v}\cdot\nabla)\mathbf{v}_0=-\frac{1}{\rho_0}\nabla\delta p-\nabla\delta\phi. \tag{5.16}$$

The perturbed version of (5.9) is easily found, since Poisson's equation is linear and ϕ_0 and ρ_0 are solutions of the unperturbed version:

$$\nabla^2 \delta\phi = 4\pi G \delta\rho. \tag{5.17}$$

Equations (5.15,5.16,5.17) 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 $\mathbf{r}$,

$$\mathbf{x} = a(t)\mathbf{r}, \tag{5.18}$$

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

$$\delta\mathbf{v} = a(t)\frac{\delta\mathbf{r}}{\delta t} \equiv a(t)\mathbf{u}. \tag{5.19}$$

The velocity $\mathbf{u}$, describes deviations from the smooth Hubble flow, and is often called the peculiar velocity. Equation (5.16) 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

$$\dot{\mathbf{u}}+2\frac{\dot{a}}{a}\mathbf{u} = -\frac{1}{\rho_0 a^2}\nabla_c\delta p - \frac{1}{a^2}\nabla_c\delta\phi. \tag{5.20}$$

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

$$\delta p = c_s^2\delta\rho, \tag{5.21}$$

where c_s is the sound speed in the system. With this extra condition, (5.20) can be rewritten as

$$\dot{\mathbf{u}}+2\frac{\dot{a}}{a}\mathbf{u} = -\frac{c_s^2}{\rho_0 a^2}\nabla_c\delta\rho - \frac{1}{a^2}\nabla_c\delta\phi. \tag{5.22}$$

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 (5.22):

$$\nabla_c\cdot\dot{\mathbf{u}}+2\frac{\dot{a}}{a}\nabla_c\mathbf{u} = -\frac{c_s^2}{\rho_0 a^2}\nabla_c^2\delta\rho - \frac{1}{a^2}\nabla_c^2\delta\phi.$$

From (5.17) in co-moving coordinates we have

$$\frac{1}{a^2}\nabla_c^2\delta\phi = 4\pi G\delta\rho,$$

and therefore

$$\nabla_c\cdot\dot{\mathbf{u}}+2\frac{\dot{a}}{a}\nabla_c\mathbf{u} = -\frac{c_s^2}{\rho_0 a^2}\nabla_c^2\delta\rho - 4\pi G\delta\rho. \tag{5.23}$$

From equation (5.15) 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 (5.23) results in

$$\frac{d^2\Delta}{dt^2} + 2\frac{\dot{a}}{a}\frac{d\Delta}{dt} = \frac{c_s^2}{\rho_0 a^2}\nabla_c^2\delta\rho + 4\pi G\delta\rho, \tag{5.24}$$

where $\Delta = \delta\rho/\rho_0$. This is the desired equation for $\delta\rho$.

We write the density perturbation as a Fourier series

$$\Delta(\mathbf{r}, t) = \sum_{\mathbf{k}} \Delta_k(t) e^{i\mathbf{k}_c \cdot \mathbf{r}}, \tag{5.25}$$

where $\mathbf{k}_c = a\mathbf{k}$ is the co-moving wave number, 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. Since equation (5.25) 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 (5.25). 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 (5.24) can be written

$$\frac{d^2\Delta_k}{dt^2} + 2\frac{\dot{a}}{a}\frac{d\Delta_k}{dt} = \Delta_k(4\pi G\rho_0 - k^2c_s^2). \tag{5.26}$$

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 $k = |\mathbf{k}|$.

5.2 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 (5.26), we see that ω must obey the dispersion relation

$$\omega^2 = c_s^2k^2 - 4\pi G\rho_0. \tag{5.27}$$

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

$$k_{\rm J} = \frac{\sqrt{4\pi G\rho_0}}{c_s}. \tag{5.28}$$

and the corresponding wave length is called the Jeans length,

$$\lambda_{\rm J} = \frac{2\pi}{k_{\rm J}} = c_s\sqrt{\frac{\pi}{G\rho_0}}. \tag{5.29}$$

For $k > k_{\rm J}$ ($\lambda < \lambda_{\rm J}$) the right-hand side of equation (5.27) 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 (5.27) 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

$$\Gamma = \sqrt{4\pi G\rho_0 - c_s^2k^2} = \left[4\pi G\rho_0\left(1 - \frac{\lambda_{\rm J}^2}{\lambda^2}\right)\right]^{1/2}. \tag{5.30}$$

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

$$p = G\rho^2 r^2$$

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.

5.3 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 (5.26) as

$$\frac{d^2\Delta_k}{dt^2} + 2\frac{\dot{a}}{a}\frac{d\Delta_k}{dt} = 4\pi G\rho_0\left(1 - \frac{\lambda_{\rm J}^2}{\lambda^2}\right)\Delta_k, \tag{5.31}$$

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

$$\frac{d^2\Delta_k}{dt^2} + \frac{4}{3t}\frac{d\Delta_k}{dt} - \frac{2}{3t^2}\Delta_k = 0. \tag{5.32}$$

We look for a solution of the form $\Delta_k = Kt^n$, where K is a constant. Inserted in equation (5.32), 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 (5.26) 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 (5.26), and they can therefore be neglected in first-order perturbation theory.

5.4 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 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 $1 + z > 4 \times 10^4 \Omega_{\mathrm{m}0} h^2$.

In the relativistic case one can show that one gets two equations expressing conservation of energy and momentum:

$$\begin{aligned} \frac{\partial \rho}{\partial t} &= -\nabla \cdot \left[\left(\rho + \frac{p}{c^2} \right) \mathbf{v} \right] & (5.33) \\ \frac{\partial}{\partial t} \left(\rho + \frac{p}{c^2} \right) &= \frac{\dot{p}}{c^2} - \left(\rho + \frac{p}{c^2} \right) (\nabla \cdot \mathbf{v}). & (5.34) \end{aligned}$$

In the special case $p = \rho c^2/3$ both equations reduce to

$$\frac{d\rho}{dt} = -\frac{4}{3} \rho (\nabla \cdot \mathbf{v}). \tag{5.35}$$

The analogue of the Euler equation turns out to be

$$\left(\rho + \frac{p}{c^2} \right) \left[\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla) \mathbf{v} \right] = -\nabla p - \left(\rho + \frac{p}{c^2} \right) \nabla \phi, \tag{5.36}$$

while the analogue of the Poisson equation is

$$\nabla^2 \phi = 4\pi G \left(\rho + \frac{3p}{c^2} \right), \tag{5.37}$$

which for $p = \rho c^2/3$ gives

$$\nabla^2 \phi = 8\pi G \rho. \tag{5.38}$$

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 (5.26):

$$\frac{d^2 \Delta_k}{dt^2} + 2 \frac{\dot{a}}{a} \frac{d\Delta_k}{dt} = \left(\frac{32\pi G \rho_0}{3} - k^2 c_s^2 \right) \Delta_k, \tag{5.39}$$

and the Jeans length in the relativistic case is therefore

$$\lambda_{\mathrm{J}} = c_s \left(\frac{3\pi}{8 G \rho_0} \right)^{1/2}, \tag{5.40}$$

where $c_s = c/\sqrt{3}$.

For modes with $\lambda \gg \lambda_{\rm J}$ equation (5.39) becomes

$$\frac{d^2\Delta_k}{dt^2} + 2\frac{\dot{a}}{a}\frac{d\Delta_k}{dt} - \frac{32\pi G\rho_0}{3}\Delta_k = 0, \tag{5.41}$$

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

$$\frac{d^2\Delta_k}{dt^2} + \frac{1}{t}\frac{d\Delta_k}{dt} - \frac{1}{t^2}\Delta_k = 0. \tag{5.42}$$

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 $\Delta_k \propto t \propto a^2 \propto (1+z)^{-2}$.

5.5 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

$$\delta\phi_k \propto \rho_0 a^2\Delta_k, \tag{5.43}$$

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 spatially flat. In particular, we have that ϕ is constant in an Einstein-de Sitter universe to first order in perturbation theory.

5.6 No significant growth while radiation dominates

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 (5.26) 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

$$\ddot{\Delta}_k + 2\frac{\dot{a}}{a}\dot{\Delta}_k - 4\pi G\rho_m\Delta_k = 0. \tag{5.44}$$

To solve this equation, it is convenient to change variable from t to a, so that

$$\begin{aligned} \frac{d}{dt} &= \frac{da}{dt}\frac{d}{da} = \dot{a}\frac{d}{da} && (5.45)\\ \frac{d^2}{dt^2} &= \frac{d}{dt}\left(\dot{a}\frac{d}{da}\right) = \dot{a}^2\frac{d^2}{da^2} + \ddot{a}\frac{d}{da}. && (5.46) \end{aligned}$$

Furthermore, we introduce

$$y = \frac{a}{a_{\rm eq}}, \tag{5.47}$$

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

$$a_{\rm eq} = \frac{\rho_{r0}}{\rho_{m0}}. \tag{5.48}$$

We also see that

$$\frac{\rho_m}{\rho_r} = \frac{\rho_{m0}a^{-3}}{\rho_{r0}a^{-4}} = \frac{a}{\rho_{r0}/\rho_{m0}} = \frac{a}{a_{\rm eq}} = y. \tag{5.49}$$

The Friedmann equations can then be written as

$$\begin{aligned} \left(\frac{\dot{a}}{a}\right)^2 &= \frac{8\pi G}{3}(\rho_m + \rho_r) = \frac{8\pi G}{3}\rho_r\left(1 + \frac{\rho_m}{\rho_r}\right)\\ &= \frac{8\pi G}{3}\rho_r(1+y), && (5.50) \end{aligned}$$

and

$$\begin{aligned} \frac{\ddot{a}}{a} &= -\frac{4\pi G}{3}(\rho_m + \rho_r + 3p_r) = -\frac{4\pi G}{3}(\rho_m + 2\rho_r)\\ &= -\frac{4\pi G}{3}\rho_r(2+y). && (5.51) \end{aligned}$$

We express d/da by d/dy:

$$\frac{d}{da} = \frac{dy}{da}\frac{d}{dy} = \frac{1}{a_{\text{eq}}}\frac{d}{dy} \tag{5.52}$$

$$\frac{d^2}{da^2} = \frac{1}{a_{\text{eq}}^2}\frac{d^2}{dy^2}. \tag{5.53}$$

Inserting all of this into equation (5.44), 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

$$\frac{d^2\Delta_k}{dy^2} + \frac{2+3y}{2y(1+y)}\frac{d\Delta_k}{dy} - \frac{3}{2y(1+y)}\Delta_k = 0. \tag{5.54}$$

By substitution one easily sees that this equation has the growing solution

$$\Delta_k \propto 1 + \frac{3}{2}y, \tag{5.55}$$

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.

5.7 The power spectrum

We have seen that we can write the general solution of equation (5.26)as a Fourier series

$$\Delta(\mathbf{x}, t) = \sum_{\mathbf{k}} \Delta_k(t)e^{-i\mathbf{k}\cdot\mathbf{x}}, \tag{5.56}$$

where

$$\Delta_k(t) = \frac{1}{V}\int \Delta(\mathbf{x},t)e^{i\mathbf{k}\cdot\mathbf{x}}d^3x. \tag{5.57}$$

Here V is some large normalization volume and

$$\frac{1}{V}\int e^{i(\mathbf{k}-\mathbf{k}')\cdot\mathbf{x}}d^3x = \delta_{\mathbf{k},\mathbf{k}'}. \tag{5.58}$$

In the limit $V \to \infty$, we can write $\Delta(\mathbf{x},t)$ as a Fourier integral

$$\Delta(\mathbf{x},t) = \frac{1}{(2\pi)^3}\int \Delta_k e^{-i\mathbf{k}\cdot\mathbf{x}}d^3k, \tag{5.59}$$

where

$$\Delta_k(t) = \int \Delta(\mathbf{x},t)e^{i\mathbf{k}\cdot\mathbf{x}}d^3x. \tag{5.60}$$

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, (5.56) or (5.59) 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 (5.26). 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. The most common view is 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 $\Delta_{\text{in}}(\mathbf{x},t)$ had a Gaussian distribution

$$p(\Delta_{\text{in}}) \propto \exp\left(-\frac{\Delta_{\text{in}}^2}{2\sigma^2}\right).$$

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 (5.56) we get

$$\Delta^2(\mathbf{x},t) = \sum_{\mathbf{k},\mathbf{k}'} \Delta_k(t)\Delta_{k'}(t)e^{-i(\mathbf{k}+\mathbf{k}')\cdot\mathbf{x}}. \tag{5.61}$$

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

$$\Delta_{-k}(t) = \Delta_k^*(t). \tag{5.62}$$

Therefore,

$$\begin{aligned}\langle\Delta^2(\mathbf{x},t)\rangle &= \sum_{\mathbf{k}}|\Delta_k(t)|^2 \\ &= \frac{1}{(2\pi)^3}\int|\Delta_k(t)|^2d^3k \equiv \frac{1}{(2\pi)^3}\int P(\mathbf{k},t)d^3k, \end{aligned}\tag{5.63}$$

where we have defined the *power spectrum* of the density fluctuations as

$$P(\mathbf{k},t) \equiv |\Delta_k(t)|^2. \tag{5.64}$$

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

$$\langle\Delta^2(\mathbf{x},t)\rangle = \frac{1}{2\pi^2}\int_0^\infty k^2P(k)dk. \tag{5.65}$$

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

$$dN_{12} = \bar{n}^2[1+\xi(\mathbf{r},t)]dV_1 dV_2, \tag{5.66}$$

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

$$\begin{aligned} dN_{12} &= \langle\rho(\mathbf{x},t)\rho(\mathbf{x}+\mathbf{r},t)\rangle dV_1 dV_2 \\ &= \rho_0^2\langle[1+\Delta(\mathbf{x},t)][1+\Delta(\mathbf{x}+\mathbf{r},t)]\rangle dV_1 dV_2 \\ &= \rho_0^2[1+\langle\Delta(\mathbf{x},t)\Delta(\mathbf{x}+\mathbf{r},t)\rangle]dV_1 dV_2, \end{aligned} \tag{5.67}$$

where we have used $\langle\Delta\rangle = 0$. We therefore see that

$$\xi(\mathbf{r},t) = \langle\Delta(\mathbf{x},t)\Delta(\mathbf{x}+\mathbf{r},t)\rangle. \tag{5.68}$$

We can now derive a relation between the correlation function and the power spectrum:

$$\begin{aligned} \xi(\mathbf{r},t) &= \langle\Delta(\mathbf{x},t)\Delta(\mathbf{x}+\mathbf{r},t)\rangle = \langle\Delta(\mathbf{x},t)\Delta^*(\mathbf{x}+\mathbf{r},t)\rangle \\ &= \langle\sum_{\mathbf{k},\mathbf{k}'}\Delta_k(t)\Delta_{k'}^*(t)e^{-i\mathbf{k}\cdot\mathbf{x}}e^{i\mathbf{k}'\cdot(\mathbf{x}+\mathbf{r})}\rangle \\ &= \sum_{\mathbf{k},\mathbf{k}'}\Delta_k(t)\Delta_{k'}^*(t)e^{-i\mathbf{k}'\cdot\mathbf{r}}\frac{1}{V}\int e^{i(\mathbf{k}'-\mathbf{k})\cdot\mathbf{x}}d^3x \\ &= \sum_{\mathbf{k}}|\Delta_k(t)|^2 e^{-i\mathbf{k}\cdot\mathbf{r}} \\ &= \frac{1}{(2\pi)^3}\int|\Delta_k(t)|^2 e^{-i\mathbf{k}\cdot\mathbf{r}}d^3k \\ &= \frac{1}{(2\pi)^3}\int P(\mathbf{k},t)e^{-i\mathbf{k}\cdot\mathbf{r}}d^3k. \end{aligned} \tag{5.69}$$

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:

$$\begin{aligned} \xi(\mathbf{r},t) = \xi(r,t) &= \frac{1}{(2\pi)^3}\int_0^{2\pi}d\phi\int_{-1}^{+1}d(\cos\theta)\int_0^\infty dk k^2 P(k,t)e^{-ikr\cos\theta} \\ &= \frac{1}{4\pi^2}\int_0^\infty dk k^2 P(k,t)\frac{1}{ikr}(e^{ikr}-e^{-ikr}) \\ &= \frac{1}{2\pi^2}\int_0^\infty dk k^2 P(k,t)\frac{\sin(kr)}{kr}, \end{aligned} \tag{5.70}$$

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 $k > 1/r$.

5.8 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

$$\frac{\Delta T}{T_0}(\theta,\phi) = \frac{T(\theta,\phi) - T_0}{T_0}, \tag{5.71}$$

and it is practical to decompose $\Delta T/T_0$ in spherical harmonics:

$$\frac{\Delta T}{T_0}(\theta,\phi) = \sum_{\ell=0}^{\infty}\sum_{m=-\ell}^{\ell} a_{\ell m} Y_{\ell m}(\theta,\phi), \tag{5.72}$$

where the spherical harmonics obey the orthogonality relation

$$\int Y^*_{\ell m} Y_{\ell' m'} d\Omega = \delta_{\ell\ell'}\delta_{mm'}. \tag{5.73}$$

The coefficients $a_{\ell m}$ are given by

$$a_{\ell m} = \int \frac{\Delta T}{T_0}(\theta,\phi) Y_{\ell m}(\theta,\phi) d\Omega. \tag{5.74}$$

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:

$$C_\ell = \frac{1}{2\ell+1}\sum_{m=-\ell}^{\ell} a^*_{\ell m} a_{\ell m} = \langle |a_{\ell m}|^2 \rangle. \tag{5.75}$$

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

$$C(\theta) = \langle \frac{\Delta T(\mathbf{n}_1)}{T_0}\frac{\Delta T(\mathbf{n}_2)}{T_0}\rangle = \frac{1}{4\pi}\sum_{\ell=0}^{\infty}(2\ell+1)C_\ell P_\ell(\cos\theta), \tag{5.76}$$

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

$$\begin{aligned} n_\gamma &\propto n_b \propto \rho_b & (5.77)\\ n_\gamma &\propto T^3, & (5.78)\end{aligned}$$

which gives $T \propto \rho_b^{1/3}$ and

$$\frac{\Delta T}{T} \equiv \Theta_0 = \frac{1}{3}\frac{\Delta\rho_b}{\rho_b} = \frac{1}{3}\Delta_b. \tag{5.79}$$

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

$$\frac{d^2\Delta_b}{dt^2} + 2\frac{\dot{a}}{a}\frac{d\Delta_b}{dt} = \text{gravitational term} - \text{pressure term}. \tag{5.80}$$

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

$$\frac{d^2\Delta_b}{dt^2} + 2\frac{\dot{a}}{a}\frac{d\Delta_b}{dt} = 4\pi G\rho_D\Delta_D - \Delta_b k^2 c_s^2. \tag{5.81}$$

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

$$\frac{d^2\Theta_0}{dt^2} = \frac{4\pi G\Delta_D\rho_D}{3} - k^2c_s^2\Theta_0. \tag{5.82}$$

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

$$\phi_k = -\frac{4\pi G \rho_D \Delta_D}{k^2}, \tag{5.83}$$

so that

$$\frac{d^2\Theta_0}{dt^2} = -\frac{1}{3}k^2\phi_k - k^2 c_s^2 \Theta_0. \tag{5.84}$$

We look at adiabatic perturbations, and the entropy is dominated by the photons,

$$S \propto T^3 V \propto \frac{T^3}{m_b/V} \propto \frac{T^3}{\rho_b} \propto \frac{\rho_r^{3/4}}{\rho_b}, \tag{5.85}$$

where m_b is the baryon mass, and we recall that $\rho_r \propto T^4$, so we have

$$\frac{\delta S}{S} = \frac{3}{4}\frac{\delta\rho_r}{\rho_r} - \frac{\delta\rho_b}{\rho_b} = 3\frac{\delta T}{T} - \frac{\delta\rho_b}{\rho_b} = 0, \tag{5.86}$$

so that

$$\Delta_b = \frac{\delta\rho_b}{\rho_b} = 3\frac{\delta T}{T} = \frac{3}{4}\frac{\delta\rho_r}{\rho_r}. \tag{5.87}$$

The speed of sound is given by

$$c_s = \left(\frac{\partial p}{\partial \rho}\right)_S^{1/2}. \tag{5.88}$$

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

$$\begin{aligned} c_s^2 &= \frac{\delta p}{\delta \rho} = \frac{\delta\rho_r c^2/3}{\delta\rho_b + \delta\rho_r} \\ &= \frac{c^2}{3}\frac{1}{1+\frac{\delta\rho_b}{\delta\rho_r}}, \end{aligned} \tag{5.89}$$

so that

$$\begin{aligned} c_s &= \frac{c}{\sqrt{3}}\left[1+\left(\frac{\delta\rho_b}{\delta\rho_r}\right)_S\right]^{-1/2} \\ &= \frac{c}{\sqrt{3}}\left(1+\frac{3}{4}\frac{\rho_b}{\rho_r}\right)^{-1/2} \\ &= \frac{c}{\sqrt{3(1+\mathcal{R})}}, \end{aligned} \tag{5.90}$$

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 (5.84) is a simple oscillator equation, and by substitution one can show that

$$\begin{aligned} \Theta_0(t) &= \left[\Theta_0(0) + \frac{(1+\mathcal{R})}{c^2}\phi_k\right]\cos(kc_s t) \\ &+ \frac{1}{kc_s}\dot{\Theta}_0(0)\sin(kc_s t) - \frac{(1+\mathcal{R})}{c^2}\phi_k \end{aligned} \tag{5.91}$$

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,

$$kc_s t_{\rm rec} = k\lambda_S, \tag{5.92}$$

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

$$\begin{aligned} \Theta_0(t_{\rm rec}) &= \left[\Theta_0(0) + \frac{(1+\mathcal{R})}{c^2}\phi_k\right]\cos(k\lambda_S) \\ &+ \frac{1}{kc_S}\dot{\Theta}_0(0)\sin(k\lambda_S) - \frac{(1+\mathcal{R})}{c^2}\phi_k. \end{aligned} \tag{5.93}$$

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,

$$\Theta_0(t_{\rm rec}) + \frac{\phi_k}{c^2} = \left[\Theta_0(0) + \frac{(1+\mathcal{R})}{c^2}\phi_k\right]\cos(k\lambda_S) - \frac{\mathcal{R}}{c^2}\phi_k. \tag{5.94}$$

The term in the angular brackets correspond to horizon-scale fluctuations, the so-called Sachs-Wolfe effect, and one can show that

$$\Theta_0(0) + \frac{\phi_k}{c^2} = \frac{\phi_k}{3c^2}, \tag{5.95}$$

and that the observed temperature fluctuations therefore can be written as

$$\left(\frac{\Delta T}{T_0}\right)_{\rm eff} = \frac{\phi_k}{3c^2}(1+3\mathcal{R})\cos(k\lambda_S) - \frac{\mathcal{R}}{c^2}\phi_k. \tag{5.96}$$

The first extremal value occurs for $k\lambda_S = \pi$, which gives

$$\left(\frac{\Delta T}{T_0}\right)_{\text{eff}} = -\frac{\phi_k}{3c^2}(1 + 6\mathcal{R}), \tag{5.97}$$

and the next one occurs for $k\lambda_S = 2\pi$, giving

$$\left(\frac{\Delta T}{T_0}\right)_{\text{eff}} = \frac{\phi_k}{3c^2} \tag{5.98}$$

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 $\Omega_{b0}h^2$.

Chapter 6

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.

6.1 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

$$\Omega(t_{\rm P}) - 1 \sim 3 \times 10^{-61}.$$

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

$$r_{\rm PH}(z_{\rm LSS}) = \int_{z_{\rm LSS}}^{\infty} \frac{c dz}{a_0 H_0 \sqrt{\Omega_{\rm m0}}(1+z)^{3/2}} = \frac{2c}{a_0 H_0 \sqrt{\Omega_{\rm m0}}}(1 + z_{\rm LSS})^{-1/2}.$$

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

$$r(z_{\rm LSS}) = \int_0^{z_{\rm LSS}} \frac{c dz}{a_0 H(z)}.$$

For a spatially flat universe with dust and a cosmological constant, one finds numerically that to a very good approximation,

$$r(z_{\mathrm{LSS}}) \approx \frac{1.94c}{a_0 H_0 \Omega_{\mathrm{m0}}^{0.4}}.$$

This gives us the angular size of the particle horizon at last scattering on the sky today as

$$\theta_{\mathrm{PH}} = \frac{r_{\mathrm{PH}}(z_{\mathrm{LSS}})}{r(z_{\mathrm{LSS}})} \sim 1.8\Omega_{\mathrm{m0}}^{-0.1} \text{ degrees.}$$

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.

6.2 Example: de Sitter space

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' (2nd edition,Cambridge University Press, 2017), 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 ESA's Planck mission: $|\Omega(t_0) - 1| \leq 0.005$. Assuming the universe was matter dominated back to $t_{\rm eq}$, we find that

$$|\Omega(t_{\rm eq}) - 1| \leq |\Omega(t_0) - 1| \left(\frac{t_{\rm eq}}{t_0}\right)^{2/3} \sim 0.005 \times \left(\frac{1.6 \times 10^{12}\ \text{s}}{4.4 \times 10^{17}\ \text{s}}\right)^{2/3} \sim 10^{-6}.$$

From there and back to the end of inflation, we take the universe to be radiation dominated, and hence

$$|\Omega(t = 10^{-34}\ \text{s}) - 1| \leq 10^{-6} \left(\frac{10^{-34}\ \text{s}}{1.6 \times 10^{12}\ \text{s}}\right) \sim 6 \times 10^{-53}.$$

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

$$e^{-2N} \sim 6 \times 10^{-53},$$

which gives $N \sim 60$.

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. Also, if vacuum energy drives the present era of accelerated expansion, its value is far too low to provide the rapid expansion required for inflation to work. For these reasons, 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.

6.3 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. The discovery of the Higgs boson, the particle corresponding to the Higgs field, at the LHC in 2012 lent some moral support to the idea behind inflation since this confirmed that fundamental scalar fields exist in Nature.

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, it can be shown that the energy density of the field is given by

$$\rho_\phi c^2 = \frac{1}{2\hbar c^3}\dot{\phi}^2 + V(\phi), \tag{6.1}$$

and the pressure by

$$p_\phi = \frac{1}{2\hbar c^3}\dot{\phi}^2 - V(\phi), \tag{6.2}$$

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

$$H^2 = \frac{8\pi G}{3c^2}\rho_\phi c^2 = \frac{8\pi G}{3c^2}\left(\frac{1}{2\hbar c^3}\dot{\phi}^2 + V(\phi)\right). \tag{6.3}$$

As the second equation to use, we will choose the adiabatic expansion equation

$$\dot{\rho}c^2 = -3H(\rho c^2 + p).$$

From equation (6.1) we get

$$\dot{\rho}_\phi c^2 = \frac{\dot{\phi}\ddot{\phi}}{\hbar c^3} + \frac{dV}{d\phi}\dot{\phi},$$

and from (6.1) and (6.2) we see that $\rho_\phi c^2 + p_\phi = \dot{\phi}^2/(\hbar c^3)$. Hence, the equation for the scalar field becomes

$$\ddot{\phi} + 3H\dot{\phi} + \hbar c^3 V'(\phi) = 0, \tag{6.4}$$

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

$$\left(\frac{dV}{d\phi}\right)^2 \ll \frac{24\pi G}{\hbar c^5}V^2 = \frac{24\pi}{E_{\rm P}^2}V^2,$$

or

$$\frac{2}{3}\frac{E_{\rm P}^2}{16\pi}\left(\frac{V'}{V}\right)^2 \ll 1.$$

It is usual to define the so-called *slow-roll parameter* ϵ by

$$\epsilon = \frac{E_{\rm P}^2}{16\pi}\left(\frac{V'}{V}\right)^2, \tag{6.5}$$

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,

$$\frac{1}{3} \frac{E_{\mathrm{P}}^2}{8\pi} \frac{V''}{V} \ll 1.$$

Defining

$$\eta = \frac{E_{\mathrm{P}}^2}{8\pi} \frac{V''}{V}, \tag{6.6}$$

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 (6.3) and (6.4) reduce to

$$\begin{aligned} H^2 &\approx \frac{8\pi G}{3c^2} V(\phi) && (6.7) \\ 3H\dot{\phi} &\approx -\hbar c^3 V'(\phi). && (6.8) \end{aligned}$$

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 (6.7) 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 (6.8) by (6.7):

$$\frac{3H\dot{\phi}}{H^2} = -\hbar c^3 \frac{3c^2}{8\pi G}\frac{V'}{V},$$

which gives

$$\frac{\dot{\phi}}{H} = -\frac{\hbar c^5}{8\pi G}\frac{V'}{V} = -\frac{E_{\rm P}^2}{8\pi}\frac{V'}{V}.$$

By inserting this in the expression for $\dot{H}$ above, we find

$$\dot{H} = -\frac{4\pi G}{3c^2}\frac{E_{\rm P}^2}{8\pi}\frac{(V')^2}{V}.$$

If we now use equation (6.7) again, we get

$$-\frac{\dot{H}}{H^2} = \frac{4\pi G}{3c^2}\frac{3c^2}{8\pi G}\frac{1}{V}\frac{E_{\rm P}^2}{8\pi}\frac{(V')^2}{V} = \frac{E_{\rm P}^2}{16\pi}\left(\frac{V'}{V}\right)^2 = \epsilon,$$

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 remain at a given time t before inflation ends. This number is defined as

$$N = \ln\left[\frac{a(t_{\rm end})}{a(t)}\right], \qquad (6.9)$$

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 (6.7) by (6.8) we get

$$N(t) = -\frac{8\pi}{E_{\rm P}^2}\int_t^{t_{\rm end}} \frac{V}{V'}\dot{\phi}dt = \frac{8\pi}{E_{\rm P}^2}\int_{\phi_{\rm end}}^{\phi} \frac{V}{V'}d\phi, \qquad (6.10)$$

where $\phi_{\rm end} = \phi(t_{\rm end})$ can be found from the criterion $\epsilon(\phi_{\rm end}) = 1$.

6.3.1 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 6.1.) 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:

$$\epsilon = \frac{E_{\rm P}^2}{16\pi}\left(\frac{V'}{V}\right)^2 = \frac{E_{\rm P}^2}{4\pi\phi^2},$$

and

$$\eta = \frac{E_{\rm P}^2}{8\pi}\frac{V''}{V} = \frac{E_{\rm P}^2}{4\pi\phi^2} = \epsilon.$$

The criterion for the SRA to be valid hence becomes

$$\phi \gg \frac{E_{\rm P}}{2\sqrt{\pi}} \equiv \phi_{\rm end},$$

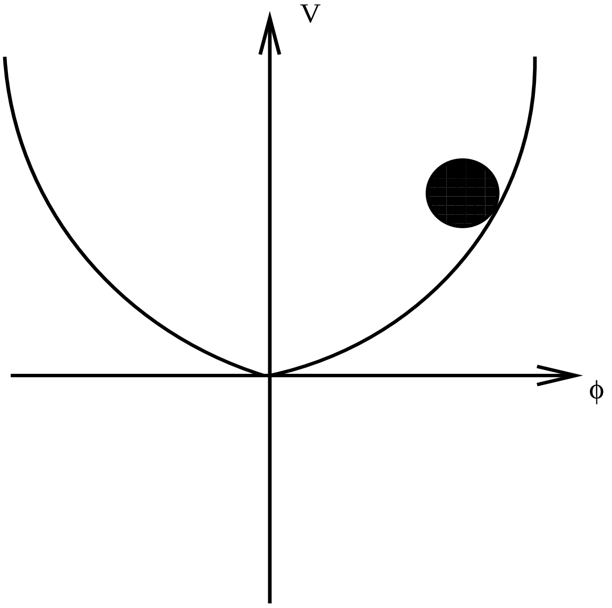

Figure 6.1: The inflaton depicted as a ball rolling down a potential well.

and inflation will be over when $\phi \sim \phi_{\mathrm{end}}$.

Inserting the potential in the SRA equations (6.7) and (6.8) gives

$$\begin{aligned} H^2 &= \frac{4\pi G}{3}\frac{m^2c^2}{(\hbar c)^3}\phi^2 = \frac{4\pi}{3}\frac{m^2c^4}{\hbar^2}\frac{\phi^2}{E_{\mathrm{P}}^2} \\ 3H\dot{\phi} &= -\frac{m^2c^4}{\hbar^2}\phi. \end{aligned}$$

Taking the square root of the first equation and inserting it in the second, we get

$$\sqrt{12\pi}\frac{mc^2}{\hbar}\frac{\dot{\phi}\phi}{E_{\mathrm{P}}} + \frac{m^2c^4}{\hbar^2}\phi = 0,$$

i.e.,

$$\dot{\phi} = -\frac{E_{\mathrm{P}}}{\sqrt{12\pi}}\frac{mc^2}{\hbar},$$

which can be trivially integrated to give

$$\phi(t) = \phi_i - \frac{mc^2E_{\mathrm{P}}}{\hbar\sqrt{12\pi}}t,$$

where for convenience we take inflation to begin at $t_i = 0$. Inserting this result in the equation for H, we get

$$H = \sqrt{\frac{4\pi}{3}}\frac{mc^2}{\hbar E_{\mathrm{P}}}\left(\phi_i - \frac{mc^2E_{\mathrm{P}}}{\hbar\sqrt{12\pi}}t\right),$$

and since $H = \dot{a}/a = da/adt$, we get

$$\int_{a_i}^{a(t)} \frac{da}{a} = \sqrt{\frac{4\pi}{3}} \frac{mc^2}{\hbar E_{\rm P}} \int_0^t \left(\phi_i - \frac{mc^2 E_{\rm P}}{\hbar\sqrt{12\pi}} t \right) dt,$$

and finally,

$$a(t) = a_i \exp\left[\sqrt{\frac{4\pi}{3}} \frac{mc^2}{\hbar E_{\rm P}} \left(\phi_i t - \frac{mc^2 E_{\rm P}}{2\hbar\sqrt{12\pi}} t^2 \right) \right].$$

We can find the total number of e-foldings produced for a given initial field value ϕ_i by using (6.10):

$$N = \frac{8\pi}{E_{\rm P}^2} \int_{\phi_{\rm end}}^{\phi_i} \frac{V d\phi}{V'} = \frac{8\pi}{E_{\rm P}^2} \int_{E_{\rm P}/\sqrt{4\pi}}^{\phi_i} \frac{1}{2} \phi d\phi = \left(\frac{\phi_i \sqrt{2\pi}}{E_{\rm P}} \right)^2 - \frac{1}{2}.$$

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

$$\phi_i = \frac{11}{2\sqrt{\pi}} E_{\rm P} \approx 3.10 E_{\rm Pl}.$$

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 $E_{\rm P}$ 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, $E_{\rm P}/l_{\rm P}^3$, 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

$$V(\phi_i) = \frac{1}{2} \frac{m^2 c^4}{(\hbar c)^3} \phi_i^2 = \frac{121}{8\pi} \frac{E_{\rm P}^2 m^2 c^4}{(\hbar c)^3}.$$

This should be compared to the Planck energy density

$$\rho_{\rm P} c^2 = \frac{c^7}{\hbar G^2},$$

and $V(\phi_i)$ will therefore be much less than $\rho_{\rm Pl} c^2$ if m satisfies

$$mc^2 \ll \left[\frac{(\hbar c)^3}{E_{\rm P} l_{\rm P}^3} \right]^{1/2} = E_{\rm P}.$$

Therefore, as long as the mass of the scalar field is much smaller than the Planck mass, we should be safe.

6.3.2 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 (6.4) 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.

6.4 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 $k_B T \sim mc^2$, and from the relationship between temperature and time in the early universe (derived in chapter 2) I find

$$k_B T = mc^2 \sim E_P \sqrt{\frac{t_P}{t}},$$

so that

$$t \sim \frac{\hbar E_{\mathrm{P}}}{m^2 c^4}.$$

The order of magnitude of the fluctuations in the energy per particle is therefore

$$\Delta E \sim \frac{\hbar}{t} \sim \frac{m^2 c^4}{E_{\mathrm{P}}},$$

and the relative fluctuations have amplitude

$$\frac{\Delta E}{E} \sim \frac{1}{mc^2}\frac{m^2 c^4}{E_{\mathrm{P}}} \sim \frac{mc^2}{E_{\mathrm{Pl}}}.$$

The energy density is given by $\rho \propto T^4 \propto E^4$, and so I find

$$\frac{\Delta\rho}{\rho} \sim \frac{d\rho}{\rho} \sim \frac{1}{E^4} 4E^3 dE \sim \frac{dE}{E} \sim \frac{\Delta E}{E},$$

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

$$\begin{aligned}
\Delta\rho &= \frac{U}{a^3} - \frac{U}{(a+\Delta a)^3} \\
&= U\left(\frac{1}{a^3} - \frac{1}{(a+\dot{a}\Delta t)^3}\right) \\
&= \frac{U}{a^3}\left(1 - \frac{1}{\left(1+\frac{\dot{a}}{a}\Delta t\right)^3}\right) \\
&\approx \frac{U}{a^3}\left[1 - \left(1 - 3\frac{\dot{a}}{a}\Delta t\right)\right] \\
&= 3H\Delta t\rho,
\end{aligned}$$

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),

$$\begin{aligned}
H^2 &= \frac{8\pi\hbar c^3}{3E_{\rm P}^2}V(\phi) \\
\dot{\phi} &= -\frac{\hbar c^3}{3H}V'(\phi).
\end{aligned}$$

If I insert these equations in the expression for $\Delta\rho/\rho$, I find

$$\begin{aligned}
\frac{\Delta\rho}{\rho} &\sim \hbar\frac{\hbar c^3}{E_{\rm P}^2}V\frac{H}{\hbar c^3 V'} \\
&\sim \frac{\hbar}{E_{\rm P}^3}\frac{V}{V'}H \\
&\sim \frac{(\hbar c)^{3/2}}{E_{\rm P}^3}\frac{V^{3/2}}{V'}.
\end{aligned}$$

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, and the ESA satellite Planck 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.

6.4.1 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. An object hit by a passing gravitational wave would show an oscillatory pattern of being tretched and compressed in the directions perpendicular to the direction of propagation of the wave. The amplitude of these waves is, however, very small. The first detection of gravitational waves was announced by the LIGO team in 2016. Their two interferometers had seen gravitational waves from two black holes merging in a galaxy 1.4 billion light years away. These waves induced a change in the length of their detectors of about one thousandth of the size of a proton!

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

$$\Phi(\vec{x}, t) = -G \int \frac{\rho(\vec{y}, t)}{|\vec{x} - \vec{y}|} d^3y. \tag{6.11}$$

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 (6.11) 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:

$$\Phi(\vec{x}, t) = -G \int \frac{\rho\left(\vec{y}, t - \frac{|\vec{x}-\vec{y}|}{c}\right)}{|\vec{x}-\vec{y}|} d^3y. \tag{6.12}$$

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 (6.12) 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

$$\Phi(\vec{x}, t) = A e^{i(\vec{k}\cdot\vec{x} - \omega t)},$$

where $\omega = c|\vec{k}|$.

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

$$Q_{\ell m} = \int d^3 x r^\ell Y^*_{\ell m}(\theta, \phi)\rho,$$

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 not 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_{P}, since this is believed to be the energy scale of quantized gravity. Thus,

$$\frac{1}{H}\frac{\Delta\Phi}{\Phi}E_{\mathrm{P}} \sim \hbar,$$

which gives

$$\frac{\Delta\Phi}{\Phi} \sim \frac{\hbar H}{E_{\mathrm{P}}} \propto (\hbar c)^{3/2}\frac{V^{1/2}}{E_{\mathrm{P}}},$$

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.

6.4.2 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

$$50 = \frac{8\pi}{E_{\rm P}^2}\int_{\phi_{\rm end}}^{\phi_*}\frac{V}{V'}d\phi.$$

We have seen that

$$\begin{aligned}\frac{\Delta\rho}{\rho} &\sim \frac{(\hbar c)^{3/2}}{E_{\rm P}^3}\frac{V^{3/2}}{V'}\\ \frac{\Delta\Phi}{\Phi} &\sim \frac{(\hbar c)^{3/2}}{E_{\rm P}^2}V^{1/2}.\end{aligned}$$

If I form the ration of these two amplitudes, I find that

$$r \equiv (\Delta\Phi/\Phi)/(\Delta\rho/\rho) \sim E_{\rm P}\frac{V'}{V} \propto \sqrt{\epsilon}.$$

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

$$\epsilon = \frac{E_{\mathrm{P}}^2}{4\pi\phi^2},$$

and that inflation ends when the field has dercreased to the value

$$\phi_{\mathrm{end}} = \frac{E_{\mathrm{P}}}{2\sqrt{\pi}}.$$

Note that we can also write

$$\epsilon = \frac{\phi_{\mathrm{end}}^2}{\phi^2}.$$

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

$$50 = \frac{8\pi}{E_{\mathrm{P}}^2}\int_{\phi_{\mathrm{end}}}^{\phi_*} \frac{V'}{V} d\phi = \frac{8\pi}{E_{\mathrm{P}}^2}\int_{\phi_{\mathrm{end}}}^{\phi_*} \frac{1}{2}\phi d\phi.$$

The integral is easily evaluated, and the resulting equation just as easliy solved with the result

$$\left(\frac{\phi_*}{\phi_{\mathrm{end}}}\right)^2 = 101,$$

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 Planck satellite has found an upper limit of $r < 0.09$, so the quadratic potential we have studied seems to be ruled out.

6.4.3 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

$$\frac{\Delta E}{E} \sim \frac{mc^2}{E_{\mathrm{P}}},$$

and

$$\frac{\Delta\rho}{\rho} \sim \frac{(\hbar c)^{3/2}}{E_{\mathrm{P}}^3}\frac{V^{3/2}}{V'}$$

we see that no specific length scale is picked out by the fluctuations. That does not exclude that the amplitude varies with length 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.

Chapter 7

Bonus material: General relativity

7.1 Why general relativity?

The most important equations in the course are arguably the Friedmann equations, which describe the evolution of a homogeneous and isotropic universe. One can derive something very similar to them from Newtonian gravity, but not without making assumptions that are strictly not warranted, and radiation (photons and other particles moving at, or close to, the speed of light) have to be inserted by hand. In my experience many students are left unconvinced and unsatisfied by this shortcut. I experimented once with starting the course with the following short introduction to general relativity and derivation of the Friedmann equations, but many of the students thought that I wasted valuable time on material that would not appear in the final exam. I can see their point, but I have decided to keep this chapter in the lecture notes for those who would like to browse through it in their leisure time.

General Relativity (GR for short) represents our best description and understanding of space, time and gravity to date. It is essential for formulating consistent cosmological models. It is a geometric theory and can be formulated in a so-called coordinate-independent way. This you will learn in the GR course in the physics department. A coordinate-independent approach requires a higher level of abstraction than our purposes in this course demand. When we apply the theory, we will always choose a specific set of coordinates. I will therefore present the theory in a more old-fashioned form.

7.2 Tensors

Consider two points P and Q with coordinates x^μ and $x^\mu + dx^\mu$, respectively, where $\mu = 0, 1, \ldots n-1$ (so that space has n dimensions). These two points define an infinitesimal vector $\vec{PQ}$ which we consider to be attached to the starting point P. The components of the vector in the x^μ system are dx^μ. What will they be in another coordinate system, say x'^μ? There will be a transformation from the first system to the latter wich can be expressed as $x'^\mu = x'^\mu(x^\nu)$. Since we consider an infinitesimal vector, we can use the chain rule for differentiation to find the coordinates in the new system:

$$dx'^\mu = \frac{\partial x'^\mu}{\partial x^\nu} dx^\nu,$$

where we have used the convention that repeated indices are summed over (we sum over all values of ν, from 0 to $n-1$). The partial derivatives are to be evaluated at the point P. The relation above for the transformation of an infinitesimal vector is the basis for the definition of what we mean by a **contravariant vector** (can also be seen as a **contravariant tensor of rank 1**). A contravariant vector is a set of quantities A^μ in the x^μ system which transform to the x'^μ system in the same way as dx^μ:

$$A'^\mu = \frac{\partial x'^\mu}{\partial x^\nu} A^\nu,$$

where the partial derivatives again are to be evaluated at the point P.

A **contravariant tensor of rank 2** is a set of n^2 quantities $T^{\mu\nu}$ associated with the point P in the x^μ system which transform in the following way:

$$T'^{\mu\nu} = \frac{\partial x'^\mu}{\partial x^\alpha} \frac{\partial x'^\nu}{\partial x^\beta} T^{\alpha\beta}.$$

Perhaps the simplest example of this are the products of the components of to contravariant vectors, $A^\mu B^\nu$. We can define contravariant tensors of arbitrarily high rank by adding more factors of $\partial x'^\mu / \partial x^\alpha$ to the transformation equation.

An important special case is a contravariant tensor of rank zero, better known as a *scalar*. Perhaps not surprisingly, a scalar, say ϕ, transforms as

$$\phi' = \phi$$

Let $\phi = \phi(x^\mu)$ be a continuous, differentiable scalar function, so that its derivatives $\partial\phi/\partial x^\mu$ exist at all points. We may consider the coordinates x^μ as functions of x'^ν and write

$$\phi = \phi(x^\mu(x'^\nu)).$$

If we now use the chain rule to differentiate with respect to x'^ν, we get

$$\frac{\partial \phi}{\partial x'^\nu} = \frac{\partial x^\mu}{\partial x'^\nu} \frac{\partial \phi}{\partial x^\mu}.$$

This equation serves as the prototype for how a **covariant vector** (also known as a **covariant tensor of rank 1**) transforms. In general a covariant vector is a set of quantities A_μ associated with the point x^μ transforming as

$$A'_\mu = \frac{\partial x^\nu}{\partial x'^\mu} A_\nu .$$

Note that x and x' have exchanged places in the partial derivative. We can go on in the same way as for contravariant tensors and define covariant tensors of higher rank. As an example, a covariant tensor of rank 2 transforms as

$$T'_{\mu\nu} = \frac{\partial x^\alpha}{\partial x'^\mu} \frac{\partial x^\beta}{\partial x'^\nu} T_{\alpha\beta} .$$

We can also construct mixed tensors. A mixed tensor of rank 3 may, for example, have one contravariant and two covariant indices and transform as

$$T'^{\mu}_{\ \nu\sigma} = \frac{\partial x'^\mu}{\partial x^\alpha} \frac{\partial x^\beta}{\partial x'^\nu} \frac{\partial x^\gamma}{\partial x'^\sigma} T^{\alpha}_{\beta\gamma} .$$

Why bother with tensors? An important reason for us is that they are important in relativity. We can catch a glimpse of their significance if we consider the tensor equation

$$X_{\mu\nu} = Y_{\mu\nu} .$$

This equation tells us that the components of the covariant tensors X and Y are equal in the coordinate system x. But then we also have

$$\frac{\partial x^\mu}{\partial x'^\alpha} \frac{\partial x^\nu}{\partial x'^\beta} X_{\mu\nu} = \frac{\partial x^\mu}{\partial x'^\alpha} \frac{\partial x^\nu}{\partial x'^\beta} Y_{\mu\nu}$$

and since we know that X and Y transform as covariant tensors of rank 2, we may conclude that $X'_{\alpha\beta} = Y'_{\alpha\beta}$. The components of X and Y are equal also in the new coordinate system. Since there is nothing special about x and x', this shows that a tensor equation is valid in **all** coordinate systems. Tensors are therefore natural objects to make use of if we want to formulate laws of nature that are valid in all reference frames.

One thing to bear in mind is that a tensor equation like

$$X^{\mu}_{\alpha\beta} = Y^{\mu\alpha}_{\beta}$$

MAKES ABSOLUTELY NO SENSE! ABSOLUTELY NONE! If you ever make a mistake like this, years will be added to your time in purgatory, I promise. I dread to think about how much time there I have ratched up. Although as mixed tensors of rank 3 they have the same number of components, they are mathematically totally different objects. X has one contravariant and two covariant indices, Y has two contravariant and one covariant. This will become more obvious when you learn what sort of geometrical objects tensors are, but for now we may note that X and Y transform differently.

If tensors are to be useful to us, we must be able to do things to them, Things like adding and subtracting two tensors, for example. One important point here, related to the previous paragraph, is that we can only add and subtract tensors of the same type. If they are, addition and subtraction are defined componentwise. For example:

$$S_{\mu\nu} = X_{\mu\nu} + Y_{\mu\nu}.$$

Symmetric and **antisymmetric** tensors are concepts that are useful from time to time. A tensor of rank 2 is symmetric if (shown here for covariant tensors) $X_{\mu\nu} = X_{\nu\mu}$, and antisymmetric if $X_{\mu\nu} = -X_{\nu\mu}$. Any tensor of rank 2 can be written as the sum of a symmetric part $X_{(\mu\nu)}$ and an antisymmetric part $X_{[\mu\nu]}$, where

$$\begin{aligned} X_{(\mu\nu)} &= \frac{1}{2}\left(X_{\mu\nu} + X_{\nu\mu}\right) \\ X_{[\mu\nu]} &= \frac{1}{2}\left(X_{\mu\nu} - X_{\nu\mu}\right) \end{aligned}$$

which you can easily check.

Another important operation is a **contraction**: From a tensor of contravariant rank $p \geq 1$ and covariant rank $q \geq 1$, we can form a tensor of rank $(p-1, q-1)$ by equating one contravariant index to one covariant index and sum over them. An example will make this clearer: From $X^{\mu}_{\nu\gamma\sigma}$ we can form the tensor $Y_{\gamma\sigma}$ by taking

$$Y_{\gamma\sigma} = X^{\mu}_{\mu\gamma\sigma}.$$

Remember here that repeated indices are summed over. Note that if we contract a tensor of rank $(1, 1)$, we get a scalar:

$$X^{\mu}_{\mu} = A.$$

The metric tensor

A particularly important tensor $g_{\mu\nu}$ of rank 2 is the **metric tensor**, or quite simply the **metric**. This is a symmetric tensor used to define the distance ds between two infinitesimally separated points x^{μ} and $x^{\mu} + dx^{\mu}$:

$$ds^2 = g_{\mu\nu}(x) dx^{\mu} dx^{\nu}.$$

It also defines the length of a vector A in the point x as

$$A^2 = g_{\mu\nu}(x) A^{\mu} A^{\nu},$$

and the scalar product between two vectors A and B,

$$AB = g_{\mu\nu}(x)A^\mu B^\nu.$$

Two vectors are said to be orthogonal if their scalar product vanishes. In both the special and the general theory of relativity it is possible for a vector which is different from the null vector to have zero length, $A^2 = 0$.

Although it is not strictly true, it is sometimes useful to think of the metric as an $n \times n$ quadratic matrix. It then has a determinant, which we denote by g, $g = \det(g_{\mu\nu})$. If $g \neq 0$, the metric has an inverse. This inverse is a contravariant tensor of rank 2, $g^{\mu\nu}$, satisfying

$$g_{\mu\alpha}g^{\alpha\nu} = \delta^\nu_\mu,$$

where $\delta^\nu_\mu = 1$ for $\mu = \nu$ equal to zero otherwise. The metric $g_{\mu\nu}$ can be used to lower indices, and its inverse $g^{\mu\nu}$ to lift indices. For example:

$$\begin{aligned} T^\nu_\mu &= g_{\mu\alpha}T^{\alpha\nu} \\ T^\mu_\nu &= g^{\mu\alpha}T_{\alpha\nu}. \end{aligned}$$

The equivalence principle

The equivalence principle is one of the starting points for the general theory of relativity. The crucial observation forming the basis for the equivalence principle is that an observer in free fall feels no gravitation field. If he releases an object from rest, the object will remain at rest with respect to him after he has let go of it. In general, there is no experiment he can make that will reveal to him that he is in a gravitational field. He can regard himself as being at rest. More precisely, although perhaps also more obscurely, we can say that *In any point in a gravitation field we can choose a frame of reference, the so-called free-fall system, defined by the fact that it moves with the acceleration a freely falling body would have had at the same point. In ths system, all the laws of physics will have the samme form as in the special theory of relativity. The exception is gravity, which vanishes locally in this system*

A couple of comments are in order:

1. This formulation of the equivalence principle implies that inertial mass m_I (which appears in Newton's 2. law) equals gravitating mass m_G (which appears in the law of gravity). If this were not the case, the observer and the object he releases would have experienced different accelerations, and they would not have remained at rest with respect to each other.

2. Note that this is valid *at a point.* It is in general not possible to find a frame of reference covering all of spacetime in which gravity vanishes everywhere.

The equivalence principle is important because it helps us with formulating relativistically correct laws: Start by analysing the situation the the free-fall system where special relativity can be applied. If we can formulate the result as a tensor equation, we then know that it will be valid in *all* reference frames.

The geodesic equation

We will now use the equivalence principle to find the equation of motion of a particle in a graviational field by starting in the free-fall frame where we can use special relativity. The particle is at a point with coordinates $\xi^\mu = (t, x, y, z)$, and we have chosen units where $c = 1$. The line element is

$$ds^2 = dt^2 - dx^2 - dy^2 - dz^2 = \eta_{\mu\nu} d\xi^\mu d\xi^\nu,$$

so the metric is the Minkowski metric $\eta_{\mu\nu} = \mathrm{diag}(1, -1, -1, -1)$. Since there are no forces acting on the particle in the free-fall system, its equation of motion according to special relativity is simply

$$\frac{d^2\xi^\mu}{d\tau^2} = 0,$$

where $d\tau^2 = ds^2$ is **proper time**, that is, time as measured on a watch following the particle. This is a tensor equation in special relativity: Under Lorentz transformations ξ is a four-vector, and τ is a scalar. But Lorentz transformations only apply between frames moving with constant relative velocity. We need an equation which is invariant under more general transformations. Let us find out what happens to the equation above under a general transformation to new coordinates x^μ. Under such a transformation we will have

$$d\xi^\mu = \frac{\partial \xi^\mu}{\partial x^\nu} dx^\nu,$$

so

$$\frac{d\xi^\mu}{d\tau} = \frac{\partial \xi^\mu}{\partial x^\nu} \frac{dx^\nu}{d\tau}.$$

We then have

$$\begin{aligned} 0 &= \frac{d^2\xi^\mu}{d\tau^2} = \frac{d}{d\tau}\left(\frac{\partial \xi^\mu}{\partial x^\nu}\frac{dx^\nu}{d\tau}\right) \\ &= \frac{\partial \xi^\mu}{\partial x^\nu}\frac{d^2x^\nu}{d\tau^2} + \frac{dx^\nu}{d\tau}\frac{d}{d\tau}\left(\frac{\partial \xi^\mu}{\partial x^\nu}\right) \\ &= \frac{\partial \xi^\mu}{\partial x^\nu}\frac{d^2x^\nu}{d\tau^2} + \frac{\partial^2 \xi^\mu}{\partial x^\nu \partial x^\rho}\frac{dx^\nu}{d\tau}\frac{dx^\rho}{d\tau} \end{aligned}$$

Next we multiply this equation by $\partial x^\sigma / \partial \xi^\mu$ and sum over μ. In the first term we find the factor

$$\frac{\partial x^\sigma}{\partial \xi^\mu}\frac{\partial \xi^\mu}{\partial x^\nu}\frac{d^2x^\nu}{d\tau^2} = \frac{\partial x^\sigma}{\partial x^\nu}\frac{d^2x^\nu}{d\tau^2} = \delta^\sigma_\nu \frac{d^2x^\nu}{d\tau^2} = \frac{d^2x^\sigma}{d\tau^2}$$

and hence

$$\frac{d^2x^\sigma}{d\tau^2}+\frac{\partial x^\sigma}{\partial\xi^\mu}\frac{\partial^2\xi^\mu}{\partial x^\nu\partial x^\rho}\frac{dx^\nu}{d\tau}\frac{dx^\rho}{d\tau}=0.$$

We rewrite this equation as

$$\frac{d^2x^\sigma}{d\tau^2}+\Gamma^\sigma_{\nu\rho}\frac{dx^\nu}{d\tau}\frac{dx^\rho}{d\tau}=0,$$

where

$$\Gamma^\sigma_{\nu\rho}=\frac{\partial x^\sigma}{\partial\xi^\mu}\frac{\partial^2\xi^\mu}{\partial x^\nu\partial x^\rho}$$

is called **the Christoffel symbol** or **the connection**. In the new coordinates, the proper time is given by

$$d\tau^2=\eta_{\mu\nu}d\xi^\mu d\xi^\nu=\eta_{\mu\nu}\frac{\partial\xi^\mu}{\partial x^\alpha}\frac{\partial\xi^\nu}{\partial x^\beta}dx^\alpha dx^\beta\equiv g_{\alpha\beta}dx^\alpha dx^\beta,$$

where the metric in the new coordinates is

$$g_{\alpha\beta}=\eta_{\mu\nu}\frac{\partial\xi^\mu}{\partial x^\alpha}\frac{\partial\xi^\nu}{\partial x^\beta}.$$

You can now convince yourself that the new equation of motion, called **the geodesic equation** is invariant under a general coordinate transformation. You need to know that the Christoffel symbol is *not* a tensor, but transforms as

$$\Gamma'^\alpha_{\beta\gamma}=\frac{\partial x'^\alpha}{\partial x^\delta}\frac{\partial x^\eta}{\partial x'^\beta}\frac{\partial x^\phi}{\partial x'^\gamma}\Gamma^\delta_{\eta\phi}+\frac{\partial x'^\alpha}{\partial x^\delta}\frac{\partial^2 x^\delta}{\partial x'^\beta\partial x'^\gamma}.$$

If we now introduce *the covariant derivative*

$$\nabla_\gamma A^\alpha=\frac{\partial A^\alpha}{\partial x^\gamma}+\Gamma^\alpha_{\beta\gamma}A^\beta,$$

(we will often write this as $A^\alpha_{;\gamma}$) we can show that $\nabla_\gamma A^\alpha$ transforms as a mixed rank-2 tensor. The trajectory of the particle is given by $x^\mu(\tau)$.

The tangent vector to the trajectory at a given point is given by $X^\mu = dx^\mu/d\tau$, which transforms as a contravariant vector. It is now a quite managable task (you should do it!) to show that the geodesic equation can be written as

$$X^\gamma\nabla_\gamma X^\alpha=0.$$

We will often use the notation $\nabla_\gamma X^\alpha = X^\alpha_{;\gamma}$, and $\frac{\partial X^\alpha}{\partial x^\gamma}=X^\alpha_{,\gamma}$. The geodesic equation can then be written in the (misleadingly) simple form

$$X^\gamma X^\alpha_{;\gamma}=0.$$

Written like this it is quite clear that the geodesic equation is a tensor equation, and therefore is valid in all frames of reference. Note that the equation is valid under general transformations, also pure coordinate transformations like a change from Cartesian to spherical coordinates. This makes it sometimes hard in general relativity to separate physical effects from effects caused simply by the choice of coordinates. This is a problem you will meet in the Cosmology II course in connection with cosmological perturbation theory, the so-called **gauge problem**. It will not trouble us in this course.

With hindsight we can now derive the geodesic equation in a much simpler way. In the free-fall frame the tangent vector to the particle's trajectory is given by $\Xi^\mu = d\xi^\mu/d\tau$, and the equation of motion can be written as

$$\frac{d}{d\tau}\left(\frac{d\xi^\mu}{d\tau}\right) = \frac{d}{d\tau}\Xi^\mu = 0.$$

Using the chain rule, we rewrite this as

$$\frac{d\xi^\nu}{d\tau}\frac{\partial \Xi^\mu}{\partial \xi^\nu} = \Xi^\nu \Xi^\mu_{,\nu} = 0.$$

Partial derivatives with respect to coordinates is not a tensorial operation, but covariant differentiation is, and in the free-fall frame these are the same, since all the Christoffel symbols vanish there. We can therefore write the equation as

$$\Xi^\nu \Xi^\mu_{;\nu} = 0.$$

We have now written the equation of motion as a tensor equation, and hence it is valid in all frames of reference.

Note that Γ and $g_{\mu\nu}$ are geometric quantities. Gravity is encoded in the geometry of spacetime, and has therefore become a geometric effect. There are no forces in the geodesic equation. Since both Γ and $g_{\mu\nu}$ are geometric quantities, it is perhaps not surprising that they are related. I state without proof that

$$\Gamma^\sigma_{\mu\nu} = \frac{1}{2} g^{\rho\sigma}\left(\frac{\partial g_{\nu\rho}}{\partial x^\mu} + \frac{\partial g_{\mu\rho}}{\partial x^\nu} - \frac{\partial g_{\mu\nu}}{\partial x^\rho}\right).$$

The metric is therefore an extremely important object in GR. If we know it, we know the geometry of spacetime, and the geometry of spacetime directs the motion of free particles.

I end this section with a note on covariant differentiation. For a mixed tensor, the covariant derivative is given by

$$\nabla_\gamma T^{\alpha\cdots}_{\beta\cdots} = \frac{\partial}{\partial x^\gamma} T^{\alpha\cdots}_{\beta\cdots} + \Gamma^\alpha_{\delta\gamma} T^{\delta\cdots}_{\beta\cdots} + \cdots - \Gamma^\delta_{\beta\gamma} T^{\alpha\cdots}_{\delta\cdots} - \cdots,$$

so that each contravariant index gives rise to a Christoffel symbol with positive sign, whereas each covariant index gives rise to one with negative sign. For a contravariant tensor of rank 2, the covariant derivative is

$$\nabla_\gamma T^{\mu\nu} = \frac{\partial}{\partial x^\gamma} T^{\mu\nu} + \Gamma^\mu_{\beta\gamma} T^{\beta\nu} + \Gamma^\nu_{\beta\gamma} T^{\mu\beta}.$$

The Newonian limit

We will now consider a particle moving slowly in a weak, static gravitational field. Remember that we use units where $c = 1$, so slow motion means

$$\frac{dx^i}{dt} \ll 1.$$

In the geodesic equation this means that all terms containing $dx^i/d\tau$ ($i = 1, 2, 3$) are negligible in comparison with the term containing$(dt/d\tau)^2$, We therefore get

$$\frac{d^2x^\sigma}{d\tau^2} + \Gamma^\sigma_{00} \left(\frac{dt}{d\tau}\right)^2 = 0.$$

The gravitational field is assumed to be weak, which should mean that the metric is not very different from the Minkowski metric of flat spacetime. We therefore write it as

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu},$$

der $|h_{\mu\nu}| \ll 1$. We can therefore neglect all terms containing more than one factor of $h_{\mu\nu}$. A static gravitational field means

$$\frac{\partial g_{\mu\nu}}{\partial t} = \frac{\partial h_{\mu\nu}}{\partial t} = 0.$$

The Christoffel symbol we need then becomes

$$\Gamma^\sigma_{00} = \frac{g^{\rho\sigma}}{2} \left(\frac{\partial g_{0\rho}}{\partial x^0} + \frac{\partial g_{0\rho}}{\partial x^0} - \frac{\partial g_{00}}{\partial x^\rho}\right) = -\frac{1}{2}\eta^{\rho\sigma}\frac{\partial h_{00}}{\partial x^\rho}.$$

For $\sigma = i = 1, 2, 3$ the geodetic equation becomes

$$\frac{d^2x^i}{d\tau^2} = \frac{\eta^{i\rho}}{2}\frac{\partial h_{00}}{\partial x^\rho}\left(\frac{dt}{d\tau}\right)^2 = -\frac{1}{2}\left(\frac{dt}{d\tau}\right)^2 \frac{\partial h_{00}}{\partial x^i}.$$

Furthermore, for $\sigma = 0$ we see that $\Gamma^0_{00} \propto \frac{\partial h_{00}}{\partial t} = 0$, so this component of the geodesic equation becomes

$$\frac{d^2t}{d\tau^2} = 0,$$

so that $dt/d\tau =$ konstant. We can then divide the equations for $\sigma = i$ by $(dt/d\tau)^2$ and get

$$\frac{d^2x^i}{dt^2} = -\frac{1}{2}\frac{\partial h_{00}}{\partial x^i}.$$

In Newtonian mechanics, the equation of motion for a particle in a static gravitational field is

$$\frac{d^2x^i}{dt^2} = -\frac{\partial \Psi}{\partial x^i},$$

so we see that by taking $h_{00} = 2\Psi$, the geodesic equation becomes the Newtonian equation of motion. In other words, in the Newtonian limit the 00 component of the metric must be $g_{00} = 1 + 2\Psi$.

Important tensors in GR

As said before, the Christoffel symbol is not a tensor. But we can use it to construct tensors that are related to the curvature of spacetime. The first we need to know about is the **Riemann tensor**

$$R^{\mu}_{\sigma\beta\alpha} = \Gamma^{\mu}_{\sigma\alpha,\beta} - \Gamma^{\mu}_{\sigma\beta,\alpha} + \Gamma^{\mu}_{\rho\beta}\Gamma^{\rho}_{\sigma\alpha} - \Gamma^{\mu}_{\rho\alpha}\Gamma^{\rho}_{\sigma\beta},$$

where I have used the notation

$$_{,\alpha} = \frac{\partial}{\partial x^{\alpha}}.$$

We obtain **the Ricci tensor** by contracting to indices in the Riemann tensor:

$$R_{\mu\nu} = \Gamma^{\alpha}_{\mu\nu,\alpha} - \Gamma^{\alpha}_{\mu\alpha,\nu} + \Gamma^{\alpha}_{\beta\alpha}\Gamma^{\beta}_{\mu\nu} - \Gamma^{\alpha}_{\beta\nu}\Gamma^{\beta}_{\mu\alpha}.$$

The Ricci scalar is then given by

$$\mathcal{R} = g^{\mu\nu}R_{\mu\nu}.$$

Finally, **the Einstein tensor** is defined as

$$G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}\mathcal{R}.$$

This tensor has the crucial property that its covariant divergence vanishes: $G^{\mu\nu}_{;\nu} = 0$.

The energy-momentum tensor

Let us briefly return to special relativity and consider a system of non-interacting particles with energy density described by the function $\rho(x)$. We let this function represent the energy density as measured by an observer moving with the particles, which motion is characterized by the four-velocity field

$$u^{\mu} = \frac{dx^{\mu}}{d\tau}.$$

From these quantities we can construct a contravariant tensor of rank 2:

$$T^{\mu\nu} = \rho u^{\mu}u^{\nu}.$$

In special relativity we have

$$u^{\mu} = \gamma(1, \vec{u}),$$

where $\vec{u} = d\vec{x}/dt$,

$$d\tau^2 = ds^2 = dt^2 - d\vec{x}^2 = dt^2(1 - u^2) = \frac{dt^2}{\gamma^2},$$

and $\gamma = (1 - u^2)^{-1/2}$. Written out as a matrix, T looks like this:

$$(T^{\mu\nu}) = \rho \begin{pmatrix} 1 & u_x & u_y & u_z \\ u_x & u_x^2 & u_x u_y & u_x u_z \\ u_y & u_x u_y & u_y^2 & u_y u_z \\ u_z & u_x u_z & u_y u_z & u_z^2 \end{pmatrix}$$

From fluid mechanics we know the continuity equation

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \vec{u}) = 0,$$

which expresses (local) conseravation of energy. With our definition of T, we can write this equation in a very compact form. By calculating $T^{0\nu}_{,\nu}$, we get

$$T^{0\nu}_{,\nu} = \frac{\partial T^{0\nu}}{\partial x^\nu} = \frac{\partial \rho}{\partial t} + \frac{\partial \rho u_x}{\partial x} + \frac{\partial \rho u_y}{\partial y} + \frac{\partial \rho u_z}{\partial z} = \frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \vec{u}).$$

Therefore, the continuity equation can be written as

$$T^{0\nu}_{,\nu} = 0.$$

Another important equation in hydrodynamics, the Navier-Stokes equation, says that for a fluid without internal pressure and external forces, we have

$$\rho \left[\frac{\partial \vec{u}}{\partial t} + (\vec{u} \cdot \nabla) \vec{u} \right] = 0.$$

The physical content of this equation is (local) conservation of momentum. One can show that with our choice of T, this equation can be written in component form as

$$T^{i\nu}_{,\nu} = 0,$$

for $i = 1, 2, 3$. We can therefore express conservation of energy and momentum for our system in a very elegant way as

$$T^{\mu\nu}_{,\nu} = 0.$$

Written like this, we see that the extension of this equation to general relativity should be

$$T^{\mu\nu}_{;\nu} = 0.$$

Because T summarizes the conservation of energy and momentum, it is called **the energy-momentum tensor**.

The case we are dealing with most of the time in cosmology is that of a **perfect fluid**. A perfect fluid is characterized by its energy density $\rho = \rho(x)$, its internal pressure $p = p(x)$, and its four-velocity field $u^\mu = dx^\mu / d\tau$.

When the pressure $p = 0$, we should regain the energy-momentum tensor of the previous section. This suggests that we should choose

$$T^{\mu\nu} = \rho u^\mu u^\nu + pS^{\mu\nu},$$

where $S^{\mu\nu}$ is a symmetric tensor. The only symmetric tensors of rank 2 associated with the fluid are $u^\mu u^\nu$ and $g^{\mu\nu}$, so the simplest choice is

$$S^{\mu\nu} = Au^\mu u^\nu + Bg^{\mu\nu},$$

where A and B are constants. We can determine these constants by taking the special relativistic limit. In this limit $g^{\mu\nu} = \eta^{\mu\nu}$, the Minkowski metric. We want, once again, $T^{\mu\nu}_{,\nu} = 0$ to express conservation of energy and momentum. By demanding that $T^{0\nu}_{,\nu} = 0$ should give us the continuity equation, and that $T^{i\nu}_{,\nu} = 0$ should give us the Navier-Stokes equation (this time with a term from internal pressure), we find (not shown here, but it is, with the famous phrase, 'straightforward but tedious') $A = 1$, $B = -1$. We therefore take

$$T^{\mu\nu} = (\rho + p)u^\mu u^\nu - pg^{\mu\nu},$$

as the energy-momentum tensor of a perfect fluid. It has, by construction, vanishing covariant divergence: $T^{\mu\nu}_{;\nu} = 0$.

The Einstein equation

We still lack a crucial ingredient: an equation that relates the geometry of spacetime to the distribution of mass and energy. This is analogous to Newtonian gravity, where Newton's law of gravity tells us the field produced by a given set of masses.

The natural point to start with is the tensor that tells us about mass and energy, $T_{\mu\nu}$. Somehow, we must relate it to a geometrical quantity. We note that T has vanishing covariant divergence, and recall that the same is the case for the purely geometric tensor $G_{\mu\nu}$, the Einstein tensor. The simplest guess we can make for the field equation is therefore that these two are proportional. The constant of proportionality can be determined by requiring that we regain the Newtonian limit when velocities are small and the gravitational field is weak and static. The result is

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

This is the Einstein equation, one of the highlighs of human intellectual history (and I say this without any irony). If you want to use units where c is not equal to one, the constant of proportionality is $8\pi G/c^4$.

Note that this equation is a *guess*. It cannot be derived stringently from the postulates of the theory. It is merely the simplest equation that we can write down. If you wish to (and in recent years several people have), you can write down more complicated field equations. The validation of the Einstein equation must come from comparing its predictions with observations. So far, there are no signs that we need a more complicated equation.

7.3 Spacetime curvature or spatial curvature, which is more important?

In popular accounts of GR, a common way to illustrate the way gravity is understood as geometry is the (in)famous ‘bowling ball on a matress’. The surface of the matress represents space, and the bowling ball represents, e.g., the Sun or Earth. The bowling ball makes a dent in the matress, and this represents spatial curvature. Marbles flung across the matress on paths which take them close to the bowling ball, will be deflected. This is taken to be an anologue to how spacetime curvature causes objects to move on curved paths.

There are obvious problems with this analogy. For example, space is three-dimensional, not two-dimensional like the surface of the matress. What happens to space above the bowling ball? If you put another mattress on top of the bowling ball, it would curve the opposite way, and it would seem that gravity above the ball is repulsive! This is clearly not the case with the gravitational fields of the Sun and Earth.

A less obvious, but deeper problem with this analogy is that for particles moving at speeds significantly below the speed of light, spatial curvature is not the main cause of gravity. The main reason why, for example, a rock falls to the ground when dropped, is how Earth’s mass causes clocks to tick at different rates at different heights. To show this, we will follow a paper by R. R. Gould in the American Journal of Physics, volume 84 (2016), page 396.

Let us look at spacetime outside Earth. We consider our planet to be a perfect sphere, and neglect its rotation. In AST1100 (now AST2000)you learnt what spacetime looks like outside a non-rotating, spherical mass distribution. The geometry is given by the Schwarzschild line element

$$ds^2 = \left(1 - \frac{2GM}{rc^2}\right) c^2 dt^2 - \frac{dr^2}{1 - \frac{2GM}{rc^2}}.$$

Note that I have left out the angular part of the line element since we will only be looking at radial motion in this section. In AST1100 you used this line element mostly to study black holes, but it is valid outside *all* spherical mass distributions.

Let us consider the situation where you hold a rock at waist height, release it and let if fall to the ground. The quantitative aspects of this case is perfectly described by Newtonian physics. If the rock starts at height h_0, it will hit the ground after a time

$$t_{\mathrm{f}} = \sqrt{\frac{2h_0}{g}},$$

where $g = 9.8\ \mathrm{ms}^{-2}$. Taking $h_0 = 1$ m, we get $t_{\mathrm{f}} = 0.45$ s. Let us turn to how the situation is described in GR.

We start by simplifying the line element. Let R be the distance from Earth's center to waist height, and let x be the distance the rock has fallen below waist height since you dropped it. The coordinate r in the line element is related to them by $r = R - x$. Now, R is for all pracitical purposes equal to the radius of Earth, so clearly $x \ll R$. We can use this fact to simplify the line element by expanding it to first order in x/R:

$$\begin{aligned} 1 - \frac{2GM}{rc^2} &= 1 - \frac{2GM}{c^2}\frac{1}{R - x} \\ &= 1 - \frac{2GM}{Rc^2}\frac{1}{1 - \frac{x}{R}} \\ &\approx 1 - \frac{2GM}{Rc^2}\left(1 + \frac{x}{R}\right) \\ &= 1 - \frac{2GM}{Rc^2} - \frac{2GM}{R^2c^2}x \\ &\approx 1 - \frac{2GM}{R^2c^2}x, \end{aligned}$$

where in the last line we have discarded a constant term because it is very, very much smaller than 1: R is essentially Earth's radius, whereas $2GM/c^2$ is Earth's Schwartzschild radius, which is of the order of 1 cm.

The same manipulations lead to

$$\frac{1}{1 - \frac{2GM}{rc^2}} \approx \frac{1}{1 - \frac{2GM}{R^2c^2}x} \approx 1 + \frac{2GM}{R^2c^2}x.$$

Clearly $dr = -dx$, so we get

$$ds^2 \approx \left(1 - \frac{2GM}{R^2c^2}x\right)c^2dt^2 - \left(1 + \frac{2GM}{R^2c^2}x\right)dx^2.$$

Now x varies between 0 (waist height) and 1 m (when the rock hits the ground). The spatial part of the line element, which measures how the length of measuring rods change, changes in the course of the fall by

$$\sqrt{\frac{2GM}{R^2c^2}xdx^2} \approx 10^{-8}\ \mathrm{m}.$$

The temporal part of the line element measures of the rates of clock vary. The fall is over in a time $dt = 0.45$ s, and in this time the change in the temporal part is

$$\sqrt{\frac{2GM}{R^2c^2}}cdt \approx 1.5 \text{ m}.$$

In other words, the warping of time changes the metric much more than the curvature of space. In the 0.45 s it takes the stone to fall 1 meter, it explores only 1 meter of spatial curvature, but $ct_{\mathrm{f}} = 1.5 \times 10^8$ meters of time warp. The latter is therefore much more influential in determining the path of the rock, and only for particles moving at, or close to the speed of light, will spatial curvature be explored to an extent that it matters.

A further indication of how spatial curvature alone is not enough to cause gravitatinal effects is described in another article in the American Journal of Physics, volume 84 (2016) page 588. I will briefly describe its main argument.

Consider the static line element

$$ds^2 = c^2dt^2 + g_{ij}dx^idx^j,$$

where the g_{ij} are functions of spatial coordinates alone, and not of time. Furthermore, we also assume that $g_{ij} = 0$ when $i \neq j$. We now go on to calculate the Christoffel symbols:

$$\begin{aligned} \Gamma^0_{\alpha\beta} &= \frac{1}{2}g^{0\nu}(g_{\alpha\nu,\beta} + g_{\beta\nu,\alpha} - g_{\alpha\beta,\nu}) \\ &= \frac{1}{2}g^{00}(g_{\alpha0,\beta} + g_{\beta0,\alpha} - g_{\alpha\beta,0}) \\ &= 0. \end{aligned}$$

$$\begin{aligned} \Gamma^i_{\alpha\beta} &= \frac{1}{2}g^{i\nu}(g_{\alpha\nu,\beta} + g_{\beta\nu,\alpha} - g_{\alpha\beta,\nu}) \\ &= \frac{1}{2}g^{ii}(g_{\alpha i,\beta} + g_{\beta i,\alpha} - g_{\alpha\beta,i}). \end{aligned}$$

From this we see that

$$\begin{aligned} \Gamma^i_{0\beta} &= \frac{1}{2}g^{ii}(g_{0i,\beta} + g_{\beta i,0} - g_{0\beta,i}) \\ &= 0 = \Gamma^i_{\beta0}, \end{aligned}$$

so the only non-vanishing Christoffel symbols are those with all the indices spatial:

$$\Gamma^i_{jk} = \frac{1}{2}g^{ii}(g_{ji,k} + g_{ki,j} - g_{jk,i}).$$

We note that this is exactly the same as for the purely spatial metric of the 3d space described by $dl^2 = g_{ij}dx^idx^j$.

The motion of particles in this spacetime is determined by the geodesic equation

$$\frac{d^2x^\sigma}{d\tau^2} + \Gamma^\sigma_{\nu\rho}\frac{dx^\nu}{d\tau}\frac{dx^\rho}{d\tau} = 0.$$

For $\sigma = 0$, we get

$$\frac{d^2t}{d\tau^2} + \Gamma^0_{\nu\rho}\frac{dx^\nu}{d\tau}\frac{dx^\rho}{d\tau} = 0.$$

Since $\Gamma^0_{\nu\rho} = 0$ for all ν and ρ, this equation becomes simply

$$\frac{d^2t}{d\tau^2} = 0,$$

which means that $t = a\tau + b$, and we can choose the constants $a = 1$, $b = 0$ so that $t = \tau$.

The spatial components of the geodesic equation are

$$\frac{d^2x^i}{d\tau^2} + \Gamma^i_{\nu\rho}\frac{dx^\nu}{d\tau}\frac{dx^\rho}{d\tau} = 0.$$

Since $\Gamma^i_{\nu\rho} \neq 0$ only when ν and ρ both are spatial indices, we have

$$\frac{d^2x^i}{d\tau^2} + \Gamma^i_{jk}\frac{dx^i}{d\tau}\frac{dx^j}{d\tau} = 0.$$

We now see that x^i = constant is a solution. This means that a particle which starts at rest, will remain at rest. If you drop a rock in this spacetime, it will not fall to the ground. In this sense, gravity is absent in this spacetime where all the curvature comes from the spatial part of the metric.

This is all well and good, but are there any examples of actual solutions of Einstein's field equation which have this property? As it turns out, there is: Einstein's first attempt at constructing a model for the Universe, published in 1917 in his first paper on cosmology. The model has the line element

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

It is static, and the spatial geometry is that of the surface of a 4-dimensional sphere. The model is unrealistic, because we know that the Universe is *not* static, it is expanding. However, it is amusing that the inventor of our best theory of how gravity works, arrived at a model of the Universe where particles don't feel gravity. This model is, essentially, special relativity on the surface of a 4-dimensional sphere.

7.4 The Friedmann equations

We have now reached the goal of this chapter. As I may or may not have said before, I have not dragged you through this basic introduction to GR to give you anything resembling a deep understading of it (I am far too shallow for that). If the previous sections have left you in a confused state, don't worry: This is my fault, and you won't be asked about GR on your final exam. I simply want to show you where the Friedmann equations come from. Now you know just enough to follow the proper, general relativistic derivation. Once you have seen it, you should feel free to forget about it.

The Friedmann equations follow from the Einstein equation in the special case where the metric is given by the Robertson-Walker line element for a homogeneous and isotropic universe, and the source for spacetime curvature is a perfect fluid. Our task in the following is to set up both sides of the equation. There is nothing particularly difficult about this, conceptually or technically. It is just a long calculation. Most of the time we are simply taking derivatives and multiplying and adding stuff. However, there are quite a lot of operations like this to be carried out, and one absent-minded mistake in one place will propagate through the whole calculation, leading to an erroneous result. When I did the calculation when preparing these notes, I had to work through it three times before I got the right result. I say this to comfort you. Later in life you may have to do similar calculations in cases where you don't already know what the final result should be. Trust me, unless you are having an especially good day, you probably won't get it right the first time. You should do the calculation at least twice. If your results agree, then you may be on to something, but redo it once more, just to be safe. Or, even better, get a friend to make the same calculation. If hers or his results agree with yours, then you are reasonably safe. Of course, if you are a complete chicken, you can use one of several Mathematica packages for tensor manipulations which will do the whole thing for you. This will only take a few seconds, and in the same time you will have lost my respect, thus accomplishing two major tasks at once. Frankly, in my opinion, you should only leave your mathematics to a computer if you are prepared to let it do your drinking and fornication for you, too.

So, let us start by writing down the Robertson-Walker line element. To save a little bit of writing, I will follow standard practice among fans of GR and use units where $c = 1$. Trust me, it is OK, and I will reinstate it at the end of the calculation. With this convention we can write

$$ds^2 = dt^2 - a^2(t)\left(\frac{dr^2}{1-kr^2} + r^2 d\theta^2 + r^2 \sin^2\theta d\phi^2\right) \tag{7.1}$$

In what follows, indices will represent coordinates with the following pairings: $x^0 = t$, $x^1 = r$, $x^2 = \theta$, $x^3 = \phi$. The relationship between the line element and the metric is $ds^2 = g_{\mu\nu}dx^\mu dx^\nu$, so we note with satisfaction that the metric is diagonal and read off the following components:

$$g_{\mu\nu} = \mathrm{diag}\left(1, -\frac{a^2}{1-kr^2}, -a^2r^2, -a^2r^2\sin^2\theta\right), \tag{7.2}$$

where I have saved some more writing by not writing the time-dependence of a explicitly. Please remember that it is still there. Since the metric is diagonal, finding its 'inverse' $g^{\mu\nu}$ is a doddle:

$$g^{\mu\nu} = \mathrm{diag}\left(1, -\frac{1-kr^2}{a^2}, -\frac{1}{a^2r^2}, -\frac{1}{a^2r^2\sin^2\theta}\right). \tag{7.3}$$

We now have to calculate the Christoffel symbols, then from them we can go on to find the Ricci tensor, the Ricci scalar, and finally the Einstein tensor.

The Christoffel symbols are given by

$$\Gamma^{\mu}_{\alpha\beta} = \frac{1}{2}g^{\mu\nu}\left(g_{\alpha\nu,\beta} + g_{\beta\nu,\alpha} - g_{\alpha\beta,\nu}\right). \tag{7.4}$$

I am not going to do all of them, only a few so you can see how it can be done, and then I will just give the final result for the rest. I start with the case $\mu = 0$:

$$\begin{aligned}\Gamma^{0}_{\alpha\beta} &= \frac{1}{2}g^{0\nu}\left(g_{\alpha\nu,\beta} + g_{\beta\nu,\alpha} - g_{\alpha\beta,\nu}\right)\\ &= \frac{1}{2}g^{00}\left(g_{\alpha0,\beta} + g_{\beta0,\alpha} - g_{\alpha\beta,0}\right)\\ &= \frac{1}{2}\left(g_{00,\beta}\delta_{\alpha0} + g_{00,\alpha}\delta_{\beta0} - g_{\alpha\beta,0}\right)\\ &= -\frac{1}{2}g_{\alpha\beta,0}.\end{aligned} \tag{7.5}$$

A brief explanation of what happened here:

- In the first line, I have just substituted $\mu = 0$ in the general expression (7.4).
- Since $g^{\mu\nu}$ is diagonal, the only term in the sum over ν that contributes, is when $\nu = 0$.
- I have inserted $g^{00} = 1$. Also, $g_{\mu\nu}$ is diagonal, so in the first term $\alpha = 0$, while in the second $\beta = 0$. I have emphasized this fact by also inserting Kronecker deltas.
- Since $g_{00} = 1$ is a constant, taking derivatives with respect to any coordinate will give zero as a result. The only term that survives is therefor the last.

Since $g_{00} = 1$, this result implies that $\Gamma^0_{00} = -\frac{1}{2}g_{00,0} = 0$, and since the metric is diagonal, we have $\Gamma^0_{0i} = \Gamma^0_{i0} = 0$, and $\Gamma^0_{ij} = \Gamma^0_{ji} = 0$ when $i \neq j$. The non-zero Christoffel symbols with upper index = 0 are

$$\begin{aligned}\Gamma^{0}_{11} &= -\frac{1}{2}g_{11,0} = -\frac{1}{2}\frac{\partial}{\partial t}\left(-\frac{a^2}{1-kr^2}\right)\\ &= \frac{a\dot{a}}{1-kr^2}\end{aligned} \tag{7.6}$$

$$\Gamma^0_{22} = -\frac{1}{2}\frac{\partial}{\partial t}(-a^2r^2) = r^2 a\dot{a} \tag{7.7}$$

$$\Gamma^0_{33} = -\frac{1}{2}\frac{\partial}{\partial t}(-a^2r^2\sin^2\theta) = r^2\sin^2\theta a\dot{a}. \tag{7.8}$$

I hope it is clear by now how to proceed. None of the operations are very difficult, but you have to keep track of the indices and make sure that you cover all cases. The burden is somewhat lessened by the fact that the Christoffel symbols are symmetric in the lower indices: $\Gamma^\mu_{\alpha\beta} = \Gamma^\mu_{\beta\alpha}$. The remaining set of non-zero Γs is

$$\Gamma^0_{11} = \frac{a\dot{a}}{1-kr^2},\ \Gamma^0_{22} = r^2 a\dot{a},\ \Gamma^0_{33} = r^2\sin^2\theta a\dot{a} \tag{7.9}$$

$$\Gamma^1_{01} = \Gamma^2_{02} = \Gamma^3_{03} = \frac{\dot{a}}{a} \tag{7.10}$$

$$\Gamma^1_{11} = \frac{kr}{1-kr^2},\ \Gamma^1_{22} = -r(1-kr^2),\ \Gamma^1_{33} = -r(1-kr^2)\sin^2\theta \tag{7.11}$$

$$\Gamma^2_{12} = \Gamma^3_{13} = \frac{1}{r},\ \Gamma^2_{33} = -\sin\theta\cos\theta,\ \Gamma^3_{23} = \cot\theta. \tag{7.12}$$

The next task is to calculate the Ricci tensor $R_{\mu\nu} = R_{\nu\mu}$. It is given in terms of the Christoffel symbols as

$$R_{\mu\nu} = \Gamma^\alpha_{\mu\nu,\alpha} - \Gamma^\alpha_{\mu\alpha,\nu} + \Gamma^\alpha_{\beta\alpha}\Gamma^\beta_{\mu\nu} - \Gamma^\alpha_{\beta\nu}\Gamma^\beta_{\mu\alpha}. \tag{7.13}$$

People with minds greater than my own often find elegant ways of calculating the components of the Ricci tensor. I usually have to resort to the painful method of calculating it term by term, component by component. I will show one example:

$$R_{00} = \Gamma^\alpha_{00,\alpha} - \Gamma^\alpha_{0\alpha,0} + \Gamma^\alpha_{\beta\alpha}\Gamma^\beta_{00} - \Gamma^\alpha_{\beta 0}\Gamma^\beta_{0\alpha}. \tag{7.14}$$

The first term vanishes, since all the Christoffel symbols involved are equal to zero. In the third term, $\Gamma^\beta_{00} = 0$ for all β, so this term also vanishes. For the two remaining terms, I find

$$\begin{aligned} -\Gamma^\alpha_{0\alpha,0} &= -\Gamma^0_{00,0} - \Gamma^1_{01,0} - \Gamma^2_{02,0} - \Gamma^3_{03,0} \\ &= -3\frac{\partial}{\partial t}\frac{\dot{a}}{a} = -3\frac{\ddot{a}a - \dot{a}^2}{a^2} \\ &= 3\left(\frac{\dot{a}}{a}\right)^2 - 3\frac{\ddot{a}}{a}, \end{aligned} \tag{7.15}$$

and

$$\begin{aligned} -\Gamma^\alpha_{\beta 0}\Gamma^\beta_{0\alpha} &= -\Gamma^0_{\beta 0}\Gamma^\beta_{00} - \Gamma^i_{\beta 0}\Gamma^\beta_{0i} \\ &= -\Gamma^i_{00}\Gamma^0_{0i} - \Gamma^i_{j0}\Gamma^j_{0i} \\ &= -\sum_i\sum_j \Gamma^i_{0i}\delta_{ij}\Gamma^i_{0i}\delta_{ij} \\ &= -\sum_i\left(\Gamma^i_{0i}\right)^2 = -3\left(\frac{\dot{a}}{a}\right)^2. \end{aligned} \tag{7.16}$$

In the first and second lines I have used that $\Gamma^{\beta}_{00} = 0$ for all β. Adding up all the terms I get

$$R_{00} = -3\frac{\ddot{a}}{a}. \tag{7.17}$$

The rest involve similarly inspiring calculations. To retain whatever sanity one might have, it is useful to recall that $R_{\mu\nu} = R_{\nu\mu}$, so that, once you have found that, e.g., $R_{12} = 0$, you can spare yourself the trouble of calculating R_{21}. Here are the final results:

$$\begin{aligned}
R_{\mu\nu} &= 0, \text{ if } \mu \neq \nu && (7.18)\\
R_{00} &= -3\frac{\ddot{a}}{a} && (7.19)\\
R_{11} &= \frac{2\dot{a}^2 + a\ddot{a} + 2k}{1 - kr^2} && (7.20)\\
R_{22} &= r^2(2\dot{a}^2 + a\ddot{a} + 2k) && (7.21)\\
R_{33} &= r^2 \sin^2\theta(2\dot{a}^2 + a\ddot{a} + 2k) && (7.22)
\end{aligned}$$

We also need the Ricci scalar,

$$R = g^{\mu\nu}R_{\mu\nu} = g^{00}R_{00} + g^{11}R_{11} + g^{22}R_{22} + g^{33}R_{33}. \tag{7.23}$$

Compared to the preceding calculations, doing this sum is a doddle, and I find:

$$R = -6\left[\left(\frac{\dot{a}}{a}\right)^2 + \frac{\ddot{a}}{a} + \frac{k}{a^2}\right]. \tag{7.24}$$

Now we have all we need to set up the left-hand side of the Einstein equation. The right-hand side is proportional to the energy-momentum tensor, and I will assume that it can be approximated by a perfect fluid. Recall that, in units where $c = 1$, the energy-momentum tensor of a perfect fluid is given by

$$T^{\mu\nu} = (\rho + p)u^{\mu}u^{\nu} - pg^{\mu\nu}, \tag{7.25}$$

where u^{μ} is the four-velocity of the fluid. The coordinates we are using are co-moving: The observer is in a reference frame which moves along with the fluid. In other words, the fluid is at rest relative to the observer. In that case, the spatial part of the four-velocity vanishes, and we have

$$u^{\mu} = (1, 0, 0, 0) = \delta^{\mu}_0, \tag{7.26}$$

and

$$u_{\mu} = g_{\mu\nu}u^{\nu} = g_{\mu\nu}\delta^{\mu}_0 = g_{\mu 0} = \delta^0_{\mu}, \tag{7.27}$$

so I get

$$T_{\mu\nu} = (\rho + p)\delta^0_{\mu}\delta^0_{\nu} - pg_{\mu\nu}. \tag{7.28}$$

The Einstein equation (with $c = 1$) was

$$G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = 8\pi G T_{\mu\nu}. \tag{7.29}$$

I am now going to add an extra term to this equation. You may recall that the form of equation (7.29) was motivated by the fact that both the left-hand side and the right-hand side have vanishing covariant divergence. It turns out that there is an extra term we can add to the left-hand side without screwing up this property. It is namely generally true that $\nabla^\mu g_{\mu\nu} = 0$, so we can add a term proportional to $g_{\mu\nu}$ to the left-hand side. Let us agree to name the constant of proportionality $\Lambda(\Lambda$ can be either positive or negative). Our complete equation is therefore

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R + \Lambda g_{\mu\nu} = 8\pi G T_{\mu\nu}. \tag{7.30}$$

Let us see what the Einstein equation can give us. I start with the case $\mu = \nu = 0$. Then

$$\begin{aligned} R_{00} - \frac{1}{2} g_{00} R + \Lambda g_{00} &= -3\frac{\ddot{a}}{a} - \frac{1}{2}(-6)\left[\left(\frac{\dot{a}}{a}\right)^2 + \frac{\ddot{a}}{a} + \frac{k}{a^2}\right] + \Lambda \\ &= 3\left(\frac{\dot{a}}{a}\right)^2 + \frac{3k}{a^2} + \Lambda, \end{aligned} \tag{7.31}$$

and

$$T_{00} = \rho + p - p = \rho, \tag{7.32}$$

so, after equating (7.31 and (7.32) and dividing through by 3, I get

$$\left(\frac{\dot{a}}{a}\right)^2 + \frac{k}{a^2} + \frac{1}{3}\Lambda = \frac{8\pi G}{3}\rho. \tag{7.33}$$

This was fun, so I will do one more: $\mu = \nu = 1$. This turns out to be a little bit messier:

$$\begin{aligned} R_{11} - \frac{1}{2} g_{11} R + \Lambda g_{11} &= \frac{2\dot{a}^2 + a\ddot{a} + 2k}{1 - kr^2} - \frac{1}{2}\frac{-a^2}{1-kr^2}(-6)\left[\left(\frac{\dot{a}}{a}\right)^2 + \frac{\ddot{a}}{a} + \frac{k}{a^2}\right] + \Lambda\frac{-a^2}{1-kr^2} \\ &= -\frac{a^2}{1-kr^2}\left[\left(\frac{\dot{a}}{a}\right)^2 + 2\frac{\ddot{a}}{a} + \frac{k}{a^2} + \Lambda\right], \end{aligned} \tag{7.34}$$

and

$$T_{11} = 0 - p g_{11} = \frac{a^2}{1-kr^2} p, \tag{7.35}$$

so after plugging these into the Einstein equation and multiplying through by $-(1 - kr^2)$, I get

$$\left(\frac{\dot{a}}{a}\right)^2 + 2\frac{\ddot{a}}{a} + \frac{k}{a^2} + \Lambda = -8\pi G p. \tag{7.36}$$

Now this equation involves both $\dot{a}$ and$\ddot{a}$, whereas the equation (7.33) just involved $\dot{a}$. I would like to get an equation which just contains $\ddot{a}$, and I can get what I want by subtracting (7.33) from (7.36). This leaves me with

$$2\frac{\ddot{a}}{a}+\frac{2}{3}\Lambda=-\frac{8\pi G}{3}(\rho+3p), \tag{7.37}$$

so, finally,

$$\frac{\ddot{a}}{a}+\frac{1}{3}\Lambda=-\frac{4\pi G}{3}(\rho+3p). \tag{7.38}$$

Equations (7.33) and (7.38) are differential equations for the scale factor a. We could try to get more equations by plugging different values of μ and ν into the Einstein equation, but it turns out that the results will always be some linear combination of the equations we already have. They are called *the Friedmann equations* after the Russian mathematician Alexander Friedmann (1888-1925) who in two papers, published in 1922 and 1924, both derived these equations and used them, for the first time, to study dynamical models of the Universe. Sadly, he didn't live long enough to see his work recieve the recognition it deserved.

The Friedmann equations contain ρ and p, and if we want to find a, we need to know how these evolve. We have assumed spatial homogeneity and isotropy, so the pressure and density cannot vary in space, but they may still be functions of time. As it turns out, we can get an equation for their time evolution from the requirement $\nabla_\nu T^{\mu\nu}=0$. I will consider the case $\mu=0$:

$$\begin{aligned}
\nabla_\nu T^{0\nu} &= \partial_\nu T^{0\nu}+\Gamma^0_{\beta\nu}T^{\beta\nu}+\Gamma^\nu_{\beta\nu}T^{0\beta}\\
&= \partial_0 T^{00}+\Gamma^0_{\nu\nu}T^{\nu\nu}+\Gamma^\nu_{0\nu}T^{00}\\
&= \frac{\partial\rho}{\partial t}+\Gamma^0_{11}T^{11}+\Gamma^0_{22}T^{22}+\Gamma^0_{33}T^{33}\\
&+ \Gamma^1_{01}T^{00}+\Gamma^2_{02}T^{00}+\Gamma^3_{03}T^{00}\\
&= \dot{\rho}+\frac{a\dot{a}}{1-kr^2}p\frac{1-kr^2}{a^2}+r^2a\dot{a}p\frac{1}{a^2r^2}+r^2\sin^2\theta p\frac{1}{a^2r^2\sin^2\theta}+3\frac{\dot{a}}{a}\rho\\
&= \dot{\rho}+3\frac{\dot{a}}{a}(\rho+p), \qquad (7.39)
\end{aligned}$$

which gives

$$\dot{\rho}+3\frac{\dot{a}}{a}(\rho+p)=0. \tag{7.40}$$

It turns out that this equation can also be derived from the two Friedmann equations, so it is not a new, independent equation. Deriving it from the vanishing of the covariant divergence of the energy-momentum tensor makes its physical meaning clear: The equation expresses the local conservation of energy.

To sum up this section, I give you again the three important equations we have derived, but now with the cs reinstated:

$$
\begin{aligned}
\left(\frac{\dot a}{a}\right)^2 + \frac{kc^2}{a^2} + \frac{1}{3}\Lambda c^2 &= \frac{8\pi G}{3}\rho && (7.41)\\
\frac{\ddot a}{a} + \frac{1}{3}\Lambda c^2 &= -\frac{4\pi G}{3}\left(\rho + 3\frac{p}{c^2}\right) && (7.42)\\
\dot\rho + 3\frac{\dot a}{a}\left(\rho + \frac{p}{c^2}\right) &= 0. && (7.43)
\end{aligned}
$$

Chapter 8

Bonus material: The cosmological constant

The expansion of the Universe seems to be accelerating. We have seen that the cosmological constant can describe this type of evolution. All observations are consistent with a model where the geometry is described by the RW line element, and where the Universe is currently dominated by non-relativistic matter (about 32 % of the energy density) and the cosmological constant (about 68 % of the energy density.) So what is the problem? This is topic of this chapter which is largely based upon a review article by J. Martin, available at arxiv.org, article number 1205.3365.

8.1 The cosmological constant problem formulated

The Einstein equation with the cosmological constant added is

$$G_{\mu\nu} + \Lambda g_{\mu\nu} = 8\pi G T_{\mu\nu} \tag{8.1}$$

where I use units where $c = 1$. Einstein thought the Λ term ruined the beauty of his original equation. In fact, he would have been wrong not to include it. The Einstein equation can be derived from a so-called action principle: The field equation above follows from demanding that a quantity known as the action should be stationary under small variations of the the metric $g_{\mu\nu}$. When writing down the action, all terms that are consistent with the underlying principles and symmetries of the theory should be included. Equation (8.1) follows from this procedure when one restricts oneself to actions which contain no higher than second-order derivatives of the metric.

All the contributions to the energy-momentum tensor on the right-hand side of equation (8.1) are, at the most fundamental level, described by quantum fields. Dark matter, radiation and baryonic matter are all states of their respective underlying fields. The lowest energy state of a quantum field is called the vacuum. It is a state where no particles are present. Quantum field theory leads, as we will see, to the conclusion that this ground state energy is not necessarily zero. The vacuum may have net energy. This vacuum should be Lorentz invariant: You should not be able to detect the vacuum by moving through it. The only tensor we have available in flat spacetime to construct the energy-momentum tensor of the vacuum, is the metric tensor $\eta_{\mu\nu}$. Thus $T^{\text{vac}}_{\mu\nu} = \rho_v \eta_{\mu\nu}$. The generalisation to curved spacetime is found by replacing η with $g_{\mu\nu}$, so

$$T^{\text{vac}}_{\mu\nu} = \rho_v g_{\mu\nu}. \tag{8.2}$$

Treating the vacuum energy separately from the other contributions to the energy-momentum tensor, I can write the Einstein equation as

$$G_{\mu\nu} + \Lambda g_{\mu\nu} = 8\pi G T_{\mu\nu} + 8\pi G \rho_v g_{\mu\nu}, \tag{8.3}$$

and I can either move the Λ term to the right-hand-side to get

$$G_{\mu\nu} = 8\pi G T_{\mu\nu} + 8\pi G \left(\rho_v - \frac{\Lambda}{8\pi G}\right) g_{\mu\nu}, \tag{8.4}$$

or the vacuum energy term to the left-hand-side and find

$$G_{\mu\nu} + (\Lambda - 8\pi G \rho_v) g_{\mu\nu} = 8\pi G T_{\mu\nu}. \tag{8.5}$$

This means that I can define an effective vacuum energy as

$$\rho_v^{\text{eff}} = \rho_v - \frac{\Lambda}{8\pi G}, \tag{8.6}$$

or an effective cosmological constant as

$$\Lambda^{\text{eff}} = \Lambda - 8\pi G \rho_v. \tag{8.7}$$

I will try to stick to the former, but the cosmological constant problem is the same either way.

Our observations are best fitted by an effective vacuum energy density amounting to about 70 % of the critical energy density. The critical energy density presently has the value

$$\rho_{c0} = 1.054 \times 10^{-5}\, h^2 \text{ GeV cm}^{-3}, \tag{8.8}$$

where h is the dimensionless Hubble constant. As we will see in the following, the expectation from quantum field theory is that the vacuum energy is many, many order of magnitudes greater than the critical density. We can get the correct value by adjusting the value of the cosmological constant to cancel most of this contribution to the effective vacuum energy density, because there is no known theoretical prediction for Λ, but this creates the cosmological constant problem:

- Why is Λ so finely adjusted to cancel (almost all of) the vacuum energy density?

This is a real conundrum, because there is no obvious reason why these two numbers should be so closely related.

In the popular (and even in the professional) literature you may have seen the vacuum energy density claimed to be ~ 120 orders of magnitude greater than the critical energy density. This corresponds to calculating the vacuum energy in a way which we will see is wrong. When done correctly, the mismatch is about 55 orders of magnitude, which is still a huge mismatch. This means that Λ must be equal to ρ_v to 55 decimal places, a strange coincidence indeed.

Many attempts have been and are being made to understand this problem better and solve it, but the starting point has to be to learn how the vacuum energy in quantum field theory is calculated.

8.2 The vacuum energy density of a scalar field

The simplest case we can consider is that of a scalar field ϕ, whose quanta are spin-0 particles. A free scalar field with mass m has the potential energy density

$$V(\phi) = \frac{1}{2}m^2\phi^2, \tag{8.9}$$

(I am now using units where both $\hbar$ and c are equal to 1) and the field is a solution of the so-called Klein-Gordon equation

$$\ddot{\phi} - \nabla^2\phi + m^2\phi = 0, \tag{8.10}$$

where dots over a quantity denotes derivatives with respect to time. Note that I am working in flat spacetime here and in the following. The solution of this equation can be written as a Fourier transform

$$\phi(t, \mathbf{x}) = \frac{1}{(2\pi)^{3/2}} \int \frac{d^3k}{\sqrt{2\omega(k)}} \left[c_{\mathbf{k}} e^{-i\omega t + i\mathbf{k}\cdot\mathbf{x}} + c^*_{\mathbf{k}} e^{i\omega t - i\mathbf{k}\cdot\mathbf{x}} \right], \tag{8.11}$$

where * denotes complex conjugation, and $\omega(k) = \sqrt{k^2 + m^2}$. This is all at the classical level. Quantisation proceeds by promoting the Fourier coefficients $c_{\mathbf{k}}$ and $c^*_{\mathbf{k}}$ to annihilation and creation operators with the commutator relation

$$[c_{\mathbf{k}}, c^\dagger_{\mathbf{k}'}] = \delta^{(3)}(\mathbf{k} - \mathbf{k}'), \tag{8.12}$$

where $\delta^{(3)}$ is the delta function in three dimensions. The annihilation and creation operators work in the same way as for a harmonic oscillator. Analogous to this more familiar system it can be shown that there is a lowest-energy state $|0\rangle$ corresponding to no quanta being present, and that $c_{\mathbf{k}}|0\rangle = 0$ for any $\mathbf{k}$.

Quantum field theory also provides an expression for the energy-momentum tensor of the field, from which the expressions for the vacuum energy density and the pressure can be read off:

$$\begin{aligned} \rho_v &= \frac{1}{(2\pi)^3}\frac{1}{2}\int d^3k\omega(k) & (8.13) \\ p_v &= \frac{1}{(2\pi)^3}\frac{1}{6}\int d^3k\frac{k^2}{\omega(k)}. & (8.14) \end{aligned}$$

Let me start by calculating the vacuum energy density:

$$\rho_v = \frac{1}{16\pi^3}\int_0^\infty 4\pi k^2 dk\sqrt{k^2+m^2} = \frac{1}{4\pi^2}\int_0^\infty dkk^2\sqrt{k^2+m^2}. \qquad (8.15)$$

This integral is badly divergent. For the case of a massless scalar field it diverges like k^4. So formally the vacuum energy is infinite. In situations where gravity is unimportant, that is to say in all particle physics experiments to date, this is not a problem. All that matters in these situations are energies relative to the vacuum, and they are well-defined. But when gravity comes into play, the situation is different. At face value, the message of the Einstein equation is that *all* sources of energy contribute on the right-hand side of it, so having an infinite contribution from the vacuum is therefore a highly non-negligible problem.

What can be done? A common argument is as follows: In equation (8.15) I allow the momentum/frequency to take on arbitrarily high values. This means that I trust my description of the field up to infinite energy. But we have reason to suspect that the description should break down before that. At the very least, we know that at some energy quantum gravity should come into play, and my description of the field in flat spacetime is invalid. With that in mind, I introduce a cut-off in the integral at the mass/energy scale M where my theory breaks down:

$$\rho_v = \frac{1}{4\pi^2}\int_0^M dkk^2\sqrt{k^2+m^2}. \qquad (8.16)$$

I introduce the dimensionless variable $x = k/m$ and get

$$\rho_v = \frac{m^4}{4\pi^2}\int_0^{M/m} dxx^2\sqrt{x^2+1}. \qquad (8.17)$$

The integral is of a form which invites a hyperbolic substitution. With $x = \sinh t$, $dx = \cosh tdt$ and $x^2 + 1 = \sinh^2 t + 1 = \cosh^2 t$, I find

$$\begin{aligned} \int x^2\sqrt{x^2+1}dx &= \int(\sinh t\cosh t)^2 dt \\ &= \int\left(\frac{1}{2}\sinh 2t\right)^2 dt = \frac{1}{4}\int\sinh^2 2tdt \\ &= \frac{1}{8}\int(\cosh 4t - 1)dt = \frac{1}{32}\sinh 4t - \frac{1}{8}t, \end{aligned}$$

where I have dropped to integration constant and have used $\sinh 2u = 2\sinh u \cosh u$ and $\sinh^2(u/2) = (\cosh u - 1)/2$. Now replace t by x:

$$t = \sinh^{-1} x = \ln(x + \sqrt{x^2+1}),$$

and

$$\sinh 4t = 2\sinh(2t)\cosh(2t) = 4\sinh t \cosh t(1{+}2\sinh^2 t) = 4x\sqrt{x^2+1}(1{+}2x^2),$$

so that the final result is

$$\int x^2\sqrt{x^2+1}dx = \frac{1}{8}x\sqrt{x^2+1}(1+2x^2) - \frac{1}{8}\ln(x+\sqrt{x^2+1}).$$

Now I have all I need to find ρ_v:

$$\begin{aligned}
\rho_v &= \frac{m^4}{4\pi^2}\int_0^{M/m} dx x^2\sqrt{x^2+1} \\
&= \frac{m^4}{4\pi^2}\left[\frac{1}{8}x\sqrt{x^2+1}(1+2x^2) - \frac{1}{8}\ln(x+\sqrt{x^2+1})\right]_0^{M/m} \\
&= \frac{m^4}{32\pi^2}\left[\frac{M}{m}\sqrt{\frac{M^2}{m^2}+1}\left(1+2\frac{M^2}{m^2}\right) - \ln\left(\frac{M}{m}+\sqrt{\frac{M^2}{m^2}+1}\right)\right] \\
&= \frac{m^4}{32\pi^2}\left[\left(\frac{M}{m}\right)^4\sqrt{1+\frac{m^2}{M^2}}\left(\frac{m^2}{M^2}+2\right) - \ln\left(\frac{M}{m}+\sqrt{\frac{M^2}{m^2}+1}\right)\right] \\
&= \frac{M^4}{16\pi^2}\left[\sqrt{1+\frac{m^2}{M^2}}\left(1+\frac{m^2}{2M^2}\right) - \frac{1}{2}\frac{m^4}{M^4}\ln\left(\frac{M}{m}+\frac{M}{m}\sqrt{1+\frac{m^2}{M^2}}\right)\right] \\
&\approx \frac{M^4}{16\pi^2}\left[\left(1+\frac{m^2}{2M^2}\right)\left(1+\frac{m^2}{2M^2}\right) - -\frac{1}{2}\frac{m^4}{M^4}\ln\left(\frac{M}{m}+\frac{M}{m}\sqrt{1+\frac{m^2}{M^2}}\right)\right] \\
&\approx \frac{M^4}{16\pi^2}\left[1+\frac{m^2}{M^2}+\mathcal{O}\left(\frac{m^4}{M^4}\right)\right] \qquad (8.18)
\end{aligned}$$

Next I evaluate the pressure:

$$p_v = \frac{1}{3}\frac{1}{4\pi^2}\int_0^M dk\frac{k^4}{\sqrt{k^2+m^2}} = \frac{m^4}{12\pi^2}\int_0^{M/m} dx\frac{x^4}{\sqrt{x^2+1}},$$

where I again have substituted $k = mx$. The indefinite integral involved here can again be evaluated using the hyperbolic substitution $x = \sinh t$:

$$\begin{aligned}
\int\frac{x^4}{\sqrt{x^2+1}}dx &= \int\frac{\sinh^4 t}{\cosh t}\cosh t dt \\
&= \int(\sinh^2 t)^2 dt = \frac{1}{4}\int(\cosh 2t - 1)^2 dt
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{4}\int(\cosh^2 2t - 2\cosh 2t + 1)dt \\
&= \frac{1}{4}\int \cosh^2 2t dt - \frac{1}{4}\sinh 2t + \frac{1}{4}t \\
&= \frac{1}{8}\int(\cosh 4t + 1)dt - \frac{1}{4}\sinh 2t + \frac{1}{4}t \\
&= \frac{1}{32}\sinh 4t + \frac{1}{8}t - \frac{1}{4}\sinh 2t + \frac{1}{4}t \\
&= \frac{3}{8}t - \frac{1}{4}\sinh 2t + \frac{1}{32}\sinh 4t, \qquad (8.19)
\end{aligned}$$

where I have used identities like $\sinh^2 u = (\cosh 2u - 1)/2$ and $\cosh^2 u = (\cosh 2u + 1)/2$. I express this result in terms of x:

$$\begin{aligned}
t &= \sinh^{-1} x = \ln(x + \sqrt{x^2+1}) \qquad (8.20) \\
\sinh 2t &= 2\sinh t \cosh t = 2x\sqrt{x^2+1} \qquad (8.21) \\
\sinh 4t &= 2\sinh 2t \cosh 2t = 4x\sqrt{x^2+1}(1+2x^2) \qquad (8.22)
\end{aligned}$$

so the final result is

$$\begin{aligned}
\int \frac{x^4}{\sqrt{x^2+1}}dx &= \frac{3}{8}\ln(x+\sqrt{x^2+1}) - \frac{1}{2}x\sqrt{x^2+1} + \frac{1}{8}x\sqrt{x^2+1}(1+2x^2) \\
&= \frac{1}{8}\left[3\ln(x+\sqrt{x^2+1}) - 4x\sqrt{x^2+1} + x\sqrt{x^2+1}(1+2x^2)\right] \\
&= \frac{1}{8}\left[x\sqrt{x^2+1}(2x^2-3) + 3\ln(x+\sqrt{x^2+1})\right]. \qquad (8.23)
\end{aligned}$$

The pressure therefore becomes

$$\begin{aligned}
p_v &= \frac{m^4}{12\pi^2}\int_0^{M/m}\frac{x^4 dx}{\sqrt{x^2+1}} \\
&= \frac{m^4}{12\pi^2}\frac{1}{8}\left[\frac{M}{m}\sqrt{\frac{M^2}{m^2}+1}\left(2\frac{M^2}{m^2}-3\right) + 3\ln\left(\frac{M}{m}+\frac{M}{m}\sqrt{1+\frac{m^2}{M^2}}\right)\right] \\
&= \frac{1}{3}\frac{m^4}{16\pi^2}\left[\frac{M}{m}\sqrt{\frac{M^2}{m^2}+1}\left(\frac{M^2}{m^2}-\frac{3}{2}\right) + \frac{3}{2}\ln\left(\frac{M}{m}+\frac{M}{m}\sqrt{1+\frac{m^2}{M^2}}\right)\right] \\
&= \frac{1}{3}\frac{M^4}{16\pi^2}\left[\sqrt{1+\frac{m^2}{M^2}}\left(1-\frac{3}{2}\frac{m^2}{M^2}\right) + \frac{3}{2}\frac{m^4}{M^4}\ln\left(\frac{M}{m}+\frac{M}{m}\sqrt{1+\frac{m^2}{M^2}}\right)\right] \\
&\approx \frac{1}{3}\frac{M^4}{16\pi^2}\left[1-\frac{m^2}{M^2}+\mathcal{O}\left(\frac{m^4}{M^4}\right)\right]. \qquad (8.24)
\end{aligned}$$

On comparison with equation (8.18) it is clear that $p_v \neq -\rho_v$. In fact, for a massless scalar field the vacuum follows the equation of state of an ultrarelativistic gas, $p_v = \rho_v/3$. The source of this problem is the fact that this calculation does not respect the Lorentz invariance of the theory. I have treated the spatial component of the four-momentum $k^\mu = (\omega, \mathbf{k})$ differently from the timelike component by introducing a cutoff on $k = \sqrt{\mathbf{k}\cdot\mathbf{k}}$. It is clear that this leads to an unphysical result, and we must look for a better way to handle the divergent integrals.

8.3 What should M be anyway?

The calculation of the vacuum energy density shown in the previous section is the traditional way of presenting the cosmological constant problem. We see that the leading term goes as M^4, so it is clearly important for the final result what we choose for the cut-off. At what energy do we expect our quantum field theory for the scalar field to break down? While it may certainly break down at even lower energies, there is a special energy where we can be certain that our theory is inadequate: The energy where quantum gravitational effects become significant, better known as the Planck energy.

By combining the three fundamental constants G (representing gravity), c (representing relativity) and $\hbar$ (representing quantum mechanics) one can construct a complete system of units, the so-called Planck, or natural, units:

$$\ell_P = \sqrt{\frac{\hbar G}{c^3}} = 1.6 \times 10^{-35} \text{ m} \tag{8.25}$$

$$T_P = \sqrt{\frac{\hbar G}{c^5}} = 0.54 \times 10^{-43} \text{ s} \tag{8.26}$$

$$M_P = \sqrt{\frac{\hbar c}{G}} = 2.2 \times 10^{-8} \text{ kg} \tag{8.27}$$

$$E_P = M_P c^2 = 1.2 \times 10^{19} \text{ GeV} \tag{8.28}$$

At first sight there seems to be no reason why these quantities should have any special significance. After all, they are just combinations of three constants which happen to give quantities with units of length, time etc. But the common wisdom is that they signify the scales where we should expect the classical description of spacetime as a continuum to break down and quantum gravity to be important. Here is an argument for this viewpoint, taken from the paper ”Six easy roads to the Planck scale” by R. J. Adler, published in 2010 in the American Journal of Physics, volume 78, page 925.

The Heisenberg uncertainty principle, which I will write in the approximate form $\Delta x \Delta p \sim \hbar$, is a consequence of the axioms of quantum mechanics. However, in his original paper Heisenberg also gave a heuristic derivation of it, based on a though experiment where he tried to measure the position and momentum of an electron by shining light on it and looking at the scattered light in a microscope. The more accurately he wanted to locate the electron, the shorter the wavelength of the light he had to use. But shorter wavelength means photons of higher energy, and therefore a greater "kick" imparted to the electron, and hence a greater uncertainty in its momentum.

For perfectly good reasons Heisenberg did not consider gravitational forces in his thought experiment. But that is what I will do now. The photon can be assigned an effective mass

$$M_{\text{eff}} = \frac{E}{c^2} = \frac{h\nu}{c^2} = \frac{h}{c\lambda}. \tag{8.29}$$

The photon will exert a gravitational force on the electron which will accelerate it and cause an additional uncertainty in the position. I am only interested in an order-of-magnitude estimate, so I allow myself to work with Newtonian gravity. The gravitational acceleration is then

$$\Delta a_g \sim \frac{GM_{\text{eff}}}{r_{\text{eff}}^2} = G\frac{h}{c\lambda}\frac{1}{r_{\text{eff}}^2}, \tag{8.30}$$

where r_{eff} is the effective range at which the interaction takes place. The uncertainty in the posistion is then of order

$$\Delta x_g \sim \Delta a_g t_{\text{eff}}^2 \sim \frac{Gh}{c\lambda}\left(\frac{t_{\text{eff}}}{r_{\text{eff}}}\right)^2, \tag{8.31}$$

where t_{eff} is the effective time over which the interaction takes place. Now, $r_{\text{eff}}/t_{\text{eff}}$ has units of speed, and the only speed naturally associated with the photon is c, so I take $r_{\text{eff}}/t_{\text{eff}} \sim c$ and find

$$\Delta x_g = \frac{Gh}{c^3}\frac{1}{\lambda} \sim \frac{\ell_P^2}{\lambda}. \tag{8.32}$$

This is presumably a sub-dominant effect, so we can estimate λ from the original version of the uncertainty principle, $\lambda \sim h/\Delta p$ and get

$$\Delta x_g \sim \ell_P^2 \frac{\Delta p}{h} \tag{8.33}$$

and add it to the uncertainty in x:

$$\Delta x \sim \frac{\hbar}{\Delta p} + \ell_P^2 \frac{\Delta p}{\hbar}, \tag{8.34}$$

where I have ignored the factor of 2π difference between h and $\hbar$ (taking $2\pi \approx 1$ is also known as "Feynman units") since, again, I am only interested in order-of-magnitude estimates. Dodgy though this derivation may seem, a similar result appears in, e.g., string theory.

The uncertainty principle without gravity allows for a precise determination of the position ($\Delta x = 0$) if we are prepared to forego all information about the momentum ($\Delta p \to \infty$). With the gravitational term, however, we see that there is a new situation. We have $\Delta x \to \infty$ as $\Delta p \to 0$, but also

$\Delta x \to \infty$ as $\Delta p \to \infty$. An absolutely precise measurement of the position is now impossible, and there is a minimum uncertainty found by taking the derivative of equation (8.34) and setting it equal to zero:

$$\frac{d(\Delta x)}{d(\Delta p)} = -\frac{\hbar}{(\Delta p)^2} + \frac{\ell_P^2}{\hbar} = 0,$$

which gives

$$\Delta p = \frac{\hbar}{\ell_P},$$

and

$$(\Delta x)_{\min} = \frac{\hbar}{\hbar/\ell_P} + \ell_P^2 \frac{\hbar/\ell_P}{\hbar} = 2\ell_P. \tag{8.35}$$

With gravity in the picture, it is no longer possible to measure positions with accuracy greater than $\sim \ell_P$, and in a sense, therefore, the Planck length is the shortest physically meaningful distance.

To resolve the smallest distance requires a probe of wavelength $\sim \ell_P$, momentum $\Delta p \sim \hbar/\ell_P$, and energy $E \sim \hbar c/\ell_P = \hbar c\sqrt{c^3/\hbar G} = \sqrt{\hbar c^5/G} = E_P$. A probe with this energy has mass $M_P = E_P/c^2$, and Schwarzschild radius

$$R_s \sim \frac{GM_P}{c^2} = \sqrt{\frac{\hbar G}{c^3}} = \ell_P, \tag{8.36}$$

So the probe has a wavelength equal to its Schwarzschild radius, and will therefore form a black hole! In our estimate of the vacuum energy it therefore makes no sense to integrate to masses higher than the Planck mass. A zero-point vibration with this energy/mass probes the shortest meaningful distance and corresponds to a black hole. To describe this situation, we clearly need a theory of quantum gravity. Hence, we should not trust quantum field theory at energies above the Planck energy, and we should use $M = M_P$ as a cut-off.

With the cutoff at the Planck mass, the leading contribution to the vacuum energy density is

$$\rho_v = \frac{M_P}{16\pi^2},$$

but this is in units where $\hbar = c = 1$, so M_P has units of energy, and is equal to the Planck energy. To compare with the observed value of the effective cosmological constant, 0.7 times the present critical density, I must convert to units of GeV cm^{-3}. I note that $\hbar c = 197.326$ MeV fm $= 1.97326 \times 10^{-16}$ GeV cm, so by dividing the vacuum energy density by $(\hbar c)^3$ I get the correct units:

$$\rho_v = \frac{(1.2 \times 10^{19}\ \text{GeV})^4}{(16\pi^2)(1.97326 \times 10^{-16}\ \text{GeV cm})^3} = 1.7 \times 10^{115}\ \text{Gev cm}^{-3}.$$

I use $h = 0.7$ and find that the ratio of the theoretical estimate of the vacuum energy to the observed value of the effective cosmological constant is

$$\frac{\rho_v}{0.7\rho_{c0}} = \frac{1.7 \times 10^{115}}{3.6 \times 10^{-6}} = 4.7 \times 10^{120},$$

a true mismatch if there ever was one.

8.4 Taming infinities respectfully: Dimensional regularization

Regardless of the choice of cut-off, my calculation of the vacuum energy density and pressure was clearly wrong. The failure to produce the correct equation of state can, as I said, be traced to the fact that introducing a cut-off in momentum violates Lorentz symmetry. Infinities appear in all quantum field theory calculations beyond the lowest order, and in the process of isolating them (called *regularization*) and taming them (called *renormalization*) it turns out to be vital to use procedures which preserve symmetries like Lorentz invariance. I will use one of the most popular regularization techniques to recalculate the vacuum energy density and pressure, the method known as *dimensional regularization.* To introduce it to you, I will first consider a simpler example from classical electromagnetism. The following section follows the paper "Regularization, renormalization, and dimensional analysis: Dimensional regularization meets freshman E&M" by F. Olness and R. Scalise, published in 2011 in the American Journal of Physics, volume 79, page 306.

I want to calculate the electrostatic potential from an infinitely long, thin line of uniformly distributed electrical charge. This means that the linear charge density $\lambda = dQ/dy$, where y is the spatial coordinate along the line, is a constant. At a perpendicular distance x from the line, the contribution to the potential from the charge element between y and $y + dy$ is given by

$$dV = \frac{1}{4\pi\epsilon_0}\frac{dQ}{r} = \frac{\lambda}{4\pi\epsilon_0}\frac{dy}{\sqrt{x^2+y^2}}, \tag{8.37}$$

so the total potential is found by integrating over the whole line to be

$$V(x) = \frac{\lambda}{4\pi\epsilon_0}\int_{-\infty}^{+\infty}\frac{dy}{\sqrt{x^2+y^2}} = \frac{\lambda}{4\pi\epsilon_0}\left[\sinh^{-1}\left(\frac{y}{x}\right)\right]_{-\infty}^{+\infty} = \infty, \tag{8.38}$$

i.e., it diverges.

Note that the potential is scale invariant, that is, it doesn't change under the transformation $x \to kx$:

$$V(kx) \quad = \quad \frac{\lambda}{4\pi\epsilon_0}\int_{-\infty}^{+\infty}\frac{dy}{\sqrt{k^2x^2+y^2}}$$

$$\begin{aligned}
&= \frac{\lambda}{4\pi\epsilon_0 k}\int_{-\infty}^{+\infty}\frac{dy}{\sqrt{x^2+(y/k)^2}} \\
&= \frac{\lambda}{4\pi\epsilon_0 k}\int_{-\infty}^{+\infty}\frac{kdz}{\sqrt{x^2+z^2}} \\
&= \frac{\lambda}{4\pi\epsilon_0}\int_{-\infty}^{+\infty}\frac{dz}{\sqrt{x^2+z^2}} \\
&= V(x) \qquad (8.39)
\end{aligned}$$

where I have substituted $z = y/k$. This, however, implies that $V(x_1) = V(x_2)$ for all x_1 and x_2, which again seems to imply that the electric field $\mathbf{E} = -\nabla V = 0$! Bear in mind, though, that V is a divergent quantity. When we are dealing with infinities it is not necessarily true that $\infty - \infty = 0$. But how do we make this subtraction? One way is to proceed like we tried to with the vacuum energy and introduce a cut-off so that V is finite, make the subtraction, and then let the cut-off approach infinity. If I take the cut-off to be L, then

$$V(x) = \frac{\lambda}{4\pi\epsilon_0}\int_{-L}^{+L}\frac{dy}{\sqrt{x^2+y^2}} = \frac{\lambda}{4\pi\epsilon_0}\ln\left(\frac{L+\sqrt{L^2+x^2}}{-L+\sqrt{L^2+x^2}}\right). \qquad (8.40)$$

Note that the translational invariance is lost, $V(kx) \neq V(x)$. In this case, this is not disastrous. I find that the electric field is

$$E(x) = -\frac{\partial V}{\partial x} = \frac{\lambda}{2\pi\epsilon_0}\frac{1}{x}\frac{L}{\sqrt{L^2+x^2}} \rightarrow_{L\rightarrow\infty} \frac{\lambda}{2\pi\epsilon_0 x}, \qquad (8.41)$$

which is the correct result (you should check this using Gauss' law.) In this case nothing bad happened even though I broke the symmetry of the problem along the way. In quantum field theory, I would not have been so lucky. Let me therefore introduce a better way of handling the infinity: Dimensional regularization.

The idea is to calculate $V(x)$ in n spatial dimensions (where n is not necessarily an integer), and then let $n \rightarrow 1$ (the situation I am considering is essentially one-dimensional). I will generalize

$$\int_{-\infty}^{+\infty} dy \equiv \int dV_1$$

to

$$\int dV_n = \int d\Omega_n \int_0^{\infty} y^{n-1}dy, \qquad (8.42)$$

where

$$\Omega_n = \int d\Omega_n = \frac{2\pi^{n/2}}{\Gamma\left(\frac{n}{2}\right)} = \frac{n\pi^{n/2}}{\Gamma\left(\frac{n}{2}+1\right)}, \qquad (8.43)$$

is the n-dimensional solid angle, and $\Gamma(x)$ is the gamma function which has the property $x\Gamma(x) = \Gamma(x+1)$. See appendix A for proof of equation (8.43).

The generalised expression for the potential is

$$V(x) = \frac{\lambda}{4\pi\epsilon 0} \int d\Omega_n \int_0^\infty \frac{y^{n-1}}{\mu^{n-1}} \frac{dy}{\sqrt{x^2+y^2}}, \tag{8.44}$$

where μ is an a so-called auxiliary length scale which I must introduce to ensure that the potential has the correct units. I use equation (8.43) to proceed:

$$\begin{aligned} V(x) &= \frac{\lambda}{4\pi\epsilon_0} \frac{n\pi^{n/2}}{\Gamma\left(\frac{n}{2}+1\right)} \frac{1}{\mu^{n-1}} \int_0^\infty \frac{y^{n-1}dy}{\sqrt{x^2+y^2}} \\ &= \frac{\lambda}{4\pi\epsilon_0} \frac{n\pi^{n/2}}{\Gamma\left(\frac{n}{2}+1\right)} \left(\frac{x}{\mu}\right)^{n-1} \int_0^\infty \frac{z^{n-1}dz}{\sqrt{1+z^2}} \end{aligned} \tag{8.45}$$

where I have substituted $y = xz$. The integral I need to evaluate can be related to the so-called beta function

$$B(p,q) = \int_0^\infty \frac{y^{p-1}}{(1+y)^{p+q}} = \int_0^1 x^{p-1}(1-x)^{q-1}dx = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)} \tag{8.46}$$

which you can read more about in appendix B. To calculate the integral, I substitute $x = z^2$, $z = x^{1/2}$, $dz = \frac{1}{2}x^{-1/2}dx$ and get

$$\begin{aligned} \int_0^\infty \frac{z^{n-1}}{\sqrt{1+z^2}}dz &= \int_0^\infty x^{(n-1)/2} \frac{1}{(1+x)^{1/2}} \frac{1}{2} x^{-1/2} dx \\ &= \frac{1}{2} \int_0^\infty \frac{x^{n/2-1}}{(1+y)^{1/2}} dy, \end{aligned} \tag{8.47}$$

Comparing with equation (8.46) I see that this corresponds to $p = n/2$, $q = 1/2 - p = (1-n)/2$, so

$$\begin{aligned} \int_0^\infty \frac{z^{n-1}}{\sqrt{1+z^2}}dz &= \frac{1}{2} B\left(\frac{n}{2}, \frac{1-n}{2}\right) \\ &= \frac{1}{2} \frac{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{1-n}{2}\right)}{\Gamma\left(\frac{n}{2}+\frac{1-n}{2}\right)} \\ &= \frac{1}{2} \frac{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{1-n}{2}\right)}{\sqrt{\pi}} \end{aligned} \tag{8.48}$$

where I have used that $\Gamma(1/2) = \sqrt{\pi}$. The potential becomes

$$\begin{aligned} V(x) &= \frac{\lambda}{4\pi\epsilon_0} \frac{n\pi^{n/2}}{\Gamma\left(\frac{n}{2}+1\right)} \frac{x^{n-1}}{\mu^{n-1}} \frac{1}{2} \frac{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{1-n}{2}\right)}{\sqrt{\pi}} \\ &= \frac{\lambda}{4\pi\epsilon_0} (\sqrt{\pi})^{n-1} \left(\frac{x}{\mu}\right)^{n-1} \Gamma\left(\frac{1-n}{2}\right), \end{aligned} \tag{8.49}$$

where I have used that $\Gamma(n/2+1) = \frac{n}{2}\Gamma(n/2)$.

The idea is now to let $n \to 1$. To do this, I write $n = 1 - 2\epsilon$, where $\epsilon \to 0$. The potentiall can then be written as

$$V(x) = \frac{\lambda}{4\pi\epsilon_0}\left(\frac{\mu^{2\epsilon}}{x^{2\epsilon}}\frac{\Gamma(\epsilon)}{\pi^{\epsilon}}\right). \tag{8.50}$$

V depends on the so-called regulator ϵ, which is dimensionless, and the auxilliary scale μ, which has dimensions of length. The translational invariance is preserved, and physical quantities will turn out to be independent of ϵ and μ.

I have now carried out what is known as *regularization* of V. What remains to be done is to subtract the divergent part of the potential, and this is known as *renormalization.* To do this, I need to expand the expression around $\epsilon = 0$. For small x, the gamma function can be expanded as (see appendix C for the derivation)

$$\Gamma(x) \approx -\gamma_E + \frac{1}{x}, \tag{8.51}$$

where $\gamma_E \approx 0.577216$ is known as Eulers constant. I can therefore write $\Gamma(\epsilon) \approx -\gamma_E + 1/\epsilon$. Furthermore, I have

$$\left(\frac{\mu^2}{x^2}\right)^{\epsilon} = \left[e^{\ln\left(\frac{\mu^2}{x^2}\right)}\right]^{\epsilon} = e^{\epsilon\ln\left(\frac{\mu^2}{x^2}\right)} \approx 1 + \epsilon\ln\left(\frac{\mu^2}{x^2}\right), \tag{8.52}$$

and

$$\pi^{-\epsilon} = (e^{\ln\pi})^{-\epsilon} = e^{-\epsilon\ln\pi} \approx 1 - \epsilon\ln\pi. \tag{8.53}$$

I can now multiply these factors to obtain V, bearing in mind that I want to take the limit $\epsilon \to 0$, so I can discard all terms of order 1 or higher in ϵ. This gives me

$$\begin{aligned} V(x) &= \frac{\lambda}{4\pi\epsilon_0}\left[1 + \epsilon\ln\left(\frac{\mu^2}{x^2}\right)\right](1 - \epsilon\ln\pi)\left(-\gamma_E + \frac{1}{\epsilon}\right) \\ &\approx \frac{\lambda}{4\pi\epsilon_0}\left[-\gamma_E + \frac{1}{\epsilon} - \ln\pi + \ln\left(\frac{\mu^2}{x^2}\right)\right] \\ &= \frac{\lambda}{4\pi\epsilon_0}\left[\frac{1}{\epsilon} + \ln\left(\frac{e^{-\gamma_E}}{\pi}\right) + \ln\left(\frac{\mu^2}{x^2}\right)\right]. \end{aligned} \tag{8.54}$$

All physical quantities are invariant under a constant shift in V: $V(x) \to V(x) + c$. I can therefore subtract the $1/\epsilon$ term, and this is what renormalization means in this example:

$$V_{\mathrm{MS}}(x) = \frac{\lambda}{4\pi\epsilon_0}\left[\ln\left(\frac{e^{-\gamma_E}}{\pi}\right) + \ln\left(\frac{\mu^2}{x^2}\right)\right], \tag{8.55}$$

where MS stands for "minimal scheme". I can also choose to subtract all constants in V, leaving me with

$$V_{\overline{\mathrm{MS}}}(x) = \frac{\lambda}{4\pi\epsilon_0}\ln\left(\frac{\mu^2}{x^2}\right), \tag{8.56}$$

where $\overline{\mathrm{MS}}$ means "modified minimal scheme". What I cannot do, however, is remove the auxilliary scale μ from the final result. It is needed to ensure that the argument of the logarithm is dimensionless. However, it will disappear when I calculate physical quantities like forces and potentail differences, for example

$$V_{\mathrm{MS}}(x_1) - V_{\mathrm{MS}}(x_2) = \frac{\lambda}{4\pi\epsilon_0}\ln\left(\frac{x_2^2}{x_1^2}\right) = V_{\overline{\mathrm{MS}}}(x_1) - V_{\overline{\mathrm{MS}}}(x_2). \tag{8.57}$$

Note also that the result does not depend on the renormalization scheme.

8.5 The vacuum energy done correctly

I now return to the vacuum energy of the scalar field to calculate the energy density using dimensional regularization. To check that I get the correct equation of state, I will also calculate the pressure.

The first thing to do is to calculate the vacuum energy density in d spacetime dimensions. The expression for it is

$$\begin{aligned}\rho_v &= \frac{\mu^{4-d}}{(2\pi)^{d-1}}\frac{1}{2}\int d^{d-1}k\omega(k) \\ &= \frac{\mu^{4-d}}{(2\pi)^{d-1}}\frac{1}{2}\int d^{d-2}\Omega\int_0^\infty \omega(k),\end{aligned} \tag{8.58}$$

where the auxilliary momentum scale μ has to be included to make sure that the energy density has the correct units. I know how to do the angular integral:

$$\int d^{d-2}\Omega = \Omega_{d-1} = \frac{(2\pi)^{(d-1)/2}}{\Gamma\left(\frac{d-1}{2}\right)}, \tag{8.59}$$

and I rewrite the momentum integral using the dimensionless variable x defined by $k = mx$:

$$\begin{aligned}\int_0^\infty dkk^{d-2}\omega(k) &= \int_0^\infty dkk^{d-2}\sqrt{k^2+m^2} \\ &= \int_0^\infty mdxm^{d-2}x^{d-2}m\sqrt{x^2+1} \\ &= m^d\int_0^\infty x^{d-2}\sqrt{x^2+1}dx \equiv m^d I.\end{aligned} \tag{8.60}$$

I evaluate the integral I by relating it to the beta function introduced in appendix B. To do this, I use the substitution $x = \tan\theta$, which gives $1+x^2 = 1/\cos^2\theta$, $dx = d\theta/\cos^2\theta$, and

$$\begin{aligned} I &= \int_0^{\pi/2} \frac{(\sin\theta)^{d-2}}{(\cos\theta)^{d-2}} \frac{1}{\cos\theta} \frac{d\theta}{\cos^2\theta} \\ &= \int_0^{\pi/2} (\sin\theta)^{d-2} (\cos\theta)^{-d-1} d\theta. \end{aligned} \tag{8.61}$$

Comparing this with equation (8.93), I find that this integral is one half times the beta function with $p = (d-1)/2$, $q = -d/2$, so

$$I = \frac{1}{2} \frac{\Gamma\left(\frac{d-1}{2}\right)\Gamma\left(-\frac{d}{2}\right)}{\Gamma\left(\frac{d}{2} - \frac{1}{2} - \frac{d}{2}\right)} = \frac{1}{2} \frac{\Gamma\left(\frac{d-1}{2}\right)\Gamma\left(-\frac{d}{2}\right)}{\Gamma\left(-\frac{1}{2}\right)}. \tag{8.62}$$

The vacuum energy is therefore given by

$$\begin{aligned} \rho_v &= \frac{\mu^{4-d}}{(2\pi)^{d-1}} \frac{1}{2} \frac{2\pi^{(d-1)/2}}{\Gamma\left(\frac{d-1}{2}\right)} \frac{1}{2} m^d \frac{\Gamma\left(\frac{d-1}{2}\right)\Gamma\left(-\frac{d}{2}\right)}{\Gamma\left(-\frac{1}{2}\right)} \\ &= \frac{\mu^4}{2(4\pi)^{(d-1)/2}} \frac{\Gamma\left(-\frac{d}{2}\right)}{\Gamma\left(-\frac{1}{2}\right)} \left(\frac{m}{\mu}\right)^d. \end{aligned} \tag{8.63}$$

Before taking the limit $d \to 4$, I will check that I get the correct equation of state, $p_v = -\rho_v$. In d spacetime dimensions the expression for the pressure becomes

$$\begin{aligned} p_v &= \frac{\mu^{4-d}}{(2\pi)^{d-1}} \frac{1}{2(d-1)} \int d^{d-1}k \frac{k^2}{\omega(k)} \\ &= \frac{\mu^{4-d}}{(2\pi)^{d-1}} \frac{1}{2(d-1)} \int d^{d-2}\Omega \int dk k^{d-2} \frac{k^2}{\omega(k)} \\ &= \frac{\mu^{4-d}}{(2\pi)^{d-1}} \frac{1}{2(d-1)} \frac{2\pi^{(d-1)/2}}{\Gamma\left(\frac{d-1}{2}\right)} \int_0^\infty \frac{k^d}{\omega(k)}. \end{aligned} \tag{8.64}$$

I rewrite the momentum integral using the substitution $k = mx$:

$$\begin{aligned} \int_0^\infty dk \frac{k^d}{\omega(k)} &= \int_0^\infty dk \frac{k^d}{\sqrt{k^2+m^2}} \\ &= \int_0^\infty m dx \frac{m^d x^d}{m\sqrt{x^2+1}} \\ &= m^d \int_0^\infty \frac{x^d}{\sqrt{x^2+1}} dx \equiv m^d I, \end{aligned} \tag{8.65}$$

and I then use the substitution $x = \tan\theta$ to relate the integral I to the beta function:

$$\begin{aligned} I &= \int_0^\infty \frac{x^d}{\sqrt{x^2+1}} dx \\ &= \int_0^{\pi/2} \frac{\left(\frac{\sin\theta)^d}{(\cos\theta)^d}\right.}{\frac{1}{\cos\theta}} \frac{d\theta}{(\cos\theta)^2} \\ &= \int_0^{\pi/2} (\sin\theta)^d (\cos\theta)^{-d-1} d\theta, \end{aligned} \tag{8.66}$$

and comparing this with equation (8.93), I see that I is one half times the beta function with $p = (d+1)/2$ and $q = -d/2$:

$$I = \frac{1}{2} \frac{\Gamma\left(\frac{d+1}{2}\right)\Gamma\left(-\frac{d}{2}\right)}{\Gamma\left(\frac{1}{2}\right)}. \tag{8.67}$$

The pressure now becomes

$$\begin{aligned} p_v &= \frac{\mu^{4-d}}{(2\pi)^{d-1}} \frac{1}{4} \frac{2\pi^{(d-1)/2}}{\frac{d-1}{2}\Gamma\left(\frac{d-1}{2}\right)} \frac{1}{2} m^d \frac{\Gamma\left(\frac{d+1}{2}\right)\Gamma\left(-\frac{d}{2}\right)}{\Gamma\left(\frac{1}{2}\right)} \\ &= \frac{\mu^4}{4(4\pi)^{(d-1)/2}} \frac{\Gamma\left(-\frac{d}{2}\right)}{\Gamma\left(-\frac{1}{2}\right)} \left(\frac{m}{\mu}\right)^d, \end{aligned} \tag{8.68}$$

where I have used the fact that $\frac{d+1}{2} = \frac{d-1}{2} + 1$ and $\Gamma(x+1) = x\Gamma(x)$. Using the latter identity again, I can write

$$\Gamma\left(\frac{1}{2}\right) = \Gamma\left(-\frac{1}{2}+1\right) = -\frac{1}{2}\Gamma\left(-\frac{1}{2}\right), \tag{8.69}$$

so I can write the pressure as

$$\begin{aligned} p_v &= \frac{\mu^4}{4(4\pi)^{(d-1)/2}} \frac{\Gamma\left(-\frac{d}{2}\right)}{-\frac{1}{2}\Gamma\left(-\frac{1}{2}\right)} \left(\frac{m}{\mu}\right)^d \\ &= -\frac{\mu^4}{2(4\pi)^{(d-1)/2}} \frac{\Gamma\left(-\frac{d}{2}\right)}{\Gamma\left(-\frac{1}{2}\right)} \left(\frac{m}{\mu}\right)^d \\ &= -\rho_v, \end{aligned} \tag{8.70}$$

which shows that dimensional regularization, in contrast to using a cutoff, reproduces the correct equation of state.

It is now time to take the limit $d \to 4$. Similarly to he example with the electrostatic potential in the previous section, I write $d = 4 - \epsilon$ and expand the vacuum energy in the small quantity ϵ. I start with the factor $\Gamma(-d/2) = \Gamma(-2+\epsilon/2)$, rewriting it in the not-so-obvious way:

$$\begin{aligned}\Gamma\left(1+\frac{\epsilon}{2}\right) &= \frac{\epsilon}{2}\Gamma\left(\frac{\epsilon}{2}\right) \\ &= \frac{\epsilon}{2}\left(\frac{\epsilon}{2}-1\right)\Gamma\left(\frac{\epsilon}{2}-1\right) \\ &= \frac{\epsilon}{2}\left(\frac{\epsilon}{2}-1\right)\left(\frac{\epsilon}{2}-2\right)\Gamma\left(\frac{\epsilon}{2}-2\right), \qquad (8.71)\end{aligned}$$

so that

$$\Gamma\left(-2+\frac{\epsilon}{2}\right) = \frac{1}{-2+\frac{\epsilon}{2}}\frac{1}{-1+\frac{\epsilon}{2}}\frac{1}{\frac{\epsilon}{2}}\Gamma\left(1+\frac{\epsilon}{2}\right). \qquad (8.72)$$

The point of rewriting the gamma function in this way is that I can now use the results from appendix C to Taylor expand it:

$$\begin{aligned}\Gamma\left(1+\frac{\epsilon}{2}\right) &\approx \Gamma(1) + \frac{\epsilon}{2}\Gamma'(1) \\ &= 1 + \frac{\epsilon}{2}\Gamma(1)\psi_1(1) \\ &= 1 - \gamma_E\frac{\epsilon}{2}. \qquad (8.73)\end{aligned}$$

I can now write

$$\begin{aligned}\Gamma\left(-2+\frac{\epsilon}{2}\right) &\approx \frac{1}{-2+\frac{\epsilon}{2}}\frac{1}{-1+\frac{\epsilon}{2}}\frac{1}{\frac{\epsilon}{2}}\Gamma\left(1+\frac{\epsilon}{2}\right) \\ &\approx -\frac{1}{2}\frac{1}{1-\frac{\epsilon}{4}}(-1)\frac{1}{1-\frac{\epsilon}{2}}\frac{2}{\epsilon}\left(1-\frac{\gamma_E}{2}\epsilon\right) \\ &\approx \frac{1}{\epsilon}\left(1+\frac{\epsilon}{4}\right)\left(1+\frac{\epsilon}{2}\right)\left(1-\frac{\gamma_E}{2}\epsilon\right) \\ &\approx \frac{1}{\epsilon}\left(1+\frac{3}{4}\epsilon\right)\left(1-\frac{\gamma_E}{2}\epsilon\right) \\ &\approx \frac{1}{\epsilon}+\left(\frac{3}{4}-\frac{\gamma_E}{2}\right) \qquad (8.74)\end{aligned}$$

where I have used the approximation $\frac{1}{1+x} \approx 1-x$ and neglected all terms of order ϵ and higher in the final result, since I eventually want to take the limit $\epsilon \to 0$. Furthermore, I write

$$\begin{aligned}(4\pi)^{-\frac{d-1}{2}} &= (4\pi)^{-3/2}(4\pi)^{\epsilon/2} \\ &= (4\pi)^{-3/2}e^{\frac{\epsilon}{2}\ln 4\pi} \\ &\approx (4\pi)^{-3/2}\left[1+\frac{\epsilon}{2}\ln 4\pi\right], \qquad (8.75)\end{aligned}$$

and

$$\begin{aligned}
\left(\frac{m}{\mu}\right)^d &= \left(\frac{m}{\mu}\right)^4 \left(\frac{m}{\mu}\right)^{-\epsilon} \\
&= \left(\frac{m}{\mu}\right)^4 e^{-\epsilon \ln(m/\mu)} \\
&\approx \left(\frac{m}{\mu}\right)^4 \left[1 - \epsilon \ln\left(\frac{m}{\mu}\right)\right]. \qquad (8.76)
\end{aligned}$$

With the additional note that $\Gamma\left(-\frac{1}{2}\right) = -2\Gamma\left(\frac{1}{2}\right) = -2\sqrt{\pi}$, I can now expand ρ_v in ϵ:

$$\begin{aligned}
\rho_v &\approx \frac{\mu^4}{2}\frac{1}{(4\pi)^{3/2}}\left[1 + \frac{\epsilon}{2}\ln 4\pi\right]\frac{1}{-2\sqrt{\pi}}\left[\frac{1}{\epsilon} + \left(\frac{3}{4} - \frac{\gamma_E}{2}\right)\right]\left(\frac{m}{\mu}\right)^4\left[1 - \epsilon \ln\left(\frac{m}{\mu}\right)\right] \\
&= -\frac{m^4}{64\pi^2}\left[\frac{2}{\epsilon} + \frac{3}{2} - \gamma_E - \ln\left(\frac{m^2}{4\pi\mu^2}\right)\right], \qquad (8.77)
\end{aligned}$$

where I again have neglected terms of order ϵ and higher. The divergence is contained in the first term in the brackets. I subtract it along with the other constant terms using the $\overline{\text{MS}}$ renormalization scheme, and find the final result for the vacuum energy density:

$$\rho_v = \frac{m^4}{64\pi^2}\ln\left(\frac{m^2}{\mu^2}\right). \qquad (8.78)$$

Two points are worth noting about this result. First of all, the vacuum energy density depends on the mass m of the particle corresponding to the scalar field as m^4, in contrast with my first attempt at calculating ρ_v which resulted in the vacuum energy density depending on the cutofff to the fourth power. As a corollary, a massless field does not contribute to the vacuum energy density. Photons have spin 1, so my calculation is not directly applicable to them, but the result turns out to have the same dependence on the field mass for both spin-1 particles and for fermions, which means, for example, that the electromagnetic field does not contribute to the vacuum energy density.

Secondly, the vacuum energy density can be positive or negative, depending on whether the mass m is greater than or smaller than the auxiliary scale μ. Its numerical value depends on μ, but only logarithmically. In the review by Jerome Martin it is suggested that the choice of μ should be guided by the fact that we infer the observed value of the vacuum energy density from observations of type Ia supernovae. With his choice, I find that the fields of the Standard Model contribute 2.6×10^{41} GeV cm^{-3} to the vacuum energy density. If this is the case, the ratio of the prediction to the observed value is "only" about 10^{47}, which means that the cosmological constant must be tuned to 47 decimal places to give the observed value of the effective vacuum energy density. This still represents a huge and unsatisfactory fine tuning, but it is significantly less than 120 orders of magnitude.

To conclude: The often-heard statement that the mismatch between theory and observations regarding the cosmological constant is 120 orders of magnitude is simply wrong. It is based on a calculation which breaks Lorentz invariance, resulting in the vacuum having the wrong equation of state. When the calculation is done correctly, the mismatch is much smaller, although still huge. The cosmological constant problem is still one of the most important unsolved problems in physics, but knowing how big the problem is, and understanding a bit more about how it arises are certainly important first steps along the path to solving it.

8.6 Appendix A: Area and volume in n dimensions

In three dimensions we know that the volume of a sphere with radius R can be written as

$$\begin{aligned} V_3 &= \int d\Omega_3 \int_0^R r^2 dr \\ &= \int_0^{2\pi} d\phi \int_0^{\pi} \sin\theta d\theta \int_0^R r^2 dr \\ &= \frac{4\pi}{3} R^3, \end{aligned} \tag{8.79}$$

and the surface area can be found by restricting the integration to points where $r = R$:

$$S_2 = \int d\Omega_n \int_0^R drr^2 \delta(r-R) = 4\pi R^2. \tag{8.80}$$

The generalization to n dimensions is

$$V_n = \int d\Omega_n \int_0^R r^{n-1} dr = \Omega_n \frac{R^n}{n} \tag{8.81}$$

$$S_{n-1} = \int d\Omega_n \int_0^R drr^{n-1}\delta(r-R) = \Omega_n R^{n-1}. \tag{8.82}$$

We see that

$$\frac{V_n}{S_{n-1}} = \frac{R}{n}, \tag{8.83}$$

and

$$\frac{dV_n}{dR} = S_{n-1}. \tag{8.84}$$

What remains to be determined is Ω_n. I can find it by using a trick (needless to say not of my own invention). I can write the volume in Cartesian coordinates as

$$V_n = \int dV_n = \int_{x^2 \leq R^2} d^n x = \frac{\Omega_n}{n} R^n, \tag{8.85}$$

and from equations (8.84) and (8.82) I have

$$dV_n = S_{n-1}dR = \Omega_n R^{n-1}dR. \tag{8.86}$$

I know that $\int_{-\infty}^{+\infty} dxe^{-x^2} = \sqrt{\pi}$, so

$$\begin{aligned}
\prod_{i=1}^{n}\int_{-\infty}^{+\infty} dx_i e^{-x_i^2} &= \pi^{n/2} \\
&= \int dV_n e^{-R^2} \\
&= \Omega_n \int_0^\infty dRR^{n-1}e^{-R^2}.
\end{aligned} \tag{8.87}$$

I now substitute $t = R^2$, $dt = 2RdR$, $R^{n-2} = t^{(n-2)/2} = t^{n/2-1}$ and get

$$\begin{aligned}
\pi^{n/2} &= \Omega_n \frac{1}{2}\int_0^\infty dtt^{n/2-1}e^{-t} \\
&= \frac{1}{2}\Omega_n \Gamma\left(\frac{n}{2}\right) \\
&= \frac{\Omega_n}{n}\frac{n}{2}\Gamma\left(\frac{n}{2}\right) \\
&= \frac{\Omega_n}{n}\Gamma\left(\frac{n}{2}+1\right)
\end{aligned} \tag{8.88}$$

where I have used the definition of the gamma function

$$\Gamma(x) = \int_0^\infty dtt^{x-1}e^{-t}, \tag{8.89}$$

and the property $x\Gamma(x) = \Gamma(x+1)$. Solving equation (8.88) for Ω_n, I find

$$\Omega_n = \frac{n\pi^{n/2}}{\Gamma\left(\frac{n}{2}+1\right)}. \tag{8.90}$$

8.7 Appendix B: The beta function

The beta function (the "B" is a capital β) is defined as

$$B(p,q) = \int_0^1 x^{p-1}(1-x)^{q-1}dx, \tag{8.91}$$

with $p > 0, q > 0$. An alternative form can be found by substituting $x = y/(1+y)$. This makes $dx = dy/(1+y)^2$, $1-x = 1/(1+y)$, and

$$\begin{aligned}
B(p,q) &= \int_0^\infty \frac{y^{p-1}}{(1+y)^{p-1}}\frac{1}{(1+y)^{q-1}}\frac{dy}{(1+y)^2} \\
&= \int_0^\infty \frac{y^{p-1}}{(1+y)^{p+q}}.
\end{aligned} \tag{8.92}$$

Yet another form can be obtained by substituting $x = \sin^2\theta$, $1 - x = \cos^2\theta$, $dx = 2\sin\theta\cos\theta$ in equation (8.91):

$$\begin{aligned} B(p,q) &= \int_0^1 x^{p-1}(1-x)^{q-1}dx \\ &= 2\int_0^{\pi/2}(\sin\theta)^{2p-2}(\cos\theta)^{2q-2}\sin\theta\cos\theta d\theta \\ &= 2\int_0^{\pi/2}(\sin\theta)^{2p-1}(\cos\theta)^{2q-1}d\theta. \end{aligned} \tag{8.93}$$

To express the beta function in terms of the gamma function, I start with the definition of the latter:

$$\Gamma(p) = \int_0^\infty t^{p-1}e^{-t}dt \tag{8.94}$$

and substitute $t = y^2$:

$$\Gamma(p) = 2\int_0^\infty y^{2p-2}e^{-y^2}ydy = 2\int_0^\infty y^{2p-1}e^{-y^2}dy. \tag{8.95}$$

I can therefore also write

$$\Gamma(q) = 2\int_0^\infty x^{2q-1}e^{-x^2}dx, \tag{8.96}$$

multiply by equation (8.95) and switch to polar coordinates in the xy plane, bearing in mind that I integrate over the first, positive quadrant:

$$\begin{aligned} \Gamma(p)\Gamma(q) &= 4\int_0^\infty\int_0^\infty x^{2q-1}y^{2p-1}e^{-(x^2+y^2)}dxdx \\ &= 4\int_0^\infty rdr\int_0^{\pi/2}d\theta(r\cos\theta)^{2q-1}(r\sin\theta)^{2p-1}e^{-r^2} \\ &= 4\int_0^\infty r^{2(p+q)-1}e^{-r^2}dr\int_0^{\pi/2}(\sin\theta)^{2p-1}(\cos\theta)^{2q-1}d\theta \\ &= \Gamma(p+q)2\int_0^{\pi/2}(\sin\theta)^{2p-1}(\cos\theta)^{2q-1}d\theta \end{aligned} \tag{8.97}$$

where I have used equation (8.95) again in the last step. But the factor after $\Gamma(p+q)$ is by equation (8.93) the beta function $B(p,q)$, so I have

$$\Gamma(p)\Gamma(q) = \Gamma(p+q)B(p,q), \tag{8.98}$$

which means

$$B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}. \tag{8.99}$$

8.8 Appendix C: The gamma function

In the example with the renormalization of the electrostatic potential and in the renormalization of the vacuum energy, I have needed an expansion of the gamma function for small values of the argument. Let me show how this expansion can be derived.

The gamma function is defined as

$$\Gamma(x) = \int_0^\infty dt t^{x-1} e^{-t}. \tag{8.100}$$

I easily find that

$$\Gamma(1) = \int_0^\infty dt e^{-t} = [-e^{-t}]_0^\infty = 1, \tag{8.101}$$

and using integration by parts it is also quite easy to show that $\Gamma(x+1) = x\Gamma(x)$:

$$\begin{aligned} \Gamma(x+1) &= \int_0^\infty t^x e^{-t} dt \\ &= [-t^x e^{-t}]_0^\infty - \int_0^\infty (-e^{-t}) x t^{x-1} dt \\ &= 0 + x \int_0^\infty t^{x-1} e^{-t} dt \\ &= x\Gamma(x). \end{aligned} \tag{8.102}$$

I know define a new function $\psi_1(x)$ (also known as the digamma function) as the logarithmic derivative of the gamma function:

$$\psi_1(x) = \frac{d(\ln \Gamma(x))}{dx} = \frac{\Gamma'(x)}{\Gamma(x)}, \tag{8.103}$$

and its value at $x = 1$ defines the Euler constant,

$$\psi_1(1) = \frac{\Gamma'(1)}{\Gamma(1)} \equiv -\gamma_E. \tag{8.104}$$

Taking the derivative of $x\Gamma(x) = \Gamma(x+1)$ I get

$$\Gamma(x) + x\Gamma'(x) = \Gamma'(x+1), \tag{8.105}$$

which gives

$$1 + \frac{x\Gamma'(x)}{\Gamma(x)} = \frac{\Gamma'(x+1)}{\Gamma(x)} = \frac{\Gamma'(x+1)}{\Gamma(x+1)/x} = \frac{x\Gamma'(x+1)}{\Gamma(x+1)}, \tag{8.106}$$

which means

$$1 + x\psi_1(x) = x\psi_1(x+1), \tag{8.107}$$

so that

$$\psi_1(x+1) = \frac{1}{x} + \psi_1(x). \tag{8.108}$$

By repeated use of the last relation I find

$$\begin{aligned}\psi_1(n+1) &= \frac{1}{n} + \psi_1(n) = \frac{1}{n} + \frac{1}{n-1} + \psi_1(n-1) \\ &= \frac{1}{n} + \frac{1}{n-1} + \frac{1}{n-2} + \cdots + \psi_1(1) \\ &= -\gamma_E + \sum_{k=1}^{n} \frac{1}{k} \end{aligned} \tag{8.109}$$

where n is an integer.

I want to take the limit $n \to \infty$, so I use Stirling's approximation (to be proven shortly) which says that for $x \gg 1$

$$\ln\Gamma(x+1) = \left(x + \frac{1}{2}\right)\ln x - x + \frac{1}{2}\ln 2\pi + \mathcal{O}(x^{-1}). \tag{8.110}$$

This gives

$$\frac{d}{dx}\ln\Gamma(x+1) = \ln x + \frac{1}{2x} + \mathcal{O}(x^{-2}), \tag{8.111}$$

which means that as x goes to infinity

$$\psi_1(x+1) = \ln x. \tag{8.112}$$

Going back to equation (8.109) I find that in the limit $n \to \infty$

$$\ln n = -\gamma_E + \sum_{k=1}^{\infty} \frac{1}{k}, \tag{8.113}$$

so I now have an expression for Euler's constant:

$$\gamma_E = \lim_{n\to\infty}\left(\sum_{k=1}^{n}\frac{1}{k} - \ln n\right) = 0.5772\ldots. \tag{8.114}$$

I now Taylor expand Γ around $x = 1$:

$$\begin{aligned}\Gamma(1+\epsilon) &= \Gamma(1) + \epsilon\Gamma'(1) + \mathcal{O}(\epsilon^2) \\ &= 1 + \epsilon\Gamma(1)\psi_1(1) + \mathcal{O}(\epsilon^2) \\ &= 1 - \gamma_E\epsilon + \mathcal{O}(\epsilon^2), \end{aligned} \tag{8.115}$$

and using $\Gamma(1+\epsilon) = \epsilon\Gamma(\epsilon)$, I finally arrive at

$$\Gamma(\epsilon) = \frac{1}{\epsilon} - \gamma_E + \mathcal{O}(\epsilon). \tag{8.116}$$

The final hole to plug is Stirling's approximation. I start by showing *Laplace's approximation*:

$$\int_a^b e^{Mf(x)}dx \approx \sqrt{\frac{2\pi}{M|f''(x_0)|}}e^{Mf(x_0)}, \tag{8.117}$$

as $M \to \infty$ if f has a global maximum at $x = x_0$ in the interval $[a, b]$. If I assume that the latter is the case, I can approximate f around x_0 as

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2, \tag{8.118}$$

and since f has a maximum at x_0, I have $f'(x_0) = 0$ and $f''(x_0) < 0$, so

$$f(x) \approx f(x_0) - \frac{1}{2}|f''(x_0)|(x - x_0)^2, \tag{8.119}$$

The assumption that the maximum is global in ensures that $e^{Mf(x_0)}$ will be more and more strongly peaked around $x = x_0$ as M increases, so that all significant contributions to the integral come from a small interval around this point. For a function of the assumed form, this assumption is accurate already for moderately large values of M. So for large M I can extend the range of the integral from $[a, b]$ to $(-\infty, +\infty)$ without changing its value significantly, since all important contributions will come from a small range of values around x_0, which lies in $[a, b]$. I can therefore write

$$\begin{aligned}
\int_a^b e^{Mf(x)}dx &\approx \int_{-\infty}^{+\infty} e^{M\left[f(x_0)-\frac{1}{2}|f''(x_0)|(x-x_0)^2\right]}dx \\
&= e^{Mf(x_0)}\int_{-\infty}^{+\infty} e^{-M|f''(x_0)|(x-x_0)^2/2}dx \\
&= \sqrt{\frac{2\pi}{M|f''(x_0)|}}e^{Mf(x_0)}, \qquad (8.120)
\end{aligned}$$

where I have used the standard result for a Gaussian integral.

I now recall that

$$\Gamma(N + 1) = N! = \int_0^\infty e^{-x}x^N dx, \tag{8.121}$$

and substitute $x = Nz$:

$$\begin{aligned}
N! &= \int_0^\infty e^{-Nz}N^N z^N N dz \\
&= N^{N+1}\int_0^\infty e^{-Nz}z^N dz \\
&= N^{N+1}\int_0^\infty e^{-Nz}e^{N\ln z}dz \\
&= N^{N+1}\int_0^\infty e^{N(\ln z - z)}dz. \qquad (8.122)
\end{aligned}$$

The function $f(z) = \ln z - z$ goes to $-\infty$ at both endpoints of the interval $[0, \infty)$, and has a global maximum where $f'(z) = \frac{1}{z} - 1 = 0$, that is, $z = z_0 = 1$. The maximum value is $f(z_0) = \ln 1 - 1 = -1$, and $f''(z_0) = -1/z_0^2 = -1$. Plugging this into Laplace's approximation gives

$$\begin{aligned}\Gamma(N+1) = N! &\approx N^{N+1}\sqrt{\frac{2\pi}{N}}e^{-N} \\ &= \sqrt{2\pi N}N^N e^{-N}, \end{aligned} \tag{8.123}$$

so I can write

$$\Gamma(x+1) \approx \sqrt{2\pi x}x^x e^{-x}, \tag{8.124}$$

for $x \gg 1$. In the final step I take the logarithm of this relation:

$$\begin{aligned}\ln\Gamma(x+1) &\approx \ln\sqrt{2\pi} + \ln x^{1/2} + \ln x^x + \ln e^{-x} \\ &= \frac{1}{2}\ln 2\pi + \frac{1}{2}\ln x + x\ln x - x \\ &= \left(x + \frac{1}{2}\right)\ln x - x + \frac{1}{2}\ln 2\pi, \end{aligned} \tag{8.125}$$

and I have derived Stirling's approximation.