Preprint typeset in JHEP style - HYPER VERSION

Michaelmas Term, 2019

Cosmology

University of Cambridge Part II Mathematical Tripos

David Tong

Department of Applied Mathematics and Theoretical Physics,
Centre for Mathematical Sciences,
Wilberforce Road,
Cambridge, CB3 OBA, UK

http://www.damtp.cam.ac.uk/user/tong/cosmo.html
d.tong@damtp.cam.ac.uk

Recommended Books and Resources

Cosmology textbooks sit in one of two camps. The introductory books do a good job of describing the expanding universe, but tend to be less detailed on the hot Big Bang and structure formation. Meanwhile, advanced books which cover these topics assume prior exposure to both general relativity and statistical mechanics. This course sits somewhere between the two.

The first two books below cover the material at an elementary level; the last three are more advanced.

- Barbara Ryden, *Introduction to Cosmology*

A clearly written book that presents an excellent, gentle introduction to the expanding universe, with subsequent chapters on thermal history and structure formation .

- Andrew Liddle *An Introduction to Modern Cosmology*

Another gentle introduction, and one that is especially good when describing expanding spacetimes. However, it becomes more descriptive, and less quantitative, as the subject progresses.

- Scott Dodelson *Modern Cosmology*
- Daniel Baumann *Cosmology*

Both of these are fantastic books: clear, detailed and comprehensive. I have a slight preference for Daniel's book, although in part this is because I'm proud of the blurb that I wrote on the back cover.

- Steven Weinberg *Cosmology*

Weinberg is one of the smarter Nobel prize winners in physics. Here he offers a scholarly account of the subject, devoid of pretty pictures and diagrams, and with a dogged refusal to draw graphs, yet full of clarity and insight.

A number of further lecture notes are available on the web. Links can be found on the course webpage: http://www.damtp.cam.ac.uk/user/tong/cosmo.html

Contents

Acknowledgements

This is an introductory course on cosmology aimed at mathematics undergraduates at the University of Cambridge. You will need to be comfortable with the basics of Special Relativity, but no prior knowledge of either General Relativity or Statistical Mechanics is assumed. In particular, the minimal amount of statistical mechanics will be developed in order to understand what we need. I have made the slightly unusual choice of avoiding all mention of entropy on the grounds that nearly all processes in the early universe are adiabatic and we can, for the most part, get by without it. (The two exceptions are a factor of 4/11 in the cosmic neutrino background and, relatedly, the number of effective relativistic species during nucleosynthesis: for each of these I've quoted, but not derived, the relevant result about entropy.)

I'm very grateful to both Enrico Pajer and Blake Sherwin for explaining many subtle (and less subtle) points to me, and to Daniel Baumann and Alex Considine Tong for encouragement. I'm supported by the Royal Society and by the Simons Foundation.

0. Introduction

All civilisations have an origin myth. We are the first to get it right.

Our origin myth goes by the name of the Big Bang theory. It is a wonderfully evocative name, but one that seeds confusion from the off. The Big Bang theory does not say that the universe started with a bang. In fact, the Big Bang theory has nothing at all to say about the birth of the universe. There is a very simple answer to the question "how did the universe begin?" which is "we don't know".

Instead our origin myth is more modest in scope. It tells us only what the universe was like when it was very much younger. Our story starts from a simple observation: the universe is expanding. This means, of course, that in earlier times everything was closer together. We take this observation and push it to the extreme. As objects are forced closer together, they get hotter. The Big Bang theory postulates that there was a time, in the distant past, when the Universe was so hot that matter, atoms and even nuclei melted and all of space was filled with a fireball. The Big Bang theory is a collection of ideas, calculations and predictions that explain what happened in this fireball, and how it subsequently evolved into the universe we see around us today.

The word "theory" in the Big Bang theory might suggest an element of doubt. This is misleading. The Big Bang theory is a theory in the same way that evolution is a theory. In other words, it happened. We know that the universe was filled with a fireball for a very simple reason: we've seen it. In fact, not only have we seen it, we have taken a photograph of it. Of course, this being science we don't like to brag about these things, so rather than jumping up and down and shouting "we've taken a fucking photograph of the fucking Big Bang", we instead wrap it up in dull technical words. We call it the cosmic microwave background radiation. We may, as a community, have underplayed our hand a little here. The photograph is shown in Figure 1 and contains a wealth of information about what the universe was like when it was much younger.

As we inch further back towards the "$t = 0$" moment, known colloquially but inaccurately as "the Big Bang", the universe gets hotter and energies involved get higher. One of the goals of cosmology is to push back in time as far as possible to get closer to that mysterious "$t = 0$" moment. Progress here has been nothing short of astonishing. As we will learn, we have a very good idea of what was happening a minute or so after the Big Bang, with detailed calculations of the way different elements are forged in the early universe in perfect agreeement with observations. As we go back further, the observational evidence is harder to come by, but our theories of particle physics give us a reasonable level of confidence back to $t = 10^{-12}$ seconds after the Big Bang. As

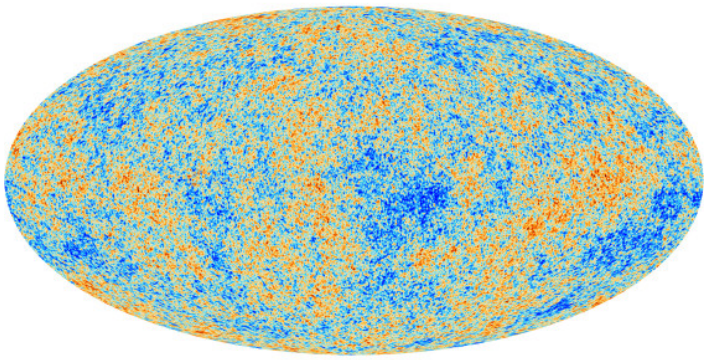

Figure 1: This is a photograph of the Big Bang.

we will see, there are also good reasons to think that, at still earlier times, there was a period of very rapid expansion in the universe known as inflation.

It feels strange to talk with any level of seriousness about the universe when it was a few minutes old, let alone at time $t < 10^{-12}$ seconds. Nonetheless, there are a number of clues surviving in the universe to tell us about these early times, all of which can be explained with impressive accuracy by applying some simple and well tested physical ideas to this most extreme of environments.

The purpose of these lectures is to tell the story above in some detail, to describe 13.8 billion years of history, starting when the Universe was just a fraction of a second old, and extending to the present day.

1. The Expanding Universe

Our goal in this section is an ambitious one: we wish to construct, and then solve, the equations that govern the evolution of the entire universe.

When describing any system in physics, the trick is to focus on the right degrees of freedom. A good choice of variables captures the essence of the problem, while ignoring any irrelevant details. The universe is no different. To motivate our choice, we make the following assumption: the universe is a dull and featureless place. To inject some gravity into this proposal, we elevate it to an important sounding principle:

> The Cosmological Principle: On the largest scales, the universe is spatially homogeneous and isotropic.

Here, *homogeneity* is the property that the universe looks identical at every point in space, while *isotropy* is the property that it looks the same in every direction. Note that the cosmological principle refers only to space. The universe is neither homogenous nor isotropic in time, a fact which underpins this entire course.

Why make this assumption? The primary reason is one of expediency: the universe is, in reality, a complicated place with interesting things happening in it. But these things are discussed in other courses and we will be best served by ignoring them. By averaging over such trifling details, we are left with a description of the universe on the very largest scales, where things are simple.

This averaging ignores little things, like my daily routine, and it is hard to imagine that these have much cosmological significance. However, it also ignores bigger things, like the distribution of galaxies in the universe, that one might think are relevant. Our plan is to proceed with the assumption of simplicity and later, in Section 3, see how we can start to add in some of the details.

The cosmological principle sounds eminently reasonable. Since Copernicus we have known that, while we live in a very special place, we are not at the centre of everything. The cosmological principle allows us to retain our sense of importance by asserting: "if we're not at the centre, then surely no one else is either". You should, however, be suspicious of any grand-sounding principle. Physics is an empirical science and in recent decades we have developed technologies to the point where the cosmological principle can be tested. Fortunately, it stands up pretty well. There are two main pieces of evidence:

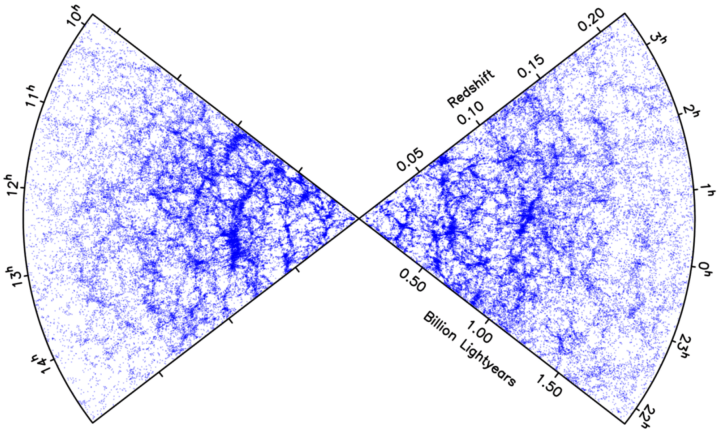

Figure 2: The distribution of galaxies in a wedge in the sky, as measured by the 2dF redshift survey. The distribution looks increasingly smooth on larger scales.

- The cosmic microwave background radiation (CMB) is the afterglow of the Big Bang, an almost uniform sea of photons which fills all of space and provides a snapshot of the universe from almost 14 billion years ago. This is important and will be discussed in more detail in Section 2.2. The temperature of the CMB is[1]

$$T_{\rm CMB} \approx 2.73 \text{ K}$$

  However, it's not quite uniform. There are small fluctuations in temperature with a characteristic scale

$$\frac{\delta T}{T_{\rm CMB}} \sim 10^{-5}$$

  These fluctuations are depicted in the famous photograph shown in Figure 1, taken by the Planck satellite. The fact that the temperature fluctuations are so small is telling us that the early universe was extremely smooth.

- A number of *redshift surveys* have provided a 3d map of hundreds of thousands of galaxies, stretching out to distances of around 2×10^9 light years. The evidence suggests that, while clumpy on small scales, the distribution of galaxies is roughly homogeneous on distances greater than $\sim 3 \times 10^8$ lightyears. An example of such a galaxy survey is shown in Figure 2.

[1]The most accurate determination gives $T_{\rm CMB} = 2.72548 \pm 0.00057$ K; See D.J. Fixsen, "*The Temperature of the Cosmic Microwave Background*", arXiv:0911.1955.

A Sense of Scale

Before we proceed, this is a good time to pause and try to gain some sense of perspective about the universe. First, let's introduce some units. The standard SI units are hopelessly inappropriate for use in cosmology. The metre, for example, is officially defined to be roughly the size of things in my house. Thinking slightly bigger, the average distance from the Earth to the Sun, also known as one *Astronomical Unit* (symbol AU), is

$$1 \text{ AU} \approx 1.5 \times 10^{11} \text{ m}$$

To measure distances of objects that lie beyond our solar system, it's useful to introduce further, farther units. A familiar choice is the lightyear (symbol ly), given by

$$1 \text{ ly} \approx 9.5 \times 10^{15} \text{ m}$$

However, a more commonly used unit among astronomers is the *parsec* (symbol pc), which is based on the observed parallax motion of stars as the Earth orbits the Sun. A parsec is defined as the distance at which a star will exhibit one arcsecond of parallax, which means it wobbles by $1/3600^{\text{th}}$ of a degree in the sky over the course of a year.

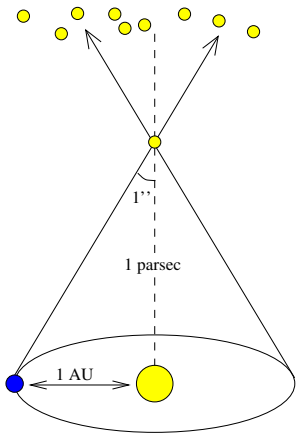

Figure 3: Not to scale.

$$1 \text{ pc} \approx 3.26 \text{ ly}$$

This provides a good unit of measurement to nearby stars. Our closest neighbour, Proxima Centauri, sits at a distance of 1.3 pc. The distance to the centre of our galaxy, the Milky Way, is around 8×10^3 pc, or 8 kpc. Our galaxy is home to around 100 billion stars (give or take) and is approximately 30 kpc across.

There are a large number of neighbouring dwarf galaxies, some of which are actually closer to us than the centre of the Milky Way. But the nearest spiral galaxy is Andromeda, which is approximatey 1 *Megaparsec* (symbol Mpc) or one million parsecs away. The megaparsec is one of the units of choice for cosmologists.

Galaxies are not the largest objects in the universe. They, in turn, gather into clusters and then superclusters and various other filamentary structures. There also appear to be enormous voids in the universe, and it seems plausible that there are more big things to find. Currently, the largest such structures appear to be a few 100 Mpc or so across.

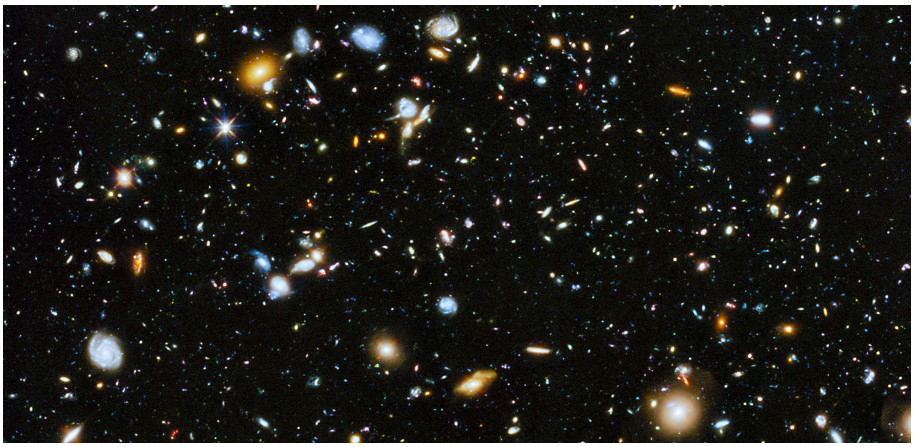

Figure 4: Hubble ultra-deep field shows around 10,000 galaxies.

All of this is to say that we have to look at very large scales before the universe appears to obey the cosmological principle, but it does finally get there. As we will see later in this course, there is a limit to how big we can go. The size of the observable universe is around

$$3000 \text{ Mpc} \approx 10^{26} \text{ m}$$

and there seems to be no way to peer beyond this. The observable universe contains, we think, around 100 billion galaxies, each of them with around 100 billion stars. It is difficult to build intuition for numbers this big, and distances this vast. Some help comes from the Hubble ultra-deep field, shown in Figure 4, which covers a couple of arcminutes of sky, roughly the same as a the tip of a pencil held out at arms length. The image shows around 10,000 galaxies, some no more than a single pixel, but each containing around 100 billion suns, each of which is likely to play host to a solar system of planets.

For more intuition about the size of the universe, we turn to the classics

> "When you're thinking big, think bigger than the biggest thing ever and then some. Much bigger than that in fact, really amazingly immense, a totally stunning size, real 'wow, that's big' time. It's just so big that by comparison, bigness itself looks really titchy. Gigantic multiplied by colossal multiplied by staggeringly huge is the sort of concept we're trying to get across here."
>
> *Douglas Adams*

1.1 The Geometry of Spacetime

The cosmological principle motivates us to treat the universe as a boring, featureless object. Given this, it's not obvious what property of the universe we have left to focus on. The answer is to be found in geometry.

1.1.1 Homogeneous and Isotropic Spaces

The fact that space (and time) can deviate from the seemingly flat geometry of our everyday experience is the essence of the theory general relativity. Fortunately, we will need very little of the full theory for this course. This is, in large part, due to the cosmological principle which allows us to focus on spatial geometries which are homogeneous and isotropic. There are three such geometries:

- **Flat Space:** The simplest homogeneous and isotropic three-dimensional space is flat space, also known as Euclidean space. We will denote it by $\mathbf{R}^3$.

  We describe the geometry of any space in terms of a *metric*. This gives us a prescription for measuring the distance between two points on the space. More precisely, we will specify the metric in terms of the *line element* ds which tells us the infinitesimal distance between two nearby points. For flat space, this is the familiar Euclidean metric

  $$ds^2 = dx^2 + dy^2 + dz^2 \tag{1.1}$$

  We'll also work in a number of other coordinates systems, such as spherical polar coordinates

  $$x = r\sin\theta\cos\phi \quad , \quad y = r\sin\theta\sin\phi \quad , \quad z = r\cos\theta \tag{1.2}$$

  with $r \in [0, \infty)$, $\theta \in [0, \pi]$ and $\phi \in [0, 2\pi)$. To compute the metric in these coordinates, we relate small changes in (r, θ, ϕ) to small changes in (x, y, z) by the Leibniz rule, giving

  $$\begin{aligned} dx &= dr\sin\theta\cos\phi + r\cos\theta\cos\phi\, d\theta - r\sin\theta\sin\phi\, d\phi \\ dy &= dr\sin\theta\sin\phi + r\cos\theta\sin\phi\, d\theta + r\sin\theta\cos\phi\, d\phi \\ dz &= dr\cos\theta - r\sin\theta\, d\theta \end{aligned}$$

  Substituting these expressions into the flat metric (1.1) gives us the flat metric in polar coordinates

  $$ds^2 = dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2) \tag{1.3}$$

- **Positive Curvature** The next homogeneous and isotropic space is also fairly intuitive: we can take a three-dimensional sphere $\mathbf{S}^3$, constructed as an embedding in four-dimensional Euclidean space $\mathbf{R}^4$

$$x^2 + y^2 + z^2 + w^2 = R^2$$

with R the radius of the sphere. The sphere has uniform positive curvature. On such a space, parallel lines will eventually meet.

We again have different choices of coordinates. One option is to retain the 3d spherical polars (1.2) and eliminate w using $w^2 = R^2 - r^2$. A point on the sphere $\mathbf{S}^3$ is then labelled by a "radial" coordinate r, with range $r \in [0, R]$, and the two angular coordinates $\theta \in [0, \pi]$ and $\phi \in [0, 2\pi)$. We can compute the metric on $\mathbf{S}^3$ by noting

$$w^2 = R^2 - r^2 \quad \Rightarrow \quad dw = -\frac{r\, dr}{\sqrt{R^2 - r^2}}$$

The metric on the sphere is then inherited from the flat metric in $\mathbf{R}^4$. We substitute the expression above into the flat metric $ds^2 = dx^2 + dy^2 + dz^2 + dw^2$ to find the metric on $\mathbf{S}^3$,

$$ds^2 = \frac{R^2}{R^2 - r^2} dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2) \tag{1.4}$$

Strictly speaking, this set of coordinates only covers half the $\mathbf{S}^3$, the hemisphere with $w \geq 0$.

Arguably a more natural set of coordinates are provided by the 4d generalisation of the spherical polar coordinates (1.2). These are defined by writing $r = R\sin\chi$, so

$$\begin{aligned} x = R\sin\chi\sin\theta\cos\phi \quad &, \quad y = R\sin\chi\sin\theta\sin\phi \\ z = R\sin\chi\cos\theta \quad &, \quad w = R\cos\chi \end{aligned} \tag{1.5}$$

Now a point on $\mathbf{S}^3$ is determined by three angular coordinates, $\chi, \theta \in [0, \pi]$ and $\phi \in [0, 2\pi)$. The metric becomes

$$ds^2 = R^2\Big[d\chi^2 + \sin^2\chi(d\theta^2 + \sin^2\theta\, d\phi^2)\Big] \tag{1.6}$$

Although we introduced the 3d sphere $\mathbf{S}^3$ by embedding it $\mathbf{R}^4$, the higher dimensional space is a crutch that we no longer need. Worse, it is a crutch that can

be quite misleading. Both mathematically, and physically, the sphere $\mathbf{S}^3$ makes sense on its own without any reference to a space in which it's embedded. In particular, should we discover that the spatial geometry of our universe is $\mathbf{S}^3$, this does not imply the physical existence of some ethereal $\mathbf{R}^4$ in which the universe is floating.

- **Negative Curvature** Our final homogeneous and isotropic space is perhaps the least familiar. It is a hyperboloid $\mathbf{H}^3$, which can again be defined as an embedding in $\mathbf{R}^4$, this time with

$$x^2 + y^2 + z^2 - w^2 = -R^2 \tag{1.7}$$

This is a space of uniform negative curvature. Parallel lines diverge on a space with negative curvature.

Once again, the metric is inherited from the embedding in $\mathbf{R}^4$, but this time with signature $(+++-)$, so $ds^2 = dx^2 + dy^2 + dz^2 - dw^2$ as befits the embedding (1.7). Using the 3d coordinates (r, θ, ϕ), we have $w^2 = r^2 + R^2$. The metric is

$$ds^2 = \frac{R^2}{R^2 + r^2} dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2) \tag{1.8}$$

Alternatively, we can write $r = R\sinh\chi$, in which case the metric becomes

$$ds^2 = R^2\Big[d\chi^2 + \sinh^2\chi(d\theta^2 + \sin^2\theta\, d\phi^2)\Big] \tag{1.9}$$

It is often useful to write these metrics in a unified form. In the (r, θ, ϕ) coordinates, we can write the general metric (1.3), (1.4) and (1.8) as

$$ds^2 = \frac{dr^2}{1 - kr^2/R^2} + r^2(d\theta^2 + \sin^2\theta\, d\phi^2) \quad \text{with } k = \begin{cases} +1 & \text{Spherical} \\ 0 & \text{Euclidean} \\ -1 & \text{Hyperbolic} \end{cases} \tag{1.10}$$

Throughout these lectures, we will use $k = -1, 0, +1$ to denote the three possible spatial geometries. Alternatively, in the coordinates (χ, θ, ϕ), the metrics (1.3), (1.6) and (1.9) can be written in a unified way as

$$ds^2 = R^2\Big[d\chi^2 + S_k^2(\chi)(d\theta^2 + \sin^2\theta\, d\phi^2)\Big] \quad \text{with } S_k(\chi) = \begin{cases} \sin\chi & k = +1 \\ \chi & k = 0 \\ \sinh\chi & k = -1 \end{cases} \tag{1.11}$$

where now χ is a dimensionless coordinate. (In flat space, we have to introduce an arbitrary, fiducial scale R to write the metric in this form.)

Global Topology

We have identified three possible spatial geometries consistent with the cosmological principle. Of these, $\mathbf{S}^3$ is a *compact* space, meaning that it has finite volume (which is $2\pi^2 R^3$). In contrast, both $\mathbf{R}^3$ and $\mathbf{H}^3$ are non-compact, with infinite volume.

In fact, it is straightforward to construct compact spaces for the $k = 0$ and $k = -1$ cases. We simply need to impose periodicity conditions on the coordinates. For example, in the $k = 0$ case we could identify the points $x^i = x^i + R^i$, $i =, 1, 2, 3$ with some fixed R^i. This results in the torus $\mathbf{T}^3$.

Spaces constructed this way are homogenous, but no longer isotropic. For example, on the torus there are special directions that bring you back to where you started on the shortest path. This means that such spaces violate the cosmological principle. More importantly, there is no observational evidence that they do, in fact, describe our universe so we will not discuss them in what follows.

1.1.2 The FRW Metric

Our universe is not three-dimensional. It is four-dimensional, with time as the forth coordinate. In special relativity, we consider the flat four-dimensional spacetime known as Minkowski space, with metric[2]

$$ds^2 = -c^2\, dt^2 + d\mathbf{x}^2$$

with c the speed of light. This metric has the property that the distance between two points in spacetime is invariant under Lorentz transformations; it is the same for all inertial observers.

The Minkowski metric is appropriate for describing physics in some small region of space and time, like the experiments performed here on Earth. But, on cosmological scales, the Minkowski metric needs replacing so that it captures the fact that the universe is expanding. This is straightforward. We replace the flat spatial metric $d\mathbf{x}^2$ with one of the three homogeneous and isotropic metrics that we met in the previous section and write

$$ds^2 = -c^2\, dt^2 + a^2(t)\left[\frac{1}{1 - kr^2/R^2}dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2)\right] \tag{1.12}$$

This is the Friedmann-Robertson-Walker, or FRW metric. The role of the dimensionless *scale factor* $a(t)$ is, as we shall see, to change distances over time.

[2]An introduction to special relativity can be found in Section 7 of the lectures on Dynamics and Relativity. There we used the metric with opposite signature $ds^2 = +c^2\, dt^2 - d\mathbf{x}^2$.

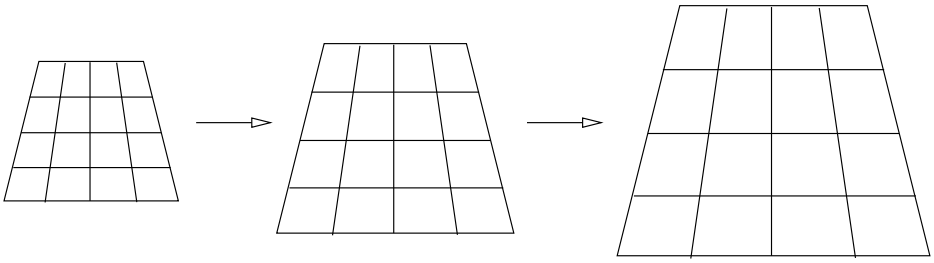

Figure 5: The expansion of the universe. The physical distance between fixed co-moving coordinates increases with time.

There is a redundancy in the description of the metric. If we rescale coordinates as $a \to \lambda a$, $r \to r/\lambda$ and $R \to R/\lambda$ then the metric remains unchanged. We use this to set the scale factor evaluated at the present time t_0 to unity,

$$a_0 = a(t_0) = 1$$

where the subscript 0 will always denote the value of a quantity evaluated today.

Consider a galaxy sitting at some fixed point point (r, θ, ϕ). We refer to the coordinates (r, θ, ϕ) (or, equivalently, (χ, θ, ϕ)) on the 3d space as *co-moving coordinates*. They are analogous to the Lagrangian coordinates used in fluid mechanics. The physical (or proper) distance between the point (r, θ, ϕ) and the origin is then

$$d_{\rm phys} = a(t) \int_0^r \frac{1}{\sqrt{1 - kr'^2/R^2}}\, dr' = a(t) R\chi \tag{1.13}$$

However, there is nothing special about the origin, and the same scaling with $a(t)$ is seen for the distance between any two points. If we choose a function $a(t)$ with $\dot{a} > 0$, then the distance between any two points is increasing. This is the statement that the universe is expanding: two galaxies, at fixed co-moving co-ordinates, will be swept apart as spacetime stretches.

Importantly, the universe isn't expanding "into" anything. Instead, the geometry of spacetime, as described by the metric (1.12), is getting bigger, without reference to anything which sits outside. Similarly, a metric with $\dot{a} < 0$ describes a contracting universe. In Section 1.2, we will introduce the tools needed to calculate $a(t)$. But first, we look at some general features of expanding, or contracting universes.

The FRW metric is not invariant under Lorentz transformation. This means that the universe picks out a preferred rest frame, described by co-moving coordinates. We

can still shift this rest frame by translations (in flat space) or rotations, but not by Lorentz boosts. Consider a galaxy which, in co-moving coordinates, traces a trajectory $\mathbf{x}(t)$. Then, in physical coordinates, the position is

$$\mathbf{x}_{\rm phys}(t) = a(t)\mathbf{x}(t) \tag{1.14}$$

The physical velocity is then

$$\mathbf{v}_{\rm phys}(t) = \frac{d\mathbf{x}_{\rm phys}}{dt} = \frac{da}{dt}\mathbf{x} + a\frac{d\mathbf{x}}{dt} = H\mathbf{x}_{\rm phys} + \mathbf{v}_{\rm pec} \tag{1.15}$$

There are two terms. The first, which is due entirely to the expansion of the universe is written in terms of the *Hubble parameter*,

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

The second term, $\mathbf{v}_{\rm pec}$, is referred to as the *peculiar velocity* and is describes the inherent motion of the galaxy relative to the cosmological frame, typically due to the gravitational attraction of other nearby galaxies.

Our own peculiar velocity is $v_{\rm pec} \approx 400$ km s^{-1} which is pretty much typical for a galaxy. Meanwhile, the present day value of the Hubble parameter is

$$H_0 \approx 70 \text{ km s}^{-1} \text{ Mpc}^{-1}$$

This is, rather misleadingly, referred to as the *Hubble constant*. Clearly there is nothing constant about it. Although, in fairness, it is pretty much the same today as it was yesterday. It is also common to see the notation

$$H_0 = 100h \text{ km s}^{-1} \text{ Mpc}^{-1} \tag{1.16}$$

and then to describe the value of the Hubble constant in terms of the dimensionless number $h \approx 0.7$. In this course, we'll simply use the notation H_0.

The Hubble parameter has dimensions of time^{-1}, but is written in the rather unusual units km s^{-1} Mpc^{-1}. This is telling us that a galaxy 1 Mpc away will be seen to be retreating at a speed of 70 km s^{-1} due to the expansion of space. For nearby galaxies, this tends to be smaller than their peculiar velocity. However, as we look further away, the expansion term will dominate. The numbers above suggest that this will happen at distances around $400/70 \approx 5$ Mpc.

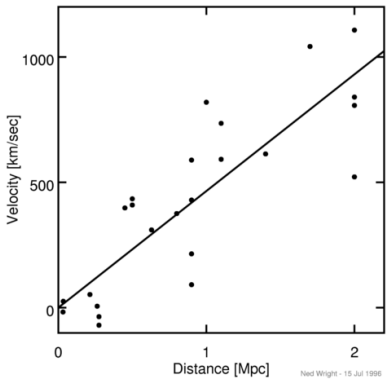

Figure 6: Hubble's original data, from 1929, with a rather optimistic straight line drawn through it.

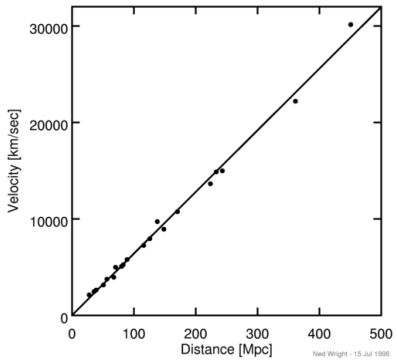

Figure 7: Data from 1996, looking out to much further distances.

If we ignore the peculiar velocities, and further assume that we can approximate the Hubble parameter $H(t)$ as the constant H_0, then the velocity law (1.15) becomes a linear relation between velocity and distance

$$\mathbf{v}_{\rm phys} = H_0\, \mathbf{x}_{\rm phys} \tag{1.17}$$

This linear relationship is referred to as Hubble's law; some data is shown in the figures[3]. At yet further distances, we would expect the time dependence of $H(t)$ to reveal itself. We will discuss this in Section 1.4.

There is no obstacle in (1.17) to velocities that exceed the speed of light, $|\mathbf{v}_{\rm phys}| > c$. This may make you nervous. However, there is no contradiction with relativity and, indeed, the entire framework that we have discussed above sits, without change, in the full theory of general relativity. The statement that "nothing can travel faster than the speed of light" is better thought of as "nothing beats light in a race". Given two objects at the same point, their relative velocity is always less than c. However, the velocity $\mathbf{v}_{\rm phys}$ is measuring the relative velocity of two objects at very distant points and, in an expanding spacetime, there is no such restriction.

1.1.3 Redshift

All our observational information about the universe comes to us through light waves and, more recently, gravitational waves. To correctly interpret what we're seeing, we need to understand how such waves travel in an expanding spacetime.

[3]Both of these plots are taken from Ned Wright's cosmology tutorial.

In a spacetime metric, light travels along null paths with $ds = 0$. In the FRW metric (1.12), light travelling in the radial direction (i.e. with fixed θ and ϕ) will follow a path,

$$c\,dt = \pm a(t)\frac{dr}{\sqrt{1-kr^2/R^2}} \tag{1.18}$$

If we place ourselves at the origin, the minus sign describes light moving towards us. Aliens on a distant planet, tuning in for the latest Buster Keaton movie, should use the plus sign.

Suppose that a distant galaxy sits stationary in co-moving coordinate r_1 and emits light at time t_1. We observe this signal at $r = 0$, at time t_0, determined by solving the integral equation

$$c\int_{t_1}^{t_0}\frac{dt}{a(t)} = \int_0^{r_1}\frac{dr}{\sqrt{1-kr^2/R^2}}$$

If the galaxy emits a second signal at time $t_1 + \delta t_1$, this is observed at $t_0 + \delta t_0$, with

$$c\int_{t_1+\delta t_1}^{t_0+\delta t_0}\frac{dt}{a(t)} = \int_0^{r_1}\frac{dr}{\sqrt{1-kr^2/R^2}}$$

The right-hand side of both of these equations is the same because it is written in co-moving coordinates. We therefore have

$$\int_{t_1+\delta t_1}^{t_0+\delta t_0}\frac{dt}{a(t)} - \int_{t_1}^{t_0}\frac{dt}{a(t)} = 0 \quad\Rightarrow\quad \frac{\delta t_1}{a(t_1)} = \frac{\delta t_0}{a(t_0)} = \delta t_0 \tag{1.19}$$

where, in the last equality, we've used the fact that we observe the signal today, where $a(t_0) = 1$. We see that the expansion of the universe means that the time difference between the two emitted signals differs from the time difference between the two observed signals. This has an important implication when applied to the wave nature of light. Two successive wave crests are separated by a time

$$\delta t_1 = \frac{\lambda_1}{c}$$

with λ_1 the wavelength of the emitted light. Similarly, the time interval between two observed wave crests is

$$\delta t_0 = \frac{\lambda_0}{c}$$

The result (1.19) tells us that the wavelength of the observed light differs from that of the emitted light,

$$\lambda_0 = \frac{a(t_0)}{a(t_1)}\lambda_1 = \frac{\lambda_1}{a(t_1)} \tag{1.20}$$

This is intuitive: the light is stretched by the expansion of space as it travels through it so that the observed wavelength is longer than the emitted wavelength. This effect is known as *cosmological redshift*. It shares some similarity with the Doppler effect, in which the wavelength of light or sound from moving sources is shifted. However, the analogy is not precise: the Doppler effect depends only on the relative velocity of the source and emitter, while the cosmological redshift is independent of $\dot{a}$, instead depending on the overall expansion of space over the light's journey time.

The *redshift parameter* z is defined as the fractional increase in the observed wavelength,

$$z = \frac{\lambda_0 - \lambda_1}{\lambda_1} = \frac{1 - a(t_1)}{a(t_1)} \quad \Rightarrow \quad 1 + z = \frac{1}{a(t_1)} \tag{1.21}$$

As this course progresses, we will often refer to times in the past in terms of the redshift z. Today we sit at $z = 0$. When $z = 1$, the universe was half the current size. When $z = 2$, the universe was one third the current size.

The redshift is something that we can directly measure. Light from far galaxies come with a fingerprint, the spectral absorption lines that reveal the molecular and atomic makeup of the stars within. By comparing the frequencies of those lines to those on Earth, it is a simple matter to extract z. As an aside, by comparing the relative positions of spectral lines, one can also confirm that atomic physics in far flung places works the same as on Earth, with no detected changes in the laws of physics or the fundamental constants of nature.

1.1.4 The Big Bang and Cosmological Horizons

We will find that all our cosmological models predict a time in the past, $t_{BB} < t_0$, where the scale factor vanishes, $a(t_{BB}) = 0$. This point is colloquially referred to as the Big Bang. The Big Bang is not a point in space, but is a point in time. It happens everywhere in space.

We can get an estimate for the age of the Universe by Taylor expanding $a(t)$ about the present day, and truncating at linear order. Recalling that $a(t_0) = 1$, we have

$$a(t) \approx 1 + H_0(t - t_0) \tag{1.22}$$

This rather naive expansion suggests that the Big Bang occurs at

$$t_0 - t_{BB} = H_0^{-1} \approx 4.4 \times 10^{17} \text{ s} \approx 1.4 \times 10^{10} \text{ years} \tag{1.23}$$

This result of 14 billion years is surprisingly close to the currently accepted value of around 13.8 billion years. However, there is a large dose of luck in this agreement, since the linear approximation (1.22) is not very good when extrapolated over the full age of the universe. We'll revisit this in Section 1.4.

Strictly speaking, we should not trust our equations at the point $a(t_{BB}) = 0$. The metric (1.12) is singular here, and any matter in the universe will be squeezed to infinite density. In such a regime, our simple minded classical equations are not to be trusted, and should be replaced by a quantum theory of matter and gravity. Despite much work, it remains an open problem to understand the origin of the universe at $a(t_{BB}) = 0$. Did time begin here? Was there a previous phase of a contracting universe? Did the universe emerge from some earlier, non-geometric form? We simply don't know.

Understanding the Big Bang is one of the ultimate goals of cosmology. In the meantime, the game is to push as far back in time as we can, using the classical (and semi-classical) theory of gravity that we trust. We will be able to reach scales $a \ll 1$, even if we can't get all the way to $a = 0$, and follow the subsequent evolution of the universe from the initial hot, dense state to the world we see today. This set of ideas, is often referred to as the *Big Bang theory*, even though it tells us nothing about the initial "Big Bang" itself.

The Size of the Observable Universe

The existence of a special time, t_{BB}, means that there is a limit as to how far we can peer into the past. In co-moving coordinates, the greatest distance $r_{\rm max}$ that we can see is the distance that light has travelled since the Big Bang. From (1.18), this is given by

$$c\int_{t_{BB}}^{t} \frac{dt'}{a(t')} = \int_0^{r_{\rm max}(t)} \frac{dr}{\sqrt{1-kr^2/R^2}}$$

The corresponding physical distance is

$$d_H(t) = a(t)\int_0^{r_{\rm max}(t)} \frac{dr}{\sqrt{1-kr^2/R^2}} = c\,a(t)\int_0^t \frac{dt'}{a(t')} \tag{1.24}$$

This is the size of the observable universe. Note that this size is not simply $c(t - t_{BB})$, which is the naive distance that light has travelled since the Big Bang. Indeed, mathematically it could be that the integral on the left-hand side of (1.24) does not converge at t_{BB}, in which case the maximum distance $r_{\rm max}$ would be infinite.

The distance d_H is sometimes referred to as the *particle horizon*. The name mimics the event horizon of black holes. Nothing inside the event horizon of a black hole can influence the world outside. Similarly, nothing outside the particle horizon can influence us today.

The Event Horizon

> "It does seem rather odd that two or more observers, even such as sat on the same school bench in the remote past, should in future, when they have followed different paths in life, experience different worlds, so that eventually certain parts of the experienced world of one of them should remain by principle inaccessible to the other and vice versa."
>
> Erwin Schrödinger, 1956

The particle horizon tells us that there are parts of the universe that we cannot presently see. One might expect that, as time progresses, more and more of spacetime comes into view. In fact, this need not be the case.

One option is that the universe begins collapsing in the future, and there is a second time $t_{BC} > t_0$ where $a(t_{BC}) = 0$. This is referred to as the Big Crunch. In this case, there is a limit on how far we can communicate before the universe comes to an end, given by

$$c\int_t^{t_{BC}} \frac{dt'}{a(t')} = \int_0^{r_{\rm max}(t)} \frac{dr}{\sqrt{1-kr^2/R^2}}$$

Perhaps more surprisingly, even if the universe continues to expand and the FRW metric holds for $t \to \infty$, then there could still be a maximum distance that we can influence. The relevant equation is now

$$c\int_t^{\infty} \frac{dt'}{a(t')} = \int_0^{r_{\rm max}(t)} \frac{dr}{\sqrt{1-kr^2/R^2}} \tag{1.25}$$

The maximum co-moving distance $r_{\rm max}$ is finite provided that the left-hand side converges. For example, this happens if we have $a(t) \sim e^{Ht}$ as $t \to \infty$. As we will see later in the course, this seems to be the most likely fate of our universe. As Schrödinger described, it is quite possible that two friends who once played together as children could move apart from each other, only to find that they've travelled too far and can never return as they are inexorably swept further apart by the expansion of the universe. It's not a bad metaphor for life.

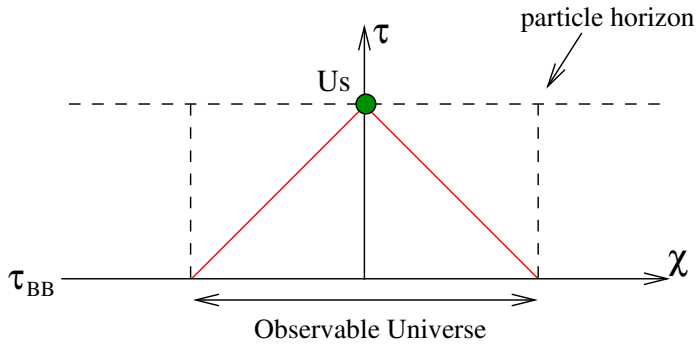

Figure 8: The particle horizon defines the size of your observable universe.

In this context, the distance $r_{\rm max}(t)$ is called the (co-moving) *cosmological event horizon*. Once again, there is the analogy with the black hole. Regions beyond the cosmological horizon are beyond our reach; if we choose to sit still, we will never see them and never communicate with them. However, there are also important distinctions. In contrast to the event horizon of a black hole, the concept of cosmological event horizon depends on the choice of observer.

Conformal Time

The properties of horizons are perhaps best illustrated by introducing a different time coordinate,

$$\tau = \int^t \frac{dt'}{a(t')} \tag{1.26}$$

This is known as *conformal time*. If we also work with the χ spatial coordinate (1.11) then the FRW metric takes the simple form

$$ds^2 = a^2(\tau)\left[-c^2d\tau^2 + R^2d\chi^2 + R^2S_k(\chi)^2(d\theta^2 + \sin^2\, d\phi^2)\right]$$

with all time dependence sitting as on overall factor outside. This has a rather nice consequence because if we draw events in the $(c\tau, R\chi)$ plane then light-rays, which travel with $ds^2 = 0$, correspond to 45° lines, just like in Minkowski space. This helps visualise the causal structure of an expanding universe.

Suppose that we sit at some conformal time τ. A signal can be emitted no earlier than τ_{BB} where the Big Bang singularity occurs. This then puts a restriction on how far we can see in space, defined to be the particle horizon

$$R\chi_{\rm ph} = c(\tau - \tau_{BB})$$

This is shown in Figure 8.

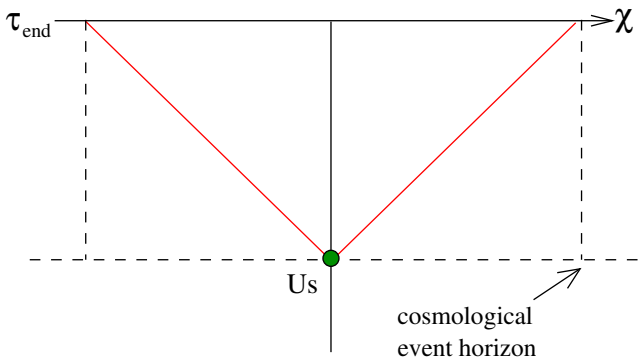

Figure 9: The cosmological event horizon defines the events you can hope to influence.

Looking forward, the issue comes because the end of the universe, at $t \to \infty$, corresponds to a finite conformal time $\tau_{\rm end}$. This means that nothing we can do will be seen beyond a maximum distance which defines the cosmological event horizon,

$$R\chi_{\rm eh} = \tau_{\rm end} - \tau$$

This is shown in Figure 9.

It turns out that conformal time is also a useful change of variable when solving the equations of cosmology. We'll see an example in Section 1.3.2.

1.1.5 Measuring Distance

These lectures are unapologetically theoretical. Nonetheless, we should ask how we know certain facts about the universe. One of the most important challenges facing observational astronomers and cosmologists is the need to accurately determine the distance to various objects in the universe. This is crucial if we are to reconstruct the history of the expansion of the universe $a(t)$.

Furthermore, there is even an ambiguity in what we mean by "distance". So far, we have defined the co-moving distance $R\chi$ and, in (1.13), the physical distance $d_{\rm phys}(t) = a(t)R\chi$. The latter is, as the name suggests, more physical, but it does not equate directly to something we can measure. Instead, $d_{\rm phys}(t)$ is the distance between two events which took place at some fixed time t, but to measure this distance, we would need to pause the expansion of the universe while we wheel out a tape measure, typically one which stretches over several megaparsecs. This, it turns out, is impractical.

Figure 10: This cow is small.

Figure 11: This cow is far away.

For these reasons, we need a more useful definition of distance and how to measure it. A useful measure of distance should involve what we actually see, and what we see is light that has travelled across the universe, sometimes for a long long time.

For objects that are reasonably close, we can use parallax, the slight wobble of a stars position caused by the Earth orbiting the Sun. The current state of the art is the Gaia satellite which can measure the parallax of sufficiently bright star to an accuracy of 2×10^{-5} arc seconds, corresponding to distances of $1/10^{\text{th}}$ of a megaparsec. While impressive this is, to quote the classics, peanuts to space. We therefore need to turn to more indirect methods.

The Luminosity Distance

One way to measure distance is to use the brightness of the object. Obviously, the further away an object is, the less bright it appears in the sky. The problem with this approach is that it's difficult to be sure if an object is genuinely far away, or intrinsically dim. It is entirely analogous to the famous problem with cows: how do we tell if they are small, or merely far away?

To resolve this degeneracy, cosmologists turn to *standard candles*. These are objects whose intrinsic brightness can be determined by other means. There are a number of candidates for standard candles, but some of the most important are:

- Cepheids are bright stars which pulsate with a period ranging from a few days to a month. This periodicity is thought to vary linearly with the intrinsic brightness of the star. These were the standard candles originally used by Hubble.

- A type Ia supernova arises when a white dwarf accretes too much matter from an orbiting companion star, pushing it over the Chandrasekhar limit (the point at which a star collapses). Such events are rare — typically a few a century in a galaxy the size of the Milky Way — but with a brightness that is comparable

to all the stars in the host galaxy. The universal nature of the Chandrasekhar limit means that there is considerable uniformity in these supernovae. What little variation there is can be accounted for by studying the "light curve", meaning how fast the supernova dims after the original burst. These supernovae were first developed as standard candles in the 1990s and resulted in the discovery of the acceleration of the universe.

- The more recent discovery of gravitational waves opens up the possibility for a *standard siren*. The gravitational waveform can be used to accurately determine the distance. When these waves arise from the collision of a neutron star and black hole (sometimes called a kilanova), the event can also be seen in the electromagnetic spectrum, allowing identification of the host galaxy.

Given a standard candle, we can be fairly sure that we know the intrinsic *luminosity* L of an object, defined as the energy emitted per unit time. We would like to determine the apparent luminosity l, defined by the energy per unit time per unit area, seen by a distant observer. In flat space, this is straightforward: at a distance d, the energy has spread out over a sphere $\mathbf{S}^2$ of area $4\pi d^2$, giving us

$$l = \frac{L}{4\pi d^2} \quad \text{in flat space} \tag{1.27}$$

The question we would like to ask is: how does this generalise in an FRW universe? To answer this, it's best to work in the coordinates (1.11), so the FRW metric reads

$$ds^2 = -c^2dt^2 + R^2\Big[d\chi^2 + S_k^2(\chi)(d\theta^2 + \sin^2\theta\, d\phi^2)\Big]$$

with

$$S_k(\chi) = \begin{cases} \sin\chi & k = +1 \\ \chi & k = 0 \\ \sinh\chi & k = -1 \end{cases}$$

There are now three things that we need to take into account. The first is that a sphere $\mathbf{S}^2$ with radius χ now has area $4\pi R^2 S_k(\chi)^2$, which agrees with our previous result in flat space, but differs when $k \neq 0$. Secondly, the photons are redshifted after their long journey. If they are emitted with frequency ν_1 then, from (1.20), they arrive with frequency

$$\nu_0 = \frac{2\pi c}{\lambda_0} = \frac{\nu_1}{1+z}$$

This lower arrival rate decreases the observed flux. Finally, the observed energy E_0 of each photon is reduced compared to the emitted energy E_1,

$$E_0 = \hbar\nu_0 = \frac{E_1}{1+z}$$

The upshot is that, in an expanding universe, the observed flux from a source with intrinsic luminosity L sitting at co-moving distance χ is

$$l = \frac{L}{4\pi R^2 S_k(\chi)^2(1+z)^2}$$

Comparing to (1.27) motivates us to define the *luminosity distance*

$$d_L(\chi) = RS_k(\chi)(1+z) \tag{1.28}$$

For a standard candle, where L is known, the luminosity distance d_L is something that can be measured. From this, and the redshift, we can infer the co-moving distance $RS_k(\chi)$. In flat space, this is simply $R\chi = r$.

Extracting H_0

Finally, we can use this machinery to determine the Hubble constant H_0. We first Taylor expand the scale factor $a(t)$ about the present day. Setting $a_0 = 1$, we have

$$a(t) = 1 + H_0(t-t_0) - \frac{1}{2}q_0 H_0^2(t-t_0)^2 + \ldots \tag{1.29}$$

Here we've introduced the second order term, with dimensionless parameter q_0. This is known as the *deceleration parameter*, and should be thought of as the present day value of the function

$$q(t) = -\frac{\ddot{a}a}{\dot{a}^2} = -\frac{\ddot{a}}{aH^2}$$

The name is rather unfortunate because, as we will learn in Section 1.4, the expansion of our universe is actually accelerating, with $\ddot{a} > 0$! In our universe, the deceleration parameter is negative: $q_0 \approx -0.5$.

First, we integrate the path of a light-ray (1.18) to get an expression for the co-moving distance χ in terms of the "look-back time" $(t_0 - t_1)$

$$\begin{aligned} R\chi = c\int_{t_1}^{t_0}\frac{dt}{a(t)} &= c\int_{t_1}^{t_0}\Big[1 - H_0(t-t_0) + \ldots\Big]dt \\ &= c(t_0-t_1)\Big[1 + \frac{1}{2}H_0(t_0-t_1) + \ldots\Big] \end{aligned} \tag{1.30}$$

Next, we get an expression for the look-back time $t_0 - t_1$ in terms of the redshift z. From (1.21), light emitted at some time t_1 suffers a redshift $1 + z = 1/a(t)$. Inverting the Taylor expansion (1.29), we have

$$z = \frac{1}{a(t_1)} - 1 \approx H_0(t_0 - t_1) + \frac{1}{2}(2 + q_0)H_0^2(t_0 - t_1)^2 + \ldots$$

We now invert this to give the "look-back time" $t_0 - t_1$ as a Taylor expansion in the redshift z. (As an aside: you could do the inversion by solving the quadratic formula, and subsequently Taylor expanding the square-root. But when inverting a power series, it's more straightforward to write an ansatz $H_0(t_0 - t_1) = A_1 z + A_2 z^2 + \ldots$, which we substitute this into the right-hand side and match terms.) We find

$$H_0(t_0 - t_1) = z - \frac{1}{2}(2 + q_0)z^2 + \ldots \tag{1.31}$$

Combining (1.30) and (1.31) gives

$$\frac{H_0 R\chi}{c} = z - \frac{1}{2}(1 + q_0)z^2 + \ldots$$

We can now substitute this into our expression for the luminosity distance (1.28). Life is easiest in flat space, where $RS_k(\chi) = R\chi$ and we find

$$d_L = \frac{c}{H_0}\left(z + \frac{1}{2}(1 - q_0)z^2 + \ldots\right)$$

This expression is valid only for $z \ll 1$. By plotting the observed d_L vs z, and fitting to this functional form, we can extract H_0 and q_0.

1.2 The Dynamics of Spacetime

We have learned that, on the largest distance scales, the universe is described by the FRW metric

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

with the history of the expansion (or contraction) of the universe captured by the function $a(t)$. Our goal now is to calculate this function.

A good maxim for general relativity is: spacetime tells matter how to move, matter tells space how to curve. We saw an example of the first statement in the previous section, with galaxies swept apart by the expansion of spacetime. The second part of the statement tells us that, in turn, the function $a(t)$ is determined by the matter, or more precisely the energy density, in the universe. Here we will first describe the kind of substances that fill the universe and then, in Section 1.2.3 turn to their effect on the expansion.

1.2.1 Perfect Fluids

The cosmological principle guides us to model the contents of the universe as a homogeneous and isotropic fluid. The lumpy, clumpy nature of galaxies that we naively observe is simply a consequence of our small perspective. Viewed from afar, we should think of these galaxies as like atoms in a cosmological fluid. Moreover, as we will learn, the observable galaxies are far from the most dominant energy source in the universe.

We treat all such sources as homogeneous and isotropic perfect fluids. This means that they are characterised by two quantities: the *energy density* $\rho(t)$ and the *pressure* $P(t)$. (If you've taken a course in fluid mechanics, you will be more used to thinking of $\rho(t)$ as the mass density. In the cosmological, or relativistic context, this becomes the total energy density.)

The Equation of State

For any fluid, there is a relation between the energy and pressure, $P = P(\rho)$, known as the *equation of state*.

We will need the equation of state for two, different kinds of fluids. Both of these fluids contain constituent "atoms" of mass m which obey the relativistic energy-momentum relation

$$E^2 = p^2c^2 + m^2c^4 \tag{1.32}$$

The two fluids come from considering this equation in two different regimes:

- Non-Relativistic Limit: $pc \ll mc^2$. Here the energy is dominated by the mass, $E \approx mc^2$, and the velocity of the atoms is $\mathbf{v} \approx \mathbf{p}/m$.
- Relativistic Limit: $pc \gg mc^2$. Now the energy is dominated by the momentum, $E \approx pc$, and the velocity of the atoms approaches the speed of light $|\mathbf{v}| \approx c$.

Suppose that there are N such atoms in a volume V. In general, these atoms will not have a fixed momentum and energy, but instead the number density $n(p)$ will be some distribution. Because the fluid is isotropic, this distribution can depend only on the magnitude of momentum $p = |\mathbf{p}|$. It is normalised by

$$\frac{N}{V} = \int_0^\infty dp\ n(p)$$

The pressure of a gas is defined to be force per unit area. For our purposes, a better definition is the flux of momentum across a surface of unit area. This is equivalent

to the earlier definition because, if the surface is a solid wall, the momentum must be reflected by the wall resulting in a force. However, the "flux" definition can be used anywhere in the fluid, not just at the boundary where there's a wall. Because the fluid is isotropic, we are free to choose this area to be the (x, y)-plane. Then, we have

$$P = \int_0^\infty dp\ v_z p_z n(p)$$

(If this is unfamiliar, an elementary derivation of this formula is given later in Section 2.1.2.) Because $\mathbf{v}$ and $\mathbf{p}$ are parallel, we can write

$$\mathbf{v} \cdot \mathbf{p} = vp = v_x p_x + v_y p_y + v_z p_z = 3 v_z p_z$$

where the final equality is ensured by isotropy. This then gives us

$$P = \frac{1}{3} \int_0^\infty dp\ vp\, n(p) \tag{1.33}$$

Now we can relate this to the energy density in the two cases. First, the non-relativistic gas. In this case, $p \approx mv$ so we have

$$P_{\rm non-rel} \approx \frac{1}{3} \int_0^\infty dp\ mv^2 n(p) = \frac{1}{3} \frac{N}{V} m \langle v^2 \rangle \tag{1.34}$$

where $\langle v^2 \rangle$ is the average square-velocity in the gas.

For cosmological purposes, our interest is in the total energy (1.32) and this is dominated by the contribution from the mass $E \approx mc^2 + \ldots$. If we relate the pressure of a non-relativistic gas to this total energy E, we have

$$P_{\rm non-rel} = \frac{NE}{3V} \frac{\langle v^2 \rangle}{c^2}$$

Since $\langle v^2 \rangle / c^2 \ll 1$, we say that the pressure of a non-relativistic gas is simply

$$P_{\rm non-rel} \approx 0$$

Note that this is the same pressure that keeps balloons afloat and your eardrums healthy: it's not really vanishing. But it is negligible when it comes to its effect on the expansion of the universe. (We will, in fact, revisit this in Section 2 where we'll see that the pressure does give rise to important phenomena in the early universe.)

Cosmologists refer to a non-relativistic gas as *dust*, a name designed to reflect the fact that it just hangs around and is boring. Examples of dust include galaxies, dark matter, and hydrogen atoms floating around and not doing much. We will also refer to dust simply as *matter*.

We can repeat this for a gas of relativistic particles with $v \approx c$ and $E \approx pc$. Now the formula for the pressure (1.33) becomes

$$P_{\rm rel} \approx \frac{1}{3}\int_0^\infty dp\ vp\, n(p) \approx \frac{1}{3}\int_0^\infty dp\ E\, n(p) = \frac{N\langle E\rangle}{3V}$$

with $\langle E\rangle$ the average energy of a particle. The energy density is $\rho = N\langle E\rangle/V$, so the relativistic gas obeys the equation of state

$$P_{\rm rel} = \frac{1}{3}\rho$$

Cosmologists refer to such a relativistic gas as *radiation*. Examples of radiation include the gas of photons known as the cosmic microwave background, gravitational waves, and neutrinos.

Most of the equations of state we meet in cosmology have the simple form

$$P = w\rho \tag{1.35}$$

for some constant w. As we have seen, dust has $w = 0$ and radiation has $w = 1/3$. We will meet other, more exotic fluids as the course progresses.

There is an important restriction on the equation of state. The speed of sound c_s in a fluid is given by

$$c_s^2 = c^2\,\frac{dP}{d\rho}$$

We will derive this formula in Section 3.1.1, but for now we simply quote it. It's important that the speed of sound is less than the speed of light. (Remember: nothing can beat light in a race.) This means that to be consistent with relativity, we must have $w \leq 1$. In fact, the more exotic substances we will meet will have $w < 0$, suggesting an imaginary sound speed. What this is really telling us is that substances with $w < 0$ do not support propagating sound waves, with perturbations decaying exponentially in time.

An Aside: The Equation of State and Temperature

In many other areas of physics, the equation of state is usually written in terms of the temperature T of a fluid. For example, the ideal gas equation relates the pressure P and volume V as

$$PV = Nk_BT \tag{1.36}$$

where N is the number of particles and k_B the Boltzmann constant. (You may have seen this written in chemist's notation $Nk_B = nR$ where n is the number of moles and R the gas constant. Our way is better.) The equations of state that we're interested in can be viewed in this way if we relate T/V to the energy density.

For example, starting from our expression, in (1.34) we derived an expression for the pressure of a non-relativistic gas: $P_{\rm non-rel} \approx Nm\langle v^2\rangle/3V$. This coincides with the ideal gas law if we relate the temperature to the average kinetic energy of an atom in the gas through

$$\frac{1}{2}m\langle v^2\rangle = \frac{3}{2}k_B T \tag{1.37}$$

We will revisit this in Section 2.1 and gain a better understanding of this result and the role played by temperature.

1.2.2 The Continuity Equation

As the universe expands, we expect the energy density (of any sensible fluid) to dilute. The way this happens is dictated by the conservation of energy, also known as the continuity equation.

A proper discussion of the continuity equation requires the machinery of general relativity. This is one of a number of places were we will revert to some simple Newtonian thinking to derive the correct equation. Such derivations are not entirely convincing, not least because it's unclear why they would be valid when applied to the entire universe. Nonetheless, they will give the correct answer. A more rigorous approach can be found in the lectures on General Relativity.

Consider a gas trapped in a box of volume V. The gas exerts pressure on the sides of the box. If the box increases in size, as shown in the figure, then the change of volume is $dV = \text{Area} \times dx$. The work done by the gas is Force $\times\, dx = (PA)dx = P\,dV$, and this reduces the internal energy of the gas. We have

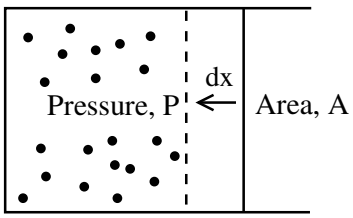

Figure 12:

$$dE = -P\,dV$$

This is a simple form of the *first law of thermodynamics*, valid for reversible or adiabatic processes. It is far from obvious that we can view the universe as a box filled with gas and naively apply this formula. Nonetheless, it happily turns out that the final result

agrees with the more rigorous GR approach so we will push ahead, and invoke the time dependent version of the first law,

$$\frac{dE}{dt} = -P\frac{dV}{dt} \tag{1.38}$$

Now consider a small region of fluid, in co-moving volume V_0. The physical volume is

$$V(t) = a^3(t)V_0 \quad \Rightarrow \quad \frac{dV}{dt} = 3a^2\dot{a}V_0$$

Meanwhile, the energy in this volume is

$$E = \rho a^3 V_0 \quad \Rightarrow \quad \frac{dE}{dt} = \dot{\rho}a^3V_0 + 3\rho a^2\dot{a}V_0$$

The first law (1.38) then becomes

$$\dot{\rho} + 3H\left(\rho + P\right) = 0 \tag{1.39}$$

This slightly unfamiliar equation is the expression of energy conservation in a cosmological setting.

Before we proceed, a warning: energy is a famously slippery concept in general relativity, and we will meet things later which, taken naively, would seem to violate energy conservation. For example, in Section 1.3.3, we will meet a fluid with equation of state $\rho = -P$. For such a fluid, $\dot{\rho} = 0$ which means that the energy density remains constant even as the universe expands. Such is the way of the world and we need to get used to it. If this makes you nervous, recall that the usual derivation of energy conservation, via Noether's theorem, holds only in time independent settings. So perhaps it's not so surprising that energy conservation takes a somewhat different form in an expanding universe.

If we specify an equation of state $P = w\rho$, as in (1.35), then we can integrate the continuity equation (1.39) to determine how the energy density depends on the scale factor. We have

$$\begin{aligned} \frac{\dot{\rho}}{\rho} = -3(1+w)\frac{\dot{a}}{a} \quad &\Rightarrow \quad \log(\rho/\rho_0) = -3(1+w)\log\, a \\ &\Rightarrow \quad \rho(t) = \rho_0 a^{-3(1+w)} \end{aligned} \tag{1.40}$$

with $\rho_0 = \rho(t_0)$ and we've used the fact that $a(t_0) = 1$.

We can look at how this behaves in simple examples. For dust (also known as matter), we have $w = 0$ and so

$$\rho_m \sim \frac{1}{a^3}$$

This makes sense. As the universe expands, the volume increases as a^3, and so the energy density decreases as $1/a^3$.

For radiation, we instead find

$$\rho_r \sim \frac{1}{a^4} \tag{1.41}$$

This also makes sense. The energy density is diluted as $1/a^3$ but, on top of this, there is also a redshift effect which shifts the frequency, and hence the energy, by a further power of $1/a$.

The fact that the energy densities of dust and radiation scale differently plays a crucial role in our cosmological history. As we shall see in Section 1.4, our current universe has much greater energy density in dust than in radiation. However, this wasn't always the case. There was a time in far past when the converse was true, with the radiation subsequently diluting away faster. We'll see other contributions to the energy density of the universe that have yet different behaviour.

1.2.3 The Friedmann Equation

> "Friedmann more than once said that his task was to indicate the possible solutions of Einstein's equations, and that the physicists could do what they wished with these solutions"
>
> Vladimir Fock, on his friend Alexander Friedmann

Finally we come to the main part of the story: we would like to describe how the perfect fluids which fill all of space affect the expansion of the universe. We start by giving the answer. The dynamics of the scale factor is dictated by the energy density $\rho(t)$ through the *Friedmann equation*

$$H^2 \equiv \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3c^2}\rho - \frac{kc^2}{R^2a^2} \tag{1.42}$$

Here R is some fixed scale, as in the FRW metric (1.12), $k = -1, 0, +1$ determines the curvature of space, and G is Newton's gravitational constant

$$G \approx 6.67 \times 10^{-11}\ \mathrm{m^3\,kg^{-1}\,s^{-2}}$$

The Friedmann equation is arguably the most important equation in all of cosmology. Taken together with the continuity equation (1.39) and the equation of state (1.35), they provide a closed system which can be solved to determine the history and fate of the universe itself.

At this point, I have a confession to make. The only honest derivation of the Friedmann equation is in the framework of General Relativity Here we can only present a dishonest derivation, using Newtonian ideas. In an attempt to alleviate the shame, I will at least be open about where the arguments are at their weakest.

First, we work in flat space, with $k = 0$. This, of course, is the natural habitat for Newtonian gravity. Nonetheless, we will see the possibility of a curvature term $-k/a^2$ in the Friedmann equation, re-emerging at the end of our derivation.

Our discussions so far prompt us to consider an infinite universe, filled with a constant matter density. That, it turns out, is rather subtle in a Newtonian setting. Instead, we consider a ball of uniform density of size L, expanding outwards away from the origin, and subsequently pretend that we can take $L \to \infty$.

Consider a particle (or element of fluid) of mass m at some position $\mathbf{x}$ with $r = |\mathbf{x}| \ll L$. It will experience the force of gravity in the form of Newton's inverse-square law. But a rather special property of this law states that, for a spherically symmetric distribution of masses, the gravitational force at some point $\mathbf{x}$ depends only on the masses at distances smaller than r and, moreover, acts as if all the mass is concentrated at the origin.

This statement is simplest to prove if we formulate the gravitational force law as a kind of Gauss' law,

$$\mathbf{F}_{\rm grav} = -m\nabla\Phi \quad \text{where} \quad \nabla^2\Phi = \frac{4\pi G}{c^2}\rho$$

with Φ the gravitational potential. The (perhaps) unfamiliar factor of c^2 in the final equation arises because, for us, ρ is the energy density, rather than mass density. We then integrate both sides over a ball V of radius x, centred at the origin. Using the kind of symmetry arguments that we used extensively in the lectures on Electromagnetism, we have

$$\int_S \nabla\Phi \cdot d\mathbf{S} = \int_V \frac{4\pi G}{c^2}\rho\, dV \quad \Rightarrow \quad \nabla\Phi(r) = \frac{GM(r)}{r^2}$$

where $M(r) = 4\pi\rho r^3/3c^2$ is the mass contained inside the ball of radius r. This means that the acceleration of the particle at $\mathbf{x}$ is given by

$$m\ddot{r} = -\frac{GmM(r)}{r^2}$$

We multiply by $\dot{r}$ and integrate. As the ball expands with $\dot{r} \neq 0$, the total mass contained with a ball of radius $r(t)$ does not change, so $\dot{M} = 0$. We then get

$$\frac{1}{2}\dot{r}^2 - \frac{GM(r)}{r} = E \tag{1.43}$$

where we recognise E as the energy (per unit mass) of the particle. Finally, we describe the position $\mathbf{x}$ of the particle in a way that chimes with our previous cosmological discussion, introducing a scale factor $a(t)$

$$\mathbf{x}(t) = a(t)\mathbf{x}_0$$

Substituting this into (1.43) and rearranging gives

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3c^2}\rho - \frac{C}{a^2} \tag{1.44}$$

where $C = -2E/|\mathbf{x}_0|^2$ is a constant. This is remarkably close to the Friedmann equation (1.42). The only remaining issue is why we should identify the constant C with the curvature kc^2/R^2. There is no good argument here and, indeed, we shouldn't expect one given that the whole Newtonian derivation took place in a flat space. It is, unfortunately, simply something that you have to suck up.

There is, however, an analogy which makes the identification $C \sim k$ marginally more palatable. Recall that a particle has reached escape velocity if its total energy $E > 0$. Conversely, if $E < 0$, the particle comes crashing back down. For us, the case of $E < 0$ means $C > 0$ which, in turn, corresponds to positive curvature. We will see in Section 1.3.2 that a universe with positive curvature will, under many circumstances, ultimately suffer a big crunch. In contrast, a negatively curved space $k < 0$ will keep expanding forever.

Clearly the derivation above is far from rigorous. There are at least two aspects that should give us pause. First, when we assumed $\dot{M} = 0$, we were implicitly restricting ourselves to non-relativistic matter with $\rho \sim 1/a^3$. It turns out that in general relativity, the Friedmann equation also holds for any other scaling (1.40) of ρ.

However, the part of the above story that should make you feel most queasy is replacing an infinitely expanding universe, with an expanding ball of finite size L. This introduces an origin into the story, and gives a very misleading impression of what the expansion of the universe means. In particular, if we dial the clock back to $a(t) = 0$ in this scenario, then all matter sits at the origin. This is one of the most popular misconceptions about the Big Bang and it is deeply unfortunate that it is reinforced by the derivation above. Nonetheless, the arguments that lead to (1.44) do provide some physical insight into the meaning of the various terms that can be hard to extract from the more formal derivation using general relativity. So let us wash the distaste from our mouths, and proceed with understanding the universe.

1.3 Cosmological Solutions

We now have a closed set of equations that describe the evolution of the universe. These are the Friedmann equation,

$$H^2 \equiv \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3c^2}\rho - \frac{kc^2}{R^2a^2} \tag{1.45}$$

the continuity equation,

$$\dot{\rho} + 3H\left(\rho + P\right) = 0$$

and the equation of state

$$P = w\rho$$

In this section, we will solve them. Our initial interest will be on a number of designer universes whose solutions are particularly simple. Then, in Section 1.4, we describe the solutions of relevance to our universe.

1.3.1 Simple Solutions

To solve the Friedmann equation, we first need to decide what fluids live in our universe. In general, there will be several different fluids. If they share the same equation of state (e.g. dark matter and visible matter) then we can, for cosmological purposes, just treat them as one. However, if the universe contains fluids with different equations of state, we must include them all. In this case, we write

$$\rho = \sum_w \rho_w$$

As we have seen in (1.40), each component scales independently as

$$\rho_w = \frac{\rho_{w,0}}{a^{3(1+w)}} \tag{1.46}$$

where $\rho_{w,0} = \rho_w(t_0)$. Substituting this into the Friedmann equation then leaves us with a tricky-looking non-linear differential equation for a.

Life is considerably simpler if we restrict attention to a flat $k = 0$ universe with just a single fluid component. In this case, using (1.46), we have

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{D^2}{a^{3(1+w)}} \tag{1.47}$$

where $D^2 = 8\pi G\rho_{w,0}/3c^2$ is a constant. The solution is

$$a(t) = \left(\frac{t}{t_0}\right)^{2/(3+3w)} \tag{1.48}$$

The various constants have been massaged into $t_0 = (\frac{3}{2}(1+w)D)^{-1}$ so that we recover our convention $a_0 = a(t_0) = 1$. There is also an integration constant which we have set to zero. This corresponds to picking the time of the Big Bang, defined by $a(t_{BB}) = 0$ to be $t_{BB} = 0$. With this choice, t_0 is identified with the age of the universe.

Let's look at this solution in a number of important cases

- Dust ($w = 0$): For a flat universe filled with dust-like matter (i.e. galaxies, or cold dark matter), we have

  $$a(t) = \left(\frac{t}{t_0}\right)^{2/3} \tag{1.49}$$

  This is known as the *Einstein-de Sitter universe* (not to be confused with either the Einstein universe or the de Sitter universe, both of which we shall meet in Section 1.3.3). The exponent 2/3 is the same 2/3 that appears in Kepler's third law: the radius R of a planet's orbit is related to its period by $R \sim T^{2/3}$. Both follow by simple dimensional analysis in Newtonian gravity.

  The Hubble constant is

  $$H_0 = \frac{2}{3}\frac{1}{t_0}$$

  If we lived in such a place, then a measurement of H_0 would immediately tell us the age of the universe $t_0 = \frac{2}{3}H_0^{-1}$. Using the observed value of $H_0 \approx 70\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ gives

  $$t_0 \approx 9 \times 10^9 \text{ years} \tag{1.50}$$

The extra factor of 2/3 brings us down from the earlier estimate of 14 billion years in (1.23) to 9 billion years. This is problematic since there are stars in the universe that appear to be older than this.

Finally note that in the Einstein-de Sitter universe the matter density scales as

$$\rho(t) = \frac{c^2}{6\pi G}\frac{1}{t^2} \tag{1.51}$$

In particular, there is a direct relationship between the age of the universe and the present day matter density. We'll revisit this relationship later.

- Radiation ($w = 1/3$): For a flat universe filled with radiation (e.g. light), we have

$$a(t) = \left(\frac{t}{t_0}\right)^{1/2}$$

Once again, there is a direct relation between the Hubble constant and the age of the universe, now given by $t_0 = \frac{1}{2}H_0^{-1}$. In a radiation dominated universe, the energy density scales as

$$\rho(t) = \frac{3c^2}{32\pi G}\frac{1}{t^2}$$

- Curvature ($w = -1/3$): We can also apply the calculation above to a universe with curvature a term, which is devoid of any matter. Indeed, the curvature term in (1.45) acts just like a fluid (1.46) with $w = -1/3$. In the absence of any further fluid contributions, the Friedmann equation only has solutions for a negatively curved universe, with $k = -1$. In this case,

$$a(t) = \frac{t}{t_0}$$

This is known as the *Milne universe*.

A Comment on Multi-Component Solutions

If the universe has more than one type of fluid (or a fluid and some curvature) then it is more tricky to write down analytic solutions to the Friedmann equations. Nonetheless, we can build intuition for these solutions using our results above, together with the observation that different fluids dilute away at different rates. For example, we have seen that

$$\rho_m \sim \frac{1}{a^3} \quad \text{and} \quad \rho_r \sim \frac{1}{a^4}$$

This means means that, in a universe with both dust and radiation (like the one we call home) there will be a period in the past, when a is suitably small, when we necessarily have $\rho_r \gg \rho_m$. As a increases there will be a time when the energy density of the two are roughly comparable, before we go over to another era with $\rho_m \gg \rho_r$. In this way, the history of the universe is divided into different epochs. When one form of energy density dominates over the other, the expansion of the universe is well-approximated by the single-component solutions we met above .

The Big Bang Revisited: A Baby Singularity Theorem

All of the solutions we met above have a Big Bang, where $a = 0$. It is natural to ask: is this a generic feature of the Friedmann equation with arbitrary matter and curvature?

Within the larger framework of general relativity, there are a number of important theorems which state that, under certain circumstances, singularities in the metric necessarily arise. The original theorems, due to Penrose (for black holes) and Hawking (for the Big Bang), are tour-de-force pieces of mathematical physics. You can learn about them next year. Here we present a simple Mickey mouse version of the singularity theorem for the Friedmann equation.

We start with the Friedmann equation, written as

$$\dot{a}^2 = \frac{8\pi G}{3c^2}\rho a^2 - \frac{kc^2}{R^2}$$

Differentiating both sides with respect to time gives

$$2\ddot{a}\dot{a} = \frac{8\pi G}{3c^2}\left(\dot{\rho}a^2 + 2\rho\dot{a}a\right) = \frac{8\pi G}{3c^2}\left(-3\dot{a}a(\rho + P) + 2\rho\dot{a}a\right)$$

where, in the second equality, we have used the continuity equation $\dot{\rho} + 3H(\rho + P) = 0$ Rearranging gives the *acceleration equation*

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3c^2}(\rho + 3P) \tag{1.52}$$

This is also known as the *Raychaudhuri equation* and will be useful in a number of places in this course. (It is a special case of the real Raychaudhuri equation, which has application beyond cosmology.) Using this result, we can prove the following:

Claim: If matter obeys the *strong energy condition*

$$\rho + 3P \geq 0 \tag{1.53}$$

then there was a singularity at a finite time t_{BB} in the past where $a(t_{BB}) = 0$. Furthermore, $t_0 - t_{BB} \leq H_0^{-1}$.

Proof: The strong energy condition immediately tells us that $\ddot{a}/a \leq 0$. This is the statement that the universe is decelerating, meaning that it must have been expanding faster in the past.

Suppose first that $\ddot{a} = 0$. In this case we must have $a(t) = H_0 t + \text{const}$. (We have used the fact that $H_0 = \dot{a}_0$ since $a_0 = 1$). This is the dotted line shown in the figure. If this is the case, the Big Bang occurs at $t_0 - t_{BB} = H_0^{-1}$. But the strong energy condition ensures that $\ddot{a} \leq 0$, so the dotted line in the figure provides an upper bound on the scale factor. In such a universe, the Big Bang must occur at $t_0 - t_{BB} \leq H_0^{-1}$. □

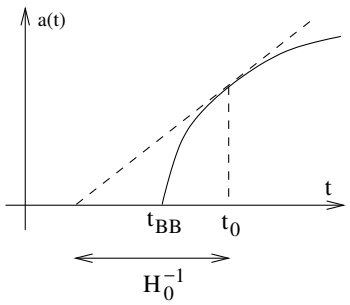

Figure 13:

The proof above is so simple because we have restricted attention to the homogeneous and isotropic FRW universe. Hawking's singularity theorem (proven in his PhD thesis) shows the necessity of a singularity even in the absence of such assumptions.

The strong energy condition is obeyed by all conventional matter, including dust and radiation. However, it's not hard to find substances which violate it, and we shall meet examples as we go along. When the strong energy condition is violated, we have an accelerating universe with $\ddot{a} > 0$. In this case, the single component solutions (1.48) still have a Big Bang singularity. However, the argument above cannot rule out the possibility of more complicated solutions which avoid this.

The Future Revisited: Cosmological Event Horizons

Recall from section 1.1.4 the idea of an event horizon: for certain universes, it may be that our friends in distant galaxies get swept away from us by the expansion of space and are lost to us forever. At a time t, the furthest distance with which we can communicate, $r_{\rm max}$ is governed by the equation (1.25)

$$c\int_t^\infty \frac{dt'}{a(t')} = \int_0^{r_{\rm max}(t)} \frac{dr}{\sqrt{1-kr^2/R^2}}$$

If the integral on the left converges then $r_{\rm max}$ is finite and there is a cosmological horizon.

When does this happen? If the late time universe is dominated by a single component with expansion given by $a \sim t^{2/(3+3w)}$ as in (1.48) then

$$\int \frac{dt}{a(t)} \sim \int \frac{dt}{t^{2/(3+3w)}} \sim t^{(3w+1)/(3w+3)}$$

For $w \geq -1/3$, the integral diverges and there is no event horizon. (In the limiting case of $w = -1/3$, the integral is replaced by $\log t$.) For $-1 \leq w < -1/3$, the integral converges and there is a horizon.

Fluids with $w < -1/3$ are precisely those which violate the strong energy condition (1.53). We learn that cosmological event horizons arise whenever the late time expansion of the universe is accelerating, rather than decelerating.

1.3.2 Curvature and the Fate of the Universe

Let's look again at a flat universe, with $k = 0$. The Friedmann equation (1.45) tells us that for such a universe to exist, something rather special has to happen, because the energy density of the universe today ρ_0 has to be precisely correlated with the Hubble constant

$$H_0^2 = \frac{8\pi G}{3c^2}\rho_0$$

We saw such behaviour in our earlier solutions. For example, this led us to the result (1.51) which relates the energy density of an Einstein-de Sitter universe to the current age of the universe.

In principle, this gives a straightforward way to test whether the universe is flat. First, you measure the expansion rate as seen in H_0. Then you add up all the energy in the universe and see if they match. In practice, this isn't possible because, as we shall see, much of the energy in the universe is invisible.

What happens if we have a universe with some small curvature and, say, a large amount of conventional matter with $w = 0$? We can think of the curvature term in the Friedmann equation as simply another contribution to the energy density, ρ_k, one which dilutes away more slowly that the matter contribution,

$$\rho_m \sim \frac{1}{a^3} \quad \text{and} \quad \rho_k \sim \frac{1}{a^2}$$

This tells us that, regardless of their initial values, if we wait long enough then the curvature of space will eventually come to dominate the dynamics.

If we start with $\rho_m > \rho_k$, then there will be a moment when the two are equal, meaning

$$\frac{8\pi G}{3c^2}\rho_m = \frac{|k|c^2}{R^2a^2}$$

For a negatively curved universe, with $k = -1$, the Friedmann equation (1.45) gives $\dot{a} > 0$. However, for a positively curved universe, with $k = +1$, we find $\dot{a} = 0$ at the moment of equality. In other words, the universe stops expanding. In fact, as we now see, such a positively curved universe subsequently contracts until it hits a big crunch.

Perhaps surprisingly, it is possible to find an exact solution to the Friedmann equation with both matter and curvature. To do this, it is useful to work in conformal time (1.26), defined by

$$\tau(t) = \int_0^t \frac{dt'}{a(t')} \quad \Rightarrow \quad \frac{d\tau}{dt} = \frac{1}{a} \tag{1.54}$$

We further define the dimensionless time coordinate $\tilde{\tau} = c\tau/R$. (In flat space, with $k = 0$, just pick a choice for R; it will drop out in what follows.) Finally, we define

$$h = \frac{a'}{a} \quad \text{with} \quad a' = \frac{da}{d\tilde{\tau}}$$

In these variables, one can check that the Friedmann equation (1.45) becomes

$$h^2 + k = \frac{8\pi G R^2}{3c^4}\rho a^2 \tag{1.55}$$

Rather than solve this in conjunction with the continuity equation, it turns out to be more straightforward to look at the acceleration equation (1.52). A little algebra shows that, for matter with $P = 0$, the acceleration equation becomes

$$h' = -\frac{4\pi G R^2}{3c^4}\rho a^2 \quad \Rightarrow \quad 2h' + h^2 + k = 0 \tag{1.56}$$

where, to get the second equation, we have simply used (1.55). Happily this latter equation is independent of ρ and we can go ahead and solve it. The solutions are:

$$h(\tilde{\tau}) = \begin{cases} \cot(\tilde{\tau}/2) & k = +1 \\ 2/\tilde{\tau} & k = 0 \\ \coth(\tilde{\tau}/2) & k = -1 \end{cases}$$

We can then solve $h = a'/a$ to derive an expression for the scale factor $a(\tilde{\tau})$ as a function of $\tilde{\tau}$,

$$a(\tilde{\tau}) = A \times \begin{cases} \sin^2(\tilde{\tau}/2) & k = +1 \\ \tilde{\tau}^2 & k = 0 \\ \sinh^2(\tilde{\tau}/2) & k = -1 \end{cases} \tag{1.57}$$

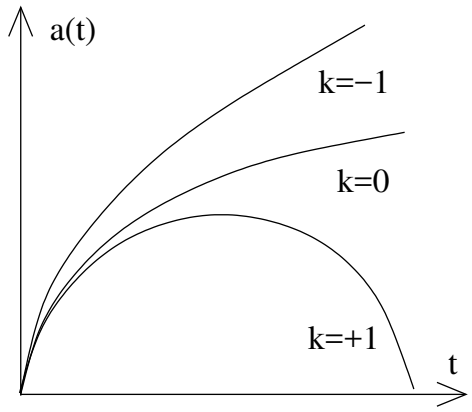

Figure 14: The FRW scale factor for a matter dominated universe with curvature.

with A an integration constant. We see that, as advertised, the positively curved $k = 1$ universe eventually re-collapses, with the Big Crunch occurring at conformal time $\tau = 2\pi R/c$. In contrast, the negatively curved $k = -1$ universe expands for ever. The flat space $k = 0$ separates these two behaviours.

Finally, we can use the solution for the scale factor to determine how conformal time (1.54) scales with our original time coordinate t,

$$t = \frac{RA}{2c} \times \begin{cases} \tilde{\tau} - \sin\tilde{\tau} & k = +1 \\ \frac{2}{3}\tilde{\tau}^3 & k = 0 \\ \sinh\tilde{\tau} - \tilde{\tau} & k = -1 \end{cases} \tag{1.58}$$

In the $k = 0$ case, this reproduces our previous result (1.49) for the expansion of the Einstein-de Sitter universe. The resulting scale factors $a(t)$ are sketched in Figure 14.

There are a couple of lessons to take from this calculation. The first is that a flat universe is dynamically unstable, rather like a pencil balancing on its tip. Any small initial curvature will grow and dominate the late time behaviour.

The second lesson comes with an important caveat. The result above suggests that a measurement of curvature of the space will tell us the ultimate fate of the universe. If we find $k = 1$, then we are doomed to suffer a Big Crunch. On the other hand, a curvature of $k = -1$ or $k = 0$ means that universe expands for ever, becoming increasingly desolate and lonely. However, this conclusion relies on the assumption that the dominant energy in the universe is matter. In fact, it's not hard to show that the conclusion is unaltered provided that all energies in the universe dilute away faster

than the curvature. However, as we will now see, there are more exotic fluids at play in the universe for which the conclusion does not hold.

1.3.3 The Cosmological Constant

The final entry in the dictionary of cosmological fluids is both the most strange and, in some ways, the most natural. A *cosmological constant* is a fluid with equation of state $w = -1$. The associated energy density is denoted ρ_Λ and obeys

$$\rho_\Lambda = -P$$

First the strange. The continuity equation (1.39) tells us that such an energy density remains constant over time: $\rho_\Lambda \sim a^0$. Naively, that would seem to violate the conservation of energy. However, as stressed previously, energy is a rather slippery concept in an expanding universe and the only thing that we have to worry about is the continuity equation (1.39) which is happily obeyed. So this is something we will just have to live with. For now, note that any universe with $\rho_\Lambda \neq 0$ will ultimately become dominated by the cosmological constant, as all other energy sources dilute away.

Now the natural. The cosmological constant is something that you've seen before. Recall that whenever you write down the energy of a system, any overall constant shift of the energy is unimportant and does not affect the physics. For example, in classical mechanics if we have a potential $V(\mathbf{x})$, then the force is $\mathbf{F} = -\nabla V$ which cares nothing about the constant term in V. Similarly, in quantum mechanics we work with the Hamiltonian H, and adding an overall constant is irrelevant for the physics. However, when we get to general relativity, it becomes time to pay the piper. In the context of general relativity, all energy gravitates, including the constant energy that we previously neglected. And the way this constant manifests itself is as a cosmological constant. For this reason, the cosmological constant is also referred to as *vacuum energy*.

Strictly speaking, ρ_Λ is the vacuum energy density, while the cosmological constant Λ is defined as

$$\rho_\Lambda = \frac{\Lambda c^2}{8\pi G}$$

so Λ has dimensions of (time)$^{-2}$. (Usually, by the time people get to describing the cosmological constant, they have long set $c = 1$, so other definitions may differ by hidden factors of c.) Here we will treat the terms "cosmological constant" and "vacuum

energy" as synonymous. In the presence of a cosmological constant and other matter, the Friedmann equation becomes

$$H^2 = \frac{8\pi G}{3c^2}\rho + \frac{\Lambda}{3} - \frac{kc^2}{R^2a^2} \tag{1.59}$$

We will shortly solve this in various cases. Before we do, we pause to ask a slightly oblique question. How does the cosmological constant appear in our Newtonian analysis of Section 1.2.3? To see this, we indulge in a little bit of answer analysis and work backwards. You can check that the steps that previously took us from Newton's law of motion (1.43) to the Friedmann equation (1.44), now require that we start with Newton's law in the form

$$\frac{1}{2}\dot{r}^2 - \frac{GM(r)}{r} - \frac{1}{6}\Lambda r^2 = E \tag{1.60}$$

In other words, the cosmological constant acts like a harmonic oscillator, with potential $V(r) = -\frac{1}{6}\Lambda r^2$. For $\Lambda > 0$ this is a an inverted harmonic oscillator and our (admittedly slightly dodgy) Newtonian analysis suggests that particles will race off to $r \to \infty$. Meanwhile, for $\Lambda < 0$ we have a standard harmonic oscillator, which suggests that particles will be trapped. We'll now see that, suitably interpreted, this is not a bad way to think about the cosmological constant.

de Sitter Space

First, consider a universe with positive cosmological constant $\Lambda > 0$. If we empty it of all other matter, so that $\rho = 0$, then we can solve the Friedmann equation for any choice of curvature $k = -1, 0, +1$ to give

$$a(t) = \begin{cases} A\cosh\left(\sqrt{\Lambda/3}\,t\right) & k = +1 \\ \exp\left(\sqrt{\Lambda/3}\,t\right) & k = 0 \\ A\sinh\left(\sqrt{\Lambda/3}\,t\right) & k = -1 \end{cases}$$

where $A^2 = 3c^2/\Lambda R^2$ for the $k = \pm 1$ solutions, and for all solutions we've made a choice of an integration constant. At large time, all of these solutions exhibit exponential behaviour, independent of the spatial curvature. In fact, it turns out (although we won't show it here) that each of these solutions describes the same spacetime, but with different coordinates that slice spacetime into space+time in different ways. (This is described in the lectures on General Relativity.) This spacetime is known as *de Sitter space*.

The $k = +1$ solution most accurately represents the geometry of de Sitter space because it uses coordinates which cover the whole spacetime. It shows a contracting phase when $t < 0$, followed by a phase of accelerating expansion when $t > 0$. The phase of exponential expansion is what was captured in our naive Newtonian perspective which suggested that particles "race off to infinity". Crucially, there is no Big Bang because there's no point in time when $a = 0$. In contrast, the $k = 0$ and $k = -1$ coordinates give a slightly misleading view of the space, because they suggest a Big Bang when $t = -\infty$ and $t = 0$ respectively. You need to work harder to show that actually this is an artefact of the choice of coordinates (a so-called "coordinate singularity") rather than anything physical. These kind of issues will be addressed in next term's course on general relativity.

To better understand this spacetime and, in particular, the existence of cosmological horizons, it is best to work with $k = +1$ and conformal time, $\tau \in (-\pi/2, +\pi/2)$, given by

$$\cos\left(\sqrt{\Lambda/3}\tau\right) = \left[\cosh\left(\sqrt{\Lambda/3}t\right)\right]^{-1}$$

You can check that $d\tau/dt = 1/\cosh(\sqrt{\Lambda/3}\,t)$, which, up to an overall unimportant scale, is the definition of conformal time (1.26). In these coordinates, the metric for de Sitter space becomes

$$ds^2 = \frac{1}{\cos^2(\sqrt{\Lambda/3}\tau)}\left[-c^2d\tau^2 + R^2d\chi^2 + R^2\sin^2\chi(d\theta^2 + \sin^2\theta\, d\phi^2)\right]$$

where we're using the polar coordinates (1.6) on the spatial $\mathbf{S}^3$. We now consider a fixed θ and ϕ and draw the remaining 2d spacetime in the $(c\tau, \chi)$ plane where $\tau \in (-\pi/2, \pi/2)$ and $\chi \in [0, \pi]$. The left-hand edge of the diagram can be viewed as the north pole of $\mathbf{S}^3$, $\chi = 0$, while the right-hand edge of the diagram is the south pole $\chi = \pi$. The purpose of this diagram is not to exhibit distances between points, because these are distorted by the $1/\cos^2\tau$ factor in front of the metric. Instead, the diagram shows only the causal structure, with 45° lines denoting light rays.

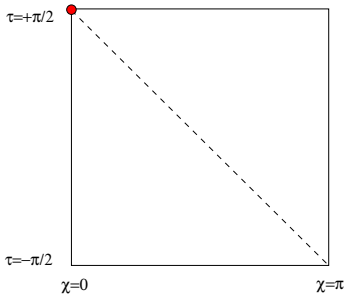

Figure 15:

Consider an observer sitting at the north pole. She has a particle horizon and an event horizon. Even if she waits forever, as shown in the figure, there will be part of the spacetime that she never sees.

Anti-de Sitter Space

We could also look at solutions with $\Lambda < 0$, again devoid of any matter so $\rho = 0$. A glance at the Friedmann equation (1.59) shows that such solutions can only exist when $k = -1$. In this case, the scale factor is given by

$$a(t) = A \sin\left(\sqrt{-\Lambda/3}\, t\right)$$

This is known as *anti-de Sitter space*. It has, as far as we can tell, no role to play in cosmology. However it has become rather important as a testing ground for ideas in quantum gravity and holography. In many ways, anti-de Sitter space acts like a gravitational box, trapping particles inside. This was suggested by the Newtonian, harmonic potential picture and will be explored more in the lectures on General Relativty. We will not discuss anti-de Sitter space further in these lectures.

Matter + Cosmological Constant

For a flat $k = 0$ universe, we can find a solution for a positive cosmological constant $\Lambda > 0$, with matter $\rho_m \sim 1/a^3$. We write the Friedmann equation as

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3c^2}\left(\rho_\Lambda + \frac{\rho_0}{a^3}\right)$$

This has the solution

$$a(t) = \left(\frac{\rho_0}{\rho_\Lambda}\right)^{1/3} \sinh^{2/3}\left(\frac{\sqrt{3\Lambda}t}{2}\right) \tag{1.61}$$

There are a number of comments to make about this. First note that, in contrast to de Sitter space, the Big Bang has unavoidably reappeared in this solution at $t = 0$ where $a(t = 0) = 0$. This, it turns out, is generic: any universe more complicated than de Sitter (like ours) has a Big Bang singularity.

The present day time t_0 is defined, as always, by $a(t_0) = 1$. There is also another interesting time, $t_{\rm eq}$, where we have matter-vacuum energy equality, so that $\rho_\Lambda = \rho_0/a^3$. This occurs when

$$\sinh\left(\frac{\sqrt{3\Lambda}t_{\rm eq}}{2}\right) = 1 \tag{1.62}$$

At late times, the solution (1.61) coincides with the de Sitter expansion $a(t) \sim e^{\sqrt{\Lambda/3}t}$, telling us that the cosmological constant is dominating as expected. Meanwhile, at early times we have $a \sim t^{2/3}$ and we reproduce the characteristic expansion of the Einstein-de Sitter universe (1.49).

An Historical Curiosity: The Einstein Static Universe

The cosmological constant was first introduced by Einstein in 1917 in an attempt to construct a static cosmology. This was over a decade before Hubble's discovery of the expanding universe.

The acceleration equation (1.52)

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3c^2}(\rho + 3P) \tag{1.63}$$

tells us that a static universe is only possible if $\rho = -3P$. Obviously this is not possible if we have only matter ρ_m with $P_m = 0$ or only a cosmological constant $\rho_\Lambda = -P_\Lambda$. But in a universe with both, we can have

$$\rho = \rho_m + \rho_\Lambda = -3P = 3\rho_\Lambda \quad \Rightarrow \quad \rho_m = 2\rho_\Lambda$$

The Friedmann equation (1.59) is then

$$H^2 = \frac{8\pi G}{3c^2}(\rho_m + \rho_\Lambda) - \frac{kc^2}{R^2a^2}$$

and the right-hand side vanishes if we take a positively curved universe, $k = +1$, with radius

$$(Ra)^2 = \frac{c^4}{8\pi G\rho_\Lambda} = \frac{c^2}{\Lambda} \tag{1.64}$$

This is the *Einstein static universe*. It is unstable. If a is a little smaller than the critical value (1.64) then $\rho_m \sim a^{-3}$ is a little larger and the acceleration equation (1.63) says that a will decrease further. Similarly, if a is larger than the critical value it will increase further.

1.3.4 How We Found Our Place in the Universe

In 1543, Copernicus argued that we do not sit at the centre of the universe. It took many centuries for us to understand where we do, in fact, sit.

Thomas Wright was perhaps the first to appreciate the true vastness of space. In 1750, he published "An original theory or new hypothesis of the universe", suggesting that the Milky Way, the band of stars that stretches across the sky, is in fact a "flat layer of stars" in which we are embedded, looking out. He further suggested that cloudy spots in the night sky, known as nebulae, are other galaxies, "too remote for even our telescopes to reach".

Wright was driven by poetry and art as much as astronomy and science and his book is illustrated by glorious pictures. His flights of fantasy led him to guesstimate that there are $3,888,000$ stars in the Milky Way, and 60 million planets. We now know, of course, that Wright's imagination did not stretch far enough: he underestimated the number of stars in our galaxy by 7 orders of magnitude.

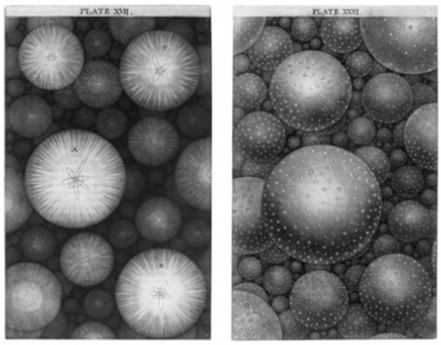

Figure 16: The wonderful imagination of Thomas Wright

Wright's suggestion that spiral nebulae are far flung galaxies, similar to our own Milky Way, was not met with widespread agreement. As late as 1920, many astronomers held that these nebulae were part of the Milky Way itself. Their argument was simple: if these were individual galaxies, or "island universes" as Kant referred to them, then they would lie at distances too vast to be credible.

The dawning realisation that our universe does indeed spread over such mind boggling distances came only with the discovery of redshifts. The American astronomer Vesto Slipher was the first to measure redshifts in 1912. He found spiral nebulae with both blueshifts and redshifts, some moving at speeds which are much too fast to be gravitationally bound to the Milky Way. Yet Slipher did not appreciate the full significance of his observations.

A number of other astronomers improved on Slipher's result, but the lion's share of the credit ended up falling into the lap of Edwin Hubble. His data, first shown in 1925, convinced everyone that the nebulae do indeed lie far outside our galaxy at distances of hundreds of kiloparsecs. Subsequently, in 1929 he revealed further data and laid claim to the law $\mathbf{v} = H\mathbf{x}$ that bears his name. For this, he is often said to have discovered the expanding universe. Yet strangely Hubble refused to accept this interpretation of his data, claiming as late as 1936 that "expanding models are definitely inconsistent with the observations that have been made".

It fell to theorists to put the pieces together. A framework in which to discuss the entire cosmos came only with the development of general relativity in 1915. Einstein himself was the first to apply relativity to the universe as a whole. In 1917, driven by a philosophical urge for an unchanging universe, he introduced the cosmological constant to apply a repulsive pressure which would counteract the gravitational attraction of

matter, resulting in the static spacetime that we met in (1.64). After Einstein's death, the physicist Gammow gave birth to the famous "biggest blunder" legend, stating

> "Einstein remarked to me many years ago that the cosmic repulsion idea was the biggest blunder he had made in his entire life."

Many other physicists soon followed Einstein. First out of the blocks was the dutch astronomer Willem de Sitter who, in 1917, published the solution that now bears his name, describing a spacetime with positive cosmological constant and no matter. de Sitter originally wrote the solution in strange coordinates, which made him think that his spacetime was static rather than expanding. He was then surprised to discover that signals between distant observers are redshifted. Both Slipher and Hubble referred to their redshift observations as the "de Sitter effect".

In St Petersburg, an applied mathematician-cum-meteorologist called Alexander Friedmann was also looking for solutions to the equations of general relativity. He derived his eponymous equation in 1922 and found a number of solutions, including universes which contracted and others which expanded indefinitely. Remarkably, at the end of his paper he pulls an estimate for the energy density of the universe out of thin air, gets it more or less right, and comes up with an age of the universe of 10 billion years. Sadly his work was quickly forgotten and three years later Friedmann died. From eating a pear. (No, really.)

The first person to understand the big picture was a Belgian, Catholic priest called Georges Lemaître. In 1927 he independently reproduced much of Friedmann's work, finding a number of further solutions. He derived Hubble's law (two years before Hubble's observations), extracting the first derivation of H_0 in the process and was, moreover, the first to connect the redshifts predicted by an expanding universe with those observed by Slipher and Hubble. For this reason, many books refer to the FRW metric as the FLRW metric. Although clearly aware of the significance of his discoveries, he chose to publish them in French in "Annales de la Société Scientifique de Bruxelles", a journal which was rather far down the reading list of most physicists. His work only became publicised in 1931 when a translation was published in the Monthly Notices of the Royal Astronomical Society, by which time much of the credit had been bagged by Hubble. Lemaître, however, was not done. Later that same year he proposed what he called the "hypothesis of the primeval atom", these days better known as the Big Bang theory. He was also the first to realise that the cosmological constant should be identified with vacuum energy.

We have not yet met R and W. The first is Howard Robertson who, in 1929, described the three homogeneous and isotropic spaces. This work was extended in 1935 by Roberston and, independently, by Arthur Walker, who proved these are the only possibilities.

Despite all of these developments, there was one particularly simple solution that had fallen through the cracks. It fell to Einstein and de Sitter to fill this gap. In 1932, when both were visitors at Caltech, they collaborated on a short, 2 page paper in which they described an expanding FRW universe with only matter. The result is the Einstein-de Sitter universe that we met in (1.49). Apparently neither thought very highly of the paper. Eddington reported a conversation with Einstein, who shrugged off this result with

> "I did not think the paper very important myself, but de Sitter was keen on it."

On hearing this, de Sitter wrote to Eddington to put the record straight,

> "You will have seen the paper by Einstein and myself. I do not myself consider the result of much importance, but Einstein seemed to think it was."

This short, unimportant paper, unloved by both authors, set the basic framework for cosmology for the next 60 years, until the cosmological constant was discovered in the late 1990s. As we will see in the next section, it provides an accurate description of the expansion of the universe for around 10 billion years of its history.

1.4 Our Universe

The time has now come to address the energy content and geometry of our own universe. We have come across a number of different entities that can contribute to the energy density of a universe. The three that we will need are

- Conventional matter, with $\rho_m \sim a^{-3}$
- Radiation, with $\rho_r \sim a^{-4}$
- A cosmological constant, with ρ_Λ constant.

We will see that these appear in our universe in somewhat surprising proportions.

Critical Density

Recall from Section 1.3.2 that in a flat universe the total energy density today must sum to match the Hubble constant. This is referred to as the *critical energy density*,

$$\rho_{\rm crit,0} = \frac{3c^2}{8\pi G} H_0^2 \tag{1.65}$$

We use this to define dimensionless *density parameters* for each fluid component,

$$\Omega_w = \frac{\rho_{w,0}}{\rho_{\rm crit,0}}$$

We have not included a subscript 0 on the density parameters but, as the definition shows, they refer to the fraction of energy observed today. Cosmologists usually specify the energy density in our Universe in terms of these dimensionless numbers Ω_w.

By design, the dimensionless density parameters sum to

$$\sum_{w=m,r,\Lambda} \Omega_w = 1 + \frac{kc^2}{R^2 H_0^2}$$

In particular, if we are to live in a flat universe then we must have $\sum_w \Omega_w = 1$. Any excess energy density, with $\sum_w \Omega_w > 1$ means that we necessarily live in a positively curved universe with $k = +1$. Any deficit in the energy, with $\sum_w \Omega_w < 1$ gives rise to a negatively curved, $k = -1$ universe.

It is sometimes useful to place the curvature term on a similar footing to the other energy densities. We define the energy density in curvature to be

$$\rho_k = -\frac{3kc^4}{8\pi G R^2 a^2}$$

and the corresponding density parameter as

$$\Omega_k = \frac{\rho_{k,0}}{\rho_{\rm crit,0}} = -\frac{kc^2}{R^2 H_0^2} \tag{1.66}$$

With these definitions, together with the scaling $\rho_w = \rho_{w,0}\, a^{-3(1+w)}$, the Friedmann equation

$$H^2 = \frac{8\pi G}{3c^2} \sum_{w=m,r,\Lambda} \rho_w - \frac{kc^2}{R^2 a^2}$$

can be rewritten in terms of the density parameters as

$$\left(\frac{H}{H_0}\right)^2 = \frac{\Omega_r}{a^4} + \frac{\Omega_m}{a^3} + \frac{\Omega_k}{a^2} + \Omega_\Lambda \tag{1.67}$$

One of the tasks of observational cosmology is to measure the various parameters in this equation.

1.4.1 The Energy Budget Today

After many decades of work, we have been able to measure the energy content of our universe fairly accurately. The two dominant components are

$$\Omega_\Lambda = 0.69 \quad \text{and} \quad \Omega_m = 0.31 \tag{1.68}$$

The cosmological constant, which we now know comprises almost 70% of the energy of our universe, was discovered in 1998. There are now two independent pieces of evidence. The first comes from direct measurement of Type Ia supernovae at large redshifts. (We saw the importance of supernovae in Section 1.1.5.) Similar data from 2003 is shown in Figure 17[4]. The 2011 Nobel prize was awarded to Perlmutter, Schmidt and Riess for this discovery.

The second piece of evidence is slightly more indirect, although arguably cleaner. The fluctuations in the cosmic microwave background (CMB) contain a wealth of information about the early universe. In combination with information from the distribution of galaxies in the universe, this provides separate confirmation of the results (1.68), as shown in Figure 18. (The label BAO in this figure refers "baryon acoustic oscillations"; we will briefly discuss these in Section 3.2.4.)

All other contributions to the current energy budget are orders of magnitude smaller. For example, the amount of energy in photons (denoted as γ) is

$$\Omega_\gamma \approx 5 \times 10^{-5} \tag{1.69}$$

Moreover, as the universe expanded and particles lost energy and slowed, they can transition from relativistic speeds, where they count as "radiation", to speeds much less than c where they count as "matter". This happened fairly recently to neutrinos, which contribute $\Omega_\nu \approx 3.4 \times 10^{-5}$.

Finally, there is no evidence for any curvature in our universe. The bound is

$$|\Omega_k| < 0.01$$

This collection of numbers, Ω_m, Ω_Λ, Ω_r and Ω_k sometimes goes by the name of the ΛCDM model, with Λ denoting the cosmological constant and CDM denoting *cold dark matter*, a subject we'll discuss more in Section 1.4.3.

[4]This data is taken from R. Knopp et al., "New Constraints on Ω_m, Ω_Λ, and w from an Independent Set of Eleven High-Redshift Supernovae Observed with HST", Astrophys.J.598:102 (2003).

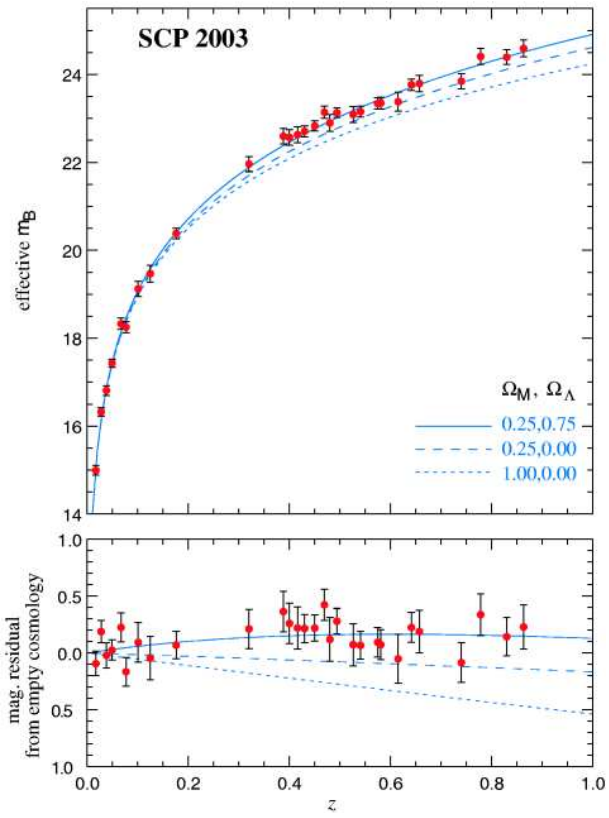

Figure 17: The redshift of a number of supernovae plotted against measured brightness. Various theoretical curves are shown for comparison.

The lack of any suggestion of curvature strongly suggests that we are living in a universe with $k = 0$. Given that the curvature of the universe is a dynamical variable and, as we have seen in Section 1.3.2, the choice of a flat universe is unstable, this is rather shocking. We will offer a putative explanation for the observed flatness in Section 1.5.

Energy and Time Scales

To convert the dimensionless ratios above into physical energy densities and time scales, we need an accurate measurement of the Hubble constant. Here there is some minor controversy. A direct measurement from Type IA supernovae gives[5]

$$H_0 = 74.0 \ \ (\pm 1.4) \ \mathrm{km\,s^{-1}Mpc^{-1}}$$

[5]The latest supernova data can be found in Riess et al., arXiv:1903.07603. Meanwhile, the final Planck results, extracting cosmological parameters from the CMB, can be found at arXiv:1807.06209.

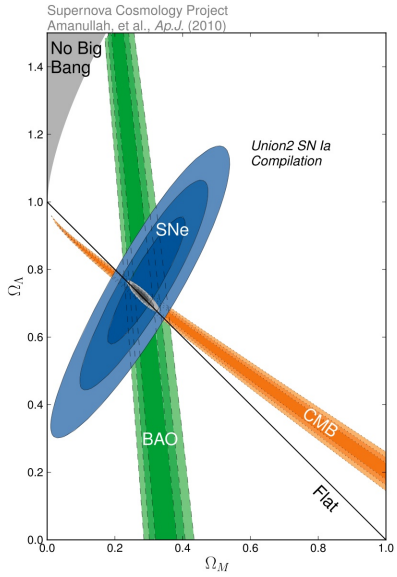

Figure 18: CMB, BAO and Supernovae results combined.

Meanwhile, analysis of the cosmic microwave background measured by the Planck satellite puts the value at

$$H_0 = 67.4 \ \ (\pm 0.5) \ \mathrm{km\,s^{-1}Mpc^{-1}}$$

The error bars suggest a 3σ discrepancy between the two measurements. Most of the community suspect that there is some systematic issue in one of the measurements, possibly in our understanding of cepheid luminosity which is used as a calibration for the supernovae results. However, it remains a possibility that there is something important and fundamental hiding in this mismatch. Here we use $H_0 \approx 70 \ \mathrm{km\,s^{-1}Mpc^{-1}}$. If we translate this into a time scale, we get

$$\frac{1}{H_0} = 4.3 \times 10^{18} \ \mathrm{s} = 1.4 \times 10^{11} \ \mathrm{years} \tag{1.70}$$

From a knowledge of the Hubble constant, and with $k = 0$, the expression (1.65) tells us that the total energy density of the universe is equal to the critical density,

$$\rho_{\mathrm{crit},0} = \frac{3c^2 H_0^2}{8\pi G} = 8.5 \ \times 10^{-10} \ \ \mathrm{kg\,m^{-1}s^{-2}} \tag{1.71}$$

A different method of calibrating supernovae distances has recently found the result $H_0 = 69.8(\pm 1.7) \ \mathrm{km\,s^{-1}Mpc^{-1}}$, in much closer agreement with the CMB data; see arXiv:1907.05922.

The corresponding mass density is

$$\frac{\rho_{\rm crit,0}}{c^2} = \frac{3H_0^2}{8\pi G} \approx 10^{-26}\ {\rm kg}\, m^{-3} \tag{1.72}$$

This is about one galaxy per cubic Mpc. Or, in more down to earth terms, one hydrogen atom per cubic metre. (Or, if you like, 10^{-68} galaxies per cubic metre!) The actual matter in the universe is, of course, fractionally less at $\rho_{m,0} = \Omega_m \rho_{\rm crit,0}$.

With the universe dominated by ρ_Λ and ρ_m, the solution (1.61), given by

$$a(t) = \left(\frac{\rho_0}{\rho_\Lambda}\right)^{1/3} \sinh^{2/3}\left(\frac{\sqrt{3\Lambda}t}{2}\right)$$

offers a good description of the expansion for much of this history. Recall that, in such a solution, the Big Bang takes place at $t = 0$ while the present day is defined by

$$\sinh^2\left(\frac{\sqrt{3\Lambda}t_0}{2}\right) = \frac{\rho_\Lambda}{\rho_0}$$

Inverting this gives the age

$$t_0 = \frac{c}{\sqrt{6\pi G\rho_\Lambda}} \sinh^{-1}\left(\sqrt{\frac{\rho_\Lambda}{\rho_0}}\right) = \frac{2}{3\sqrt{\Omega_\Lambda}H_0} \sinh^{-1}\left(\sqrt{\frac{\Omega_\Lambda}{\Omega_0}}\right)$$

The various factors almost cancel out, leaving us with an age which is very close to the naive estimate (1.23)

$$t_0 \approx 0.96 \times \frac{1}{H_0} \approx 1.4 \times 10^{10}\ {\rm years}$$

We can also calculate the age at which the vacuum energy was equal to the energy in matter (1.62). We get

$$t_{\rm eq} = \frac{2}{3\sqrt{\Omega_\Lambda}H_0} \sinh^{-1}(1) \approx 0.7 \times \frac{1}{H_0} \approx 0.98 \times 10^{10}\ {\rm years}$$

or about 4 billion years ago. To put this in perspective, the Earth is around 4.5 billion years old, and life started to evolve (at least) 3.5 billion years ago. In the grand scheme of things, equality between matter energy density and the cosmological constant occurred very recently.

Throughout these lectures, we will often use redshift z, rather than years, to refer to the time at which some event happened. Recall the the redshift is defined as (1.21)

$$1 + z = \frac{1}{a}$$

This means that at redshift z,the universe was $1/(1+z)^{\text{th}}$ its present size. This has the advantage that it's very easy to compute certain numbers in terms of z. For example, the equality of the cosmological constant and matter occurred when $\rho_m = \rho_\Lambda$ which, in terms of todays fractional energy density ,means that $\Omega_m/a^3 = \Omega_\Lambda$. Plugging in the numbers gives $z = 0.3$.

Matter-Radiation Equality

Today, radiation is an almost negligible part of the total energy density. However, this wasn't always the case. Because $\rho_r \sim 1/a^4$, as we go backwards in time the energy density in radiation grows much faster than matter, with $\rho \sim 1/a^3$, or the cosmological constant. We can ask: when do we have matter-radiation equality? In terms of redshift this requires

$$\frac{\Omega_m}{a^3} = \frac{\Omega_r}{a^4}$$

Here there is a small subtlety because neutrinos transition from relativistic to non-relativistic during this period. If we include the present day neutrino density as radiation, then we have $\Omega_r \approx 8.4 \times 10^{-5}$, which gives matter-radiation equality at $z \approx 3700$. A more accurate assessment gives

$$z_{\text{eq}} \approx 3400 \tag{1.73}$$

We can translate this into years. The universe was matter dominated for most of the time since $z = 3400$, with the cosmological constant becoming important only (relatively) recently. If we work with $a(t) = (t/t_0)^{2/3}$ as befits a matter-dominated universe, then we can trace back the evolution from the present day to get a rough estimate for the time of matter-radiation equality to be

$$t_{\text{eq}} = \frac{t_0}{(1 + z_{\text{eq}})^{3/2}} \approx 70,000 \text{ years}$$

A more accurate calculation gives

$$t_{\text{eq}} \approx 50,000 \text{ years}$$

Prior to this, the universe was radiation dominated.

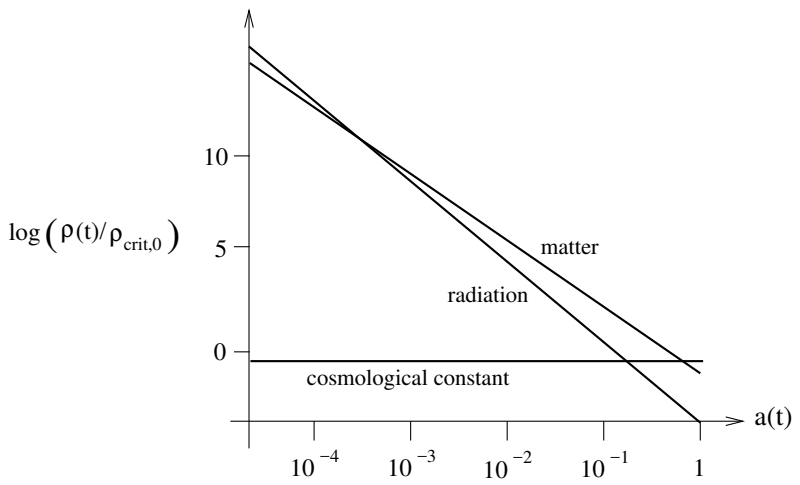

Figure 19: The evolution of the energy densities in our universe.

A plot of the evolution of the three kinds of energy is shown in Figure 19.

1.4.2 Dark Energy

For a number of observational cosmologists, who had long been wrestling with the difficulty of reconciling the early age (1.50) of a matter dominated universe with the lifetime of stars, the discovery of the cosmological constant came as a welcome relief. However, for the more theoretical minded physicists, it was something of a bombshell.

In the comfortable world of classical physics, there is no mystery to the cosmological constant. It is, as we have seen, simply the constant energy term that we previously neglected. However, our fundamental theories of physics are quantum. And here there is a problem, because they provide a way to estimate the size of the cosmological constant Λ.

Even before we put in any numbers, it's obvious that it's going to be a challenge to predict Λ from any underlying, quantum theory. That's simply because of the order of magnitudes. Recall that Λ has dimensions of $(\text{time})^{-2}$ and is given approximately by $\Lambda \sim H_0^2$. As we saw in (1.70), this time scale is measured in billions of years,

$$\Lambda \approx \frac{1}{(10^{11} \text{ years})^2}$$

That's a rather long time in anyone's book. But it's an especially long time from the perspective of fundamental particles, where time scales are typically measured in fractions of a second. Before we put in any numbers at all, it's clear that if we try

to derive the cosmological constant using, say, the Standard Model of particle physics then we're never going to get the right answer! You're surely going to get a much larger cosmological constant associated to a microscopic timescale.

Having convinced ourselves that no calculation of this kind can possibly work, let's go ahead and do it anyway, just to see how badly we fail. The story is usually told in terms of the relevant energy scales, rather than time scales. Taking the critical energy density to be (1.71), the observed vacuum energy density is $\rho_\Lambda \approx 6 \times 10^{-10}$ J m^{-3}. However, a more natural unit of energy is not the joule, but the *electron volt*, with $1 \text{ J} \approx 6.2 \times 10^{18}$ eV. In these units,

$$\rho_\Lambda = 3.7 \times 10^9 \text{ eV m}^{-3}$$

Perhaps more surprisingly, our preferred unit of inverse length is also the electron volt! To convert from one to the other, we use the fundamental constants of nature, $\hbar c \approx 2.0 \times 10^{-7}$ eV m. Putting this together, gives

$$\hbar^3 c^3 \rho_\Lambda \approx (10^{-3} \text{ eV})^4$$

Usually, this is written in natural units, with $\hbar = c = 1$, so that

$$\rho_\Lambda \approx (10^{-3} \text{ eV})^4$$

What are our expectations for the vacuum energy? Our fundamental laws of physics are written in framework called *quantum field theory*. All quantum field theories have a term, analogous to the $+\frac{1}{2}\hbar\omega$ ground state energy of the harmonic oscillator, which contributes to the vacuum energy of the universe. However, in contrast to the harmonic oscillator, in quantum field theory the ground state energy gets contributions from all possible frequencies. Taken at face value, this integral over frequencies would appear to diverge.

To make sense of this divergence, we need to embrace a little humility. Our theories have not been tested to arbitrarily high energy scales, and surely break down at some point. The best we can say at present is that the theories make sense up to the scales tested at the LHC, which operates at energies

$$M_{\text{LHC}} \sim 1 \text{ TeV} = 10^{12} \text{ eV}$$

With this conservative estimate for the validity of our theories, the most "natural" value for the vacuum energy arising from quantum field theory

$$\rho_{\text{QFT}} = (10^{12} \text{ eV})^4 = 10^{60} \rho_\Lambda$$

This is not particularly close to the observed value. It is, moreover, a ridiculous number that makes no sense in the cosmological context. Such a universe would not be conducive to forming nuclei or atoms, let alone galaxies and life. The huge discrepancy between the expected value of ρ_{QFT} and the observed value of ρ_Λ is known as the *cosmological constant problem*.

Physicists with masochistic tendencies will try to make the situation look even worse. There is some minimal, circumstantial evidence that the framework of quantum field theory holds up to the Planck scale $M_{pl} = \sqrt{\hbar c/8\pi G}$ which corresponds to the energy $M_{\rm pl}c^2 \approx 10^{19}$ GeV. In this case, we would get $\rho_{QFT} = 10^{122}\rho_\Lambda$. I'm not sure this way of stating things is particularly helpful.

The value of ρ_{QFT} is not a precise prediction of quantum field theory, but rather a ballpark figure for the natural energy scale of the theory. We are always free to just add a further arbitrary constant to the energy of the theory. In that case, there are two contributions

$$\rho_\Lambda = \rho_{QFT} + \rho_{\rm constant}$$

Apparently, the two contributions on the right must add up to give the observed value of ρ_Λ. We call this *fine-tuning*. As presented above, it looks fairly absurd: two numbers of order 10^{60} (or higher) have to coincide in the first 60 digits, but differ in the $61^{\rm st}$, leaving behind a number of order 1.

It is quite possible that there is some missing principle that we've failed to grasp that makes fine tuning less silly than it first appears. The task of finding such a mechanism is made considerably harder when we realise that there have been a number of times in the history of the universe when ρ_{QFT} abruptly changed while, presumably, $\rho_{\rm constant}$ did not. This occurs at a *phase transition*. For example, the QCD phase transition, where quarks which were once free became trapped in protons and neutrons, took place in the early universe. At this moment, there was a change $\Delta\rho_{QFT} \sim (100\ {\rm MeV})^4$. Still earlier, the electroweak phase transition, where the Higgs boson kicks in and gives mass to fundamental particles, should have resulted in a change of $\Delta\rho_{QFT} \sim (100\ {\rm GeV})^4$. In other words, any putative cancellation mechanism must conspire to give a tiny cosmological constant ρ_Λ at the end of the life of the universe, not at the beginning.

Given these difficulties, most physicists in the $20^{\rm th}$ century buried their heads in the sand and assumed that there must be some deep principle that sets the cosmological constant to zero. No such principle was found. In the $21^{\rm st}$ century, we have a much harder job. We would like a deep principle that sets the late-time cosmological constant to $\rho_\Lambda \sim (10^{-3}\ {\rm eV})^4$. Needless to say, we haven't found that either.

If this wasn't bad enough, there is yet another issue that we should confront. The value of the current vacuum energy is remarkably close to the energy in matter. Why? As illustrated in Figure 19, these energy densities scale very differently and we would naively expect that they differ by orders of magnitude. Why are Ω_m and Ω_Λ so very close today? This is known as the *coincidence problem*. We have no good explanation.

The A-Word

As we saw above, a naive application of quantum field theory suggests a ludicrous value for the cosmological constant, one that results in an expansion so fast that not even atoms have a chance to form from their underlying constituents. Given this, we could ask the following question: what is the maximum value of the cosmological constant that still allows complex structures to evolve? For example, what is the maximum allowed value of Λ that allows galaxies to form?

It turns out that the upper bound on Λ depends on the strength of the initial seeds from which the galaxies grew. At very early times, there are small variations $\delta\rho$ in the otherwise homogeneous universe. As we will discuss in more detail in Section 3, in our universe these seeds have size $\delta\rho/\rho \sim 10^{-5}$. Let us fix this initial condition, and then ask again: how big can the cosmological constant be?

We will present this calculation in Section 3.3.4. The answer is quite striking: the scale of the vacuum energy is pretty much the maximum it could be. If ρ_Λ were bigger by an order of magnitude or so, then no galaxies would form, presumably making it rather more difficult for life to find a comfortable foothold in the universe.

What to make of this observation? One possibility is to shrug and move on. Another is to weave an elaborate story. Suppose that our observable universe is part of a much larger structure, a "multiverse" in which different domains exhibit different values of the fundamental parameters, or perhaps even different laws of physics. In this way, the cosmological constant is not a fundamental parameter which we may hope to predict, but rather an environmental parameter, no different from, say, the distance between the Earth and the Sun. We should not be shocked by its seemingly small value because, were it any higher, we wouldn't be around to comment on it. Such reasoning goes by the name of the *anthropic principle*.

The anthropic explanation for the cosmological constant may be correct. But, in the absence of any testable predictions, discussions of this idea rapidly descend into a haze of sophomoric tedium. Trust me: there are better things to do with your life. (Like find a proper explanation.)

A Rebranding: Dark Energy

Given our manifest befuddlement about all things Λ, it is prudent to wonder if ρ_Λ is actually a cosmological constant at all. Perhaps it is some other form of fluid, with an equation of state $w \approx -1$, rather than precisely $w = -1$. More interesting, it may be a fluid whose equation of state evolves over time. (We will meet behaviour like this in Section 1.5.) I stress that there are no compelling theoretical reasons to believe that this is the case, and nor does it alleviate the need to explain why ρ_{QFT} does not gravitate. Nonetheless, this is clearly an area where we are totally at sea and we should be open to such possibilities. For these reasons, the mysterious 70% of the energy in the universe is often referred to as *dark energy*.

1.4.3 Dark Matter

Our embarrassing ignorance of the universe we call home is further illustrated if we delve a little deeper into the $\Omega_m = 0.31$ energy in matter. Of this, the amount that we understand is

$$\Omega_B \approx 0.05 \tag{1.74}$$

This is the energy in matter made from atoms in the periodic table. The B in Ω_B stands for "baryons", which are protons and neutrons. This is appropriate because the mass in electrons is negligible in comparison.

The remaining matter energy is in the form of *cold dark matter*,

$$\Omega_{CDM} \approx 0.26$$

This is stuff that we have not (yet?) created here on Earth. The "cold" refers to the fact that it is non-relativistic today and, moreover, has been so for some time.

We know very little about this dark matter. We do not know if it is a single species of particle, or many. We do not know if it consists of several decoupled sectors, or just one. Given the wonderful complexity that lurks in Ω_B, it seems reasonable to assume that there is still rather a lot to learn about Ω_{CDM}.

Here we simply describe some of the evidence for the existence of dark matter. To do this, we need to construct methods to determine the mass of the large objects, such as galaxies or clusters of galaxies. These are small enough for us to ignore the expansion of the universe so, for the rest of this section, we will work in flat space.

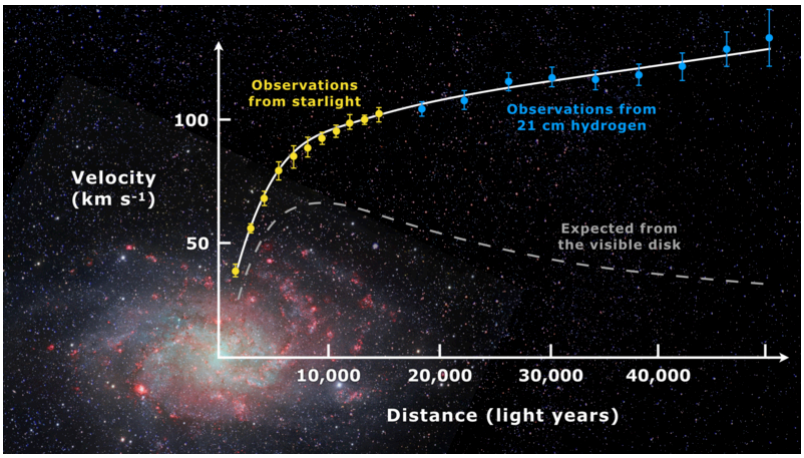

Figure 20: The rotation curve of galaxy M33. Image taken from Wikipedia.

Galaxy Rotation Curves

At the galactic scale, rotation curves provide a clean way to measure mass. This method was pioneered by Vera Rubin and her collaborator Kent Ford in the 1960s and 70s.

For a quick and dirty sketch of the idea, we will assume spherical symmetry. A quick glance at a typical spiral galaxy shows this is a poor approximation, at least for the visible matter, but it will suffice to get the basic idea across. The centrifugal acceleration of a star, orbiting at distance r from the galactic centre, must be provided by the gravitational force,

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

where $M(r)$ is the mass enclosed inside a sphere of radius r. We learn that we expect the rotational speed to vary as

$$v(r) = \sqrt{\frac{GM(r)}{r}}$$

Far from the bulk of the galaxy, we would expect that $M(r)$ is constant, so the velocity drops off as $v \sim \sqrt{1/r}$. This is not what is observed. The rotation speeds can be measured from the edge of the galaxy by studying interstellar gas, in particular the 21cm line of hydrogen. (The origin of this line was discussed in the Atomic Physics section of the Lectures on Topics in Quantum Mechanics..) One finds that the rotation remains more or less constant very far from what appears to be the edge of the galaxy. This suggests that the mass continues to grow as $M(r) \sim r$ far from the observable galaxy. This is known as the dark matter halo.

The Virial Theorem and Galaxy Clusters

The *virial theorem* offers a clever method of weighing a collection of objects that are far away.

Virial Theorem: A collection of N particles, with masses m_i and positions $\mathbf{x}_i$, interact through a gravitational potential

$$V = \sum_{i<j} V_{ij} = \sum_{i<j} -\frac{Gm_i m_j}{|\mathbf{x}_i - \mathbf{x}_j|} \tag{1.75}$$

We will assume that the system is gravitationally bound, and that the positions $\mathbf{x}_i$ and velocities $\dot{\mathbf{x}}_i$ are bounded for all time. We will further assume that the time average of the kinetic energy T and potential energy V are well defined. Then

$$\overline{T} = -\frac{1}{2}\overline{V}$$

where the bar denotes time average (a quantity we will define more precisely below).

Proof: We start by defining something akin to the moment of inertia,

$$I = \frac{1}{2}\sum_i m_i \mathbf{x}_i \cdot \mathbf{x}_i \quad \Rightarrow \quad \dot{I} = \sum_i \mathbf{p}_i \cdot \mathbf{x}_i \tag{1.76}$$

with $\mathbf{p}_i$ the momentum of the i^{th} particle. The quantity $\dot{I}$ is known as the *virial*. Note that, in contrast to the potential V, both I and $\dot{I}$ depend on our choice of origin. The correct choice is to pick this origin to be the centre of mass. The time derivative of the virial is

$$\ddot{I} = \sum_i \dot{\mathbf{p}}_i \cdot \mathbf{x}_i + \sum_i \mathbf{p}_i \cdot \dot{\mathbf{x}}_i = \sum_i \mathbf{F}_i \cdot \mathbf{x}_i + 2T$$

where, in the second equality, we have used the definition of kinetic energy T and Newton's force law $\mathbf{F}_i = \dot{\mathbf{p}}_i$. The force $\mathbf{F}_i$ on the i^{th} particle is determined by the potential V_{ij} by

$$\begin{aligned}
\mathbf{F}_i = -\sum_{j\neq i} \nabla_i V_{ij} \quad \Rightarrow \quad \sum_i \mathbf{F}_i \cdot \mathbf{x}_i &= -\sum_{i<j} \nabla_i V_{ij} \cdot \mathbf{x}_i - \sum_{j<i} \nabla_i V_{ij} \cdot \mathbf{x}_i \\
&= -\sum_{i<j} \nabla_i V_{ij} \cdot \mathbf{x}_i - \sum_{i<j} \nabla_j V_{ji} \cdot \mathbf{x}_j \\
&= -\sum_{i<j} \nabla_i V_{ij} \cdot (\mathbf{x}_i - \mathbf{x}_j)
\end{aligned}$$

where, in the second step, we simply swapped the dummy indices i and j and, in the third step, we used $V_{ij} = V_{ji}$ and $\nabla_i V_{ij} = -\nabla_j V_{ij}$. But now we can use the explicit form of the potential (1.75) to find

$$-\sum_{i<j} \nabla_i V_{ij} \cdot (\mathbf{x}_i - \mathbf{x}_j) = \sum_{i<j} V_{ij} = V$$

We learn that the time variation of the virial is

$$\ddot{I} = V + 2T$$

At this point we take the time average, defined by

$$\overline{X} = \lim_{t\to\infty} \frac{1}{t} \int_0^t X(t')\, dt'$$

The time average of all these quantities is assumed to be well-defined. But,

$$\overline{\frac{d\dot{I}}{dt}} = \lim_{t\to\infty} \frac{\dot{I}(t) - \dot{I}(0)}{t} = 0$$

Note that the last step follows only if the virial (1.76) is measured relative to the centre of mass, otherwise the positions $\mathbf{x}_i$ will have a drift linear in time. We're left with the promised virial theorem $\overline{V} + 2\overline{T} = 0$. □

As an aside: the virial theorem also holds in other contexts. For example, a proof using the variational method can be found in the Lectures on Topics in Quantum Mechanics.

The virial theorem can be used to estimate the mass of any collection of objects that satisfy the assumptions of the theorem. Roughly speaking, this holds when the objects have reached something akin to thermodynamic equilibrium. In 1933, Zwicky used this technique to estimate the mass of the Coma cluster, shown in the figure, a conglomerate of a few thousand galaxies.

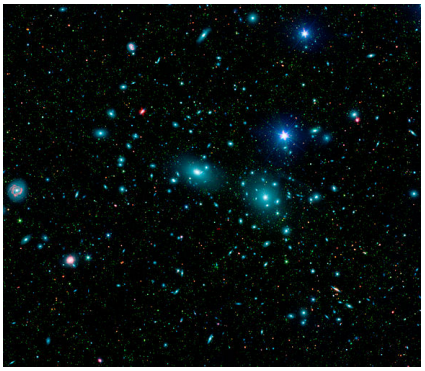

Figure 21: Coma cluster.

We will make a few simplifying assumptions. First we will assume that there are N galaxies, all of the same mass m. (We can do better, but this will serve our purposes.) Second, we will assume that the system is "self-averaging", which

means that the average over many galaxies is a proxy for averaging over time so that, for example,

$$\bar{T} \approx \langle T \rangle = \frac{1}{2N} \sum_{i=1}^{N} m v_i^2$$

This has the advantage that we don't need to wait several billion years to perform the time average. The virial theorem then reads

$$2\langle T \rangle = m\langle v^2 \rangle \approx \langle V \rangle \approx \frac{1}{2} G m^2 N \left\langle \frac{1}{r} \right\rangle$$

where $\langle 1/r \rangle$ is the average inverse distance between galaxies and, in the last step, we have replaced $N - 1$ with N. This then gives an expression for the total mass of the galaxy cluster,

$$Nm \approx \frac{2\langle v^2 \rangle}{G\langle 1/r \rangle} \tag{1.77}$$

The right-hand-side contains quantities that we can measure, giving us an estimate for the mass of the cluster. (Strictly speaking, we can measure v_{redshift}, the velocity in the line of sight. If we further assume spherical symmetry, we have $\langle v^2 \rangle = 3\langle v^2_{\text{redshift}} \rangle$.)

There is a much simpler way to compute the mass in each galaxy: simply count the number of stars. In practice, this is done by measuring the luminosity. This provides two very different ways to determine the mass and we can compare the two. One typically finds that the virial mass is greater than the luminosity mass by a factor of a couple of hundred. The difference is made up by what Zwicky referred to as *Dunkle Materie*, or dark matter.

(An aside: Zwicky was viewed by his peers as a genius and a bit of a prick. He referred to his enemies as "spherical bastards" because, no matter what direction you looked at them, they were still bastards.)

Other Evidence

There are a number of other pieces of evidence, all of which consistently point to the existence of dark matter. The mathematics underlying these requires more than just Newtonian dynamics so, for now, we will replace the maths with some pretty pictures.

- Gravitational Lensing: A classic prediction of general relativity is that light bends as it passes heavy objects. Furthermore, the image gets distorted, a phenomenon known as *gravitational lensing*. Sometimes this happens in a spectacular fashion,

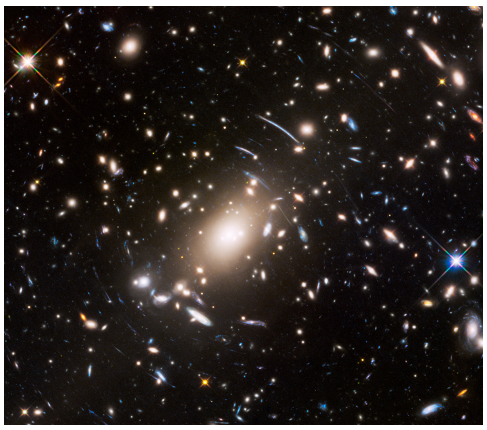

Figure 22: Abell S1063 cluster.

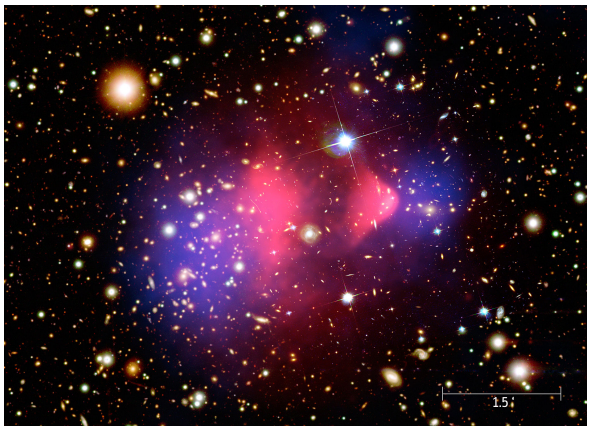

Figure 23: The bullet cluster.

as shown in the picture on the left, where the image of a background galaxy is distorted into the blue arcs by the cluster in the foreground. Even small distortions of this kind allow us accurately determine the mass of the cluster in the foreground. You will not be surprised to hear that the mass greatly exceeds that seen in visible matter.

The bullet cluster, shown in the right-hand figure, provides a particularly dramatic example of gravitational lensing. This picture shows two sub-clusters of galaxies which are thought to have previously collided. There are three types of matter shown in the picture: stars which you can see, hot gas which is observed in x-rays and is shown in pink, and the distribution of mass detected through gravitational lensing shown in blue. The stars sit cleanly in two distinct sub-clusters because individual galaxies have little chance of collision. In contrast, most of the baryonic matter sits in clouds of hot gas which interact fairly strongly as the clusters collide, slowing the gas and leaving it displaced from the stars as shown in the figure. But most of the matter, as detected through gravitational lensing, is dark and this, like the galaxies, has glided past each other seemingly unaffected by the collision. The interpretation is that dark matter interacts weakly, both with itself and with baryonic matter.

- BBN: The observations described above show clearly that on the scale of both galaxies and clusters of galaxies there is more matter than can be detected by electromagnetic radiation. This alone is not sufficient to tell us that dark matter must be composed of some new unknown particle. For example, it could be in the form of failed stars ("jupiters"). There is, however, compelling evidence that this is not the case, and dark matter is something more exotic.

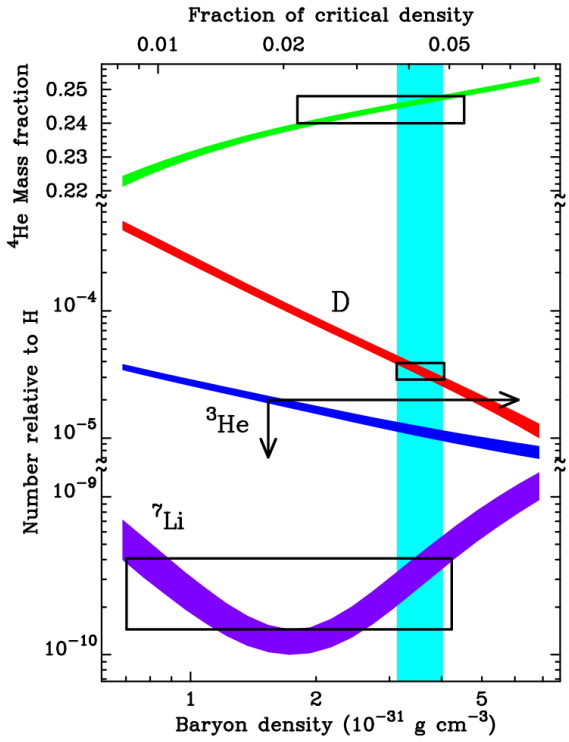

Figure 24: The relative abundance of light elements forged in the early universe, as a function of the overall baryon density.

The primary evidence comes from *Big Bang nucleosynthesis* (BBN), an impressively accurate theory of how the light elements were forged in the early universe. It turns out that the relative abundance of different elements depends on the total amount of baryon matter. In particular, the relative abundance of deuterium changes rapidly with baryon density. This is shown[6] in Figure 24, with the horizontal turquiose bar fixed by observations of relative abundance. (The black boxes show the relative abundance of each element, with error bars, and the corresponding constraint on the baryon density.) This tells us that the total amount of baryonic matter is just a few percent of the total energy density. We will describe some aspects of BBN in Section 2.5.3.

- Structure formation: The CMB tells us that the very early universe was close to homogeneous and isotropic, with fluctuations in the energy of the order of

[6]This figure is taken from Burles, Nollett and Turner, *Big-Bang Nucleosynthesis: Linking Inner Space and Outer Space*", astro-ph/99033.

$\delta\rho/\rho \sim 10^{-5}$. Yet today, these tiny fluctuations have grown into the clusters, galaxies and stars that we see around us. How did this happen?

It turns out that there this can not be achieved by baryonic matter alone. In the fireball of the Big Bang, baryonic matter is coupled to photons and these provide a pressure which suppresses gravitational collapse. This collapse can only proceed after the fireball cools and photons decouple, an event which takes place around 300,000 years after the Big Bang. This does not leave enough time to form the universe we have today. Dark matter, however, has no such constraints. It decouples from the photons much earlier, and so its density perturbations can start to grow, forming gravitational wells into which visible matter can subsequently fall. We will tell this story in Section 3.

- CMB: As we mentioned above, baryonic matter and dark matter behave differently in the early universe. Dark matter is free to undergo gravitational collapse, while baryonic matter is prevented from doing so by the pressure of the photons. These differences leave their mark on the fireball, and this shows up in the fluctuations etched in the microwave background. This too will be briefly described in Section 3.

1.5 Inflation

We have learned that our universe is a strange and unusual place. The cosmological story that emerged above has a number of issues that we would like to address. Some of these – most notably those related to dark matter and dark energy – have yet to be understood. But there are two puzzles that do have a compelling solution, known as *cosmological inflation*. The purpose of this section is to first describe the puzzles, and then describe the solution.

1.5.1 The Flatness and Horizon Problems

The first puzzle is one we've met before: our universe shows no sign of spatial curvature. We can't say for sure that it's exactly flat but observations bound the curvature to be $|\Omega_k| < 0.01$. A universe with no curvature is a fixed point of the dynamics, but it is an *unstable* fixed point, and any small amount of curvature present in the early universe should have grown over time. At heart, this is because the curvature term in the Friedmann equation scales as $1/a^2$ while both matter and radiation dilute much faster, as $1/a^3$ and $1/a^4$ respectively.

Let's put some numbers on this. We will care only about order of magnitudes. We ignore the cosmological constant on the grounds that it has been irrelevant for much of

the universe's history. As we saw in Section 1.4.1, for most of the past 14 billion years the universe was matter dominated. In this case,

$$\frac{\rho_k(t)}{\rho_m(t)} = \frac{\rho_{k,0}}{\rho_{m,0}}\, a \quad \Rightarrow \quad \Omega_k(t) = \frac{\Omega_{k,0}}{\Omega_{m,0}} \frac{\Omega_m(t)}{1+z}$$

where, for once, we have defined time-dependent density parameters $\Omega_w(t)$ and, correspondingly, added the subscript $\Omega_{m,0}$ to specify the fractional density today. This formula holds all the way back to matter-radiation equality at $t = t_{\rm eq}$ where $\Omega_m(t_{\rm eq}) \approx 1/2$ (the other half made up by radiation) and $z \approx 3000$. Using the present day value of $\Omega_{k,0}/\Omega_{m,0} \lesssim 10^{-2}$, we must have

$$|\Omega_k(t_{\rm eq})| \leq 10^{-6}$$

At earlier times, the universe is radiation dominated. Now the relevant formula is

$$\frac{\rho_k(t)}{\rho_r(t)} = \frac{\rho_{k,\rm eq}}{\rho_{r,\rm eq}} \frac{a^2}{a_{\rm eq}^2} \quad \Rightarrow \quad \Omega_k(t) = \frac{\Omega_k(t_{\rm eq})}{\Omega_r(t_{\rm eq})} \frac{(1+z_{\rm eq})^2}{(1+z)^2} \Omega_r(t)$$

We can look, for example, at the flatness of the universe during Big Bang nucleosynthesis, a period which we understand pretty well. As we will review in Section 2, this took place at $z \approx 4 \times 10^8$. Here, the curvature must be

$$|\Omega_k(t_{\rm BBN})| \leq 10^{-16}$$

We have good reason to trust our theories even further back to the electroweak phase transition at $z \approx 10^{15}$. Here, the curvature must be

$$|\Omega_k(t_{\rm EW})| \leq 10^{-30}$$

These are small numbers. Why should the early universe be flat to such precision? This is known as the *flatness problem*.

The second puzzle is even more concerning. As we have mentioned previously, and will see in more detail in Section 2, the universe is filled with radiation known as the cosmic microwave background (CMB). This dates back to 300,000 years after the Big Bang when the universe cooled sufficiently for light to propagate.

The CMB is almost perfectly uniform and isotropic. No matter which direction we look, it has the same temperature of 2.725 K. However, according to the standard cosmology that we have developed, these different parts of the sky sat outside each others particle horizons at the time the CMB was formed. This concept is simplest to see in conformal time, as shown in Figure 25.

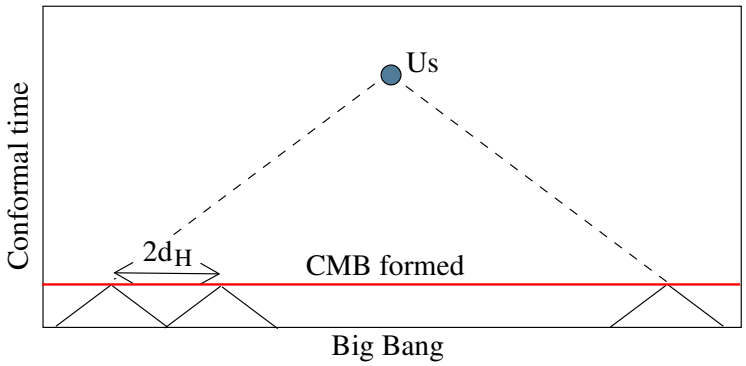

Figure 25: The horizon problem: different regions of the CMB are causally disconnected at the time it was formed.

We can put some numbers on this. For a purely matter-dominated universe, with $a(t) = (t/t_0)^{2/3}$, the particle horizon (1.24) at time t is defined by

$$d_H(t) = c\, a(t) \int_0^t \frac{dt'}{a(t')} = 3ct$$

We use $H(t) = 2/3t = H_0/a(t)^{3/2}$ to write this as

$$d_H(z) = \frac{2cH_0^{-1}}{(1+z(t))^{3/2}} \tag{1.78}$$

We will see in Section 2.3 that the CMB is formed when $z \approx 1100$. We would like to know how large the particle horizon (1.78) looks in the sky today. In the intervening time, the distance scale $d_H(z)$ has been stretched by the expansion of the universe to $(1+z)d_H(z)$. Meanwhile, this should be compared to the particle horizon today which is $d_H(t_0) = 2cH_0^{-1}$. From this, we learn that the distance $d_H(z)$ today subtends an angle on the sky given by

$$\theta \approx \frac{(1+z)d_H(z)}{d_H(t_0)} \approx \frac{1}{\sqrt{1100}} \approx 0.03 \text{ rad} \quad \Rightarrow \quad \theta \approx 1.7^\circ$$

Assuming the standard cosmology described so far, patches of the sky separated by more than $\sim 1.7^\circ$ had no causal contact at the time the CMB was formed. We would naively expect to see significant variations in temperature over the sky on this scale, but instead we see the same temperature everywhere we look. It is very hard to envisage how different parts of the universe could have reached thermal equilibrium without ever being in causal contact. This is known as the *horizon problem*.

Ultimately, the two problems above are both concerned with the initial conditions in the universe. We should be honest and admit that we're not really sure what the rules of the game are here. If you're inclined to believe in a creator, you might find it plausible that she simply stipulated that the universe was absolutely flat, with constant energy density everywhere in space at some initial time $t = \epsilon$. It's not the kind of explanation that scientists usually find compelling, but you might think it has a better chance to convince in this context.

However, there is a more nuanced version of the horizon problem which makes the issue significantly more acute, and renders the "God did it" explanation significantly less plausible. Somewhat ironically, this difficulty arises when we appreciate that the CMB is not completely uniform after all. It contains tiny, but important anisotropies. There are small fluctuations in temperature at about 1 part in 10^5. Furthermore, there are also patterns in the polarisation of the of the light in the CMB. And, importantly, the polarisation and temperature patterns are correlated. These correlations – which go by the uninspiring name of "TE correlations" – are the kind of thing that arises through simple dynamical processes in the early universe, such as photons scattering off electrons. But observations reveal that there are correlations over patches of the sky that are as large as 5°.

These detailed correlations make it more difficult to appeal to a creator without sounding like a young Earth creationist, arguing that the fossil record was planted to deceive us. Instead, the observations are clearly telling us that there were dynamical processes taking place in the early universe but, according to our standard FRW cosmology, these include dynamical processes that somehow connect points that were not in causal contact. This should make us very queasy. If we want to preserve some of our most cherished ideas in physics – such as locality and causality – it is clear that we need to do something that changes the causal structure of the early universe, giving time for different parts of space to communicate with each other.

1.5.2 A Solution: An Accelerating Phase

There is a simple and elegant solution to both these problems. We postulate that the very early universe underwent a period of accelerated expansion referred to as *inflation*. Here "very early" refers to a time before the electroweak phase transition, although we cannot currently date it more accurately than this. An accelerating phase means

$$a(t) \sim t^n \quad \text{with } n > 1 \tag{1.79}$$

Alternatively, we could have a de Sitter-type phase with $a(t) \sim e^{H_{\rm inf}t}$ with constant $H_{\rm inf}$. This is exactly the kind of accelerating phase that we are now entering due to

the cosmological constant. However, while the present dark energy is $\rho_\Lambda \sim (10^{-3}\text{ eV})^4$, the dark energy needed for inflation is substantially larger, with $\rho_{\text{inflation}} \geq (10^3\text{ GeV})^4$ and, in most models, closer to $(10^{15}\text{ GeV})^4$.

Let's see why such an inflationary phase would solve our problems. First, the horizon problem. The particle horizon is defined as (1.24),

$$d_H(t) = c\,a(t)\int_0^t \frac{dt'}{a(t')}$$

It is finite only if the integral converges. This was the case for a purely matter (or radiation) dominated universe, as we saw in (1.78). But, for $a(t) \sim t^n$ we have

$$\int_0^t \frac{dt'}{a(t')} \sim \int_0^t \frac{dt'}{t'^n} \to \infty \quad \text{if } n > 1$$

This means that an early accelerating phase buys us (conformal) time and allows far flung regions of the early universe to be in causal contact.

An inflationary phase also naturally solves the flatness problem. An inflationary phase of the form (1.79) must be driven by some background energy density that scales as

$$\rho_{\text{inf}} \sim \frac{1}{a^{2/n}}$$

which, for $n > 1$, clearly dilutes away more slowly than the curvature $\rho_k \sim 1/a^2$. This means that, with a sufficiently long period of inflation, the spatial curvature can be driven as small as we like. Although we have phrased this in terms of energy densities, there is a nice geometrical intuition that underlies this: if you take any smooth, curved manifold and enlarge it, then any small region looks increasingly flat.

This putative solution to the flatness problem also highlights the pitfalls. In the inflationary phase, the curvature ρ_k will be driven to zero but so too will the energy in matter ρ_m and radiation ρ_r. Moreover, we'll be left with a universe dominated by the inflationary energy density ρ_{inf}. To avoid this, the mechanism that drives inflation must be more dynamic than the passive fluids that we have considered so far. We need a fluid that provides an energy density ρ_{inf} for a suitably long time, allowing us to solve our problems, but then subsequently turns itself off! Or, even better, a fluid that subsequently converts its energy density into radiation. Optimistic as this may seem, we will see that there is a simple model that does indeed have this behaviour.

How Much Inflation Do We Need?

We will focus on the horizon problem. For simplicity, we will assume that the early universe undergoes an exponential expansion with $a(t) \sim e^{H_{\rm inf}t}$. Suppose that inflation lasts for some time T. If, prior to the onset of inflation, the physical horizon had size d_I then, by the end of inflation, this region of space has been blown up to $d_F = e^{H_{\rm inf}T} d_I$. We quantify the amount of inflation by $N = H_{\rm inf}T$ which we call the number of *e-folds*.

Subsequently, scales that were originally at d_I grow at a more leisurely rate as the universe expands. If the end of inflation occurred at redshift $z_{\rm inf}$, then

$$d_{\rm now} = e^N (1 + z_{\rm inf}) d_I$$

We will see that $z_{\rm inf}$ is (very!) large, and we lose nothing by writing $1 + z_{\rm inf} \approx z_{\rm inf}$. The whole point of inflation is to ensure that this length scale $d_{\rm now}$ is much larger than what we can see in the sky. This is true, provided

$$d_{\rm now} \gg c H_0^{-1} \quad \Rightarrow \quad e^N > \frac{c}{H_0 d_I} \frac{1}{z_{\rm inf}}$$

Clearly, to determine the amount of inflation we need to specify both when inflation ended, $z_{\rm inf}$, and the size of the horizon prior to inflation, d_I. We don't know either of these, so we have to make some guesses. A natural scale for the initial horizon is $d_I = c H_{\rm inf}^{-1}$, which gives

$$e^N > \frac{H_{\rm inf}}{H_0} \frac{1}{z_{\rm inf}}$$

Post-inflation, the expansion of the universe is first dominated by radiation with $H \sim 1/a^2$, and then by matter with $H \sim 1/a^{3/2}$. Even though the majority of the time is in the matter-dominated era, the vast majority of the expansion takes place in the radiation dominated era when energy densities were much higher. So we write $H_{\rm inf}/H_0 \sim (1 + z_{\rm inf})^2$. We then have

$$e^N > \left(\frac{H_{\rm inf}}{H_0} \right)^{1/2} = z_{\rm inf}$$

It remains to specify $H_{\rm inf}$ or, equivalently, $z_{\rm inf}$.

We don't currently know $H_{\rm inf}$. (We will briefly mention a way in which this can be measured in future experiments in Section 3.5.) However, as we will learn in Section 2, we understand the early universe very well back to redshifts of $z \sim 10^8 - 10^9$. Moreover, we're fairly confident that we know what's going on back to redshifts of $z \sim 10^{15}$ since

this is where we can trust the particle physics of the Standard Model. The general expectation is that inflation took place at a time before this, or

$$z_{\rm inf} > 10^{15} \quad \Rightarrow \quad N > 35$$

Recall that $H_0 \approx 10^{-18}$ s^{-1}, so if inflation took place at $z \approx 10^{15}$ then the Hubble scale during inflation was $H_{\rm inf} = 10^{12}\ s^{-1}$. In this case, inflation lasted a mere $T \sim 10^{-11}\ s$. These are roughly the time scales of processes that happen in modern particle colliders.

Many models posit that inflation took place much earlier than this, at an epoch where the early universe is getting close to Planckian energy scales. A common suggestion is

$$z_{\rm inf} \sim 10^{27} \quad \Rightarrow \quad N > 62$$

in which case $H_{\rm inf} \sim 10^{36}$ s^{-1} and $T \sim 10^{-35}$ s. This is an extraordinarily short time scale, and corresponds to energies way beyond anything we have observed in our puny experiments on Earth.

Most textbooks will quote around 60 e-foldings as necessary. For now, the take-away message is that, while there are compelling reasons to believe that inflation happened, there is still much we don't know about the process including the scale $H_{\rm inf}$ at which it occurred.

1.5.3 The Inflaton Field

Our theories of fundamental physics are written in terms of fields. These are objects which vary in space and time. The examples you've met so far are the electric and magnetic fields $\mathbf{E}(\mathbf{x}, t)$ and $\mathbf{B}(\mathbf{x}, t)$.

The simplest (and, so far, the only!) way to implement a transient, inflationary phase in the early universe is to posit the existence of a new field, usually referred to as the *inflaton*, $\phi(\mathbf{x}, t)$. This is a "scalar field", meaning that it doesn't have any internal degrees of freedom. (In contrast, the electric and magnetic fields are both vectors.)

The dynamics of this scalar field are best described using an action principle. In particle mechanics, the action is an integral over time. But for fields, the action is an integral over space and time. We'll first describe this action in flat space, and subsequently generalise it to the expanding FRW universe.

In Minkowski spacetime, the action takes the form

$$S = \int d^3x\, dt \ \left[\frac{1}{2}\dot{\phi}^2 - \frac{c^2}{2}\nabla\phi\cdot\nabla\phi - V(\phi)\right] \tag{1.80}$$

Here $V(\phi)$ is a potential. Different potentials describe different physical theories. We do not yet know the form of the inflationary potential, but it turns out that many do the basic job. (More detailed observations do put constraints on the form the potential can take as we will see in Section 3.5.) Later, when we come to solve the equations of motion, we will work with the simplest possible potential

$$V(\phi) = \frac{1}{2}m^2\phi^2 \tag{1.81}$$

The action (1.80) is then the field theory version of the harmonic oscillator. In the language of quantum field theory, m is called the *mass* of the field. (It is indeed the mass of a particles that arise when the field is quantised.)

The equations of motion for ϕ follow from the principle of least action. If we vary $\phi \to \phi + \delta\phi$, then the action changes as

$$\begin{aligned}\delta S &= \int d^3x\, dt \ \left[\dot{\phi}\,\delta\dot{\phi} - c^2\nabla\phi\cdot\nabla\delta\phi - \frac{\partial V}{\partial\phi}\delta\phi\right] \\ &= \int d^3x\, dt \ \left[-\ddot{\phi} + c^2\nabla^2\phi - \frac{\partial V}{\partial\phi}\right]\delta\phi\end{aligned}$$

where, in the second line, we have integrated by parts and discarded the boundary terms. Insisting that $\delta S = 0$ for all variations $\delta\phi$ gives the equation of motion

$$\ddot{\phi} - c^2\nabla^2\phi + \frac{\partial V}{\partial\phi} = 0$$

This is known as the *Klein-Gordon equation*. It has the important property that it is Lorentz covariant.

We want to generalise the action (1.80) to describe a scalar field in a homogenous and isotropic FRW universe. For simplicity, we restrict to the case of a $k = 0$ flat universe. This is a little bit unsatisfactory since we're invoking inflation in part to explain the flatness of space. However, it will allow us to keep the mathematics simple, without the need to understand the full structure of fields in curved spacetime. Hopefully, by the end you will have enough intuition for how scalar fields behave to understand that they will, indeed, do the promised job of driving the universe to become spatially flat.

In flat space, the FRW metric is simply

$$ds^2 = -c^2 dt^2 + a^2(t)\, d\mathbf{x}^2$$

The scale factor $a(t)$ changes the spatial distances. This results in two changes to the action (1.80): one in the integration over space, and the other in the spatial derivatives. We now have

$$S = \int d^3x\, dt\; a^3(t) \left[\frac{1}{2}\dot{\phi}^2 - \frac{c^2}{2a^2(t)} \nabla\phi \cdot \nabla\phi - V(\phi) \right] \tag{1.82}$$

Before we compute the equation of motion for ϕ, we first make a simplification: because we're only interested in spatially homogeneous solutions we may as well look at fields which are constant in space, so $\nabla\phi = 0$ and $\phi(\mathbf{x}, t) = \phi(t)$. We then have

$$S = \int d^3x\, dt\; a^3(t) \left[\frac{1}{2}\dot{\phi}^2 - V(\phi) \right] \tag{1.83}$$

Varying the action now gives

$$\delta S = \int d^3x\, dt\; a^3(t) \left[\dot{\phi}\, \delta\dot{\phi} - \frac{\partial V}{\partial \phi}\delta\phi \right] = \int d^3x\, dt\; \left[-\frac{d}{dt}\left(a^3 \dot{\phi} \right) - a^3 \frac{\partial V}{\partial \phi} \right] \delta\phi$$

Insisting that $\delta S = 0$ for all $\delta\phi$ again gives the equation of motion, but now there is an extra term because, after integration by parts, the time derivative also hits the scale factor $a(t)$. The equation of motion in an expanding universe is therefore

$$\ddot{\phi} + 3H\dot{\phi} + \frac{\partial V}{\partial \phi} = 0 \tag{1.84}$$

In the analogy with the harmonic oscillator, the extra term $3H\dot{\phi}$ looks like a friction term. It is sometimes referred to as *Hubble friction* or *Hubble drag*.

We also need to understand the energy density $\rho_{\rm inf} \equiv \rho_\rho$ associated to the inflaton field ϕ since this will determine the evolution of $a(t)$ through the Friedmann equation. There is a canonical way to compute this (through the stress-energy tensor) but the answer turns out to be what you would naively guess given the action (1.83), namely

$$\rho_\phi = \frac{1}{2}\dot{\phi}^2 + V(\phi) \tag{1.85}$$

The resulting Friedmann equation is then

$$H^2 = \frac{8\pi G}{3c^2} \left(\frac{1}{2}\dot{\phi}^2 + V(\phi) \right) \tag{1.86}$$

We will shortly solve the coupled equations (1.84) and (1.86). First we can ask: what kind of fluid is the inflaton field? To answer this, we need to determine the pressure. This follows straightforwardly by looking at

$$\dot{\rho}_\phi = \left(\ddot{\phi} + \frac{\partial V}{\partial \phi}\right)\dot{\phi} = -3H\dot{\phi}^2$$

Comparing to the continuity equation (1.39), $\dot{\rho} + 3H(\rho + P) = 0$, we see that the pressure must be

$$P_\phi = \frac{1}{2}\dot{\phi}^2 - V(\phi) \tag{1.87}$$

Clearly, this doesn't fit into our usual classification of fluids with $P = w\rho$ for some constant w. Instead, we have something more dynamical and interesting on our hands.

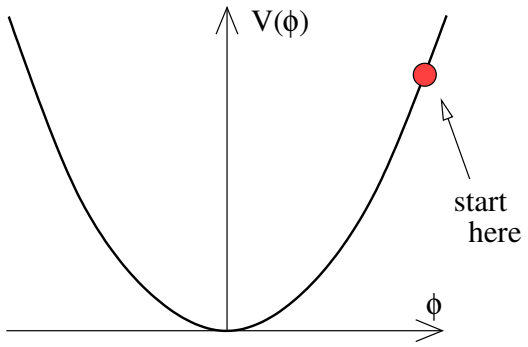

Figure 26: The inflationary scalar rolling down the potential $V(\phi)$.

Slow Roll Solutions

We want to solve the coupled equations (1.84) and (1.86). In particular, we're looking for solutions which involve an inflationary phase. Taking the time derivative of (1.86), we have

$$2H\left(\frac{\ddot{a}}{a} - H^2\right) = \frac{8\pi G}{3c^2}\left(\ddot{\phi} + \frac{\partial V}{\partial \phi}\right)\dot{\phi} = -\frac{8\pi G}{c^2}H\dot{\phi}^2$$

where, in the second equality, we have used (1.84). Rearranging gives

$$\frac{\ddot{a}}{a} = -\frac{8\pi G}{3c^2}\left(\dot{\phi}^2 - V(\phi)\right)$$

which we recognise as the Raychaudhuri equation (1.52). We see that we get an inflationary phase only when the potential energy dominates the kinetic energy, $V(\phi) > \dot{\phi}^2$. Indeed, in the limit that $V(\phi) \gg \dot{\phi}^2$, the relationship between the energy (1.85) and pressure (1.87) becomes $P_\phi \approx -\rho_\phi$, which mimics dark energy.

Now we can get some idea for the set-up. We start with a scalar field sitting high on some potential, as shown in Figure 26 with $\dot{\phi}$ small. This will give rise to inflation. As the scalar rolls down the potential, it will pick up kinetic energy and we will exit the inflationary phase. The presence of the Hubble friction term in (1.84) means that the scalar can ultimately come to rest, rather than eternally oscillating backwards and forwards.

Let's put some equations on these words. We assume that $V(\phi) \gg \frac{1}{2}\dot{\phi}^2$, a requirement that is sometimes called the *slow-roll condition*. The Friedmann equation (1.86) then becomes

$$H^2 \approx \frac{8\pi G}{3c^2} V(\phi) \tag{1.88}$$

Furthermore, if inflation is to last a suitably long time, it's important that the scalar does not rapidly gain speed. This can be achieved if the Hubble friction term dominates in equation (1.84), so that $\ddot{\phi} \ll H\dot{\phi}$. In the context of the harmonic oscillator, this is the over-damped regime. The equation of motion is then

$$3H\dot{\phi} \approx -\frac{\partial V}{\partial \phi} \tag{1.89}$$

These are now straightforward to solve. For concreteness, we work with the quadratic potential $V = \frac{1}{2}m^2\phi^2$. Then the solutions to (1.88) and (1.89) are

$$H = \alpha\phi \quad \text{and} \quad \dot{\phi} = -\frac{m^2}{3\alpha} \quad \text{with } \alpha^2 = \frac{4\pi G m^2}{3c^2}$$

Integrating the second equation gives

$$\phi(t) = \phi_0 - \frac{m^2}{3\alpha}t$$

where we have taken the scalar field to start at some initial value ϕ_0 at $t = 0$. We can now easily integrate the $H = \alpha\phi$ equation to get an expression for the scale factor,

$$a(t) = a(0)\, \exp\left[\frac{2\pi G}{c^2}(\phi_0^2 - \phi(t)^2)\right] \tag{1.90}$$

This is a quasi-de Sitter phase of almost exponential expansion.

This solution remains valid provided that the condition $V(\phi) \gg \dot{\phi}^2$ is obeyed. The space will cease to inflate when $V(\phi) \approx \dot{\phi}^2$, which occurs when $\phi^2(t_{\rm end}) \approx 2m^2/(3\alpha)^2$. By this time, the universe will have expanded by a factor of

$$\frac{a(t_{\rm end})}{a(0)} \approx \exp\left[\frac{2\pi G\phi_0^2}{c^2} - \frac{1}{3}\right]$$

We see that, by starting the scalar field higher up the potential, we can generate an exponentially large expansion.

1.5.4 Further Topics

There is much more to say about the physics of inflation. Here we briefly discuss a few important topics, some of which are fairly well understood, and some of which remain mysterious or problematic.

Reheating

By the end of inflation, the universe is left flat but devoid of any matter or radiation. For this to be a realistic mechanism, we must find a way to transfer energy from the inflaton field into more traditional forms of matter. This turns out to be fairly straightforward, although we are a long way from a detailed understanding of the process. Roughly speaking, if the inflaton field is coupled to other fields in nature, then these will be excited as the inflaton oscillates around the minimum of its potential. This process is known as *reheating*. Afterwards, the standard hot Big Bang cosmology can start.

Dark Energy or Cosmological Constant?

Inflation is a period of dynamically driven, temporary, cosmic acceleration in the very early universe. Yet, as we have seen, the universe is presently entering a second stage of comic acceleration. How do we know that this too isn't driven by some underlying dynamics and will, again, turn out to be temporary? The answer is: we don't. It is not difficult to cook up a mathematical model in which the cosmological constant is set to zero by hand and the current acceleration is driven using some scalar field. Such models go by the unhelpful name of *quintessence*.

Quintessence models are poorly motivated and do nothing to solve the fine-tuning problems of the cosmological constant. In fact, they are worse. First, we have to set the genuine cosmological constant to zero (and we have no reason to do so) and then we have to introduce a new scalar field which, to give the observed acceleration, must have an astonishingly small mass of order $m \sim 10^{-33}\ eV$.

Such models look arbitrary and absurd. And yet, given our manifest ignorance about the cosmological constant, it is perhaps best to keep a mildly open mind. The smoking gun would be to measure an equation of state $P = w\rho$ for the present day dark energy which differs from $w = -1$.

Initial Conditions

For the idea of inflation to fly, we must start with the scalar field sitting at some point high up the potential. It is natural to ask: how did it get there?

One possibility is that the initial value of the scalar field varies in space. The regions where the scalar are biggest then inflate the most, and all traces of the other regions are washed away beyond the horizon. These kind of ideas raise some thorny issues about the nature of probabilities in an inflationary universe (or multiverse) and are poorly understood. Needless to say, it seems very difficult to test such ideas experimentally.

A More Microscopic Underpinning?

Usually when we introduce a scalar field in physics, it is an approximation to something deeper going on underneath. For example, there is a simple theory of superconductivity, due to Landau and Ginsburg, which invokes a scalar field coupled to the electromagnetic field. This theory makes little attempt to justify the existence of the scalar field. Only later was a more microscopic theory of superconductivity developed — so-called BCS theory — in which the scalar field emerges from bound pairs of electrons. Many further examples, in which scalar fields are invoked to describe everything from water to magnets, can be found in the lectures on Statistical Field Theory.

This raises a question: is the scalar field description of inflation an approximation to something deeper going on underneath? We don't know the answer to this.

Quantum Fluctuations

Although inflation was first introduced to solve the flatness and horizon problems, its greatest triumph lies elsewhere. As the scalar field rolls down the potential, it suffers small quantum fluctuations. These fluctuations are swept up in the expansion of the universe and stretched across the sky where, it is thought, they provide the seeds for the subsequent formation of structure in the universe. These fluctuations are responsible for the hot and cold spots in the CMB which, in turn, determine where matter clumps and galaxies form. In Section 3.5 we will look more closely at this bold idea.

2. The Hot Universe

As we wind the clock back towards the Big Bang, the energy density in the universe increases. As this happens, particles interact more and more frequently, and the state of the universe is well approximated by a hot fluid in equilibrium. This is sometimes referred to as the *primeval fireball* of the Big Bang. The purpose of this section is to introduce a few basic properties of this fireball.

It is worth sketching the big picture. First we play the movie in reverse. As we go back in time, the Universe becomes hotter and hotter and things fall apart. Running the movie forward, the Universe cools and various objects form.

For example, there is an important event, roughly 300,000 years after the Big Bang, when atoms form for the first time. Prior to this, the temperature was higher than the 13.6 eV binding energy of hydrogen, and the electrons were stripped from the protons. This moment in time is known as *recombination* and will be described in Section 2.3. (Obviously a better name would simply be "combination" since the electrons and protons combined for the first time, but we don't get to decide these things). This is a key moment in the history of the universe. Prior to this time, space was filled with a charged plasma through which light is unable to propagate. But when the electrons and protons form to make (mostly) neutral hydrogen, the universe becomes transparent. The cosmic microwave background, which will be discussed in Section 2.2, dates from this time.

At yet earlier times, the universe was so hot that nuclei fail to cling together and they fall apart into their constituent protons and neutrons. This process – which, running forwards in time is known as nucleosynthesis – happens around 3 minutes after the Big Bang and is understood in exquisite detail. We will describe some of the basic reactions in Section 2.5.3.

As we continue to trace the clock further back, the universe is heated to extraordinary temperatures, corresponding to the energies probed in particle accelerators and beyond. Taking knowledge from particle physics, even here we have a good idea of what happens. At some point, known as the QCD phase transition, protons and neutrons melt, dissolving into a soup of their constituents known as the quark-gluon plasma. Earlier still, at the electroweak phase transition, the condensate of the Higgs boson melts. Beyond this, we have little clear knowledge but there are still other events that we know must occur. The purpose of this chapter is the tell this story.

2.1 Some Statistical Mechanics

Our first task is to build a language that allows us to describe stuff that is hot. We will cherry pick a few key results that we need. A much fuller discussion of the subject can be found in the lectures on Statistical Physics and the lectures on Kinetic Theory.

Ideas such as heat and temperature are not part of the fundamental laws of physics. There is no such thing, for example, as the temperature of a single electron. Instead, these are examples of *emergent phenomena*, concepts which arise only when a sufficiently large number of particles are thrown together. In domestic situations, where we usually apply these ideas, large means $N \sim 10^{23}$ particles. As we will see, in the cosmological setting N can be substantially larger.

When dealing with such a large number of particles, we need to shift our point of view. The kinds of things that we usually discuss in classical physics, such as the position and momentum of each individual particle, no longer hold any interest. Instead, we want to know coarse-grained properties of the system. For example, we might like to know the probability that a particle chosen at random has a momentum $\mathbf{p}$. In what follows, we call this probability distribution $f(\mathbf{p}; t)$.

Equilibrium

In general, the distribution $f(\mathbf{p}, t)$ will be very complicated. But patience brings rewards. If we wait a suitably long time, the individual particles will collide with each other, transferring energy and momentum among themselves until, eventually, any knowledge about the initial conditions is effectively lost. The resulting state is known as *equilibrium* and is described by a time-independent probability distribution $f(\mathbf{p})$. In equilibrium, the constituent particles are flying around in random directions. But, if you focus only on the coarse-grained probability distribution, everything appears calm.

Equilibrium states are characterised by a number of macroscopic quantities. These will be dealt with in detail in the Statistical Physics course, but here we summarise some key facts.

The most important characteristic of an equilibrium state is *temperature*. This is related to the average energy of the state in a way that we will make precise below. The reason that temperature plays such an important role is due to the following property: suppose that we have two different systems, each individually in equilibrium, one at temperature T_1 and the other at temperature T_2. We then bring the two systems together and allow them to exchange energy. If $T_1 = T_2$, then the two systems remain unaffected by this, and the combined system is in equilibrium. In contrast, if $T_1 \neq$

T_2, then a net energy will flow from the hotter system to the colder system, and the combined system will eventually settle down to a new equilibrium state at some intermediate temperature. Two systems which have the same temperature are said to be in *thermal equilibrium*.

Other kinds of equilibria are also possible. One that we will meet later in this section arises when two systems are able to exchange particles. Often we will be interested in this when one type of particle can transmute into another. In this case, we characterise the system by another quantity known as the *chemical potential*. (The name comes from chemical reactions although, in this course, will be more interested in processes in atomic or particle physics.) The chemical potential has the property that if two systems have the same value then, when brought together, there will not be a net transfer of particles from one system to the other. In this case, the systems are said to be in *chemical equilibrium*.

2.1.1 The Boltzmann Distribution

For now we will focus on states in thermal equilibrium. The thermal properties of a state are closely related to its energy which, in turn, is related to the momentum of the constituent particles. This means that understanding thermal equilibrium is akin to understanding the momentum distribution $f(\mathbf{p})$ of particles. We will see a number of examples of this in what follows.

A microscopic understanding of thermal equilibrium was first provided by Boltzmann. It turns out that the result is somewhat easier to state in the language of quantum mechanics, although it also applies to the classical world. Consider a system with discrete energy eigenstates $|n\rangle$, each with energy E_n. In *thermal equilibrium* at temperature T, the probability that the system sits in the state $|n\rangle$ is given by the *Boltzmann distribution*,

$$p(n) = \frac{e^{-E_n/k_BT}}{Z} \tag{2.1}$$

Here k_B is the *Boltzmann constant*, defined to be

$$k_B \approx 1.381 \times 10^{-23}\ \mathrm{J\,K^{-1}}$$

This fundamental constant provides a translation between temperatures and energies. Meanwhile Z is simply a normalisation constant designed to ensure that

$$\sum_n p(n) = 1 \quad \Rightarrow \quad Z = \sum_n e^{-E_n/k_BT}$$

This normalisation factor Z has its own name: it is called the *partition function* and it plays a starring role in most treatments of statistical mechanics. For our purposes, it will suffice to keep Z firmly in the background.

It is possible to derive the Boltzmann distribution from more elementary principles. (Such a derivation can be found in the lectures on Statistical Physics.) Here, we will simply take the distribution (2.1) to be the definition of both thermal equilibrium and the temperature.

The Boltzmann distribution gives us some simple intuition for the meaning of thermal equilibrium. We see that the any state with $E_n \ll k_B T$ has a more or less equal chance of being occupied, while any state with $E_n \gg k_B T$ has a vanishingly small chance of being occupied. In this way $k_B T$ sets the characteristic energy scale of the system.

We'll see many variations of the Boltzmann distribution in what follows. It gets tedious to keep writing $1/k_B T$. For this reason we define

$$\beta = \frac{1}{k_B T}$$

We will be careless in what follows and also refer to β as "temperature": obviously it is actually (proportional to) the inverse temperature. The Boltzmann distribution then reads

$$p(n) = \frac{e^{-\beta E_n}}{Z}$$

Above, we mentioned the key property of temperature: it determines whether two systems sit in thermal equilibrium. We should check that this is indeed obeyed by the Boltzmann distribution. Suppose that we have two systems, A and B, both at the same temperature β, but with different microscopic constituents, meaning that their energy levels differ. If we bring the two systems together, we expect that the combined system also sits in a Boltzmann distribution at temperature β. Happily, this is indeed the case. To see this note that we have independent probability distributions for A and B, so the combined probability distribution is given by

$$p(n,m) = \frac{e^{-\beta E_n^A}}{Z_A}\frac{e^{-\beta E_m^B}}{Z_B} = \frac{e^{-\beta(E_n^A + E_m^B)}}{Z_A Z_B}$$

But this is again of the Boltzmann form. The denominator $Z_A Z_B$ can be written as

$$Z_A Z_B = \left(\sum_n e^{-\beta E_n^A}\right)\left(\sum_m e^{-\beta E_m^B}\right) = \sum_{n,m} e^{-\beta E_n^A} e^{-\beta E_m^B} = \sum_{n,m} e^{-\beta(E_n^A + E_m^B)}$$

where we recognise this final expression as Z_{A+B}, the partition function of the combined system $A+B$. This had to be the case to ensure that the joint probability distribution $p(n,m)$ is correctly normalised.

It's worth re-iterating what we have learned. You might think that if we combined two systems, separately in equilibrium, then there would be no energy transfer from one to the other if the average energies coincide, i.e. $\langle E^A\rangle = \langle E^B\rangle$, with

$$\langle E\rangle = \frac{1}{Z}\sum_n E_n e^{-\beta E_n}$$

However, this is not the right criterion. As we have seen above, the average energies of the two systems can be very different. It is the temperatures that must coincide.

2.1.2 The Ideal Gas

As our first application of the Boltzmann distribution, consider a gas of non-relativistic particles, each of mass m. We will assume that there are no interactions between these particles, so the energy of each is given by

$$E = \frac{1}{2}mv^2 \tag{2.2}$$

Before we proceed, I should mention a subtlety. We've turned off interactions in order to make our life simpler. Yet, from our earlier discussion, it should be clear that interactions are crucial if we are ever going to reach equilibrium, since this requires a large number of collisions to share energy and momentum between particles! This is one of many annoying and fiddly issues that plague the fundamentals of statistical mechanics. We will argue this subtlety away by pretending that the interactions are strong enough to drive the system to equilibrium, but small enough to ignore when describing equilibrium. Obviously this is unsatisfactory. We can do better, but it is more work. (See, for example, the discussion of the interacting gas in the lectures on Statistical Physics or the derivation of the approach to equilibrium in the lectures on Kinetic Theory.) We will also see this issue rear its head in a physical context in Section 2.3.4 when we discuss the phenomenon of decoupling in the early universe.

We consider a gas of particles. We'll assume that each particle is independent of the others, and focus on the state of a just single particle, specified by the momentum $\mathbf{p}$ or, equivalently, the velocity $\mathbf{v} = \mathbf{p}/m$. If the momentum is continuous (or finely spaced) we should talk about the probability that the velocity lies in some some volume d^3v centred around $\mathbf{v}$. We denote the probability distribution as $f(\mathbf{v})\, d^3v$. The Boltzmann

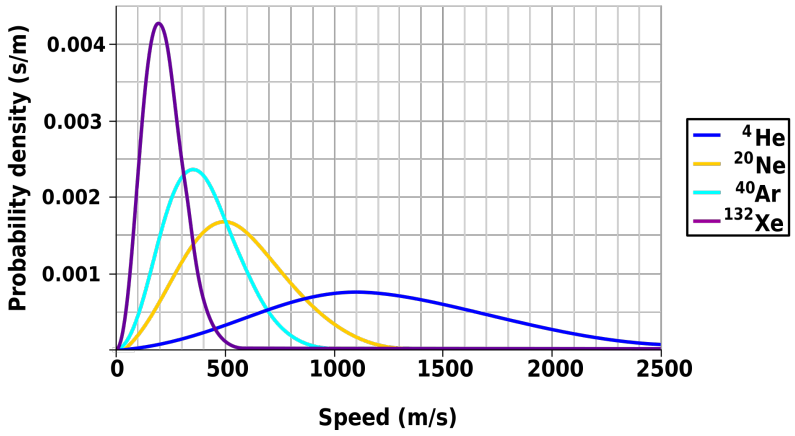

Figure 27: The distribution of the speeds of various molecules at $T = 25$ C. (Image taken from Wikipedia.)

distribution (2.1) tells us that this is

$$f(\mathbf{v})\, d^3v = \frac{e^{-\beta m v^2/2}}{Z}\, d^3v \tag{2.3}$$

where Z is a normalisation factor that we will determine shortly.

Our real interest lies in the speed $v = |\mathbf{v}|$. The corresponding speed distribution $f(v)\, dv = f(\mathbf{v})\, d^3v$ is

$$f(v)dv = \frac{4\pi v^2}{Z} e^{-\beta m v^2/2}\, dv \tag{2.4}$$

Note that we have an extra factor of $4\pi v^2$ when considering the probability distribution over speeds v, as opposed to velocities $\mathbf{v}$. This reflects the fact that there's "more ways" to have a high velocity than a low velocity: the factor of $4\pi v^2$ is the area of the sphere swept out by a velocity vector $\mathbf{v}$.

We require that

$$\int_0^\infty dv\ f(v) = 1 \quad \Rightarrow \quad Z = \left(\frac{2\pi k_B T}{m}\right)^{3/2}$$

Finally, we find the probability that the particle has speed between v and $v + dv$ to be

$$f(v)\, dv = 4\pi v^2 \left(\frac{m}{2\pi k_B T}\right)^{3/2} e^{-mv^2/2k_B T}\, dv \tag{2.5}$$

This is known as the *Maxwell-Boltzmann distribution*. It tells us the distribution of the speeds of gas molecules in this room.

Pressure and the Equation of State

We can use the Maxwell-Boltzmann distribution to compute the pressure of a gas. The pressure arises from the constant bombardment by the underlying atoms and can be calculated with some basic physics. Consider a wall of area A that lies in the (y, z)-plane. Let n denote the density of particles (i.e. $n = N/V$ where N is the number of particles and V the volume). In some short time interval Δt, the following happens:

- A particle with velocity $\mathbf{v}$ will hit the wall if it lies within a distance $\Delta L = |v_x|\Delta t$ of the wall *and* if it's travelling towards the wall, rather than away. The number of such particles with velocity centred around $\mathbf{v}$ is

$$\frac{1}{2}nA|v_x|\Delta t\, d^3v$$

  with a factor of $1/2$ picking out only those particles that travel in the right direction.

- After each such collision, the momentum of the particle changes from p_x to $-p_x$, with p_y and p_z left unchanged. As before, this holds only for the initial $p_x > 0$. We therefore write the impulse imparted by each particle as $2|p_x|$.

- This impulse is equated with $F_x\Delta t$ where F_x is the force on the wall. The force arising from particles with velocity in the region d^3v about $\mathbf{v}$ is

$$F_x\Delta t = \left(\frac{1}{2}nA|v_x|\Delta t\, d^3v\right) \times 2|p_x| \quad \Rightarrow \quad F_x = nAv_xp_x\, d^3v$$

  where we dropped the modulus signs on the grounds that the sign of the momentum p_x is the same as the sign of the velocity v_x.

- The pressure on the wall is the force per unit area, $P = F_x/A$. We learn that the pressure from those particles with velocity in the region of $\mathbf{v}$ is

$$P = nv_xp_x\, d^3v$$

  At this stage we invoke isotropy of the gas, which means that $\mathbf{v}\cdot\mathbf{p} = v_xp_x + v_yp_y + v_zp_z = 3v_xp_x$. We therefore have

$$P = \frac{n}{3}\mathbf{v}\cdot\mathbf{p}\, d^3v \tag{2.6}$$

The last stage is to integrate over all velocities, weighted with the probability distribution. In the final form (2.6), the pressure is related to the speed v rather than the

(component of the) velocity v_x. This means that we can use the Maxwell-Boltzmann distribution over speeds (2.5) and write

$$P = \frac{1}{3}\int dv\; n\,\mathbf{v}\cdot\mathbf{p}\,f(v) \tag{2.7}$$

This coincides with our earlier result (1.33) (albeit using slightly different notation for the probability distributions).

The expression (2.7) holds for both relativistic and non-relativistic systems, a fact that we will make use of later. For now, we care only for the non-relativistic case with $\mathbf{p} = m\mathbf{v}$. Here we have

$$P = \frac{4\pi n}{3}\left(\frac{m}{2\pi k_B T}\right)^{3/2}\int dv\; mv^4\, e^{-mv^2/2k_BT}$$

The integral is straightforward: it is given by

$$\int_0^\infty dx\; x^4 e^{-ax^2} = \frac{3}{8}\sqrt{\frac{\pi}{a^5}}$$

Using this, we find a familiar friend

$$P = nk_BT$$

This is the equation of state for an ideal gas.

We can also calculate the average kinetic energy. If the gas contains N particles, the total energy is

$$\langle E\rangle = \frac{N}{2}m\langle v^2\rangle = N\int_0^\infty dv\;\frac{1}{2}mv^2 f(v) = \frac{3}{2}Nk_BT \tag{2.8}$$

This confirms the result (1.37) that we met when we first introduced non-relativistic fluids.

2.2 The Cosmic Microwave Background

The universe is bathed in a sea of thermal radiation, known as the *cosmic microwave background*, or the CMB. This was the first piece of evidence for the hot Big Bang – the idea that the early universe was filled with a fireball – and remains one of the most compelling. In this section, we describe some of the basic properties of this radiation.

2.2.1 Blackbody Radiation

To start, we want to derive the properties of a thermal gas of photons. Such a gas in known, unhelpfully, as *blackbody radiation*.

The state of a single photon is specified by its momentum $\mathbf{p} = \hbar\mathbf{k}$, with $\mathbf{k}$ the *wavevector*. The energy of the photon is given by

$$E = pc = \hbar\omega$$

where $\omega = ck$ is the (angular) frequency of the photon.

Blackbody radiation comes with a new conceptual ingredient, because the number of photons is not a conserved quantity. This means that when considering the possible states of the gas, we should include states with an arbitrary number of photons. We do this by stating how many photons $N(\mathbf{p})$ sit in the state $\mathbf{p}$.

In thermal equilibrium, we will not have a definite number of photons $N(\mathbf{p})$, but rather some probability distribution over the number of photons, Focussing on a fixed state $\mathbf{p} = \hbar\mathbf{k}$, the average number of particles is dictated by the Boltzmann distribution

$$\langle N(\mathbf{p})\rangle = \frac{1}{Z}\sum_{n=0}^{\infty} ne^{-\beta n\hbar\omega} \quad \text{with } Z = \sum_{n=0}^{\infty} e^{-\beta n\hbar\omega}$$

We can easily do both of these sums. Defining $x = e^{-\beta\hbar\omega}$, the partition function is given by

$$Z = \sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$$

Meanwhile the numerator of $\langle N(\mathbf{p})\rangle$ takes the form

$$\sum_{n=0}^{\infty} nx^n = x\sum_{n=0}^{\infty} nx^{n-1} = x\frac{dZ}{dx} = \frac{x}{(1-x)^2}$$

We learn that the average number of particles with momentum $\mathbf{p}$ is

$$\langle N(\mathbf{p})\rangle = \frac{1}{e^{\beta\hbar\omega} - 1} \tag{2.9}$$

For $k_BT \ll \hbar\omega$, the number of photons is exponentially small. In contrast, when $k_BT \gg \hbar\omega$, the number of photons grows linearly as $\langle N(\mathbf{p})\rangle \approx k_BT/\hbar\omega$.

Density of States

Our next task is to determine the average number of photons $\langle N(\omega)\rangle$ with given energy $\hbar\omega$. To do this, we must count the number of states $\mathbf{p}$ which have energy $\hbar\omega$.

It's easier to count objects that are discrete rather than continuous. For this reason, we'll put our system in a square box with sides of length L. At the end of the calculation, we can happily send $L \to \infty$. In such a box, the wavevector is quantised: it takes values

$$k_i = \frac{2\pi q_i}{L} \qquad q_i \in \mathbf{Z}$$

This is true for both a classical wave or a quantum particle; in both cases, an integer number of wavelengths must fit in the box.

Different states are labelled by the integers q_i. When counting, or summing over such states, we should therefore sum over the q_i. However, for very large boxes, so that L is much bigger than any other length scale in the game, we can approximate this sum by an integral,

$$\sum_q \approx \frac{L^3}{(2\pi)^3}\int d^3k = \frac{4\pi V}{(2\pi)^3}\int_0^\infty dk\ k^2 \tag{2.10}$$

where $V = L^3$ is the volume of the box. The formula above counts all states. But the final form has a simple interpretation: the number of states with the magnitude of the wavevector between k and $k + dk$ is $4\pi V k^2/(2\pi)^3$. Note that the $4\pi k^2$ term is reminiscent of the $4\pi v^2$ term that appeared in the Maxwell-Boltzmann distribution; both have the same origin.

We would like to compute the number of states with frequency between ω and $\omega + d\omega$. For this, we simply use

$$\omega = ck \quad \Rightarrow \quad \frac{4\pi V}{(2\pi)^3}\int dk\ k^2 = \frac{4\pi V}{(2\pi c)^3}\int d\omega\ \omega^2$$

This tells us that the number of states with frequency between ω and $\omega + d\omega$ is $4\pi V\omega^2/(2\pi c)^3$.

There is one final fact that we need. Photons come with two polarisation states. This means that the total number of states is twice the number above. We can now combine this with our earlier result (2.9). In thermal equilibrium, the average number of photons with frequency between ω and $\omega + d\omega$ is

$$\langle N(\omega)\rangle\, d\omega = 2 \times \frac{4\pi V}{(2\pi c)^3}\frac{\omega^2}{e^{\beta\hbar\omega} - 1}\, d\omega$$

We usually write this in terms of the number density $n = N/V$. Moreover, we will be a little lazy and drop the expectation value $\langle n\rangle$ signs. The distribution of photons in a

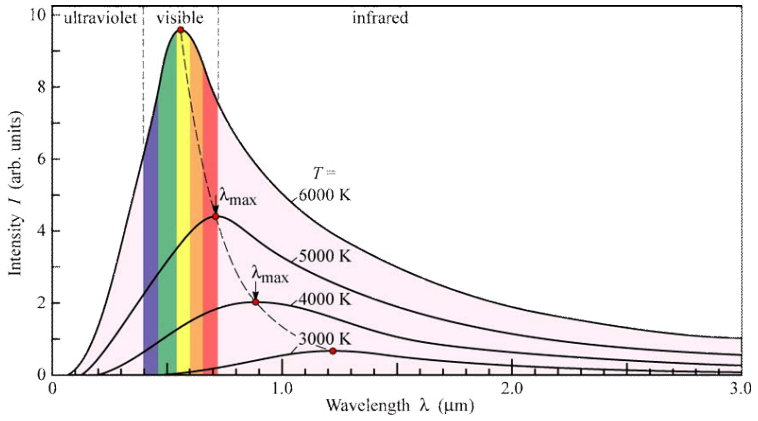

Figure 28: The distribution of colours at various temperatures.

thermal bath is then written as

$$n(\omega) = \frac{1}{\pi^2 c^3} \frac{\omega^2}{e^{\beta\hbar\omega} - 1} \tag{2.11}$$

This is the *Planck blackbody distribution*. For a fixed temperature, $\beta = 1/k_B T$, the distribution tells us how many photons of a given frequency – and hence, of a given colour – are present. The distribution peaks in visible light for temperatures around 6000 K, which is the temperature of the surface of the Sun. (Presumably the Sun evolved to be at exactly the right temperature so that our eyes can see it. Or something.)

The Equation of State

We now have all the information that we need to compute the equation of state. First the energy density. This is straightforward: we just need to integrate

$$\rho = \int_0^\infty d\omega \; \hbar\omega n(\omega) \tag{2.12}$$

Next the pressure. We can import our previous formula (2.7), now with $\mathbf{v} \cdot \mathbf{p} = \hbar ck = \hbar\omega$. But this gives precisely the same integral as the energy density; it differs only by the overall factor of 1/3,

$$P = \frac{1}{3}\rho$$

This, of course, is the relativistic equation of state that we used when describing the expanding universe.

Finally, we can actually do the integral (2.12). In fact, there's a couple of quantities of interest. The energy density is

$$\rho = \frac{\hbar}{\pi^2 c^3}\int_0^\infty d\omega\ \frac{\omega^3}{e^{\beta\hbar\omega}-1} = \frac{(k_B T)^4}{\pi^2\hbar^3 c^3}\int_0^\infty dy\ \frac{y^3}{e^y-1}$$

Meanwhile, the total number density is

$$n = \int_0^\infty d\omega\ n(\omega) = \frac{1}{\pi^2 c^3}\int_0^\infty d\omega\ \frac{\omega^2}{e^{\beta\hbar\omega}-1} = \frac{(k_B T)^3}{\pi^2\hbar^3 c^3}\int_0^\infty dy\ \frac{y^2}{e^y-1}$$

Both of these integrals take a similar form. Here we just quote the general result without proof:

$$I_n = \int_0^\infty dy\ \frac{y^n}{e^y-1} = \Gamma(n+1)\zeta(n+1) \tag{2.13}$$

The Gamma function is the analytic continuation of the factorial function to the real numbers; when evaluated on the integers it gives $\Gamma(n+1) = n!$. Meanwhile, the Riemann zeta function is defined, for $\mathrm{Re}(s) > 1$, as $\zeta(s) = \sum_{q=1} q^{-s}$. It turns out that $\zeta(4) = \pi^4/90$, giving us $I_3 = \pi^4/15$. In contrast, there is no such simple expression for $\zeta(3) \approx 1.20$. It is sometimes referred to as *Apéry's constant*. A derivation of (2.13) can be found in Section 3.5.3 of the lectures on Statistical Physics.

We learn that the energy density is

$$\rho = \frac{\pi^2}{15\hbar^3 c^3}(k_B T)^4 \tag{2.14}$$

Meanwhile, the total number density is

$$n = \frac{2\zeta(3)}{\pi^2\hbar^3 c^3}(k_B T)^3 \tag{2.15}$$

Notice, in particular, that the number density of photons varies with the temperature. This will be important in what follows.

2.2.2 The CMB Today

The universe today is filled with a sea of photons, the cosmic microwave background. This is the afterglow of the fireball that filled the universe in its earliest moments. The frequency spectrum of the photons is a perfect fit to the blackbody spectrum, with at a temperature

$$T_{\mathrm{CMB}} = 2.726\ \pm\ 0.0006\ \mathrm{K} \tag{2.16}$$

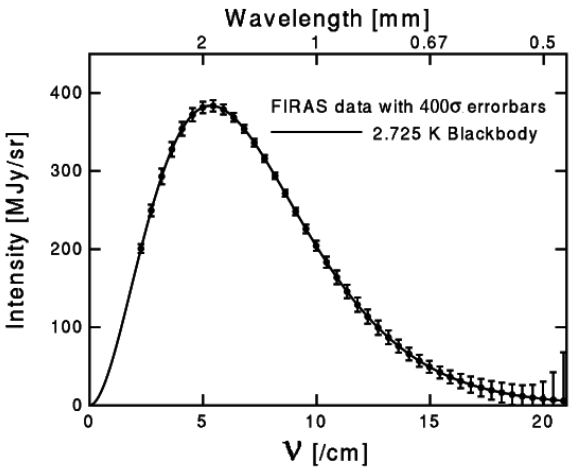

Figure 29: The blackbody spectrum of the CMB, measured in 1990 by the FIRAS detector on the COBE satellite. The error bars have been enlarged by a factor of 400 just to help you see them.

This spectrum is shown in Figure 29. There are small, local deviations in this temperature at the level of

$$\frac{\Delta T}{T_{\rm CMB}} \sim 10^{-5}$$

These fluctuations will be discussed further in Section 3.4.

From the temperature (2.16), we can determine the energy density and number density in photons. From (2.14), the energy density is given by

$$\rho_\gamma \approx 4.3 \times 10^{-14}\ {\rm kg\,m^{-1}}s^{-2}$$

We can compare this to the critical energy density (1.71), $\rho_{\rm crit,0} = 8.5\times 10^{-10}\ {\rm kg\,m^{-1}}s^{-2}$ to find

$$\Omega_\gamma = \frac{\rho_\gamma}{\rho_{\rm crit,0}} \approx 5 \times 10^{-5}$$

This is the value (1.69) that we quoted previously. There are, of course, further photons in starlight, but they are dwarfed in both energy and number by the CMB.

From (2.15), the number density of CMB photons is

$$n_\gamma = 4 \times 10^8\ {\rm m}^{-3} = 400\ {\rm cm}^{-3}$$

We can compare this to the number of baryons (i.e. protons and neutrons). The density of baryons is (1.74) $\Omega_B \approx 0.05$, so the total mass in baryons is

$$\rho_B \approx \Omega_B \rho_{\rm crit,0} \approx 4 \times 10^{-11}\ {\rm kg\,m^{-1}}s^{-2}$$

The mass of the proton and neutron are roughly the same, at $m_p \approx 1.7 \times 10^{-27}$ kg. This places the number density of baryons as

$$n_B = \frac{\rho_B}{m_p c^2} \approx 0.3\ {\rm m}^{-3}$$

We see that there are many more photons in the universe than baryons: the ratio is

$$\eta \equiv \frac{n_B}{n_\gamma} \approx 10^{-9} \tag{2.17}$$

This is one of the fundamental numbers in cosmology. As we will see, this ratio has been pretty much constant since the first second or so after the Big Bang and plays a crucial role in both nucleosynthesis (the formation of heavier nuclei) and in recombination (the formation of atoms). We do not, currently, have a good theoretical understanding of where this number fundamentally comes from: it is something that we can only derive from observation.

The CMB is a Relic

There is an important twist to the story above. We have computed the expected distribution of photons in thermal equilibrium, and found that it matches perfectly with the spectrum of the cosmic microwave background. The twist is that the CMB is *not* in equilibrium!

Recall that equilibrium is a property that arises when particles are constantly interacting. Yet the CMB photons have barely spoken to anyone for the past 13 billion years. The occasional photon may bump into a planet, or an infra-red detector fitted to a satellite, but most just wend their merry way through the universe, uninterrupted.

How then did the CMB photons come to form a perfect equilibrium spectrum? The answer is that this dates from a time when the photons were interacting frequently with matter. Fluids like this, that have long since fallen out of thermal equilibrium, but nonetheless retain their thermal character, are called *relics*.

There are a couple of questions that we would like to address. The first is: when were the photons last interacting and, hence, last genuinely in equilibrium? This is called the time of *last scattering*, $t_{\rm last}$ and we will compute it in Section 2.3 below. The second question is: what happened to the distribution of photons subsequently?

We start by answering the second of these questions. Once the photons no longer interact, they are essentially free particles. As the universe expands, each photon is redshifted as explained in Section 1.1.3. This means that the wavelength is stretched and, correspondingly, the frequency is decreased as the universe expands.

$$\lambda(t) = \lambda_{\rm last}\frac{a(t)}{a(t_{\rm last})} \quad \Rightarrow \quad \omega(t) = \omega_{\rm last}\frac{a(t_{\rm last})}{a(t)} \tag{2.18}$$

At the same time, the number of photons is diluted by a factor of $\big(a(t_{\rm last})/a(t)\big)^3$ as the universe expands. Putting these two effects together, an initial blackbody distribution (2.11) will, if left alone, evolve as

$$n(\omega_{\rm last}; T_{\rm last}, t)d\omega_{\rm last} = \frac{1}{\pi^2 c^3}\left(\frac{a(t_{\rm last})}{a(t)}\right)^3 \frac{\omega_{\rm last}^2}{e^{\beta\hbar\omega_{\rm last}}-1}d\omega_{\rm last}$$

The $1/a^3$ dilution factor is absorbed into the frequency in the ω^2 and $d\omega$ terms. But not in the exponent. However, the resulting distribution can be put back into blackbody form if we think of the temperature as time dependent

$$n(\omega; T, t)d\omega = \frac{1}{\pi^2 c^3}\frac{\omega(t)^2}{e^{\beta(t)\hbar\omega(t)}-1}d\omega(t)$$

where the $\beta(t) = 1/k_B T(t)$, with the time varying temperature

$$T(t) = T_{\rm last}\frac{a(t_{\rm last})}{a(t)} \tag{2.19}$$

We see that, left alone, a blackbody distribution will keep the same overall form, but with the temperature scaling as $T \sim 1/a$.

This means that, if we can figure out the temperature $T_{\rm last}$ when the photons were last in equilibrium, then we can immediately determine the redshift at which this occurred $1 + z_{\rm last} = a(t_{\rm last})^{-1}$. We'll compute both of these in Section 2.3.

2.2.3 The Discovery of the CMB

In 1964, two radio astronomers, Arno Penzias and Robert Wilson, got a new toy. The microwave horn antenna was originally used by their employers, the Bell telephone company, for satellite communication. Now Penzias and Wilson hoped to do some science with it, measuring the radio noise emitted in the direction away from the plane of the galaxy.

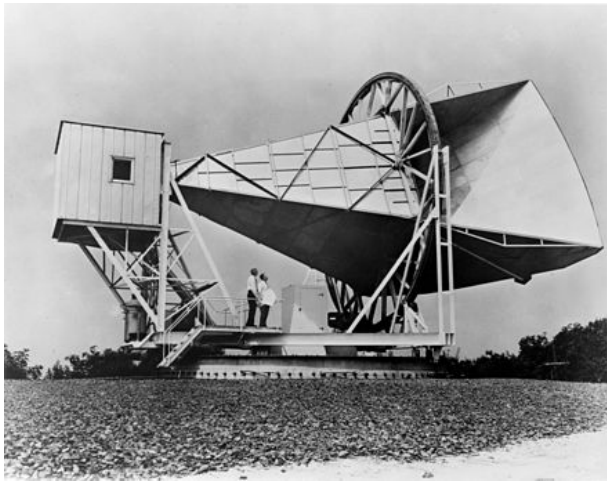

Figure 30: The Holmdel radio antenna at Bell Telephone Laboratories.

To their surprise, they found a background noise which did not depend on the direction in which they pointed their antenna. Nor did it depend on the time of day or the time of a year. Taken seriously, this suggested that the noise was a message from the wider universe.

There was, however, an alternative, more mundane explanation. Maybe the noise was coming from the antenna itself, some undiscovered systematic effect that they had failed be take into account. Indeed, they soon found a putative source of the noise: a pair of pigeons had taken roost and deposited what Penzias called "a white dielectric material" over much of the antenna. They removed this material (and shot the pigeons), but the noise remained. What Penzias and Wilson had on their hands was not pigeon shit, but one of the great discoveries of the twentieth century: the afterglow of the Big Bang itself, with a temperature that they measured to lie between 2.5 K and 4.5 K. In 1965 they published their result with the attention-grabbing title: "A Measurement of Excess Antenna Temperature at 4080 Mc/s".

Penzias and Wilson were not unaware of the significance of their finding. In the year since they first found the noise, they had done what good scientists should always do: they talked to their friends. They were soon put in touch with the group in nearby Princeton where Jim Peebles, a theoretical cosmologist, had recently predicted a background radiation with a temperature of a few degrees, based on the idea of nucleosynthesis in the very early universe (an idea we will describe in Section 2.5.3). Meanwhile, three experimental colleagues, Dicke, Roll and Wilkinson had cobbled together a small antenna in the hope of searching for this radiation. These four scientists

wrote a companion paper, outlining the importance of the discovery. In 1978, Penzias and Wilson were awarded the Nobel prize. It took another 39 years before Peebles gained the same recognition.

In fact there had been earlier predictions of the CMB. In the 1940s, Gammow together with Alpher and Herman suggested that the early universe began only with neutrons and, through somewhat dodgy calculations, concluded that there should be a background radiation at 5 K. Later other scientists, including Zel'dovich in the Soviet Union, and Hoyle and Taylor in England, used nucleosynthesis to predict the existence of the CMB at a few degrees. Yet none of these results were taken sufficiently seriously to search for the signal before Penzias and Wilson made their serendipitous discovery.

Detecting the CMB was just the beginning of the story. The radiation is not, it turns out, perfectly uniform but contains small anisotropies. These contain precious information about the make-up of the universe when it was much younger. A number of theorists, including Harrison, Zel'dovich, and Peebles and Yu, predicted that these anisotropies could be observed at a level of 10^{-4} to 10^{-5}. These were finally detected by the NASA COBE satellite in the early 1990s. Since then a number of ground based telescopes, including BOOMERanG and MAXIMA, and a two full sky maps from the satellites WMAP and Planck, have mapped out the CMB in exquisite detail. We will describe these anisotropies in Section 3.

2.3 Recombination

We've learned that the CMB is a relic, with its perfect blackbody spectrum a remnant of an earlier, more intense time in the universe, when the photons were in equilibrium with matter. We would like to gain a better understanding of this time.

Photons interact with electric charge. Nowadays, the vast majority of matter in the universe is in the form of neutral atoms, and electrons interact only with the charged constituents of the atoms. Such interactions are relatively weak. However, there was a time in the early universe when the temperature was so great that electrons and protons could no longer bind into neutral atoms. Instead, the universe was filled with a plasma. In this era, the matter and photons interacted strongly and were in equilibrium.

The CMB that we see today dates from this time. Or, more precisely, from the time when electrons and protons first bound themselves into neutral hydrogen, emitting a photon in the process

$$e^- + p^+ \leftrightarrow H + \gamma \tag{2.20}$$

The moment at which this occurs is called *recombination*. As the arrows illustrate, this process can happen in both directions.

Interactions like (2.20) involve one particle type transmuting into a different type. This means that the number of, say, hydrogen atoms is not fixed but fluctuating. We need to introduce a new concept that allows us to deal with such situations. This concept is the *chemical potential*.

2.3.1 The Chemical Potential

The chemical potential offers a slight generalisation of the Boltzmann distribution which is useful in situations where the number of particles in a system is not fixed. It was, as the name suggests, originally introduced to describe chemical reactions but we will re-purpose it to describe atomic reactions like (2.20) (and, later, nuclear reactions).

Although our ultimate goal is to describe atomic reactions, we can first introduce the chemical potential in a more mundane setting. Suppose that we have a fixed number of atoms N in a box of size V. If we focus attention on some large, fixed sub-volume $V' \subset V$, then we would expect the gas in V' to share the same macroscopic properties, such as temperature and pressure, as the whole gas in V. But particles can happily fly in and out of V' and the total number in this region is not fixed. Instead, there is some probability distribution which has the property that the average number density coincides with N/V.

In this situation, it's clear that we should consider states of all possible particle number in V'. There is a possibility, albeit a very small one, that V' contains no particles at all. There is also a small possibility that it contains all the particles.

If we work in the language of quantum mechanics, each state $|n\rangle$ in the system can be assigned both an energy E_n and a particle number N_n. Correspondingly, equilibrium states are characterised by two macroscopic properties: the temperature T and the chemical potential μ. These are defined through the generalised Boltzmann distribution

$$p(n) = \frac{e^{-\beta(E_n - \mu N_n)}}{\mathcal{Z}} \tag{2.21}$$

where $\mathcal{Z} = \sum_n e^{-\beta(E_n - \mu N_n)}$ is again the appropriate normalisation factor. In the language of statistical mechanics, this is referred to as the *grand canonical ensemble*.

Clearly, the distribution has the same exponential form as the Boltzmann distribution. This is important. We learned in Section 2.1.1 that two isolated systems which sit at the same temperature will remain in thermal equilibrium when brought together,

meaning that there will be no transfer of energy from one system to the other. Exactly the same argument tells us that if two isolated systems have the same chemical potential then, when brought together, there will be no net flux of particles from one system to the other. In this case, we say that the systems are in *chemical equilibrium*.

Notice that the requirement for equilibrium is *not* that the number densities of the systems are equal: it is the chemical potentials that must be equal. This is entirely analogous to the statement that it is temperature, rather than energy density, that determines whether systems are in thermal equilibrium.

We'll see examples of how to wield the chemical below, but before we do it's worth mentioning a few issues.

- In general, we can introduce a different chemical potential for every conserved quantity in the system. This is because conserved quantities commute with the Hamiltonian, and so it makes sense to label microscopic states by both the energy and a further quantum number. One familiar example is electric charge Q. Here, the corresponding chemical potential is voltage.

  This leads to an almost-contradictory pair of statements. First, we can only introduce a chemical potential for any conserved quantity. Second, the purpose of the chemical potential is to allow this conserved quantity to fluctuate! If you're confused about this, then think back to the volume $V' \subset V$, or to the meaning of voltage in electromagnetism, both of which give examples where these statements hold.

- The story above is very similar to our derivation of the Planck blackbody distribution for photons. There too we labeled states by both energy and particle number, but we didn't introduce a chemical potential. What's different now? This is actually a rather subtle issue. Ultimately it is related to the fact that we ignore interactions while simultaneously pretending that they are crucial to reach equilibrium. As soon as we take these interactions into account, the number of photons is not conserved so we can't label states by both energy and photon number. This is what prohibits us from introducing a chemical potential for photons. In contrast, we can introduce a chemical potential in situations where particle number (or some other quantity) is conserved even in the presence of interactions.

2.3.2 Non-Relativistic Gases Revisited

For our first application of the chemical potential, we're going to re-derive the ideal gas equation. At first sight, this will appear to be only a more complicated derivation

of something we've seen already. The pay-off will come only in Section 2.3.3 where we will understand recombination and the atomic reaction (2.20).

We consider non-relativistic particles, with energy

$$E_{\mathbf{p}} = \frac{p^2}{2m}$$

As with our calculation of photons, we now consider states that have arbitrary numbers of particles. We choose to specify these states by stating how many particles $n_{\mathbf{p}}$ have momentum $\mathbf{p}$. For each choice of momentum, the number of particles[7] can be $n_{\mathbf{p}} = 0, 1, 2, \ldots$. The generalised Boltzmann distribution (2.21) then tells us that the average number of particles with momentum $\mathbf{p}$ is

$$\langle N(\mathbf{p})\rangle = \frac{1}{\mathcal{Z}_{\mathbf{p}}}\sum_{n_{\mathbf{p}}=0}^{\infty} n_{\mathbf{p}} e^{-\beta(n_{\mathbf{p}}E_{\mathbf{p}}-\mu n_{\mathbf{p}})}$$

where the normalisation factor (or, in fancy language, the grand canonical partition function) is given by the geometric series

$$\mathcal{Z}_{\mathbf{p}} = \sum_{n=0}^{\infty} e^{-\beta n_{\mathbf{p}}(E_{\mathbf{p}}-\mu)} = \frac{e^{\beta(E_{\mathbf{p}}-\mu)}}{e^{\beta(E_p-\mu)}-1}$$

This is exactly the same calculation as we saw for photons in Section 2.2.1, but with the additional minor complication of a chemical potential. Note that computing $\mathcal{Z}_{\mathbf{p}}$ allows us to immediately determine the expected number of particles since we can write

$$\langle N(\mathbf{p})\rangle = \frac{1}{\beta}\frac{\partial}{\partial\mu}\log\mathcal{Z}_{\mathbf{p}} = \frac{1}{e^{\beta(E_{\mathbf{p}}-\mu)}-1} \tag{2.22}$$

This is known as the *Bose-Einstein distribution* and will be discussed further in Section 2.4.

To compute the average total number of particles, we simply need to integrate over all momenta $\mathbf{p}$. We must include the density of states, but this is identical to the calculation we did for photons, with the result (2.10). The total average number of particles is then

$$N = \frac{V}{(2\pi\hbar)^3}\int d^3p\; N(\mathbf{p})$$

[7]Actually, there is a subtlety here: I am implicitly assuming that the particles are bosons. We'll look at this more closely in Section 2.4.

where we've been a little lazy and dropped the $\langle\cdot\rangle$ brackets on $N(\mathbf{p})$. We usually write this in terms of the particle density $n = N/V$,

$$n = \frac{1}{(2\pi\hbar)^3}\int d^3p\; N(\mathbf{p}) = \frac{4\pi}{(2\pi\hbar)^3}\int_0^\infty dp\; \frac{p^2}{e^{-\beta\mu}e^{\beta p^2/2m}-1} \tag{2.23}$$

where, in the second equality, we have chosen to integrate using spherical polar coordinates, picking up a factor of 4π from the angular integrals and a factor of p^2 in the Jacobian for our troubles. We have also used the explicit expression $E_{\mathbf{p}} = p^2/2m$ for the energy in the distribution.

At this stage, we have an annoying looking integral to do. To proceed, let's pick a value of the chemical potential μ such that $e^{-\beta\mu} \gg 1$. (We'll see what this means physically below.) We can then drop the -1 in the denominator and approximate the integral as

$$n \approx \frac{4\pi}{(2\pi\hbar)^3}e^{\beta\mu}\int_0^\infty dp\; p^2\, e^{-\beta p^2/2m} = \left(\frac{mk_BT}{2\pi\hbar^2}\right)^{3/2} e^{\beta\mu} \tag{2.24}$$

Let's try to interpret this. Read naively, it seems to tell us that the number density of particles depends on the temperature. But that's certainly *not* what happens for the gas in this room, where ρ and P depend on temperature but the number density $n = N/V$ is fixed. We can achieve this by taking the chemical potential μ to also depend on temperature. Specifically, we wish to describe a gas with fixed n , then we simply invert the equation above to get an expression for the chemical potential

$$e^{\beta\mu} = \left(\frac{2\pi\hbar^2}{mk_BT}\right)^{3/2} n \tag{2.25}$$

Before we proceed, we can use this result to understand what the condition $e^{-\beta\mu} \gg 1$, that we used to do the integral, is forcing upon us. Comparing to the expression above, it says that the number density is bounded above by

$$n \ll \left(\frac{mk_BT}{2\pi\hbar^2}\right)^{3/2}$$

This is sensible. It's telling us that the ideal gas can't be too dense. In particular, the average distance between particles should be much larger than the length scale set by $\lambda = \sqrt{2\pi\hbar^2/mk_BT}$. This is the average de Broglie wavelength of particles at temperature T. If n is increased so that the separation between particles is comparable to λ then quantum effects kick in and we have to return to our original integral (2.23) and make a different approximation to do the integral and understand the physics. (This path will lead to the beautiful phenomenon of Bose-Einstein condensation, but it is a subject for a different course.)

We can now calculate the energy density and pressure. Once again, taking the limit $e^{-\beta\mu} \gg 1$, the energy density is given by

$$\begin{aligned}\rho &= \frac{1}{(2\pi\hbar)^3}\int d^3p\; E_{\mathbf{p}}\, N(\mathbf{p}) \\ &\approx \frac{4\pi}{(2\pi\hbar)^3}e^{\beta\mu}\int_0^\infty dp\; \frac{p^4}{2m}e^{-\beta p^2/2m} = \frac{3}{2}nk_BT\end{aligned}$$

This is a result that we have met before (2.8). Meanwhile, we can use our expression (2.7) to compute the pressure,

$$\begin{aligned}P &= \frac{1}{(2\pi\hbar)^3}\int d^3p\; \frac{\mathbf{v}\cdot\mathbf{p}}{3}\, N(\mathbf{p}) \\ &= \frac{4\pi}{(2\pi\hbar)^3}e^{\beta\hat{\mu}}\int_0^\infty dp\; \frac{p^4}{3m}\, e^{-\beta p^2/2m} = nk_BT\end{aligned}$$

Again, this recovers the familiar ideal gas equation.

So far, the chemical potential has not bought us anything new. We have simply recovered old results in a slightly more convoluted framework in which the number of particles can fluctuate. But, as we will now see, this is exactly what we need to deal with atomic reactions.

2.3.3 The Saha Equation

We would like to consider a gas of electrons and protons in equilibrium at some temperature. They have the possibility to combine and form hydrogen, which we will think of as an atomic reaction, akin to the chemical reactions that we met in school. It is

$$e^- + p^+ \leftrightarrow H + \gamma$$

The question we would like to ask is: what proportion of the particles are hydrogen, and what proportion are electron-proton pairs?

To simplify life, we will assume that the hydrogen atom forms in its ground state, with a binding energy

$$E_{\rm bind} \approx 13.6 \text{ eV}$$

In fact, this turn out to be a bad assumption! We explain why at the end of this section.

Naively, we would expect hydrogen to ionize when we reach temperatures of $k_BT \approx E_{\rm bind}$. It's certainly true that for temperature $k_BT \gg E_{\rm bind}$, the electrons can no longer cling on to the protons, and any hydrogen atom is surely ripped apart. However, it will ultimately turn out that hydrogen only forms at temperatures significantly lower than $E_{\rm bind}$.

We'll treat each of the massive particles – the electron, proton and hydrogen atom – in a similar way to the non-relativistic gas that we met in Section 2.3.2. There will, however, be two differences. First, we include the rest mass energy of the atoms, so each particle has energy

$$E_{\mathbf{p}} = mc^2 + \frac{p^2}{2m}$$

This will be useful as we can think of the binding energy $E_{\rm bind}$ as the mass difference

$$(m_e + m_p - m_H)c^2 = E_{\rm bind} \approx 13.6 \text{ eV} \tag{2.26}$$

Secondly, each of our particles comes with a number g of internal states. The electron and proton each have $g_e = g_p = 2$ corresponding to the two spin states, referred to as "spin up" and "spin down". (These are analogous to the two polarisation states of the photon that we included when discussing blackbody radiation.) For hydrogen, we have $g_H = 4$; the electron and proton spin can either be aligned, to give a spin 0 particle, or anti-aligned to give 3 different spin 1 states.

With these two amendments, our expression for the number density (2.24) of the different species of particles is given by

$$n_i = g_i \left(\frac{m_i k_B T}{2\pi\hbar^2}\right)^{3/2} e^{-\beta(m_i c^2 - \mu_i)} \tag{2.27}$$

Note that the rest mass energy mc^2 in the energy can be absorbed by a constant shift of the chemical potential.

Now we can use the chemical potential for something new. We require that these particles are in chemical equilibrium. This means that there is no rapid change from $e^- + p^+$ pairs into hydrogen, or vice versa: the numbers of electrons, protons and hydrogen are balanced. This is ensured if the chemical potentials are related by

$$\mu_e + \mu_p = \mu_H \tag{2.28}$$

This follows from our original discussion of what it means to be in chemical equilibrium. Recall that if two isolated systems have the same chemical potential then, when brought together, there will be no net flux of particles from one system to the other. This mimics the statement about thermal equilibrium, where if two isolated systems have the same temperature then, when brought together, there will be no net flux of energy from one to the other.

There is no chemical potential for photons because they're not conserved. In particular, in addition to the reaction $e^- + p^+ \leftrightarrow H + \gamma$ there can also be reactions in which the binding results in two photons, $e^- + p^+ \leftrightarrow H + \gamma + \gamma$, which is ultimately why it makes no sense to talk about a chemical potential for photons. (Some authors write this, misleadingly, as $\mu_\gamma = 0$.)

We can use the condition for chemical equilibrium (2.28) to eliminate the chemical potentials in (2.27) to find

$$\frac{n_H}{n_e n_p} = \frac{g_H}{g_e g_p}\left(\frac{m_H}{m_e m_p}\frac{2\pi\hbar^2}{k_B T}\right)^{3/2} e^{-\beta(m_H - m_e - m_p)c^2} \tag{2.29}$$

In the pre-factor, it makes sense to approximate $m_H \approx m_p$. However, in the exponent, the difference between these masses is crucial; it is the binding energy of hydrogen (2.26). Finally, we use the observed fact that the universe is electrically neutral, so

$$n_e = n_p$$

We then have

$$\frac{n_H}{n_e^2} = \left(\frac{2\pi\hbar^2}{m_e k_B T}\right)^{3/2} e^{\beta E_{\rm bind}} \tag{2.30}$$

This is the *Saha equation*.

Our goal is to understand the fraction of electron-proton pairs that have combined into hydrogen. To this end, we define the *ionisation fraction*

$$X_e = \frac{n_e}{n_B} \approx \frac{n_e}{n_p + n_H}$$

where, in the second equality, we're ignoring neutrons and higher elements. (We'll see in Section 2.5.3 that this is a fairly good approximation.) Since $n_e = n_p$, if $X_e = 1$ it means that all the electrons are free. If $X_e = 0.1$, it means that only 10% of the electrons are free, the remainder bound inside hydrogen.

Using $n_e = n_p$, we have $1 - X_e = n_H/n_B$ and so

$$\frac{1 - X_e}{X_e^2} = \frac{n_H}{n_e^2} n_B$$

The Saha equation gives us an expression for n_H/n_e^2. But to translate this into the fraction X_e, we also need to know the number of baryons. This we take from observation. First, we convert the number of baryons into the number of photons, using (2.17),

$$\eta = \frac{n_B}{n_\gamma} \approx 10^{-9}$$

Here we need to use the fact that $\eta \approx 10^{-9}$ has remained constant since recombination. Next, we use the fact that photons sit at the same temperature as the electrons, protons and hydrogen because they are all in equilibrium. This means that we can then use our earlier expression (2.15) for the number of photons

$$n_\gamma = \frac{2\zeta(3)}{\pi^2\hbar^3c^3}(k_BT)^3$$

Combining these gives our final answer

$$\frac{1-X_e}{X_e^2} = \eta\,\frac{2\zeta(3)}{\pi^2}\left(\frac{2\pi k_BT}{m_ec^2}\right)^{3/2} e^{\beta E_{\rm bind}} \tag{2.31}$$

Suppose that we look at temperature $k_BT \sim E_{\rm bind}$, which is when we might naively have thought recombination takes place. We see that there are two very small numbers in the game: the factor of $\eta \sim 10^{-9}$ and k_BT/m_ec^2, where the electron mass is $m_ec^2 \approx 0.5$ MeV $= 5\times 10^5$ eV. These ensure that at $k_BT \sim E_{\rm bind}$, the ionisation fraction X_e is very close to unity. In other words, nearly all the electrons remain free and unbound. In large part this is of the enormous number of photons, which mean that whenever a proton and electron bind, one can still find sufficient high energy photons in the tail of the blackbody distribution to knock them apart.

Recombination only takes place when the $e^{\beta E_{\rm bind}}$ factor is sufficient to compensate both the η and k_BT/m_ec^2 factors. Clearly recombination isn't a one-off process; it happens continuously as the temperature varies. As a benchmark, we'll calculate the temperature when $X_e = 0.1$, so 90% of the electrons are sitting happily in their hydrogen homes. From (2.31), we learn that this occurs when $\beta E_{\rm bind} \approx 45$, or

$$k_BT_{\rm rec} \approx 0.3 \text{ eV} \quad \Rightarrow \quad T_{\rm rec} \approx 3600 \text{ K}$$

This corresponds to a redshift of

$$z_{\rm rec} = \frac{T_{\rm rec}}{T_0} \approx 1300$$

This is significantly later than matter-radiation equality which, as we saw in (1.73), occurs at $z_{\rm eq} \approx 3400$. This means that, during recombination, the universe is matter dominated, with $a(t) \sim (t/t_0)^{2/3}$. We can therefore date the time of recombination to,

$$t_{\rm rec} \approx \frac{t_0}{(1+z_{\rm rec})^{3/2}} \approx 300,000 \text{ years}$$

After recombination, the constituents of the universe have been mostly neutral atoms. Roughly speaking this means that the universe is transparent and photons can propagate freely. We will look more closely at this statement a little more closely below.

Mea Culpa

The full story is significantly more complicated than the one told above. As we have seen, at the time of recombination the temperature is much lower than the 13.6 eV binding energy of the 1s state of hydrogen. This means that whenever a 1s state forms, it emits a photon which has significantly higher energy that the photons in thermal bath. The most likely outcome is that this high energy photon hits a different hydrogen atom, splitting it into its constituent proton and electron, resulting in no net change in the number of atoms! Instead, recombination must proceed through a rather more tortuous route.

The hydrogen atom doesn't just have a ground state: there are a whole tower of excited states. These can form without emitting a high energy photon and, indeed, at these low temperatures the thermal bath of photons is in equilibrium with the tower of excited states of hydrogen. There are then two, rather inefficient processes, which populate the 1s state. The 2s state decays down to 1s by emitting two photons (to preserve angular momentum), neither of which have enough energy to re-ionize other atoms. Alternatively, the 2p state can decay to 1s, emitting a photon whose energy is barely enough to excite another hydrogen atom out of the ground state. If this photon experiences redshift, then it can no longer do the job and we increase the number of atoms in the ground state. More details can be found in the book by Weinberg. These issues do not greatly change the values of $T_{\rm rec}$ and $z_{\rm rec}$ that we computed above.

2.3.4 Freeze Out and Last Scattering

Photons interact with electric charge. After electrons and protons combine to form neutral hydrogen, the photons scatter much less frequently and the universe becomes transparent. After this time, the photons are essentially decoupled.

Similar scenarios play out a number of times in the early universe: particles, which once interacted frequently, stop talking to their neighbours and subsequently evolve without care for what's going on around them. This process is common enough that it is worth exploring in a little detail. As we will see, at heart it hinges on what it means for particle to be in "equilibrium".

Strictly speaking, an expanding universe is a time dependent background in which the concept of equilibrium does not apply. In most situations, such a comment would be rightly dismissed as the height of pedantry. The expansion of the universe does not, for example, stop me applying the laws of thermodynamics to my morning cup of tea. However, in the very early universe this can become an issue.

For a system to be in equilibrium, the constituent particles must frequently interact, exchanging energy and momentum. For any species of particle (or pair of species) we can define the *interaction rate* Γ. A particle will, on average, interact with another particle in a time $t_{\rm int} = 1/\Gamma$. It makes sense to talk about equilibrium provided that the universe hasn't significantly changed in the time $t_{\rm int}$. The expansion of the universe is governed by the Hubble parameter, so we can sensibly talk about equilibrium provided

$$\Gamma \gg H$$

In contrast, if $\Gamma \ll H$ then by the time particles interact the universe has undergone significant expansion. In this case, thermal equilibrium cannot be maintained.

For many processes, both the interaction rate and temperature scale with time, but in different ways. The result is that particles retain equilibrium at early times, but decouple from the thermal bath at late time. This decoupling occurs when $\Gamma \approx H$ and is known as *freeze out*.

We now apply these ideas to photons, where freeze out also goes by the name of *last scattering*. In the early universe, the photons are scattered primarily by the electrons (because they are much lighter than the protons) in a process known as *Thomson scattering*

$$e + \gamma \to e + \gamma$$

The scattering is elastic, meaning that the energy, and therefore the frequency, of the photon is unchanged in the process. For Thomson scattering, the interaction rate is given by

$$\Gamma = n_e \sigma_T c$$

where σ_T is the cross-section, a quantity which characterises the strength of the scattering. We computed the cross-section for Thomson scattering in the lectures on Electromagnetism (see Section 6.3.1 of these lectures) where we showed it was given by

$$\sigma_T = \frac{\mu_0^2 e^4}{6\pi m_e^2 c^2} \approx 6 \times 10^{-30}\ {\rm m}^2$$

Note the dependence on the electron mass m_e; the corresponding cross-section for scattering off protons is more than a million times smaller.

Last scattering occurs at the temperature $T_{\rm last}$ such that $\Gamma(T_{\rm last}) \approx H(t_{\rm last})$. We can express the interaction rate by replacing the number density of electrons with the number density of photons,

$$\Gamma(T_{\rm last}) = n_B X_e(T_{\rm last})\sigma_T c = \eta\sigma_T \frac{2\zeta(3)}{\pi^2\hbar^3c^2}\,(k_B T_{\rm last})^3\, X_e(T_{\rm last}) \tag{2.32}$$

Meanwhile, we can trace back the current value of the Hubble constant, through the matter dominated era, to last scattering. Meanwhile, to compute $H(T_{\rm last})$, we use the formula (1.67)

$$\left(\frac{H}{H_0}\right)^2 = \frac{\Omega_r}{a^4} + \frac{\Omega_m}{a^3} + \frac{\Omega_k}{a^2} + \Omega_\Lambda$$

Evaluated at recombination, radiation, curvature and the cosmological constant are all irrelevant, and this formula becomes

$$\left(\frac{H}{H_0}\right)^2 \approx \frac{\Omega_m}{a^3}$$

Using the fact that temperature scales as $T \sim 1/a$, we then have

$$H(T_{\rm last}) = H_0\sqrt{\Omega_m}\left(\frac{T_{\rm last}}{T_0}\right)^{3/2}$$

Equating this with (2.32) gives

$$X_e(T_{\rm last})(k_B T_{\rm last})^{3/2} = \frac{\pi^2\hbar^3c^2}{2\zeta(3)}\frac{H_0\sqrt{\Omega_m}}{\eta\sigma_T(k_B T_0)^{3/2}}$$

Using (2.31) to solve for $X_e(T_{\rm last})$ (which is a little fiddly) we find that photons stop interacting with matter only when

$$X_e(T_{\rm last}) \approx 0.01$$

We learn that the vast majority of electrons must be housed in neutral hydrogen, with only 1% of the original electrons remaining free, before light can happily travel unimpeded. This corresponds to a temperature

$$k_B T_{\rm last} \approx 0.27 \text{ eV} \quad \Rightarrow \quad T_{\rm last} \approx 3100 \text{ K}$$

and, correspondingly, a time somewhat after recombination,

$$z_{\rm last} = \frac{T_{\rm last}}{T_0} \approx 1100 \quad \Rightarrow \quad t_{\rm last} = \frac{t_0}{(1+z_{\rm last})^{3/2}} \approx 350,000 \text{ years}$$

After this time, the universe becomes transparent. The cosmic microwave background is a snapshot of the universe from this time.

2.4 Bosons and Fermions

To better understand the physics of the Big Bang, there is one last topic from statistical physics that we will need to understand. This follows from a simple statement: quantum particles are indistinguishable. It's not just that the particles look the same: there is a very real sense in which there is no way to tell them apart.

Consider a state with two identical particles. Now swap the positions of the particles. This doesn't give us a new state: it is exactly the same state as before (at least up to a minus sign). This subtle effect plays a key role in thermal systems where we're taking averages over different states. The possibility of a minus sign is important, and means that quantum particles come in two different types, called *bosons* and *fermions*.

Consider a state with two identical particles. These particles are called *bosons* if the wavefunction is symmetric under exchange of the particles.

$$\psi(\mathbf{x}_1, \mathbf{x}_2) = \psi(\mathbf{x}_2, \mathbf{x}_1)$$

The particles are *fermions* if the wavefunction is anti-symmetric

$$\psi(\mathbf{x}_1, \mathbf{x}_2) = -\psi(\mathbf{x}_2, \mathbf{x}_1)$$

Importantly, if you try to put two fermions on top of each other then the wavefunction vanishes: $\psi(\mathbf{x}, \mathbf{x}) = 0$. This is a reflection of the *Pauli exclusion principle* which states that two or more fermions cannot sit in the same state. For both bosons and fermions, if you do the exchange twice then you get back to the original state.

There is a deep theorem – known as the spin-statistics theorem – which states that the type of particle is determined by its spin (an intrinsic angular momentum carried by elementary particles). Particles that have integer spin are bosons; particles that have half-integer spin are fermions.

Examples of spin 1/2 particles, all of which are fermions, include the electron, the various quarks, and neutrinos. Furthermore, protons and neutrons (which, roughly speaking, consist of three quarks) also have spin 1/2 and so are fermions.

The most familiar example of a boson is the photon. It has spin 1. Other spin 1 particles include the W and Z-bosons (responsible for the weak nuclear force) and gluons (responsible for the strong nuclear force). The only elementary spin 0 particle is the Higgs boson. Finally, the graviton has spin 2 and is also a boson.

While this exhausts the elementary particles, the ideas that we develop here also apply to composite objects like atoms. These too are either bosons or fermions. Since the number of electrons is always equal to the number of protons, it is left to the neutrons to determine the nature of the atom: an odd number of neutrons and it's a fermion; an even number and it's a boson.

2.4.1 Bose-Einstein and Fermi-Dirac Distributions

The generalised Boltzmann distribution (2.21) specifies the probability that we sit in a state $|n\rangle$ with some fixed energy E_n and particle number N_n.

In what follows, we will restrict attention to non-interacting particles. In this case, there is a simple way to construct the full set of states $|n\rangle$ starting from the single-particle Hilbert space. The state of a single particle is specified by its momentum $\mathbf{p} = \hbar\mathbf{k}$. (There may also be some extra, discrete internal degrees of freedom like polarisation or spin; we'll account for these later.) We'll denote this single particle state as $|\mathbf{p}\rangle$. For a relativistic particle, the energy is

$$E_{\mathbf{p}} = \sqrt{m^2c^4 + p^2c^2} \tag{2.33}$$

To specify the full multi-particle state $|n\rangle$, we need to say how many particles $n_{\mathbf{p}}$ occupy the state $|\mathbf{p}\rangle$. The possible values of $n_{\mathbf{p}}$ depend on whether the underlying particle is a boson or fermion:

$$\begin{aligned} \text{Bosons}: &\quad n_{\mathbf{p}} = 0, 1, 2, \ldots \\ \text{Fermions}: &\quad n_{\mathbf{p}} = 0, 1 \end{aligned}$$

In our previous discussions of blackbody radiation in Section 2.2.1 and the non-relativistic gas in Section 2.3.2, we did the counting appropriate for bosons. This is fine for blackbody radiation, since photons are bosons, but was an implicit assumption in the case of a non-relativistic gas.

The other alternative is a fermion. For these particles, the Pauli exclusion principle says that a given single-particle state $|\mathbf{p}\rangle$ is either empty or occupied. But you can't put more than one fermion there. This is entirely analogous to the way the periodic table is constructed in chemistry, by filling successive shells, except now the states are in momentum space. (A better analogy is the way a band is filled in solid state physics as described in the lectures on Quantum Mechanics.) For bosonic particles, there is no such restriction: you can pile up as many as you like.

Now we can compute some quantities, like the average particle number and average energy. We deal with bosons and fermions in turn

For bosons, the calculation is exactly the same as we saw in Section 2.3.2. For a given momentum $\mathbf{p}$, the average number of photons is

$$\langle N(\mathbf{p})\rangle = \frac{1}{\mathcal{Z}_{\mathbf{p}}}\sum_{n_{\mathbf{p}}=0}^{\infty} n_{\mathbf{p}} e^{-\beta(n_{\mathbf{p}}E_{\mathbf{p}}-\mu n_{\mathbf{p}})} = \frac{1}{\beta}\frac{\partial}{\partial\mu}\log\mathcal{Z}_{\mathbf{p}}$$

where the normalisation factor is given by the geometric series

$$\mathcal{Z}_{\mathbf{p}} = \sum_{n=0}^{\infty} e^{-\beta n_{\mathbf{p}}(E_{\mathbf{p}}-\mu)} = \frac{e^{\beta(E_{\mathbf{p}}-\mu)}}{e^{\beta(E_p-\mu)}-1}$$

As in the previous section, we will be a little lazy and drop the expectation value, so $\langle N(\mathbf{p})\rangle \equiv N(\mathbf{p})$. Then we have

$$N(\mathbf{p}) = \frac{1}{e^{\beta(E_{\mathbf{p}}-\mu)}-1} \tag{2.34}$$

This is known as the *Bose-Einstein distribution*.

For fermions, the calculation is easier still. We can have only $n_{\mathbf{p}} = 0$ or 1 particles in a given state $|\mathbf{p}\rangle$ so the average occupation number is

$$N(\mathbf{p}) = \frac{1}{\mathcal{Z}_p}\sum_{n_{\mathbf{p}}=0,1} n_{\mathbf{p}} e^{-\beta(n_{\mathbf{p}}E_{\mathbf{p}}-\mu n_{\mathbf{p}})} \quad \text{with} \quad \mathcal{Z}_{\mathbf{p}} = \sum_{n_{\mathbf{p}}=0,1} e^{-\beta(n_{\mathbf{p}}E_{\mathbf{p}}-\mu n_{\mathbf{p}})}$$

Again, keeping the $\langle\cdot\rangle$ expectation value signs implicit, we have

$$N(\mathbf{p}) = \frac{1}{e^{\beta(E_{\mathbf{p}}-\mu)}+1} \tag{2.35}$$

This is the *Fermi-Dirac distribution*.

For both bosons and fermions, the calculation of the density of states (2.10) proceeds as before, so that if we integrate over all possible momenta, it should be weighted by

$$\frac{4\pi V}{(2\pi\hbar)^3}\int d^3p$$

with the pre-factor telling us how quantum states are in a small region d^3p.

If we include the degeneracy factor g, which tells us the number of internal states of the particle, the number density $n = N/V$ is given by

$$n = \frac{g}{(2\pi\hbar)^3}\int d^3p\; N(\mathbf{p}) \tag{2.36}$$

Similarly, the energy density is

$$\rho = \frac{g}{(2\pi\hbar)^3}\int d^3p\; E_{\mathbf{p}}\, N(\mathbf{p}) \tag{2.37}$$

and the pressure (2.7) is

$$P = \frac{g}{(2\pi\hbar)^3}\int d^3p\; \frac{\mathbf{v}\cdot\mathbf{p}}{3}\, N(\mathbf{p}) \tag{2.38}$$

We'll now apply these in various examples.

The Non-Relativistic Gas Yet Again

In Section 2.3.2, we computed various quantities of a non-relativistic gas, so that the energy of each particle is

$$E_{\mathbf{p}} = \frac{p^2}{2m}$$

When we evaluated various quantities using the chemical potential approach, we implicitly assumed that the constituent atoms of the gas were bosons so, for example, our expression for the expression for the number density (2.23),

$$n_{\text{boson}} = \frac{g}{(2\pi\hbar)^3}\int d^3p\; N(\mathbf{p}) = \frac{4\pi g}{(2\pi\hbar)^3}\int_0^\infty dp\; \frac{p^2}{e^{-\beta\mu}e^{\beta p^2/2m}-1}$$

If, instead, we have a gas comprising of fermions then we should replace this expression with

$$n_{\text{fermion}} = \frac{g}{(2\pi\hbar)^3}\int d^3p\; N(\mathbf{p}) = \frac{4\pi g}{(2\pi\hbar)^3}\int_0^\infty dp\; \frac{p^2}{e^{-\beta\mu}e^{\beta p^2/2m}+1}$$

We can then ask: how does the physics change?

If we focus on the high temperature regime of non-relativistic gases, the answer to this question is: very little! This is because we evaluate these integrals using the approximation $e^{-\beta\mu} \gg 1$, and we can immediately drop the ± 1 in the denominator. This means that both bosons and fermions give rise to the same ideal gas equation.

We do start to see small differences in the behaviour of the gases if we expand the integrals to the next order in $e^{\beta\mu}$. We see much larger differences if we instead study the integrals in a very low-temperature limit. These stories are told in the lectures on Statistical Physics but they hold little cosmological interest.

Instead, the difference between bosons and fermions in cosmology is really only important when we turn to very high temperatures, where the gas becomes relativistic.

2.4.2 Ultra-Relativistic Gases

As we will see in the next section, as we go further back in time, the universe gets hot. Really hot. For any particle, there will be a time such that

$$k_B T \gg 2mc^2$$

In this regime, particle-anti-particle pairs can be created in the fireball. When this happens, both the mass and the chemical potential are negligible. We say that the particles are *ultra-relativistic*, with their energy given approximately as

$$E_{\mathbf{p}} \approx pc$$

just as for a massless particle. We can use our techniques to study the behaviour of gases in this regime.

We start with ultra-relativistic bosons. We work with vanishing chemical potential, $\mu = 0$. (This will ensure that we have equal numbers of particles an anti-particles. The presence of a chemical potential results in a preference for one over the other, and will be explored in Examples Sheet 3.) The integral (2.36) for the number density gives

$$n_{\text{boson}} = \frac{4\pi g}{(2\pi\hbar)^3}\int dp\ \frac{p^2}{e^{\beta pc}-1} = \frac{gI_2}{2\pi^2\hbar^3c^3}(k_BT)^3$$

while the energy density is

$$\rho_{\text{boson}} = \frac{4\pi g}{(2\pi\hbar)^3}\int dp\ \frac{p^3c}{e^{\beta pc}-1} = \frac{gI_3}{2\pi^2\hbar^3c^3}(k_BT)^4$$

where we've used the definition (2.13) of the integral

$$I_n = \int_0^\infty dy\ \frac{y^n}{e^y-1} = \Gamma(n+1)\zeta(n+1)$$

In both cases, the integrals coincide with those that we met for blackbody radiation

Meanwhile, for fermions we have

$$n_{\text{fermion}} = \frac{4\pi g}{(2\pi\hbar)^3}\int dp\ \frac{p^2}{e^{\beta pc}+1} = \frac{gJ_2}{2\pi^2\hbar^3c^3}(k_BT)^3$$

and

$$\rho_{\rm fermion} = \frac{4\pi g}{(2\pi\hbar)^3}\int dp\ \frac{p^3 c}{e^{\beta pc}+1} = \frac{gJ_3}{2\pi^2\hbar^3c^3}(k_BT)^4$$

where, this time, we get the integral

$$J_n = \int_0^\infty dy\ \frac{y^n}{e^y+1} = \int_0^\infty dy\ \left[\frac{y^n}{e^y-1} - \frac{2y^n}{e^{2y}-1}\right] = \left(1-\frac{1}{2^n}\right) I_n$$

The upshot of these calculations is that the number density is

$$n = \frac{g\zeta(3)}{\pi^2\hbar^3c^3}(k_BT)^3 \times \begin{cases} 1 \text{ for bosons} \\ \frac{3}{4} \text{ for fermions} \end{cases}$$

and the energy density is

$$\rho = \frac{g\pi^2}{30\,\hbar^3c^3}(k_BT)^4 \times \begin{cases} 1 \text{ for bosons} \\ \frac{7}{8} \text{ for fermions} \end{cases}$$

The differences are just small numerical factors but, as we will see, these become important in cosmology.

Ultimately, we will be interested in gases that contain many different species of particles. In this case, it is conventional to define the effective number of relativistic species in thermal equilibrium as

$$g_\star(T) = \sum_{\rm bosons} g_i + \frac{7}{8}\sum_{\rm fermions} g_i \tag{2.39}$$

As the temperature drops below a particle's mass threshold, $k_BT < m_ic^2$, this particle is removed from the sum. In this way, the number of relativistic species is both time and temperature dependent. The energy density from all relativistic species is then written as

$$\rho = g_\star\,\frac{\pi^2}{30\,\hbar^3c^3}(k_BT)^4 \tag{2.40}$$

To calculate $g_\star$ in different epochs, we need to know the matter content of the Standard Model and, eventually, the identity of dark matter. We'll make a start on this in the next section.

2.5 The Hot Big Bang

We have seen that for the first 300,000 years or so, the universe was filled with a fireball in which photons were in thermal equilibrium with matter. We would like to understand what happens to this fireball as we dial the clock back further. This collection of ideas goes by the name of the hot Big Bang theory.

2.5.1 Temperature vs Time

It turns out, unsurprisingly, that the fireball is hotter at earlier times. This is simplest to describe if we go back to when the universe is radiation dominated, at $z > 3400$ or $t < 50,000$ years. Here, the energy density scales as (1.41),

$$\rho \sim \frac{1}{a^4}$$

We can compare this to the thermal energy density of photons, given by (2.14)

$$\rho = \frac{\pi^2}{15\hbar^3 c^3}(k_B T)^4$$

To see that the temperature scales inversely with the scale factor

$$T \sim \frac{1}{a} \tag{2.41}$$

This is the same temperature scaling that we saw for the CMB after recombination (2.19). Indeed, the underlying arguments are also the same: the energy of each photon is blue-shifted as we go back in time, while their number density increases, resulting in the $\rho \sim 1/a^4$ behaviour. The difference is that now the photons are in equilibrium. If they are disturbed in some way, they will return to their equilibrium state. In contrast, if the photons are disturbed after recombination they will retain a memory of this.

What happens during the time $1100 < z < 3400$, before recombination but when matter was the dominant energy component? First consider a universe with only non-relativistic matter, with number density n. The energy density is

$$\rho_m = nmc^2 + \frac{1}{2}nmv^2$$

The first term drives the expansion of the universe and is independent of temperature. The second term, which we completely ignored in Section 1 on the grounds that it is negligible, depends on temperature. This was computed in (2.8) and is given by $\frac{1}{2}nmv^2 = \frac{3}{2}nk_B T$.

As the universe expands, the velocity of non-relativistic particles is red-shifted as $v \sim 1/a$. (This is hopefully intuitive, but we have not actually demonstrated this previously. We will derive this redshift in Section 3.1.3.) This means that, in a universe with *only* non-relativistic matter, we would have

$$T \sim \frac{1}{a^2}$$

So what happens when we have both matter and radiation? We would expect that the temperature scaling sits somewhere between $T \sim 1/a$ and $T \sim 1/a^2$. In fact, it is entirely dominated by the radiation contribution. This can be traced to the fact that there are many more photons that baryons; $\eta = n_B/n_\gamma \approx 10^{-9}$. A comparable ratio is expected to hold for dark matter. This means that the photons, rather than matter, dictate the heat capacity of the thermal bath. The upshot is that the temperature scales as $T \sim 1/a$ throughout the period of the fireball. Moreover, as we saw in Section 2.2, the temperature of the photons continues to scale as $T \sim 1/a$ even after they decouple.

Doing a Better Job

The formula $T \sim 1/a$ gives us an approximate scaling. But we can do better.

We start with the continuity equation (1.39) for relativistic matter, with $P = \rho/3$, is

$$\dot{\rho} = -3H(\rho + P) = -4H\rho \tag{2.42}$$

But for ultra-relativistic gases, we know that the energy density is given by (2.43), have

$$\rho = g_\star \frac{\pi^2}{30\,\hbar^3 c^3}(k_B T)^4 \tag{2.43}$$

where $g_\star$ is the effective number of relativistic degrees of freedom (2.39). Differentiating this with respect to time, and assuming that $g_\star$ is constant, we have

$$\dot{\rho} = \frac{4\dot{T}}{T}\rho \quad \Rightarrow \quad \dot{T} = -HT$$

where the second expression comes from (2.42). This is just re-deriving the fact that $T \sim 1/a$. However, now we have use the Friedmann equation to determine the Hubble parameter in the radiation dominated universe,

$$H^2 = \frac{8\pi G}{3c^2}\rho = A(k_B T)^4 \quad \text{with } A = \frac{8\pi^3 G}{90\,\hbar^3 c^5}\, g_\star$$

This leaves us with a straightforward differential equation for the temperature,

$$k_B\dot{T} = -\sqrt{A}(k_BT)^3 \quad \Rightarrow \quad t = \frac{1}{2\sqrt{A}}\frac{1}{(k_BT)^2} + \text{constant} \tag{2.44}$$

We choose to set the integration constant to zero. This means that the temperature diverges as we approach the Big Bang singularity at $t = 0$. All times will be measured from this singularity.

To turn this into something physical, we need to make sense of the morass of fundamental constants in A. The presence of Newton's constant is associated with a very high energy scale known as the *Planck mass* with the corresponding *Planck energy*,

$$M_{\rm pl}c^2 = \sqrt{\frac{\hbar c^5}{8\pi G}} \approx 2.4 \times 10^{21}\ \text{MeV}$$

Meanwhile, the value of Planck's constant is

$$\hbar \approx 6.6 \times 10^{-16}\ \text{eV s} = 6.6 \times 10^{-22}\ \text{MeV s}$$

These combine to give

$$\hbar M_{\rm pl}c^2 \approx 1.6\ \text{MeV}^2\, s$$

Putting these numbers into (2.44) gives is an expression that tells us the temperature T at a given time t,

$$\left(\frac{t}{1\,\text{second}}\right) \approx \frac{2.4}{g_\star^{1/2}}\left(\frac{1\,\text{MeV}}{k_BT}\right)^2 \tag{2.45}$$

Ignoring the constants of order 1, we say that the universe was at a temperature of $k_BT = 1\,\text{MeV}$ approximately 1 second after the Big Bang.

As an aside: most textbooks derive the relationship (2.45) by assuming conservation of entropy (which, it turns out, ensures that $g_\star T^3a^3$ is constant). The derivation given above is entirely equivalent to this.

To finish, we need to get a handle on the effective number of relativistic degrees of freedom $g_\star$. In the very early universe many particles were relativistic and $g_\star$ is bigger. As the universe cools, it goes through a number of stages where $g_\star$ drops discontinuously as the heavier particle become non-relativistic.

For example, when temperatures are around $k_BT \sim 10^6$ eV $\equiv 1$ MeV, the relativistic species are the photon (with $g_\gamma = 2$), three neutrinos and their anti-neutrinos (each with $g_\nu = 1$) and the electron and positron (each with $g_e = 2$). The effective number of relativistic species is then

$$g_\star = 2 + \frac{7}{8}(3 \times 1 + 3 \times 1 + 2 + 2) = 10.75 \tag{2.46}$$

As we go back in time, more and more species contribute. By the time we get to $k_BT \sim 100$ GeV, all the particles of the Standard Model are relativistic and contribute $g_\star = 106.75$.

In contrast, as we move forward in time, $g_\star$ decreases. Considering only the masses of Standard Model particles, one might naively think that, as electrons and positrons annihilate and become non-relativistic, we're left only with photons, neutrinos and anti-neutrinos. This would give

$$g_\star = 2 + \frac{7}{8}(3 + 3) = 7.25$$

Unfortunately, at this point one of many subtleties arises. It turns out that the neutrinos are very weakly interacting and have already decoupled from thermal equilibrium by the time electrons and protons annihilate. When the annihilation finally happens, the bath of photons is heated while the neutrinos are unaffected. We can still use the formula (2.43), but we need an amended definition of $g_\star$ to include the fact that neutrinos and electrons are both relativistic, but sitting at different temperatures. For now, I will simply give the answer:

$$g_\star \approx 3.4 \tag{2.47}$$

I will very briefly explain where this comes from in Section 2.5.4.

A Longish Aside on Neutrinos

Why do neutrinos only contribute 1 degree of freedom to (2.46) while the electron has 2? After all, they are both spin-$\frac{1}{2}$ particles. To explain this, we need to get a little dirty with some particle physics.

First, for many decades we thought that neutrinos are massless. In this case, the right characterisation is not spin, but something called *helicity*. Massless particles necessarily travel at the speed of light; their spin is aligned with their direction of travel. If the spin points in the same direction as the momentum, then it is said to be right-handed; if it points in the opposite direction then it is said to be left-handed. It

is a fact that we've only ever observed neutrinos with left-handed helicity and it was long believed that the right-handed neutrinos simply do not exist. Similarly, we've only observed anti-neutrinos with right-handed helicity; there appear to be no left-handed anti-neutrinos. If this were true, we would indeed get the $g = 1$ count that we saw above.

However, we now know that neutrinos do, in fact, have a very small mass. Here is where things get a little complicated. Roughly speaking, there are two different kinds of masses that neutrinos could have: they are called the *Majorana mass* and the *Dirac mass*. Unfortunatey, we don't yet know which of these masses (or combination of masses) the neutrino actually has, although we very much hope to find out in the near future.

The Majorana mass is the simplest to understand. In this scenario, the neutrino is its own anti-particle. If the neutrino has a Majorana mass then what we think of as the right-handed anti-neutrino is really the same thing as the right-handed neutrino. In this case, the counting goes through in the same way, but we drape different words around the numbers: instead of getting $1 + 1$ from each neutrino + anti-neutrino, we instead get 2 spin states for each neutrino, and no separate contribution from the anti-neutrino.

Alternatively, the neutrino may have a Dirac mass. In this case, it looks much more similar to the electron, and the correct counting is 2 spin states for each neutrino, and another 2 for each anti-neutrino. Here is where things get interesting because, as we will explain in Section 2.5.3, we know from Big Bang nucleosynthesis that the count (2.46) of $g_\star = 10.75$ was correct a few minutes after the Big Bang. For this reason, it must be the case that 2 of the 4 degrees of freedom interact very weakly with the thermal bath, and drop out of equilibrium in the very early universe. Their energy must then be diluted relative to everything else, so that it's negligible by the time we get to nucleosynthesis. (For example, there are various phase transitions in the early universe that could dump significant amounts of energy into half of the neutrino degrees of freedom, leaving the other half unaffected.)

2.5.2 The Thermal History of our Universe

The essence of the hot Big Bang theory is simply to take the temperature scaling $T \sim 1/a$ and push it as far back as we can, telling the story of what happens along the way.

As we go further back in time, more matter joins the fray. For some species of particles, this is because the interaction rate is sufficiently large at early times that it couples to the thermal bath. For example, there was a time when both neutrinos and (we think) dark matter were in equilibrium with the thermal bath, before both underwent freeze out.

For other species of particle, the temperatures are so great (roughly $k_B T \approx 2mc^2$) that particle-anti-particle pairs can emerge from the vacuum. For example, for the first six seconds after the Big Bang, both electrons and positrons filled the fireball in almost equal numbers.

The goal of the Big Bang theory is to combine knowledge of particle physics with our understanding of thermal physics to paint an accurate picture for what happened at various stages of the fireball. A summary of some of the key events in the early history of the universe is given in the following table. In the remainder of this section, we will tell some of these stories.

What	When (t)	When (z)	When (T)
Inflation	10^{-36} s ?	10^{28} ?	?
Baryogenesis	?	?	?
Electroweak phase transition	10^{-12} s	10^{15}	10^{22} K
QCD phase transition	10^{-6} s	10^{12}	10^{16} K
Dark Matter Freeze-Out	?	?	?
Neutrino Decoupling	1 second	6×10^9	10^{10} K
$e^- e^+$ Annihilation	6 second	2×10^9	5×10^9 K
Nucleosynthesis	3 minutes	4×10^8	10^9 K
Matter-Radiation Equality	50,000 years	3400	8700 K
Recombination	$\sim 300,000$ years	1300	3600 K
Last Scattering	350,000 years	1100	3100 K
Matter-Λ Equality	10^{10} years	0.4	3.8 K
Today	1.4×10^{10} years	0	2.7 K

2.5.3 Nucleosynthesis

One of the best understood processes in the Big Bang fireball is the formation of deuterium, helium and heavier nuclei from the thermal bath of protons and neutrons. This is known as *Big Bang nucleosynthesis*. It is a wonderfully delicate calculation, that involves input from many different parts of physics. The agreement with observation could fail in a myriad of ways, yet the end result agrees perfectly with the observed abundance of light elements. This is one of the great triumphs of the Big Bang theory.

Full calculations of nucleosynthesis are challenging. Here we simply offer a crude sketch of the formation of deuterium and helium nuclei.

Neutrons and Protons

Our story starts at early times, $t \ll 1$ second, when the temperature reached $k_B T \gg 1$ MeV. The mass of the electron is

$$m_e c^2 \approx 0.5 \text{ MeV}$$

so at this time the thermal bath contains many relativistic electron-positron pairs. These are in equilibrium with photons and neutrinos, both of which are relativistic, together with non-relativistic protons and neutrons. Equilibrium is maintained through interactions mediated by the weak nuclear force

$$n + \nu_e \;\leftrightarrow\; p + e^- \quad , \quad n + e^+ \;\leftrightarrow\; p + \bar{\nu}_e$$

These reactions arise from the same kind of process as beta decay, $n \to p + e^- + \bar{\nu}_e$.

The chemical potentials for electrons and neutrinos are vanishingly small. Chemical equilibrium then requires $\mu_n = \mu_p$, and the ratio of neutron to proton densities can be calculated using the equation (2.24) for a non-relativistic gas,

$$\frac{n_n}{n_p} = \left(\frac{m_n}{m_p}\right)^{3/2} e^{-\beta(m_n - m_p)c^2}$$

The proton and neutron have a very small mass difference,

$$m_n c^2 \approx 939.6 \text{ Mev}$$
$$m_p c^2 \approx 938.3 \text{ MeV}$$

This mass difference can be neglected in the prefactor, but is crucial in the exponent. This gives the ratio of protons to neutrons while equilibrium is maintained

$$\frac{n_n}{n_p} \approx e^{-\beta \Delta mc^2} \quad \text{with} \quad \Delta mc^2 \approx 1.3 \text{ MeV}$$

For $k_B T \gg \Delta mc^2$, there are more or less equal numbers of protons and neutrons. But as the temperature falls, so too does the number of neutrons.

However, the exponential decay in neutron number does not continue indefinitely. At some point, the weak interaction rate will drop to $\Gamma \sim H$, at which point the neutrons freeze out, and their number then remains constant. (Actually, this last point isn't quite true as we will see below but let's run with it for now!)

The interaction rate can be written as $\Gamma = n\sigma v.$ where σ is the cross-section. At this point, I need to pull some facts about the weak force out of the hat. The cross-section varies as temperature as $\sigma v \sim G_F T^2$ with $G_F \approx 1.2 \times 10^{-5}$ GeV^{-2} a constant that characterises the strength of the weak force. Meanwhile, the number density scales as $n \sim T^3$. This means that $\Gamma \sim T^5$.

The Hubble parameter scales as $H \sim 1/a^2 \sim T^2$ in the radiation dominated epoch. So we do indeed expect to find $\Gamma \gg H$ at early times and $\Gamma \ll H$ at later times. It turns out that neutrons decouple at the temperature

$$k_B T_{\rm dec} \approx 0.8 \text{ MeV}$$

Putting this into (2.45), and using $g_\star \approx 3.4$, we find that neutrons decouple around

$$t_{\rm dec} \approx 2 \text{ seconds}$$

after the Big Bang.

At freeze out, we are then left with a neutron-to-proton ratio of

$$\frac{n_n}{n_p} \approx \exp\left(-\frac{1.3}{0.8}\right) \approx \frac{1}{5}$$

In fact, this isn't the end of the story. Left alone, neutrons are unstable to beta decay with a half life of a little over 10 minutes. This means that, after freeze out, the number density of neutrons decays as

$$n_n(t) \approx \frac{1}{5} n_p(t_{\rm dec}) e^{-t/\tau_n} \tag{2.48}$$

where $\tau_n \approx 880$ second. If we want to do something with those neutrons (like use them to form heavier nuclei) then we need to hurry up: the clock is ticking.

Deuterium

Ultimately, we want to make elements heavier than hydrogen. But these heavier nuclei contain more than two nucleons. For example, the lightest is ^{3}He which contains two protons and a neutron. But the chance of three particles colliding at the same time to form such a nuclei is way too small. Instead, we must take baby steps, building up by colliding two particles at a time.

The first such step is, it turns out, the most difficult. This is the step to *deuterium*, or heavy hydrogen, consisting of a bound state of a proton and neutron that forms through the reaction

$$p + n \;\leftrightarrow\; D + \gamma$$

The binding energy is

$$E_{\rm bind} = m_n + m_p - m_D \approx 2.2 \text{ MeV}$$

Both the proton and neutron have spin 1/2, and so have $g_n = g_p = 2$. In deuterium, the spins are aligned to form a spin 1 particle, with $g_D = 3$. The fraction of deuterium is then determined by the Saha equation (2.29), using the same arguments that we saw in recombination

$$\frac{n_D}{n_n n_p} = \frac{3}{4}\left(\frac{m_D}{m_n m_p}\frac{2\pi\hbar^2}{k_B T}\right)^{3/2} e^{\beta E_{\rm bind}}$$

Approximating $m_n \approx m_p \approx \frac{1}{2}m_D$ in the pre-factor, the ratio of deuterium to protons can be written as

$$\frac{n_D}{n_p} \approx \frac{3}{4} n_n \left(\frac{4\pi\hbar^2}{m_p\, k_B T}\right)^{3/2} e^{\beta E_{\rm bind}}$$

We calculated the time-dependent neutron density n_n in (2.48). We will need this time-dependent expression soon, but for now it's sufficient to get a ballpark figure and, in this vein, we will simply approximate the number of neutrons as

$$n_n \approx n_p \approx \eta\, n_\gamma$$

The baryon-to-photon ratio has not had the opportunity to significantly change between nucleosynthesis and the present day, so we have $\eta \approx 10^{-9}$. (The last time it changed was when electrons and positrons annihilated, with $e^- + e^+ \to \gamma + \gamma$.) Using the expression $n_\gamma \approx (k_B T/c)^3$ from (2.15) for the number of photons, we then have

$$\frac{n_D}{n_p} \approx \eta \left(\frac{k_B T}{m_p c^2}\right)^{3/2} e^{\beta E_{\rm bind}} \tag{2.49}$$

We see that we only get an appreciable number of deuterium atoms when the temperature drops to a suitably small value. This delay in deuterium formation is mostly due to the large number of photons as seen in the factor η. These same photons are responsible for the delay in hydrogen formation 300,000 years later: in both cases, any putative bound state is quickly broken apart as it is bombarded by high-energy photons at the tail end of the blackbody distribution.

Solving (2.49), we find that $n_D/n_p \sim 1$ only when $\beta E_{\rm bind} \approx 35$, or

$$k_B T \lesssim 0.06 \text{ MeV}$$

Importantly, this is after the neutrinos have decoupled. Using (2.45), again with $g_\star \approx 3.4$, we find that deuterium begins to form at

$$t \approx 360 \text{ seconds}$$

This is around six minutes after the Big Bang. Fortunately (for all of us), six minutes is not yet the 10.5 minutes that it takes neutrons to decay. But it's getting tight. Had the details been different so that, say, it took 12 minutes rather than 6 for deuterium to form, then we would not be around today to tell the tale. Building a universe is, it turns out, a delicate business.

Helium and Heavier Nuclei

Heavier nuclei have significantly larger binding energies. For example, the binding energy for ^{3}He is 7.7 MeV, while for ^{4}He it is 28 MeV. In perfect thermal equilibrium, these would be present in much larger abundancies. However, the densities are too low, and time too short, for these nuclei to form in reactions involving three of more nucleons coming together. Instead, they can only form in any significant levels after deuterium has formed. And, as we saw above, this takes some time. This is known as the *deuterium bottleneck*.

Once deuterium is present, however, there is no obstacle to forming helium. This happens almost instantaneously through

$$D + p \leftrightarrow {}^3\text{He} + \gamma \quad , \quad {}^3He + D \leftrightarrow {}^4\text{He} + p$$

Because the binding energy is so much higher, all remaining neutrons rapidly bind into ^{4}He nuclei. At this point, we use the time-dependent form for the neutron density (2.48) which tells us that the number of remaining neutrons at this time is

$$\frac{n_n}{n_p} = \frac{1}{5} e^{-360/880} \approx 0.13$$

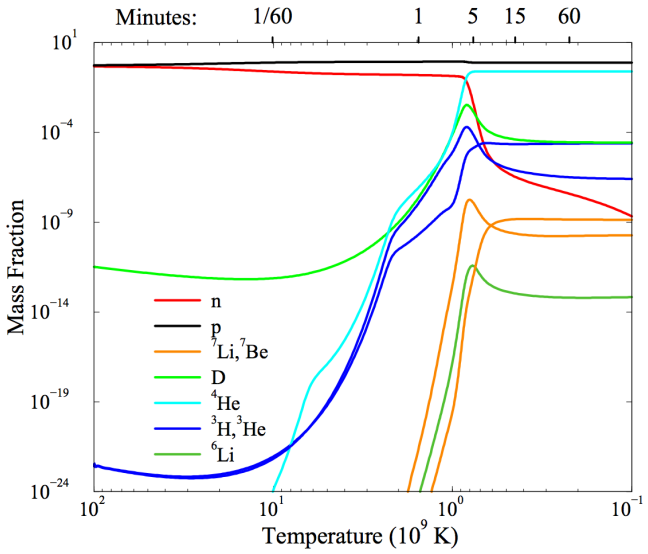

Figure 31: The abundance of light nuclei in the early universe.

Since each ^{4}He atom contains two neutrons, the ratio of helium to hydrogen is given by

$$\frac{n_{\rm He}}{n_{\rm H}} = \frac{n_n/2}{n_p - n_n} \approx 0.07$$

A helium atom is four times heavier than a hydrogen atom, which means that roughly 25% of the baryonic mass sits in helium, with the rest in hydrogen. This is close to the observed abundance.

Only trace amounts of heavier elements are created during Big Bang nucleosynthesis. For each proton, approximately 10^{-5} deuterium nuclei and 10^{-5} ^{3}He nuclei survive. Astrophysical calculations show that this a million times greater than the amount that can be created in stars. There are even smaller amounts of ^{7}Li and ^{7}Be, all in good agreement with observation.

The time dependence of the abundance of various elements is shown[8] in Figure 31. You can see the red neutron curve start to drop off as the neutrons decay, and the abundance of the other elements rising as finally the deuterium bottleneck is overcome.

[8]This figure is taken from Burles, Nollett and Turner, *Big-Bang Nucleosynthesis: Linking Inner Space and Outer Space*", astro-ph/99033.

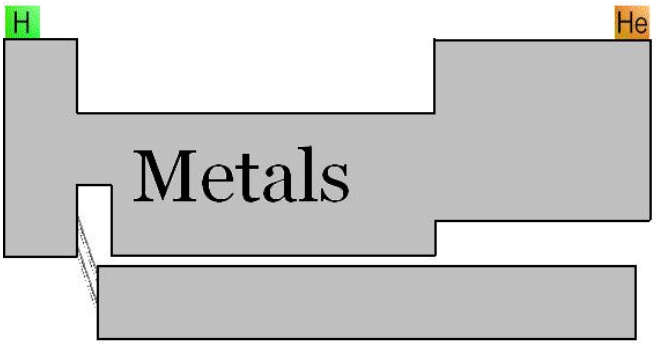

Figure 32: The elements, according to cosmologists.

Any heavier elements arise only much later in the evolution of the universe when they are forged in stars. Because of this, cosmologists have developed their own version of the periodic table, shown in the Figure 32. It is, in many ways, a significant improvement over the one adopted by atomic and condensed matter physicists.

Dependence on Cosmological Parameters

The agreement between the calculated and observed abundancies provides strong support for the seemingly outlandish idea that we know what we're talking about when the universe was only a few minutes old. The results depend in detail on a number of specific facts from both particle physics and nuclear physics.

One input into the calculation is particularly striking. The time at which deuterium finally forms is determined by the equation (2.45) which, in turn, depends on the number of relativistic species $g_\star$. If there are more relativistic species in thermal equilibrium with the heat bath then the deuterium bottleneck is overcome sooner, resulting in a larger fraction of helium. Yet, the contribution from the light Standard Model degrees of freedom (i.e. the photon and neutrinos) gives excellent agreement with observation.

This puts strong constraints on the role of dark matter in the early universe. Given its current prominence, we might naively have thought that the relativistic energy density in the early universe would receive a significant contribution from dark matter. The success of Big Bang nucleosynthesis tells us that this is not the case. Either there are no light particles in the dark sector (so the dark sector is dark even if you live there) or hot dark particles fell out of equilibrium long before nucleosynthesis took place and so sit at a much lower temperature when the all action is happening.

2.5.4 Further Topics

There are many more stories to tell about the early universe. These lie beyond the scope of this course, but here is a taster. Going back in time, we have...

Electron-Positron Annihilation

Prior to nucleosynthesis, the fireball included both electrons and positrons. At around $k_B T \sim 1$ MeV these annihilate, injecting energy into the thermal bath of photons and slightly raising their temperature.

We can give an estimate for this. Prior to annihilation, the photons and electron-positron pairs were in equilibrium, giving

$$g_\star = 2 + \frac{7}{8}(2+2) = \frac{11}{2}$$

Afterwards, there are only photons with

$$g_\star = 2$$

So far, we haven't looked closely at what happens when $g_\star$ changes. This is because we need an important concept that we haven't yet introduced: entropy. This is discussed in detail in the lectures on Statistical Physics where we show that the entropy of an ultra-relativistic gas is proportional to $g_\star T^3 a^3$.

The annihilation of electron-positron pairs is an *adiabatic process*, which means that entropy is conserved. Since $g_\star$ decreases by a factor of 11/4, this means that $T^3 a^3$ increases by 11/4. Or

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

There is one last twist to the story. The electrons and positrons do not all annihilate. There must be a very slight excess of electrons that are left over at the end. This, of course, is the stuff we're made of.

Before annihilation, the number of electron-positron pairs was the same order of magnitude as the number of photons, and these have persisted to the present day. Meanwhile, electric neutrality of the universe ensures that the number of left over electrons is comparable to the number of baryons currently in the universe. This means that the slight excess of electrons in the early universe must be roughly equal to the famous ratio $\eta \sim 10^{-9}$ of baryons to photons in the present day. In other words, in the early universe there was one extra electron for every billion electron-positron pairs. Understanding the origin of this imbalance is the goal of baryogenesis and is briefly described below.

Neutrino Decoupling

Neutrinos are very weakly interacting. They decouple from the thermal bath at temperatures of $T \sim 1$ MeV. Neutrinos have masses $m_\nu c^2$ between an meV and an eV (the exact masses are not well known) so they are very relativistic when they decouple. Like the photons after recombination, neutrinos preserve their relativistic distribution even after they decouple.

However, in contrast to the photons, neutrinos do not get the energy boost from electron-positron annihilation. This means that their temperature after this event lags behind the photon temperature, with

$$T_\nu = \left(\frac{4}{11}\right)^{1/3} T_\gamma$$

This relation persists to the present day. It is expected that there is a cosmic neutrino background filling the universe, with a temperature $T \approx (4/11)^{1/3}\, T_{\rm CMB} \approx 1.9$ K. This has not yet been observed although there is an experiment, currently in the design phase, which aims to detect it.

When we have relativistic species that sit at different temperatures, we need to revisit our formula for the effective number of degrees of freedom $g_\star$. We can continue to write the total energy density as

$$\rho = g_\star \frac{\pi^2}{30\,\hbar^3 c^3}(k_B T)^4$$

if we now define $g_\star$ to be the sum over all relativistic species, whether or not they are in equilibrium,

$$g_\star(T) = \sum_{\rm bosons} g_i \left(\frac{T_i}{T}\right)^4 + \frac{7}{8}\sum_{\rm fermions} g_i \left(\frac{T_i}{T}\right)^4$$

where T_i is the temperature of each species. In particular, after $e^- e^+$ annihilation, when nucleosynthesis occurs, the relativistic species are photons and electrons with

$$g_\star = 2 + \frac{7}{8}\left[2 \times 3 \times \left(\frac{4}{11}\right)^4\right] \approx 3.4$$

This is the value quoted in (2.47) and used when discussing nucleosynthesis.

QCD Phase Transition

At a temperature of $k_B T \approx 150$ MeV, protons and neutrons melt. They disassociate into a soup of quarks and gluons, known as the quark-gluon plasma. This state of matter has been created at particle accelerators here on Earth, freeing the quarks from their nucleon prison for the first time in 13.8 billion years.

Electroweak Phase Transition

In the Standard Model, fundamental particles such as the electron and quarks get their mass from the Higgs mechanism. Above $k_B T \approx 100$ GeV, this mechanism ceases to work. At this point all particles in the Standard Model are massless and in thermal equilibrium.

Dark Matter Freeze Out

Clearly there are many things we don't know about dark matter. We don't, for example, know if it has any interactions with the stuff we're made of. We can, however, make some assumptions and see where it leads us.

One of the most popular candidates for dark matter is a stable, massive particle that interacts only weakly with itself and with the Standard Model. These are known as weakly interacting massive particles, or WIMPs. Nearly all theories that go beyond the Standard Model predict such objects.

If the particle interacts even weakly with the Standard Model then there will be a time when dark matter is in equilibrium with the thermal bath. As the temperature lowers, the dark matter will freeze out. With very little input — just the mass and cross-section of the dark matter — we can then compute the expected abundance of dark matter seen today.

Here something nice happens. If we take the mass to be around $M_X \sim 100$ GeV, which is the energies probed by the LHC, and the cross-section to be $\sigma v \sim G_F$, which is the strength of the weak nuclear force, then we do indeed find the observed abundance of dark matter. With an overblown rhetorical flourish, this coincidence is known as the *WIMP miracle*. It was one of the reasons for optimism that it might be possible to create dark matter at the LHC. Needless to say, this was not borne out. Furthermore, a slew of impressive experiments, designed to directly detect passing dark matter, have so-far, offered only null results. While WIMPs remain a possible candidate for dark matter, there is no compelling observation beyond the coincidence above to suggest they are intimately tied to the weak force at the 100 GeV scale.

Baryogenesis

The universe contains lots of matter but very little anti-matter. How did this asymmetry come to be?

One possibility is that it is an initial condition on the universe. Another is that the universe started with equal amounts of matter and anti-matter, but somehow a small dynamical shift took place that preferred one over the other. This latter process is known as *baryogenesis*.

We don't have an established theory of baryogenesis; whatever caused it must lie beyond the Standard Model. Nonetheless, there are criteria, known as the *Sakharov conditions* that must be obeyed for baryogenesis to occur:

- The first criterion is the most obvious: baryon number cannot be a conserved quantity. Here "baryon number" refers to baryons minus anti-baryons. In a symmetric universe, this starts off as zero. We want it to end up non-zero.

  In the Standard Model, baryon number is conserved. (In fact, strictly speaking $B - L$ is conserved where B is baryon number and L is lepton number, but this is a story for another day.) But it is straightforward to cook up interactions at higher energy scales which violate baryon number.

- There is a symmetry known as CP which, roughly speaking, says that particles and anti-particles behave the same. This too must be violated for baryogenesis to occur, since particles should be favoured over anti-particles.

  In fact, CP is violated in the Standard Model. It's not clear if this is sufficient, or if further CP violation is needed in the interaction beyond the Standard Model.

- The final criterion is the least obvious: the early universe must deviate from thermal equilibrium. This is needed so that the interactions in one direction differ from the interactions running in reverse.

  A deviation from thermal equilibrium occurs when the universe undergoes a first order phase transition. (You can read more about phase transitions in the lectures on Statistical Physics and Statistical Field Theory.) The electroweak phase transition appears to be a fairly smooth crossover, which is not violent enough to do the job. For baryogenesis to occur, we most likely need a different phase transition early in the universe.

There are many models of baryogenesis, but currently no smoking gun experiment or observation to determine which, if any, is correct.

3. Structure Formation

Until now, we have discussed a universe which is perfectly homogeneous and isotropic. But that is not the universe we live in. Instead, our universe contains interesting objects which clump together, bound by the gravitational force, from planets and stars, to galaxies and clusters of galaxies. We would like to understand how these objects form.

The stakes become somewhat higher when we realise that the early universe was very much smoother than the one we live in today. Of course, there were no galaxies and planets, or even atoms, in the early fireball. But nor were there significant variations in the energy density. This can be clearly seen in the CMB, which has an almost uniform temperature T but exhibits tiny fluctuations on the scale

$$\frac{\delta T}{T} \approx 10^{-5}$$

We can compare this to the world we see around us today. As we learned in Section 1.4, the average energy density in the universe $\rho_{\rm crit,0}$, corresponds to about 1 hydrogen atom per cubic metre. But this hides the fact that most of this matter is contained in gravitationally bound objects. A measure analogous to $\delta T/T$ can be found by comparing the typical energy density contained in a galaxy to the average $\rho_{\rm crit,0}$: this turns out to be

$$\frac{\rho_{\rm galaxy}}{\rho_{\rm crit}} \approx 10^6$$

We see that the universe, like many of us, has become significantly more lumpy as it aged. The primary purpose of this section is to understand how this occurred: how did the small fluctuations seen in the CMB grow, ultimately resulting in the wondrous array clusters and galaxies that we see in the night sky. This process is known as *structure formation*.

There is also a second, more ambitious, purpose to this section, which is to trace the perturbations backwards in time. Ultimately, we would like to understand where the small fluctuations $\delta T/T$ seen in the CMB came from. Since these fluctuations grow to give rise to all the structure in the universe, this is really a rephrasing of one of the biggest questions of them all: where did we come from? We will see that when we evolve the fluctuations backwards in time, they take a very simple form, providing crucial information about what the universe looked like in its very earliest moments. Ultimately, in Section 3.5, we will offer an answer to this big question. We will argue that all the structure in the universe, including us, owes its existence to fluctuations of quantum fields, fluctuations that took place in the first few fractions of a second after the Big Bang.

3.1 Density Perturbations

In this section, we will assume that there are some small perturbations in the energy density of the universe. We will not (yet) ask where these perturbations came from. Instead, we will be interested in their fate. In particular, under what circumstances do they grow, and when do they fade away?

We start by considering non-relativistic matter. As we have seen, in our universe this is primarily dark matter. We know very little about the interactions of dark matter, but there is a wonderful universality in physics which tells us that, on suitably large distances, any substance can be described by the equations of fluid mechanics. This, then, will be our tool of choice: fluid mechanics applied on cosmological scales.

Our goal is to start with a homogeneous and isotropic fluid, and then see what happens when it is perturbed. First we need to specify our variables which, in contrast to earlier sections, now depend on both space and time. The standard variables of fluid mechanics are

- number density $n(\mathbf{x}, t)$. More precisely, we will be interested in the mass density. For now, we will consider a fluid made of a single type of particle of mass m, so the mass density is simply $mn(\mathbf{x}, t)$. For a non-relativistic fluid, the mass dominates the energy density which is given by $\rho(\mathbf{x}, t) = mn(\mathbf{x}, t)c^2$.
- Pressure $P(\mathbf{x}, t)$. As discussed in Section 1.2.1, non-relativistic fluids have $P \ll \rho$.
- Velocity $\mathbf{u}(\mathbf{x}, t)$.

Next, we need the relevant equations of fluid mechanics. These depend on the context. Ultimately, we want to understand fluids which gravitate in an expanding universe. However, we're going to build up slowly and introduce one ingredient at a time.

3.1.1 Sound Waves

First, we're going to consider fluids that don't experience gravity and live in a static spacetime. These fluids are described by three equations. The first is the *continuity equation* which captures the conservation of particles

$$\frac{\partial n}{\partial t} + \nabla \cdot (n\mathbf{u}) = 0 \tag{3.1}$$

This tells us that the particle density in some region can only change if it flows away due its velocity $\mathbf{u}$.

The second is the *Euler equation*, which can be viewed as Newton's "F=ma" for a continuous system,

$$mn\left(\frac{\partial}{\partial t}+\mathbf{u}\cdot\nabla\right)\mathbf{u}=-\nabla P \tag{3.2}$$

The left-hand side interpreted as mass $\times$ acceleration, while the pressure $-\nabla P$ on the right-hand side provides the force.

The last of our equations is the equation of state which, for now, we leave general as

$$P=P(n,T) \tag{3.3}$$

For much of this section, we will use ideal gas equation, $P=nk_BT$, which is the appropriate equation of state for a non-relativistic fluid. In time, we will also apply these ideas to other fluids.

The simplest solution to these equations describes a static fluid with $\mathbf{u}=0$ and constant density and pressure

$$n=\bar{n}\quad\text{and}\quad P=\bar{P}$$

This is a homogeneous and isotropic fluid. We take this to be our background and look at small perturbations. We take $\mathbf{u}$ to be small, and write

$$n(\mathbf{x},t)=\bar{n}+\delta n(\mathbf{x},t)\quad\text{and}\quad P(\mathbf{x},t)=\bar{P}+\delta P(\mathbf{x},t)$$

The equations (3.1) and (3.2) are linearised to give

$$\frac{\partial(\delta n)}{\partial t}=-\nabla\cdot(\bar{n}\mathbf{u})\quad\text{and}\quad m\bar{n}\frac{\partial\mathbf{u}}{\partial t}=-\nabla\delta P$$

We can combine these to find,

$$\frac{\partial^2(\delta n)}{\partial t^2}=-\bar{n}\nabla\cdot\frac{\partial\mathbf{u}}{\partial t}=\frac{1}{m}\nabla^2\delta P$$

At this point, we need to invoke the equation of state, relating P to n. It will be useful to give a new name to the quantity $\partial P/\partial n$: we write it as

$$\frac{\partial P}{\partial n}=mc_s^2 \tag{3.4}$$

We can then relate $\delta P=mc_s^2\delta n$ to find that the density perturbations obey

$$\left(\frac{\partial^2}{\partial t^2}-c_s^2\nabla^2\right)\delta n=0 \tag{3.5}$$

This is the *wave equation*. As its name suggests, its solutions are waves of the form

$$\delta n(\mathbf{x}, t) = A(\mathbf{k}) \cos(\omega t - \mathbf{k} \cdot \mathbf{x}) + B(\mathbf{k}) \sin(\omega t - \mathbf{k} \cdot \mathbf{x}) \tag{3.6}$$

We call these *sound waves*. The solution key property of the solution is the *wavevector* $\mathbf{k}$ which determines the direction of travel of the wave and the wavelength $\lambda = 2\pi/|\mathbf{k}|$. The frequency ω of the wave is given by

$$\omega = c_s |\mathbf{k}|$$

The proportionality constant c_s, defined in (3.4), has the interpretation of the speed of sound. (We'll compute examples below.) Finally, $A(\mathbf{k})$ and $B(\mathbf{k})$ are two, arbitrary integration constants. Because the wave equation is linear, we can add together as many solutions of the form (3.6) as we like, with different integration constants $A(\mathbf{k})$ and $B(\mathbf{k})$. In this way, we can build up wavepackets with different profiles.

In what follows, we will often write the solution (3.6) in complex form,

$$\delta n = C(\mathbf{k}) \exp(i(\omega t - \mathbf{k} \cdot \mathbf{x}))$$

for some complex C. This is standard, albeit inaccurate notation. Obviously the number density δn should be real. But because the wave equation is linear, we can always just take the real part of the right-hand-side to get a solution. This form of the solution is more useful simply because it's quicker to write exponentials rather than cos and sin.

Speed of Sound of a Non-Relativistic Fluid

Throughout Section 1, we treated the equation of state of a non-relativistic fluid as $P = 0$. What this really means is that $P \ll \rho$, where ρ is the energy density, mostly due to the rest mass of the fluid.

The equation for the sound speed (3.4) can alternatively be written in terms of the energy density $\rho = mnc^2$, as

$$\frac{\partial P}{\partial \rho} = \frac{c_s^2}{c^2} \tag{3.7}$$

Using $P = 0$ suggests that $c_s = 0$ for a non-relativistic fluid. But what this is really telling is simply that

$$c_s \ll c \tag{3.8}$$

This makes sense. The sound speed is related to the speed of the constituent particles in the fluid. In a non-relativistic fluid, this is necessarily much less than the speed of light.

In fact, we can do better and compute the sound speed as a function of temperature (which, itself, is related to the speed of the constituent particles). For the ideal gas, the equation of state is

$$P = nk_BT$$

It's tempting to simply differentiate $\partial P/\partial n$, with T fixed, to determine the speed of sound. But that's a little too hasty: as P and n vary, it is quite possibly that T varies as well.

To understand how this works, we need a little physical input. The energy of the ideal gas with some fixed total number of atoms $N = nV$ is (2.8)

$$E = \frac{3}{2}Nk_BT$$

If the volume changes, then the energy should change by the work done

$$\begin{aligned} dE = -P\,dV \quad &\Rightarrow \quad \frac{3}{2}Nk_B\,dT = -P\,dV \\ &\Rightarrow \quad \frac{3}{2}\frac{dT}{T} = -\frac{dV}{V} \end{aligned}$$

where, in the second line, we've used the equation of state. Integrating this expression tells us that $T^{3/2}V$ is constant. Alternatively, using $n = N/V$, we learn that $Tn^{-2/3}$ is constant. Such changes are referred to as *adiabatic*. (Underlying this is the statement that entropy is conserved for adiabatic changes; you can learn more about this in the lectures on Statistical Physics.) This means that

$$\left.\frac{\partial T}{\partial n}\right|_{\text{adiabatic}} = \frac{2}{3}\frac{T}{n}$$

The speed of sound (3.4) should be computed under the assumption of such an adiabatic change. We then have

$$\left.\frac{\partial P}{\partial n}\right|_{\text{adiabatic}} = \frac{5}{3}k_BT$$

From this, we compute the speed of sound in an ideal gas to be

$$c_s = \sqrt{\frac{5k_BT}{3m}}$$

Before we proceed, I will briefly mention another, more rigorous, approach to get this result. We could treat $T(\mathbf{x}, t)$ as a new dynamical field with its own equation of motion. This equation of motion turns out to be

$$\left(\frac{\partial}{\partial t} + \mathbf{u}\cdot\nabla\right) T + \frac{2T}{3}\nabla\cdot\mathbf{u} = 0$$

A full derivation of this needs the Boltzmann equation, and can be found in the lectures on Kinetic Theory. It's straightforward to check that this equation combines with the continuity equation (3.1) to ensure that $Tn^{-2/3}$ is indeed constant along flow lines.

Sound Speed of a Relativistic Fluid

So far, our discussion of fluid dynamics was focussed on non-relativistic fluids. However, it should come as no surprise to learn that we will also be interested in relativistic fluids in the context of cosmology.

Much of the discussion above goes through for general fluids if the equations are phrased in terms of the energy density ρ instead of the number density n. In particular, the sound speed can be computed using (3.7). For a relativistic fluid, with $P = \rho/3$, we have

$$c_s^2 = \frac{1}{3}c^2 \tag{3.9}$$

This time we see that the speed of sound is tied to the speed of light. Again, this is to be expected: in a relativistic fluid, any constituent particles are flying around at close to the speed of light. The difference in the speed of sound between a non-relativistic fluid (3.8) and a relativistic fluid (3.9) will prove to be one of the important ingredients in the story of structure formation.

3.1.2 Jeans Instability

Our next step is to add in the effects of gravity. For now we will keep ourselves in a static spacetime. The continuity equation (3.1) remains unchanged. However, the Euler equation (3.2) picks up an extra term on the right-hand-side due to the gravitational field Φ experienced by the fluid,

$$mn\left(\frac{\partial}{\partial t} + \mathbf{u}\cdot\nabla\right)\mathbf{u} = -\nabla P - mn\nabla\Phi \tag{3.10}$$

This gravitational field is determined by the matter in the fluid in the usual manner

$$\nabla^2\Phi = 4\pi G m n \tag{3.11}$$

We want to consider a homogeneous solution as before, with constant $n = \bar{n}$ and $P = \bar{P}$ and $\nabla\Phi = 0$. There's a small problem with this as a starting point: it doesn't obey the Poisson equation (3.11)! This is a famously fiddly aspect of the following derivation, one that stems from the fact that there is no infinite, static self-gravitating fluid. For now, we simply bury our head in the sand and ignore this issue, an approach which is sometimes known as the *Jeans' swindle*. But, for once, this approach will be rewarded: in the next section, we will consider perturbations in an expanding universe where this issue is resolved.

We now perturb the constant background. We require that the perturbed gravitational potential $\Phi + \delta\Phi$ obeys

$$\nabla^2\delta\Phi = 4\pi G m\,\delta n$$

The same linearisation that we saw previously now shows that the wave equation (3.5) is deformed to

$$\left(\frac{\partial^2}{\partial t^2} - c_s^2\nabla^2 - 4\pi G m\bar{n}\right)\delta n = 0$$

This is again solved by the ansatz

$$\delta n = C(\mathbf{k})\,\exp\left(i(\omega t - \mathbf{k}\cdot\mathbf{x})\right)$$

but now with the frequency and wavevector related by

$$\begin{aligned}\omega^2 &= c_s^2k^2 - 4\pi G m\bar{n}\\ &= c_s^2(k^2 - k_J^2)\end{aligned}$$

Equations of this type, which relate the frequency to the wavenumber, are referred to as *dispersion relations*. In the second line we have defined the *Jeans wavenumber* k_J,

$$k_J = \sqrt{\frac{4\pi G m\bar{n}}{c_s^2}}$$

The qualitative properties of the solution now depend on the wavenumber $\mathbf{k}$. For small wavelengths, or large wavenumbers $k > k_J$, the solutions oscillate as before. These are sound waves. However, when the wavelengths are large, $k < k_J$ then the gravitational background becomes important. Here the frequency is imaginary, which has the interpretation that perturbations $\delta n \sim e^{i\omega t}$ grow or decay exponentially. For $k \ll k_J$ we have

$$\delta n \sim e^{\pm t/\tau} \quad \text{with} \quad \tau \approx \sqrt{\frac{1}{4\pi G m\bar{n}}} \tag{3.12}$$

We learn that long wavelength perturbations no longer oscillate as sound waves. Instead, any perturbation that has a size larger than the *Jeans length*,

$$\lambda_J = \frac{2\pi}{k_J} = c_s\sqrt{\frac{\pi}{G m\bar{n}}} \tag{3.13}$$

will typically grow exponentially quickly due to the effect of gravity. This is known as the *Jeans' instability*.

The derivation above also gives us a clue as to the physical mechanism for the Jeans instability. It comes from attempting to balance the pressure and gravitational terms in the Euler equation (3.10). Consider an over-dense spherical region of radius R. In the absence of any pressure, this region would collapse with a time-scale τ given in (3.12). In a fluid, this collapse is opposed by the pressure. But the build-up of pressure is not instantaneous; it takes time given roughly by

$$t_{\rm pressure} \sim \frac{R}{c_s}$$

When R is small, $t_{\rm pressure} < \tau$ and the build-up of pressure stops the collapse and we get oscillating motion that we interpret as sound waves. In contrast, if R is large we have $\tau < t_{\rm pressure}$, there is no time for the pressure to build. In this case, the system suffers the Jeans instability and is susceptible to gravitational collapse.

3.1.3 Density Perturbations in an Expanding Space

Finally, we want to consider something more cosmological: the growth of density perturbations, interacting with gravity, in an expanding space with scale factor $a(t)$. Throughout we will work with flat space. (This means "$k = 0$" in the notation of Section 1. However, in this Section we use k to denote the wavenumber of perturbations.)

We need to revisit our equations once more. Consider a particle tracing out a trajectory $\mathbf{x}(t)$ in co-moving coordinates. The physical coordinates $\mathbf{r}(t)$ (called $\mathbf{x}_{\rm phys}$ in (1.14)) is given by

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

The physical velocity of a particle is

$$\mathbf{u} = \dot{\mathbf{r}} = H\mathbf{r} + \mathbf{v} \tag{3.14}$$

with $\mathbf{v} = a\dot{\mathbf{x}}$. In what follows, we will need to jump between physical and co-moving coordinates. The spatial derivatives are related simply by

$$\nabla_{\mathbf{r}} = \frac{1}{a}\nabla_{\mathbf{x}} \tag{3.15}$$

The temporal derivatives are a little more subtle since they differ depending on whether we keep $\mathbf{r}(t)$ fixed or $\mathbf{x}(t)$ fixed. In particular

$$\begin{aligned}\left.\frac{\partial}{\partial t}\right|_{\mathbf{r}} &= \left.\frac{\partial}{\partial t}\right|_{\mathbf{x}} + \left.\frac{\partial \mathbf{x}}{\partial t}\right|_{\mathbf{r}} \cdot \nabla_{\mathbf{x}} \\ &= \left.\frac{\partial}{\partial t}\right|_{\mathbf{x}} + \left.\frac{\partial (a^{-1}\mathbf{r})}{\partial t}\right|_{\mathbf{r}} \cdot \nabla_{\mathbf{x}} \\ &= \left.\frac{\partial}{\partial t}\right|_{\mathbf{x}} - H\mathbf{x} \cdot \nabla_{\mathbf{x}} \end{aligned} \tag{3.16}$$

Now we come to the equations describing the fluid. The equations that we dealt with previously should be viewed as given in terms of physical coordinates $\mathbf{r}$. However, it will turn out that subsequent calculations are somewhat easier if done in co-moving coordinates. We just have to translate from one to the other.

The Continuity Equation Revisited

The continuity equation (3.1) should be viewed in physical coordinates and so, in our new notation, reads

$$\left.\frac{\partial n}{\partial t}\right|_{\mathbf{r}} = -\nabla_{\mathbf{r}} \cdot (n\mathbf{u})$$

Changing to co-moving coordinates, it then becomes

$$\left(\left.\frac{\partial}{\partial t}\right|_{\mathbf{x}} - H\mathbf{x} \cdot \nabla_{\mathbf{x}} \right) n = -\frac{1}{a} \nabla_{\mathbf{x}} \cdot (n\mathbf{u})$$

In what follows, we drop the subscript $\mathbf{x}$ on everything; ∇ will always mean $\nabla_{\mathbf{x}}$ and $\frac{\partial}{\partial t}$ will always mean $\left.\frac{\partial}{\partial t}\right|_{\mathbf{x}}$.

We can make contact with the story of Section 1. Following (3.14), we write the velocity of the fluid as

$$\mathbf{u}(\mathbf{x}, t) = Ha\mathbf{x}(t) + \mathbf{v}(\mathbf{x}, t) \tag{3.17}$$

and the continuity equation becomes

$$\frac{\partial n}{\partial t} + 3Hn + \frac{1}{a} \nabla \cdot (n\mathbf{v}) = 0 \tag{3.18}$$

where we've used $\nabla \cdot \mathbf{x} = 3$. This form makes it clear that if we restrict to solutions in which $\mathbf{v} = 0$, so the velocity of the fluid simply follows the expansion of spacetime,

then we recover our earlier continuity equation (1.39), specialised to the case of non-relativistic matter,

$$\frac{\partial n}{\partial t} = -3Hn$$

(Recall that the energy density is given by $\rho = m\bar{n}c^2$.) This has the familiar solution

$$n(t) \sim \frac{1}{a^3} \tag{3.19}$$

which simply tells us that the number density dilutes as the universe expands.

Now we perturb the fluid,

$$\begin{aligned} n(\mathbf{x}, t) &= \bar{n}(t) + \delta n(\mathbf{x}, t) \\ &= \bar{n}(t)\Big[1 + \delta(\mathbf{x}, t)\Big] \end{aligned}$$

where $\bar{n}(t)$ is a spatially homogeneous density evolving as (3.19) and, in the second line, we've defined

$$\delta = \frac{\delta n}{\bar{n}} = \frac{\delta \rho}{\bar{\rho}}$$

The perturbation δ is referred to as the *density contrast*.

Let's now see what conditions the continuity equation (3.18) imposes on these perturbations. It reads

$$\begin{aligned} \frac{\partial}{\partial t}(\bar{n}\delta) + 3H\bar{n}\delta &= -\frac{1}{a}\nabla\left[\bar{n}(1+\delta)\mathbf{v}\right] \\ &= -\frac{\bar{n}}{a}\nabla\cdot\mathbf{v} + \mathcal{O}(\mathbf{v}\delta) \end{aligned}$$

We drop the second term on the grounds that it is non-linear in the small quantities δ and $\mathbf{v}$. Using the fact that the background density $\bar{n}$ evolves as (3.19), this equation reduces to the simple requirement

$$\dot{\delta} = -\frac{1}{a}\nabla\cdot\mathbf{v} \tag{3.20}$$

This is the first of our perturbed equations.

The Euler and Poisson Equations Revisited

Next up we need to deal are the Euler equation (3.10) and Poisson equation (3.11). The Euler equation as written in (3.10) should again be viewed in physical coordinates,

$$mn\left(\left.\frac{\partial}{\partial t}\right|_{\mathbf{r}} + \mathbf{u}\cdot\nabla_{\mathbf{r}}\right)\mathbf{u} = -\nabla_{\mathbf{r}}P - mn\nabla_{\mathbf{r}}\Phi$$

After substituting in (3.17), (3.15) and (3.16), this becomes

$$mna\left(\frac{\partial}{\partial t} + \frac{\mathbf{v}}{a}\cdot\nabla\right)\mathbf{u} = -\nabla P - mn\nabla\Phi \tag{3.21}$$

where, as previously, the lack of any subscript on the derivatives means that they are taken holding $\mathbf{x}$ fixed. A similar, but simpler, story also holds for the Poisson equation. In physical coordinates, this is

$$\nabla^2\Phi = 4\pi G mn$$

In co-moving coordinates, it becomes

$$\nabla^2\Phi = 4\pi G mna^2 \tag{3.22}$$

The background $\mathbf{u} = Ha\mathbf{x}$, with $\mathbf{v} = 0$, solves the Euler equation provided that we take $\nabla\bar{P} = 0$ and $\Phi = \bar{\Phi}$ such that

$$\nabla\bar{\Phi} = -\ddot{a}a\mathbf{x} \quad \Rightarrow \quad \nabla^2\bar{\Phi} = -3\ddot{a}a \tag{3.23}$$

This is now perfectly compatible with the Poisson equation; indeed, the two combine to give the

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}m\bar{n}$$

But this is precisely the acceleration equation (1.52) that we met previously. Note that we didn't assume the Friedmann equation anywhere in this derivation. Nonetheless, we find the acceleration equation (which, recall, is the time derivative of the Friedmann equation) emerging as a consistency condition on our analysis! This isn't as miraculous as it may first appear. Our derivation of the Friedmann equation in Section 1.2.3 involved only Newtonian gravity, which is the same physics we have invoked here. However, in one particular sense, the current derivation using fluids is a considerable improvement on the derivation in Section 1.2.3, because we didn't have to make the misleading assumption that there is an origin of the universe from which all matter is expanding. Instead, the fluid treatment allows us to understand the expansion of a genuinely homogeneous and infinite universe.

For our immediate purposes, the thing that should make us most happy is that we no longer have to worry about the Jeans' swindle; our spatially homogeneous background satisfies the equations of motion as it should. At heart, the Jeans's swindle was telling us that a spatially homogeneous fluid is inconsistent in Newtonian gravity. But it is perfectly consistent if we allow for an expanding universe, with the gravitational potential $\bar{\Phi}$ for a homogeneous fluid driving the expansion of space, just as we learned in Section 1.

Next, we perturb around the background. We write $P = \bar{P} + \delta P$ and $\Phi = \bar{\Phi} + \delta\Phi$ and, as before, $\mathbf{u} = Ha\mathbf{x} + \mathbf{v}$. The linearised Euler equation reads

$$m\bar{n}a\left(\dot{\mathbf{v}} + H\mathbf{v}\right) = -\nabla\delta P - m\bar{n}\nabla\delta\Phi \tag{3.24}$$

where we've used the fact that $(\mathbf{v}\cdot\nabla)\mathbf{x} = \mathbf{v}$. If we drop the pressure and gravitational perturbation on the right-hand side, this equation tells us that $\dot{\mathbf{v}} = -H\mathbf{v}$, so the peculiar velocities redshift as $\mathbf{v} \sim 1/a$. This can be viewed as a consequence of Hubble friction, which slows the peculiar velocities as the universe expands.

Finally, the linearised Poisson equation is

$$\nabla^2\delta\Phi = 4\pi G m\bar{n}a^2\delta \tag{3.25}$$

Now we combine our three linearised equations (3.20), (3.24) and (3.25). Take the time derivative of (3.20) to get

$$\ddot{\delta} = \frac{H}{a}\nabla\cdot\mathbf{v} - \frac{1}{a}\nabla\cdot\dot{\mathbf{v}} = -H\dot{\delta} - \frac{1}{a}\nabla\dot{\mathbf{v}}$$

Take the gradient of (3.24) to get

$$\begin{aligned} m\bar{n}a\left(\nabla\cdot\dot{\mathbf{v}} - Ha\dot{\delta}\right) &= -\nabla^2\delta P - m\bar{n}\nabla^2\delta\Phi \\ &= -m\bar{n}\left(c_s^2\nabla^2\delta + 4\pi G m\bar{n}a^2\delta\right) \end{aligned}$$

where in the second line we've used $\delta P = mc_s^2\delta n = mc_s^2\bar{n}\delta$ and the Poisson equation (3.25). We now combine these two results to get a single equation telling us how the density perturbation δ evolves in an expanding spacetime

$$\ddot{\delta} + 2H\dot{\delta} - c_s^2\left(\frac{1}{a^2}\nabla^2 + k_J^2\right)\delta = 0 \tag{3.26}$$

where k_J is the physical Jeans wavenumber given, as before, by $k_J^2 = 4\pi G m\bar{n}/c_s^2$.

The most important addition from the expanding space time is the friction-like term proportional $2H\dot{\delta}$. This is referred to as *Hubble friction* or *Hubble drag*. (We saw an analogous term when discussing inflation in Section 1.5.)

To solve (3.26), it is simplest to find by working in Fourier space. We define

$$\delta(\mathbf{k},t) = \int d^3x\; e^{i\mathbf{k}\cdot\mathbf{x}}\,\delta(\mathbf{x},t)$$

where we are adopting the annoying but standard convention that the function $\delta(\mathbf{x})$ and its Fourier transform $\delta(\mathbf{k})$ are distinguished only by their argument. Since $\mathbf{x}$ is the co-moving coordinate, $\mathbf{k}$ is the co-moving wavevector.

The advantage of working in Fourier space is that the equation (3.26) decomposes into a separate equation for each value of $\mathbf{k}$,

$$\ddot{\delta}(\mathbf{k},t) + 2H\dot{\delta}(\mathbf{k},t) + c_s^2\left(\frac{k^2}{a^2} - k_J^2\right)\delta(\mathbf{k},t) = 0 \tag{3.27}$$

The slightly unusual factor of a in the final term arises because k is the co-moving wavenumber and so k/a is the physical wavenumber, but k_J refers to the physical Jeans wavenumber. Our challenge now is to solve this equation.

3.1.4 The Growth of Perturbations

Solutions to (3.27) have different behaviour depending on whether the perturbations have small or large wavelength compared to the Jeans' wavelength $\lambda_J = 2\pi/k_J$.

Small wavelength modes have $k/a \gg k_J$. Here, the equation (3.27) is essentially that of a damped harmonic oscillator, with the expanding universe providing the friction term $H\dot{\delta}$. The solutions are oscillating sound-waves with the Hubble friction leading to an ever-decreasing amplitude.

If structure is to ultimately form in the universe, we need to find solutions that grow over time. These are supplied by the long-wavelength modes, with $k/a < k_J$, which suffer from the Jeans' instability. However, as we shall now see, the details of the Jeans' instability are altered in an expanding universe.

In what follows, we will see that there are two length scales at play for the growing modes. One is the Jeans' length scale $\lambda_J = 2\pi/k_J$,

$$\lambda_J = c_s\sqrt{\frac{\pi}{Gm\bar{n}}} = c_s c\sqrt{\frac{\pi}{G\bar{\rho}}} \tag{3.28}$$

Only modes with $\lambda > \lambda_J$ will grow. The other relevant physical length scale is that set by the expansion of the universe,

$$d_H \approx cH^{-1} = c^2\sqrt{\frac{3}{8\pi G\bar{\rho}}} \tag{3.29}$$

This is called the *apparent horizon*. In the standard FRW cosmology (with ordinary matter or radiation) the apparent horizon coincides with the particle horizon, defined in (1.24). In such a situation, it would make little sense to talk about perturbations with wavelength $\lambda > d_H$. This is because the Fourier mode of a perturbation is a coherent wave and causality would appear to prohibit the formation of such perturbations on distances greater than d_H since there has been no time for light, or anything else, to cross this distance since the Big Bang.

This, however, is exactly the problem that is resolved by a period of inflation in the very early universe. The whole point of inflation is to stretch the particle horizon so that it sits way outside the apparent horizon. Indeed, we will see that perturbation modes with wavelengths $\lambda > d_H$ play an important role in the story of structure formation in our universe, strongly implying that a period of inflation is needed. In what follows, we will refer to the apparent horizon (3.29) simply as the "horizon".

Matter Perturbations in a Matter Dominated Universe

For non-relativistic fluids, the Jeans' length (3.28) always sits well within the horizon (3.29),

$$c_s \ll c \quad \Rightarrow \quad \lambda_J \ll cH^{-1}$$

This means that the perturbations which suffer the Jeans' instability include both sub-horizon and super-horizon wavelengths.

Af the wavelength of the mode is sufficiently long, so $k/a \ll k_J$, then we can approximate (3.27) as

$$\ddot{\delta}(\mathbf{k}) + 2H\dot{\delta}(\mathbf{k}) - \frac{4\pi G\bar{\rho}}{c^2}\,\delta(\mathbf{k}) = 0 \tag{3.30}$$

Here we've left the t argument in $\delta(\mathbf{k}, t)$ implicit, but kept the $\mathbf{k}$ argument because it tells us that the is the mode is in Fourier space rather than real space.

In a matter dominated universe, $a \sim t^{2/3}$ so $H = 2/3t$. The third term is also related to the Hubble parameter through Friedmann equation $H^2 = 8\pi G\bar{\rho}/3c^2$. We then have

$$\ddot{\delta}(\mathbf{k}) + \frac{4}{3t}\dot{\delta}(\mathbf{k}) - \frac{2}{3t^2}\delta(\mathbf{k}) = 0$$

Substituting in the power-law ansatz $\delta(\mathbf{k}, t) \sim t^n$, we find two solutions, one decaying and one growing

$$\delta(\mathbf{k}, t) \sim \begin{cases} t^{2/3} \sim a \\ t^{-1} \sim a^{-3/2} \end{cases} \tag{3.31}$$

We see that the expansion of the universe slows down the rate at which objects undergo gravitational collapse, with the Hubble damping turning the exponential growth of the Jeans instability (3.12) into a power-law, one that scales linearly with the size of the universe.

Radiation Perturbations in a Radiation Dominated Universe

Although we have derived the perturbation equation (3.27) for non-relativistic fluids, it is not too difficult to modify them in a plausible way to give us an understanding of the perturbations in other fluid components. Here we will be interested in radiation, but things are clearer if we work with the general equation of state

$$P = w\rho$$

and only later restrict to $w = 1/3$.

We will work with the energy density $\rho(\mathbf{x}, t)$, rather than the number density $n(\mathbf{x}, t)$; for a non-relativistic fluid, they are related by $\rho = mnc^2$. We need to go through each of our original equations – continuity, Euler, and Poisson – and ask how they change for a general fluid. We will motivate each of these changes, but not derive them.

First, the continuity equation (3.18): this gets replaced by

$$\frac{\partial \rho}{\partial t} + 3(1+w)H\rho + \frac{1}{a}(1+w)\nabla \cdot (\rho \mathbf{v}) = 0$$

The equation of state parameter w appears twice. The first of these is unsurprising, since it guarantees that this equation reduces to our previous continuity equation (1.39) when $\mathbf{v} = 0$. We won't derive the presence of the $(1 + w)$ factor in the final term, but it arises in a similar way to the first.

The Euler equation (3.21) remains unchanged. Somewhat more subtle is the relativistic generalisation of the gravitational potential. In general relativity, both energy density and pressure gravitate. It turns out that the Poisson equation (3.22) should be replaced by

$$\nabla^2\Phi = \frac{4\pi G}{c^2}(1+3w)\rho a^2 \tag{3.32}$$

There is, in fact, a clue in the discussion above that strongly hints at this form. Recall that we avoided the Jeans swindle in an expanding spacetime by relating the gravitational potential to the acceleration in (3.23). The Poisson equation then became equivalent to the acceleration equation (1.52) which, in general, reads

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3c^2}(1+3w)\rho$$

We see the same distinctive factor of $(1+3w)$ appearing here.

Repeating the same steps as previously, the perturbation equation (3.27) is replaced by

$$\ddot{\delta}(\mathbf{k}) + 2H\dot{\delta}(\mathbf{k}) + c_s^2(1+w)\left(\frac{k^2}{a^2} - (1+3w)k_J^2\right)\delta(\mathbf{k}) = 0 \tag{3.33}$$

with $k_J = 2\pi/\lambda_J$ is the physical wavenumber, defined as in (3.28). It differs from the non-relativistic Jeans' length only by the expression for the speed of sound c_s.

Let's now restrict to radiation with $w = 1/3$. We know from (3.9) that the speed of sound for a relativistic fluid is

$$c_s^2 = \frac{1}{3}c^2$$

This means that there is no parametric separation between the Jeans length (3.28) and the horizon (3.29). Instead, we have $\lambda_J \approx cH^{-1}$. Any perturbation that lies inside the horizon does not grow. Instead, the pressure of the radiation causes the perturbation to oscillate as a sound wave.

Outside the horizon it's a different story. In a radiation dominated universe, $a \sim t^{1/2}$ so $H = 1/2t$. For wavenumbers $k/a \ll k_J$, the equation (3.33) governing perturbations becomes

$$\ddot{\delta}_r(\mathbf{k}) + \frac{1}{t}\dot{\delta}_r(\mathbf{k}) - \frac{1}{t^2}\delta_r(\mathbf{k}) = 0$$

Once again, substituting in the power-law ansatz $\delta(\mathbf{k}, t) \sim t^n$, we find two solutions, one decaying and one growing

$$\delta_r(\mathbf{k}, t) \sim \begin{cases} t & \sim a^2 \\ t^{-1} & \sim a^{-2} \end{cases} \tag{3.34}$$

We lean that perturbations in the density of radiation grow outside the horizon. Indeed, they grow faster than the linear growth (3.31) seen in the matter dominated era.

Matter Perturbations in a Radiation Dominated Universe

We could also ask about density perturbations of matter in a radiation dominated universe. As we've seen, the Jeans length for matter is well within the horizon (because $c_s \ll c$). In a universe with multiple energy components ρ_i, matter perturbations δ_m with $k/a \ll k_J$ are described by a modified version of (3.30),

$$\ddot{\delta}_m(\mathbf{k}) + 2H\dot{\delta}_m(\mathbf{k}) - \frac{4\pi G}{c^2}\sum_i \bar{\rho}_i \delta_i(\mathbf{k}) = 0 \tag{3.35}$$

This final term can be traced to the gravitational potential $\delta\Phi$, which receives contributions from all energy sources. However, on sub-horizon scales, we have seen that the radiation perturbation does not grow, so we can set $\delta_r(\mathbf{k}) \approx 0$. Meanwhile, in the radiation dominated phase $\bar{\rho}_r \gg \bar{\rho}_m$ and so we can also ignore the $\bar{\rho}_m \delta_m(\mathbf{k})$ which will be sub-dominant to the $H\dot{\delta}_m(\mathbf{k})$ term. Using $H = 1/2t$, we have

$$\ddot{\delta}_m(\mathbf{k}) + \frac{1}{t}\dot{\delta}_m(\mathbf{k}) \approx 0 \quad \Rightarrow \quad \delta_m(\mathbf{k}, t) \sim \begin{cases} \log t \sim \log a \\ \text{constant} \end{cases} \tag{3.36}$$

We learn that, during the radiation dominated era, the matter perturbations inside the horizon grow only logarithmically. This slow growth occurs because the expansion of the universe is faster in the radiation dominated phase than in the matter dominated phase. A logarithmic increase is rather pathetic and it means that significant growth in sub-horizon scale perturbations gets going when we hit the matter dominated era at $z \approx 3400$

In contrast, the matter perturbations with wavelength larger than the horizon obey the same equation as the radiation perturbations in the radiation dominated era. (The term $\bar{\rho}_r \delta_r$ in (3.35) cannot now be neglected.) This means that those modes outside the horizon grow as $\delta \sim a^2$ as seen in (3.34).

The Cosmological Constant

We can also ask how matter perturbations grow in a universe dominated by a cosmological constant. We again use the perturbation equation (3.35). It is not possible to have perturbations δ_Λ. (This is what the "constant" in cosmological constant means!) Once again, $\bar{\rho}_m$ is negligible, so we have

$$\ddot{\delta}_m + 2H\dot{\delta}_m \approx 0$$

where, from Section 1.3.3, $H = \sqrt{\Lambda/3}$. The solutions are now

$$\delta_m \sim \begin{cases} \text{constant} \\ e^{-2Ht} \sim a^{-2} \end{cases}$$

We learn that in a universe dominated by a cosmological constant, there is no growth of perturbations. In other words, dark energy kills any opportunity to form galaxies. We will revisit this in Section 3.3.4.

3.1.5 Validity of the Newtonian Approximation

Everything we've done in this section relies on the perturbation equation (3.27), which was derived for non-relativistic matter using Newtonian gravity. However, as we stressed in Section 1.2, a proper description of expanding spacetime requires general relativity. So should we trust the Newtonian approximation?

We should be able to trust our equations on small length scales, for the simple reason that general relativity reduces to Newtonian gravity in this regime. However, when we get to perturbations whose length is comparable to the horizon d_H, we should be more nervous, since it seems plausible that the perturbations feel the curvature of spacetime in a way that our Newtonian approximation misses.

The only way to know if the Newtonian perturbation equation (3.27) is valid is to roll up our sleeves and perform the correct, general relativistic perturbation theory. This is a somewhat painful exercise that you will be given the opportunity to embrace in next year's Part III cosmology course. There you will learn that the question "is our equation (3.27) valid?" has a short answer and a long answer.

The short answer is: yes.

The long answer is substantially more subtle. It turns out that the matter perturbation in general relativity is not diffeomorphism invariant, which means that the answer you get depends on the coordinates you use. This is bad. Indeed, one of the main philosophical lessons of general relativity is that the coordinates you use should not matter one iota. Moreover, this issue is particularly problematic for super-horizon perturbations with $\lambda \gg d_H$, and an important part of the relativistic approach is to understand the right, diffeomorphism invariant quantity to focus on. For most choices of coordinates (so called "gauges") it turns out that the Poisson equation (3.25) is not valid on super-horizon scales. There is, however, a choice of coordinates – conformal Newtonian gauge – where the Poisson equation holds even on super-horizon scales and this is the one we are implicitly choosing[9]. All of this is to say that you can trust the physics that we've derived here, but if you should be careful when comparing to analogous results derived in a general relativistic setting where the answers may look different because, although the symbols are the same, they refer to subtly different objects.

3.1.6 The Transfer Function

There are a number of different questions that we could try to answer now. We posed one such question in the introduction to this section: can we compute the overall growth of the density perturbations to explain how we got from $\delta T/T \sim 10^{-5}$ in the CMB to the world we see around us. This, it turns out, requires some more discussion which we postpone to Section 3.3.2. Instead, we will ask about the relative growth of density perturbations of different wavenumber k.

We are interested in the perturbations of the matter, since this is what we're ultimately made of. If the density perturbations remain sufficiently small, so that the linearised analysis developed above holds then the linear analysis of this section remains valid which, in particular, means that each $\delta(\mathbf{k})$ evolves independently, as seen in, for example, (3.27). The evolution of a perturbation of a given wavevector $\mathbf{k}$ from an initial time t_i to the present can be distilled into a *transfer function* $T(k)$, defined as

$$\delta(\mathbf{k}, t_0) = T(k)\, \delta(\mathbf{k}, t_i) \tag{3.37}$$

The initial time t_i is usually taken to be early, typically just after the end of inflation.

[9]More details can be found in the paper by Chisari and Zaldarriaga, "*Connection between Newtonian simulations and general relativity*, arXiv:1101.3555.

Key to understanding the physics is the question of when perturbations enter the horizon. Recall that, in physical coordinates, the apparent horizon is $d_H \approx c/H$ as in (3.29). It is simplest, however, to work in co-moving coordinates, where the apparent horizon is

$$\chi_H = \frac{c}{aH}$$

In the radiation dominated era, $a \sim t^{1/2}$ and so $H \sim 1/a^2$. In the matter dominated era, $a \sim t^{2/3}$ and so $H \sim 1/a^{3/2}$. In both cases, the co-moving horizon increases over time

$$\chi_H \sim \begin{cases} a & \text{radiation domination} \\ a^{1/2} & \text{matter domination} \end{cases} \tag{3.38}$$

The intuition behind this is that, as the universe expands, there is more that one can see and, correspondingly, the co-moving horizon grows.

The co-moving wavevector $\mathbf{k}$ remains unchanged over time. (This is the main advantage of working with the co-moving wavevector in the previous section. In contrast, the physical wavevector is $\mathbf{k}_{\rm phys} = \mathbf{k}/a$ shrinks over time as the physical wavelength $\lambda_{\rm phys} = 2\pi/k_{\rm phys} = 2\pi a/k$ is stretched by the expansion of the universe.) This means that, for each $\mathbf{k}$, there will be a time when the corresponding perturbation enters the horizon. It matters whether the time of entry occurs during the radiation or matter dominated eras.

At the time of matter-radiation equality (which occurred around $z = 3400$), modes with wavenumber $k_{\rm eq}$ have just entered the horizon, where

$$k_{\rm eq} = \frac{2\pi}{c}(aH)_{\rm eq}$$

Modes larger than this (i.e. with $k < k_{\rm eq}$) will then enter the horizon in the matter dominated era. Modes smaller than this (i.e. with $k > k_{\rm eq}$) will have entered the horizon during the radiation dominated era. Let's look at each of these in turn.

First the long wavelength modes with $k < k_{\rm eq}$. These were outside the horizon during the radiation era where δ grew as a^2, as seen in (3.34). As the universe entered the matter dominated era, the growth slows to $\delta \sim a$, as seen in (3.31). This means that, starting from an initial time t_i, they evolve to their present day value

$$\delta(\mathbf{k}, t_0) = \left(\frac{a_{\rm eq}}{a_i}\right)^2 \frac{a_0}{a_{\rm eq}}\, \delta(\mathbf{k}, t_i) \qquad \text{for } k < k_{\rm eq} \tag{3.39}$$

We learn that each mode grows by an amount independent of $\mathbf{k}$.

Things are more interesting for the short wavelength modes with $k > k_{\rm eq}$ which enter the horizon during the radiation era. Before entering the horizon, such modes grow as $\delta \sim a^2$ as in (3.34). However, when they enter the horizon, their growth slows to the logarithmic growth seen in (3.36). For our purposes, this is effectively constant. The growth only resumes when the universe enters the matter dominated era. This means that

$$\begin{aligned}\delta(\mathbf{k}, t_0) &= \left(\frac{a_{\rm enter}}{a_i}\right)^2 \frac{a_0}{a_{\rm eq}}\, \delta(\mathbf{k}, t_i) \\ &= \left(\frac{a_{\rm enter}}{a_{\rm eq}}\right)^2 \times \left[\left(\frac{a_{\rm eq}}{a_i}\right)^2 \frac{a_0}{a_{\rm eq}}\right] \delta(\mathbf{k}, t_i) \qquad \text{for } k > k_{\rm eq}\end{aligned} \tag{3.40}$$

The factor in square brackets is the same, constant amount (3.39) that the long wavelength modes grew by. However, the amplitude is suppressed by the factor of $a^2_{\rm enter}/a^2_{\rm eq}$, reflecting the fact that growth stalled during the radiation dominated era. For a given mode $\mathbf{k}$, the scale factor at horizon entry is given by

$$k = \frac{2\pi}{c}(aH)_{\rm enter}$$

Using $a \sim t^{1/2}$ in the radiation era, we have $H = 1/2t \sim 1/a^2$ so a given scale k enters the horizon at $k \sim (aH)_{\rm enter} \sim 1/a_{\rm enter}$. We can then write $(a_{\rm enter}/a_{\rm eq})^2 = k^2_{\rm eq}/k^2$.

Finally, all of this can be packaged into the transfer function (3.37). Assuming that the perturbations remain sufficiently small, so the linearised analysis is valid, the transfer function can be found in (3.39) and (3.40), we have

$$T(k) \sim \text{constant} \times \begin{cases} 1 & k < k_{\rm eq} \\ k^{-2} & k > k_{\rm eq} \end{cases} \tag{3.41}$$

We will make use of this shortly.

3.2 The Power Spectrum

It should be obvious that we're not going to understand the density perturbations in the early universe to enough accuracy to predict the location of, say, my mum's house. Or even the location of the Milky Way galaxy. Instead, if we want make progress then we must lower our ambitions. We will need to develop a statistical understanding of the distribution of galaxies in the universe.

To this end, we consider various averages of the density contrast,

$$\delta(\mathbf{x},t) = \frac{\delta\rho(\mathbf{x},t)}{\bar{\rho}(t)}$$

By construction, the spatial average of δ itself at a given time vanishes,

$$\langle\delta(\mathbf{x},t)\rangle = 0$$

The first non-trivial information lies in the correlation function, defined by the spatial average

$$\xi(|\mathbf{x}-\mathbf{y}|,t) = \langle\delta(\mathbf{x},t)\,\delta(\mathbf{y},t)\rangle \tag{3.42}$$

Our old friend, the cosmological principle, is implicit in the left-hand side where we have assumed that the universe is statistically homogeneous and isotropic, so that the function $\xi(\mathbf{x},\mathbf{y},t)$ depends only on $|\mathbf{x}-\mathbf{y}|$. The correlation function $\xi(r,t)$ tells us the likelihood that, at time t, two galaxies are separated by a distance r.

Further statistical information about $\delta(\mathbf{x})$ can be distilled into higher correlation functions, such as $\langle\delta\delta\delta\rangle$ However, in what follows we will limit ourselves to understanding the correlation function $\xi(r,t)$.

In Section 3.1, we learned that the evolution of the density perturbations is best described in momentum space,

$$\delta(\mathbf{k},t) = \int d^3x\; e^{i\mathbf{k}\cdot\mathbf{x}}\,\delta(\mathbf{x},t) \tag{3.43}$$

The correlation function in momentum space is given by

$$\begin{aligned}
\langle\delta(\mathbf{k},t)\,\delta(\mathbf{k}',t)\rangle &= \int d^3x\,d^3y\; e^{i\mathbf{k}\cdot\mathbf{x}+i\mathbf{k}'\cdot\mathbf{y}}\,\langle\delta(\mathbf{x},t)\,\delta(\mathbf{y},t)\rangle \\
&= \int d^3x\,d^3y\; e^{i\mathbf{k}\cdot\mathbf{x}+i\mathbf{k}'\cdot\mathbf{y}}\,\xi(r,t) \\
&= \int d^3r\,d^3y\; e^{i\mathbf{k}\cdot\mathbf{r}+i(\mathbf{k}+\mathbf{k}')\cdot\mathbf{y}}\,\xi(r,t) \\
&= (2\pi)^3\,\delta_D^3(\mathbf{k}+\mathbf{k}')\int d^3r\; e^{i\mathbf{k}\cdot\mathbf{r}}\,\xi(r,t)
\end{aligned} \tag{3.44}$$

where, in the second line, we've defined $\mathbf{r} = \mathbf{x}-\mathbf{y}$. The Dirac delta-function $\delta_D^3(\mathbf{k}+\mathbf{k}')$ reflects the underlying (statistical) translation invariance. (Note that I've added a

subscript D on the Dirac delta $\delta_D^3(\mathbf{k})$ to distinguish it from the density contrast $\delta(\mathbf{k}, t)$!) The remaining function is called the *power spectrum*,

$$P(k,t) = \int d^3r \; e^{i\mathbf{k}\cdot\mathbf{r}}\, \xi(r,t)$$

This is the three-dimensional Fourier transform of the correlation function. If we work in spherical polar coordinates, chosen so that $\mathbf{k}\cdot\mathbf{r} = kr\cos\theta$, then we have

$$\begin{aligned} P(k,t) &= \int_0^{2\pi} d\phi \int_{-1}^{+1} d(\cos\theta) \int_0^\infty dr \; r^2 e^{ikr\cos\theta} \xi(r,t) \\ &= 2\pi \int_0^\infty \frac{r^2}{ikr} \left[e^{ikr} - e^{-ikr}\right] \xi(r,t) \\ &= \frac{4\pi}{k} \int_0^\infty dr \; r\sin(kr)\, \xi(r,t) \end{aligned} \tag{3.45}$$

The spatial correlation function $\xi(r)$ can be measured by averaging over many galaxies in the sky. (We'll say more about this in Section 3.2.5.) Meanwhile, the power spectrum $P(k)$ is the most natural theoretical object to consider. The formula above relates the two.

3.2.1 Adiabatic, Gaussian Perturbations

To describe the structure of galaxies in our universe, we introduce a probability distribution for $\delta(k)$. The idea is that averages computed from the distribution will coincide with the spatial average which leads to $\xi(r)$ and, relatedly, $P(k)$.

There are two basic questions that we need to address:

- What is the initial probability distribution?
- How did this probability distribution subsequently evolve?

If we understand both of these well enough, we should be able to compare our results to the distribution of galaxies observed in the sky. We start by describing the initial probability distribution. We then see how it evolves in Section 3.2.3.

It may seem daunting to guess the form of the initial perturbations. However, the universe is kind to us and the observational evidence suggests that these perturbations take the simplest form possible. (We will offer an explanation for this in Section 3.5.)

First, the perturbations of each fluid component are correlated. In particular, the perturbation in any non-relativistic matter, such as baryons and cold dark matter, is the same: $\delta_B = \delta_{CDM}$. Furthermore, perturbations in matter and perturbations in radiation are related by

$$\delta_m = \frac{3}{4}\delta_r \tag{3.46}$$

Perturbations of this kind are called *adiabatic*.

It may seem like a minor miracle that the perturbations in all fluids are correlated in this way. What's really happening is that there is an initial perturbation in the gravitational potential (or, in the language of general relativity, in the metric) which, in turn, imprints itself on each of the fluids in the same way.

Logically, we could also have initial perturbations of the form $\delta\rho_m = -\delta\rho_r$. These are referred to as *isocurvature perturbations* because the net perturbation $\delta\rho = \delta\rho_m + \delta\rho_r = 0$ gives no change to the local curvature of spacetime. There is no hint of these isocurvature perturbations in our universe.

Since we have adiabatic perturbations, we need only specify a probability distribution for a single component, which we take to be $\delta \equiv \delta_m$. We take this distribution to be a simple Gaussian

$$\text{Prob}\Big[\delta(\mathbf{k})\Big] = \frac{1}{\sqrt{2\pi P(k)}} \exp\left(-\frac{\delta(\mathbf{k})^2}{2P(k)}\right) \tag{3.47}$$

This expression holds for each $\mathbf{k}$ independently. This means that there is no correlation between perturbations with different wavelengths. This is an assumption, and one that can be tested since it means that, at least initially, all higher point correlation functions are determined purely in terms of the one-point and two-point functions. For example, $\langle\delta\delta\delta\rangle = 0$.

Note that the power spectrum $P(k)$ arises in this distribution in the guise of the variance. This ensures that the two-point function is indeed given by

$$\langle\delta(\mathbf{k}, t_i)\,\delta(\mathbf{k}', t_i)\rangle = (2\pi)^3\,\delta_D^3(\mathbf{k} + \mathbf{k}')\,P(k) \tag{3.48}$$

It remains only to specify the form of the power spectrum $P(k)$ for these initial perturbations. These are usually taken to have the power-law form

$$P(k) = Ak^n \tag{3.49}$$

for constants A and n. The exponent n is called the *spectral index*.

A power-law $P \sim k^n$ gives rise to a real space correlation function $\xi(r) \sim 1/r^{n+3}$. (Actually, one must work a little harder to make sense of the inverse Fourier transform (3.45) at high k, or small r.) The choice $n = 0$ is what we would get if we sprinkle points at random in space; it is sometimes referred to as *white noise*. (We'll build more intuition for this in Section 3.2.2 below.) Meanwhile, any $n < -3$ means that $\xi(r) \to \infty$ as $r \to \infty$, so the universe gets more inhomogeneous at large scales, in contradiction to the cosmological principle. We'd like to ask: what choice of spectral index n describes our universe?

The Harrison-Zel'dovich Spectrum

A particularly special choice for the initial power spectrum is

$$n = 1$$

This is known as the *Harrison-Zel'dovich* power spectrum (named after Harrison, Zel'dovich, and Peebles and Yu). It is special for two reason. First, and most importantly, it turns out to be almost (but not quite!) the initial spectrum of density perturbations in our universe. Second, it also has a special mathematical property.

To explain this mathematical property, we need some new definitions. We start by some simple dimensional analysis. The original perturbation $\delta(\mathbf{x}) = \delta\rho/\rho$ was dimensionless, so after a Fourier transform (3.43) the perturbation $\delta(\mathbf{k})$ has dimension [length]^3. The delta-function $\delta_D^3(\mathbf{k})$ also has dimension $[k]^{-3} = [\text{length}]^3$ which means that the power spectrum $P(k)$ also has dimension [length]^3. It is often useful to define the dimensionless power spectrum

$$\Delta(k) = \frac{4\pi k^3\, P(k)}{(2\pi)^3} \tag{3.50}$$

The factors of 2 and π are conventional. Because $\Delta(k)$ is dimensionless, it makes sense to say that, for example, $\Delta(k)$ is a constant. Unfortunately, as you can see, this does not give rise to the Harrison-Zel'dovich spectrum.

However, we can also look at fluctuations in other quantities. In particular, rather than talk about perturbations in the density, ρ, we could instead talk about perturbations in the gravitational potential: $\Phi(\mathbf{x}) = \bar{\Phi}(\mathbf{x}) + \delta\Phi(\mathbf{x})$. The two are related by the Poisson equation (3.32)

$$\nabla^2\delta\Phi = \frac{4\pi G}{c^2}(1+3w)\bar{\rho}a^2\delta \quad \Rightarrow \quad -k^2\delta\Phi(\mathbf{k}) = \frac{4\pi G}{c^2}(1+3w)\bar{\rho}a^2\delta(\mathbf{k}) \tag{3.51}$$

We can then construct the power spectrum of gravitational perturbations

$$\langle \delta\Phi(\mathbf{k})\, \delta\Phi(\mathbf{k}')\rangle = (2\pi)^3\, \delta_D^3(\mathbf{k}+\mathbf{k}')\, P_\Phi(k) \tag{3.52}$$

and the corresponding dimensionless gravitational power spectrum

$$\Delta_\Phi = \frac{4\pi k^3\, P_\Phi(k)}{(2\pi)^3}$$

The Poisson equation (3.51) tells us that there's a simple relationship between $P_\Phi(k)$ and $P(k)$, namely

$$P_\Phi(k) \propto k^{-4} P(k) \tag{3.53}$$

where the proportionality factor hides the various constants arising from the Poisson equation. We can write this as

$$P(k) \propto k^4 P_\Phi(k) \propto k\Delta_\Phi$$

We see that the Harrison-Zel'dovich spectrum arises if the initial gravitational perturbations are independent of the wavelength, in the sense that Δ_Φ = constant. Such fluctuations are said to be *scale invariant*. We will see that such scale invariant perturbations in the gravitational potential are a good description of our universe, and hold an important clue to what was happening at the very earliest times. We will see what this clue is telling us in Section 3.5. First, however, it will be useful to pause to build some intuition for these different probability distributions.

3.2.2 Building Intuition For Gaussian Distributions

The discussion above can be bafflingly formal when you first meet it. At this stage, it's useful to build some intuition for what the different power spectra look like and, in particular, why $P_\Phi \sim 1/k^3$ corresponds to a scale invariant distribution.

To visualise what's going on, we'll ultimately show some pictures of distributions in $d = 2$ spatial dimensions. But, for now, let's keep the spatial dimension d arbitrary. We'll focus on the probability distribution of some scalar field $\Phi(\mathbf{x})$ which, in the cosmological context, you should think of as the gravitational perturbation $\delta\Phi$. However, for the purposes of our discussion, $\Phi(\mathbf{x})$ could be any scalar field. The Fourier transform is

$$\Phi(\mathbf{k}) = \int d^d x\; e^{i\mathbf{k}\cdot\mathbf{x}} \Phi(\mathbf{x})$$

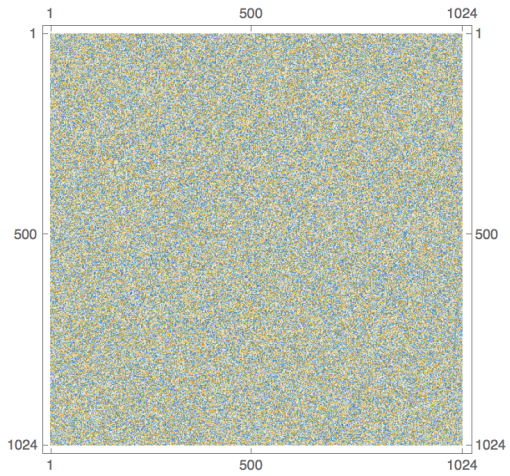

Figure 33: The white noise distribution in $d = 2$ dimensions with $n = 4$.

and we will ask that this takes values drawn from a Gaussian probability distribution of the form (3.52),

$$\langle \Phi(\mathbf{k})\Phi(\mathbf{k}')\rangle = (2\pi)^d\, \delta^d(\mathbf{k}+\mathbf{k}')\, P_\Phi(k)$$

where $\delta^d(\mathbf{k}-\mathbf{k}')$ is the usual d-dimensional delta function. The question that we'd like to ask is: what does such a distribution mean for $\Phi(\mathbf{x})$ and, in particular, how does the choice of power spectrum $P_\Phi(k)$ affect it?

In position space, the two-point correlation function is given by the Fourier transform of the power spectrum,

$$\begin{aligned}\langle \Phi(\mathbf{x})\Phi(\mathbf{y})\rangle &= \int \frac{d^dk}{(2\pi)^d}\int \frac{d^dk'}{(2\pi)^d} e^{-i\mathbf{k}\cdot\mathbf{x}-i\mathbf{k}'\cdot\mathbf{y}}\langle \Phi(\mathbf{k})\Phi(\mathbf{k}')\rangle \\ &= \int \frac{d^dk}{(2\pi)^d} e^{-i\mathbf{k}\cdot(\mathbf{x}-\mathbf{y})} P_\Phi(k) \end{aligned} \tag{3.54}$$

We'll now look at what this means for a power spectrum of the form

$$P_\Phi(k) = k^{n-4} \tag{3.55}$$

for various choices of integer n. (The exponent here is chosen to match our previous conventions.)

Obviously, in cosmology we're interested in $d = 3$ spatial dimensions. However, below we'll plot distributions in $d = 2$ dimensions. The key physics is the same but, as we'll

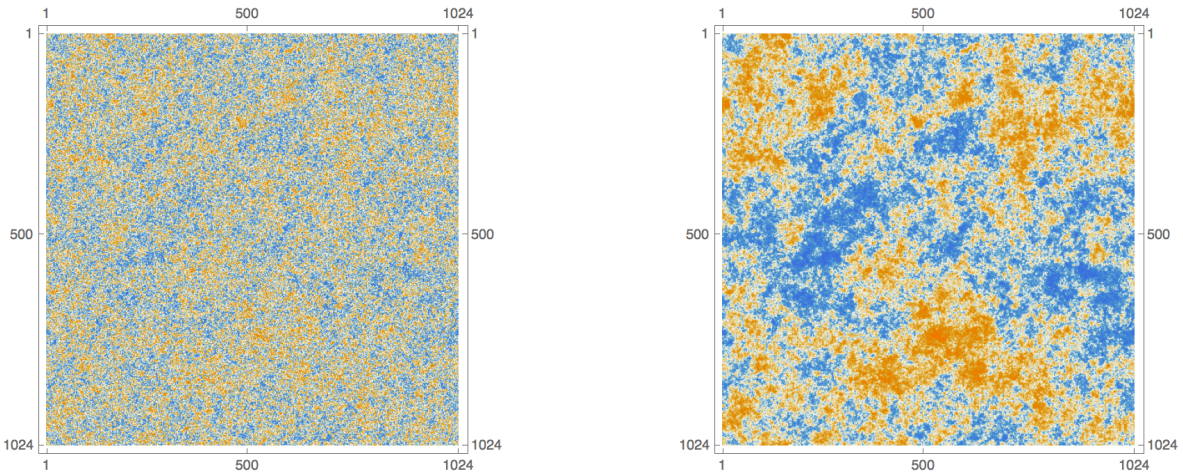

Figure 34: The distribution in $d = 2$ dimensions with $n = 3$, corresponding to $P_\Phi \sim 1/k$ (on the left) and $n = 2$, corresponding to $P_\Phi \sim 1/k^2$ (on the right). The latter is scale invariant in two dimensions.

see, occurs for different values of n. We start with constant power spectrum, or $n = 4$ in the convention of (3.55). Here we have

$$n = 4 \quad \Rightarrow \quad \langle \Phi(\mathbf{x})\Phi(\mathbf{y})\rangle \sim \delta^d(\mathbf{x}-\mathbf{y})$$

This means that there's no correlation between the value of Φ at different points. A typical configuration of $\Phi(\mathbf{x})$ is shown[10] in Figure 33. A distribution like this, with no correlation between neighbouring points, is known as *white noise*. (There's a perennial confusion here: white noise for Φ and white noise for the density perturbation occur for different values of n because the two distributions are related by a power of k^4.)

Now we look at what happens as we decrease n. For $n = 3$, corresponding to $P_\Phi(k) \sim k^{-1}$, the correlation between neighbouring points becomes stronger. A typical distribution is shown on the left in Figure 34. We see that if the field takes a particular value at some point $\mathbf{x}$, there is now an increased likelihood that it takes similar values at neighbouring points.

This likelihood increases further as we lower n. The distribution for $n = 2$ is shown on the right in Figure 34. This distribution is rather special since it gives $P_\Phi(k) \sim 1/k^2$

[10] All the images of distributions were created using Garrett Goon's publicly available mathematica script. Operationally, this starts with the white noise of Figure 33, Fourier transforms to momentum space, multiplies the resulting distribution by $P(k)$, and then Fourier transforms back. A clear and detailed account of this can be found on Garrett's webpage https://garrettgoon.com/gaussian-fields/.

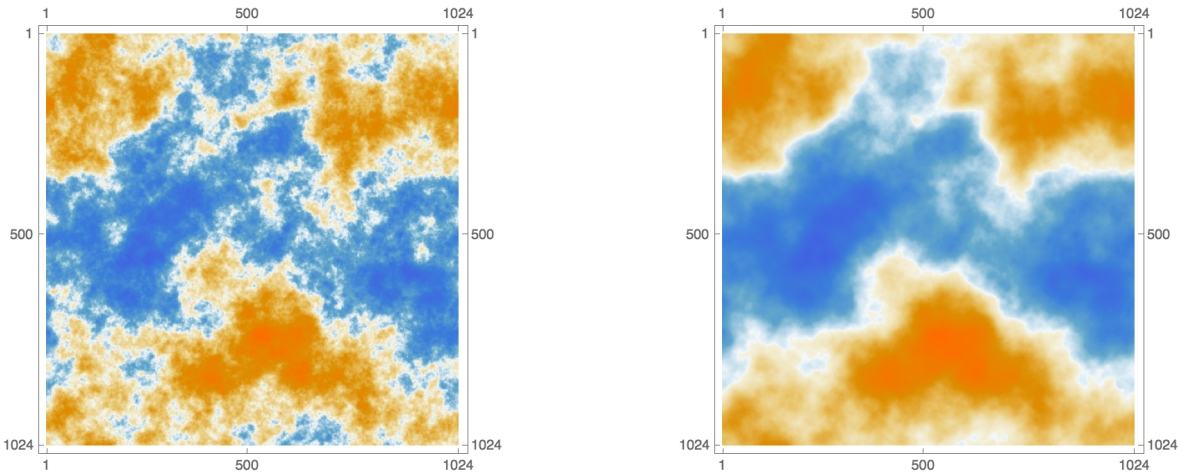

Figure 35: The distribution in two dimensions with $n = 1$ corresponding to $P_\Phi(k) = 1/k^3$ (on the left) and $n = 0$ corresponding to $P_\Phi(k) = 1/k^4$ (on the right).

and, in d spatial dimensions, the distribution $1/k^d$ is scale invariant. This means that the correlation between any two points is independent of the distance between those points! To see this, we simply need to rescale the correlation function (3.54) to find

$$\langle \Phi(\lambda \mathbf{x})\Phi(\lambda \mathbf{y})\rangle = \int \frac{d^d\mathbf{k}}{(2\pi)^d}\frac{e^{-i\lambda \mathbf{k}\cdot(\mathbf{x}-\mathbf{y})}}{k^d} = \langle \Phi(\mathbf{x})\Phi(\mathbf{y})\rangle$$

where the final equality holds by redefining $\mathbf{k}' = \lambda\mathbf{k}$ to remove λ from the exponent, and then noting that the factors of λ cancel between the measure factor d^dk and the $1/k^d$ in the power spectrum.

We can decrease n still further, to find configurations in which the spatial correlation increases. Examples for $n = 1$ and $n = 0$ are shown in Figure 35.

3.2.3 The Power Spectrum Today

The Gaussian distribution (3.47) holds at some initial time t_i, which we take to be a very early time, typically just after inflation. As we have seen, the subsequent evolution of the density perturbations is described by the transfer function

$$\delta(\mathbf{k}, t_0) = T(k)\,\delta(\mathbf{k}, t_i)$$

We computed this for non-relativistic matter in (3.41); it is

$$T(k) \sim \text{constant} \times \begin{cases} 1 & k < k_{\rm eq} \\ k^{-2} & k > k_{\rm eq} \end{cases}$$

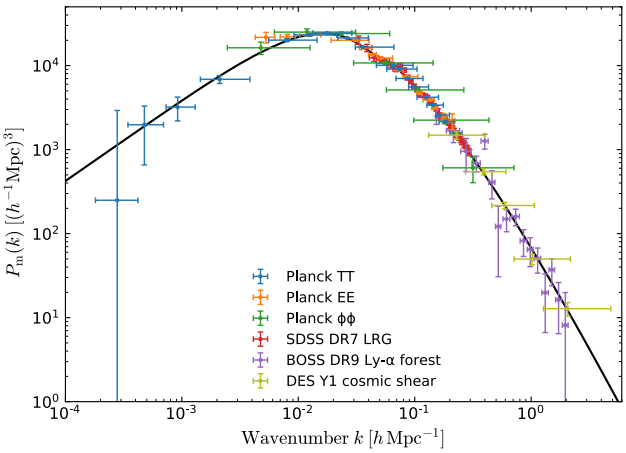

Figure 36: The observed matter power spectrum.

In general, each fluid component will have a separate transfer function, so that the adiabatic form of the initial perturbations (3.46) gets ruined as the universe evolves.

Provided that this linear analysis is valid, the distribution of fluctuations remains Gaussian, and only the power spectrum $P(k)$ changes. From the relation $P \sim \langle \delta\delta \rangle$, we have

$$P(k; t_0) = T^2(k)\, P(k; t_i)$$

As the density perturbations get large, linear perturbation theory breaks down and the evolution becomes non-linear. In this situation, perturbations with different wavevector $\mathbf{k}$ start to interact and the simple Gaussian distribution no longer holds. If we want to get a good handle on the late time universe, filled with galaxies and clusters, we must ultimately understand this non-linear behaviour. We'll start to explore this in Section 3.3 but, for now, we will content ourselves with the simple linear evolution.

If we start with the power-law spectrum $P \sim k^n$, then it subsequently evolves to

$$P(k) \sim \begin{cases} k^n & k < k_{\rm eq} \\ k^{n-4} & k > k_{\rm eq} \end{cases} \tag{3.56}$$

with the turnover near $ak \approx ak_{\rm eq} \sim 0.01$ Mpc^{-1}. A more careful analysis shows that the turnover at $k = k_{\rm eq}$ happens rather gradually.

We can now compare these expectations with the observed matter power spectrum. Data taken from a number of different sources, is shown[11] in Figure 36. At very large scales (small k) the data is taken from the CMB; we will discuss this further in Section 3.4. Longer wavelength structures are seen through various methods of measuring of structure in the universe today. One finds that the data fits very well with the initial Harrison-Zel'dovich power-law spectrum $n = 1$. More accurate observations reveal, a slight deviation from the perfect Harrison-Zel'dovich spectrum. Both large scale structure[12], and CMB measurements (which are discussed briefly in the next section) give

$$n \approx 0.97$$

The fact that perturbations in the early universe are almost, but not quite, described by the Harrison-Zel'dovich spectrum is an important clue for what was happening in the very early universe. A precise scale invariant Harrison-Zel'dovich spectrum is telling us that there must have been some symmetry in the early universe; the deviation is telling us that there was some dynamics taking place which breaks this symmetry. We will describe this more in Section 3.5.

3.2.4 Baryonic Acoustic Oscillations

There is a time in the early universe, bounded by redshifts

$$1100 \lesssim z \lesssim 3400$$

when the expansion was dominated by matter, but hydrogen had not yet formed. As we saw in Section 2, in this epoch protons, electrons and photons were in thermal equilibrium. In such a photon-baryon fluid, the speed of sound is determined by the photons rather than the matter, so $c_s \approx c/\sqrt{3}$. This means that the effective Jeans length for baryonic matter is much greater than the corresponding length for dark matter.

The consequence is that dark matter and baryonic matter behave differently in this epoch. Density perturbations in dark matter, which long ago decoupled from the photons, start to grow as $\delta \sim a$ as in (3.31). Meanwhile, density perturbations in

[11] This plot is taken from the Planck 2018 results, "*Overview and the cosmological legacy of Planck*, arXiv:1807.06205.

[12] For example, the paper *The one-dimensional Ly-alpha forest power spectrum from BOSS* by N. Palanque-Delabrouille et. al. arXiv:1306.5896 finds $n = 0.97 \pm 0.02$. Meanwhile, the Planck collaboration, in arXiv:1807.06211, quotes $n_s = 0.9649 \pm 0.0042$.

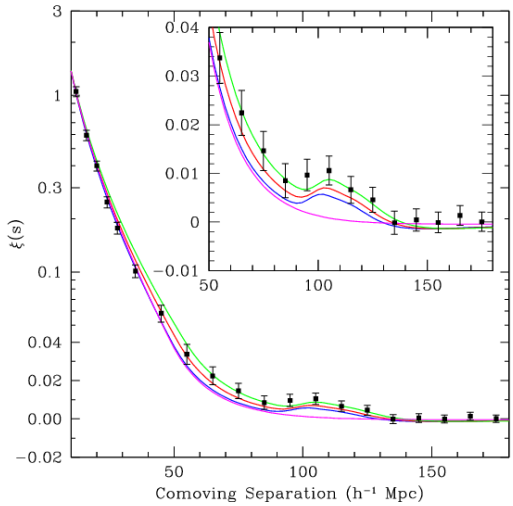

Figure 37: Baryonic acoustic oscillations seen in the distribution of galaxies.

baryonic matter are supported by the pressure from the photons and, at least on sub-horizon scale, oscillate. The resulting sound waves in the baryon-photon fluid are known as *baryonic acoustic oscillations*.

There are two important consequences of this. The immediate consequence is that dark matter has a head start in structure formation, with density perturbations starting to grow at $z \approx 3400$. By the time the baryons decouple at $z = 1100$, there are already well-established gravitational wells which act as seeds, expediting the formation of the baryonic structures that we call galaxies.

The second consequence is more subtle. At recombination, the photons stream away from the sound waves they have helped create. But the baryons are frozen in place, a remnant of this earlier time. The sound waves contain regions in which the baryons are more compressed, and regions in which they are more rarified, with the wavelength determined by the horizon at decoupling,

$$d_H \sim \frac{cH_0^{-1}}{(1+z)^{3/2}} \approx 0.1 \text{ Mpc}$$

using $cH_0^{-1} \approx 4 \times 10^3$ Mpc and $z \approx 1100$. In the subsequent evolution of the universe, these waves were stretched by a factor of $z \approx 1100$, leaving a faint imprint on the clustering of matter seen today, where there is an excess in galaxies separated by a distance

~ 150 Mpc. The effects of these baryonic acoustic oscillations in the distribution of galaxies was first observed in 2005; the correlation function is shown in Figure 37[13].

3.2.5 Window Functions and Mass Distribution

In this section, we've understood some of the mathematical properties of $\delta(\mathbf{x}, t)$. But, so far, we've not actually discussed how one might go about measuring such an object. And, as we now explain, there is a small subtlety.

Recall that $\delta(\mathbf{x})$ is a density contrast. But a density is, of course, energy per unit volume. Mathematically, there is no difficulty in defining the density at a point $\mathbf{x}$? But how do we construct $\delta(\mathbf{x})$ from observations? In particular, what volume do we divide by?!

At heart, this comes back to our initial discussion of the cosmological principle. If we observe many galaxies, each localised at some point $\mathbf{X}_i$, then the universe looks far from homogeneous. The same is true for any fluid if we look closely enough. But our interest is in a more coarse-grained description.

To this end, we introduce a *window function* which we denote as $W(\mathbf{x}; R)$. The purpose of this function is to provide a way to turn the observed density $\delta(\mathbf{x})$ into something that is smooth, and varies on length scales $\sim R$. We construct the smoothed density contrast as

$$\delta(\mathbf{x}; R) = \int d^3x' \; W(\mathbf{x} - \mathbf{x}'; R)\, \delta(\mathbf{x}') \tag{3.57}$$

In Fourier space, we have

$$\begin{aligned}
\delta(\mathbf{k}; R) &= \int d^3x \; e^{i\mathbf{k}\cdot\mathbf{x}}\, \delta(\mathbf{x}) \\
&= \int d^3x\, d^3x' \; e^{i\mathbf{k}\cdot\mathbf{x}}\, W(\mathbf{x} - \mathbf{x}'; R)\, \delta(\mathbf{x}') \\
&= \int d^3x\, d^3x' \; e^{i\mathbf{k}\cdot(\mathbf{x}-\mathbf{x}')}\, W(\mathbf{x} - \mathbf{x}'; R)\, e^{i\mathbf{k}\cdot\mathbf{x}'} \delta(\mathbf{x}') \\
&= \int d^3y\, d^3x' \; e^{i\mathbf{k}\cdot\mathbf{y}}\, W(\mathbf{y}; R)\, e^{i\mathbf{k}\cdot\mathbf{x}'} \delta(\mathbf{x}') \\
&= \tilde{W}(\mathbf{k}; R)\, \delta(\mathbf{k})
\end{aligned}$$

This is the statement that a convolution integral, like (3.57), in real space becomes a product in Fourier space.

[13]This data is taken from D. J. Eisenstein *et al.* [SDSS Collaboration], "*Detection of the Baryon Acoustic Peak in the Large-Scale Correlation Function of SDSS Luminous Red Galaxies,*" Astrophys. J. **633**, 560 (2005), astro-ph/0501171.

There is no canonical choice of window function. But there are sensible choices. These include:

- The Spherical Top Hat. This is a sharp cut-off in real space, given by

$$W(\mathbf{x}, R) = \frac{1}{V} \times \begin{cases} 1 & |\mathbf{x}| \leq R \\ 0 & |\mathbf{x}| > R \end{cases} \quad \text{with } V = \frac{4\pi}{3} R^3$$

  In Fourier space, this becomes

$$\tilde{W}(\mathbf{k}; R) = \frac{3}{(kR)^3} \Big[\sin kR - kr \cos kR \Big] \tag{3.58}$$

  Note that the Fourier transform $\tilde{W}(\mathbf{k}; R) = \tilde{W}(kR)$; this will be true of all our window functions.

- The Sharp k Filter: This is a sharp cut-off in momentum space

$$\tilde{W}(kR) = \begin{cases} 1 & kR \leq 1 \\ 0 & kR > 1 \end{cases} \tag{3.59}$$

  It looks more complicated in real space,

$$W(\mathbf{x}; R) = \frac{1}{2\pi^2 r^3} \left[\sin(r/R) - \frac{r}{R} \cos(r/R) \right]$$

- The Gaussian: This provides a smooth cut-off in both position and momentum space,

$$W(\mathbf{x}; R) = \frac{1}{(2\pi)^{3/2} R^3} \exp\left(\frac{-r^2}{2R^2} \right)$$

  which, in Fourier space, retains its Gaussian form

$$\tilde{W}(kR) = \exp\left(-\frac{k^2 R^2}{2} \right)$$

Note that, in each case, $\tilde{W}(kR = 0) = 1$. Different window functions may be better suited to different measurements or calculations. We now provide an example.

The Mass Distribution

We now use the window function technology to address a simple question: what is the distribution of masses contained within a sphere of radius R?

For each of the window functions, we can define the average mass $M(R)$ inside a sphere of radius R. It is

$$\bar{M}(R) = \frac{1}{c^2}\int d^3x\ W(\mathbf{x};R)\bar{\rho}(\mathbf{x})$$

where $\bar{\rho}(\mathbf{x})$ is the average density in the universe. Since $\bar{\rho}$ is constant, this is

$$\bar{M}(R) = \frac{4\pi R^3\bar{\rho}}{3}\gamma \tag{3.60}$$

where γ can be found by integrating each of the three window functions. Three short calculations show

$$\gamma = \begin{cases} 1 & \text{Top Hat} \\ 9\pi/2 & k-\text{Filter} \\ 3\sqrt{\pi/2} & \text{Gaussian} \end{cases}$$

Next, we want to look at deviations from the average. The smoothed mass distribution is related to the smoothed density contrast by

$$M(\mathbf{x};R) = \bar{M}(R)(1+\delta(\mathbf{x};R))$$

So we can also interpret the smoothed density contrast as

$$\delta(\mathbf{x};R) = \frac{\delta M(\mathbf{x};R)}{\bar{M}(R)}$$

where $\delta M(\mathbf{x};R) = M(\mathbf{x};R) - \bar{M}(R)$. The variance in the mass distribution is then

$$\sigma^2(M) = \langle\delta^2(\mathbf{x};R)\rangle$$

This depends on both the choice of window function and, more importantly, on the scale R at which we do the smoothing. Using our definition (3.57), this is

$$\sigma^2(M) = \int d^3x'\,d^3x''\ W(\mathbf{x}-\mathbf{x}';R)W(\mathbf{x}-\mathbf{x}'';R)\,\langle\delta(\mathbf{x}')\delta(\mathbf{x}'')\rangle \tag{3.61}$$

We introduced the two-point correlation function in (3.42),

$$\xi(r) = \langle\delta(\mathbf{x}+\mathbf{r})\,\delta(\mathbf{x})\rangle = \int\frac{d^3k}{(2\pi)^3}\ e^{-i\mathbf{k}\cdot\mathbf{r}}\,P(k)$$

where, following Section 3.2, we've written this in terms of the power spectrum $P(k)$. We then have

$$\sigma^2(M) = \int\frac{d^3k}{(2\pi)^3}\int d^3x'\,d^3x''\,W(\mathbf{x}-\mathbf{x}';R)W(\mathbf{x}-\mathbf{x}'';R)\,e^{-i\mathbf{k}\cdot(\mathbf{x}'-\mathbf{x}'')}\,P(k)$$

But the integrations over spatial coordinates now conspire to turn the window functions into their Fourier transform. We're left with

$$\sigma^2(M) = \int \frac{d^3k}{(2\pi)^3}\, \tilde{W}^2(kR)\, P(k) = \frac{1}{2\pi^2}\int dk\; k^2\, \tilde{W}(kR)\, P(k)$$

Note that, as we smooth on smaller scales, so $kR \to 0$, we have $\tilde{W}(kR) \to 1$ and, correspondingly, $\sigma^2(R) \to \sigma^2$. This is what we would wish for a variance $\sigma^2(R)$ which is smoothed on scales R.

Now recall the power spectrum from (3.56),

$$P(k) \sim \begin{cases} k^n & k < k_{\rm eq} \\ k^{n-4} & k > k_{\rm eq} \end{cases}$$

where observations of galaxy distributions give $n \approx 0.97$. At this point, it is simplest to use the sharp k-filter window function (3.59). At the largest scales, where $P(k) \sim k^n$, we then have

$$\sigma^2(M) \sim \int_0^{1/R} dk\; k^{2+n} \sim \frac{1}{R^{3+n}} \sim \frac{1}{M^{(n+3)/3}}$$

where, in the final scaling, we've used (3.60). If we have $n < -3$, we would have increasingly large mass fluctuations on large scales. This would violate our initial assumption of the cosmological principle. Fortunately, we don't live in such a universe.

Meanwhile, on shorter scales we have $P(k) \sim k^{n-4}$. Here we have

$$\sigma^2(M) \sim \int_0^{1/R} dk\; k^{n-2} \sim \frac{1}{R^{n-1}} \sim \frac{1}{M^{(n-1)/3}}$$

For $n = 1$, this becomes logarithmic scaling.

What Cosmologists Measure

As a final aside: observational cosmologists quote the fundamental parameter

$$\sigma_8^2 := \frac{1}{2\pi^2}\int dk\; k^2\, \tilde{W}^2(kR)\, P(k) \tag{3.62}$$

Here $P(k)$ is the evolved linear power spectrum that we described in Section 3.2. Meanwhile, the window function $\tilde{W}(kR)$ is taken to be the top hat (3.58), evaluated at the scale $R = 8h^{-1}$ Mpc where galactic clusters are particularly rich. (Here $h \approx 0.7$ characterises the Hubble parameter, as defined in (1.16).) Until now, we've mostly focussed on the k-dependence of $P(k)$. The variable σ_8 characterises its overall magnitude. Larger values of σ_8 imply more fluctuations, and so structure formation started earlier. For what it's worth, the current measured value is $\sigma_8 \approx 0.8$.

3.3 Nonlinear Perturbations

So far, we have relied on perturbation theory to describe the growth of density fluctuations, working with the linearised equations. But this is only tenable when the fluctuations are small. As they grow to size $\delta\rho \approx \bar{\rho}$, or $\delta \approx 1$, perturbation theory breaks down. At this point, we must solve the full coupled equations in an expanding FRW universe. This is difficult.

There are a number of ways to proceed. At some point, we simply have to resort to difficult and challenging numerical simulations. However, there is a rather simple toy model which captures some of the relevant physics.

3.3.1 Spherical Collapse

For convenience, we will work with an the Einstein-de Sitter universe, filled only with dust, so $\Omega_m = 1$. This means that the average density is equal to the critical density, $\bar{\rho}(t) = \rho_{\rm crit}(t)$.

At some time t_i, when the average density is $\bar{\rho}_i$, we create a density perturbation. To do this, consider a spherical region of radius R_i, centred about some point which we take to be the origin. Take the matter within this region and compress it into a smaller spherical region of radius $r_i < R_i$, with constant density

$$\rho_i = \bar{\rho}_i(1 + \delta_i)$$

We will initially take δ_i to be small but, in contrast to previous sections, we won't assume that it remains small for all time. Instead, we will follow its evolution as it grows.

Between r_i and R_i, there is then a gap with no matter. The mass contained in the spherical region $r < r_i$ is

$$M_i c^2 = \frac{4\pi}{3} R_i^3 \bar{\rho}_i = \frac{4\pi}{3} r_i^3 \rho_i = \frac{4\pi}{3} r_i^3 \bar{\rho}_i (1 + \delta_i)$$

Furthermore, the total mass in the perturbation remains constant at M_i, even as all the other variables, $\bar{\rho}$, δ and the edge of the over-dense region r evolve in time.

We would like to understand how this density perturbation evolves. To do this, we can revert to the simple Newtonian argument that we used in Section 1.2.3 when first deriving the Friedmann equation. Recall that, for a spherically symmetric distribution of masses, the gravitational potential at some point r depends only on the mass contained inside r and does not depend at all on the mass outside. Consider a particle

at some radius r, either inside or outside the over-dense region. The conservation of energy for this particle reads

$$\frac{1}{2}\dot{r}^2 - \frac{GM(r)}{r} = E \tag{3.63}$$

where $M(r)$ is the mass contained within the radius r and is constant: by mass conservation $M(r)$ doesn't change as r evolves. Meanwhile E is also a constant (and is identified with energy divided by the mass of a single particle).

We can now apply this formula to particles both inside and outside the over-dense region. First we look at the particles outside, with $r(t_i) \geq R_i$. For these particles, the mass $M(r)$ is the same as it was before we perturbed the distribution, so they carry on as before. But our starting point was an Einstein-de Sitter universe with critical energy density, which corresponds to $E = 0$. Integrating (3.63) gives

$$r(t) = \left(\frac{9GM(r)}{2}\right)^{1/3} t^{2/3} \qquad \text{if } r(t_i) > R_i \tag{3.64}$$

with $M(r)$ constant. This is the usual expansion of a flat, matter dominated universe. The average energy density is

$$\bar{\rho}(t) = \frac{M(r)c^2}{(4\pi/3)r^3(t)} = \frac{c^2}{6\pi G}\frac{1}{t^2} \tag{3.65}$$

which reproduces the usual time evolution of the critical energy density (1.51).

In contrast, inside the over-dense region (i.e when $r(t_i) \leq r_i$), we have $E < 0$. This means that the over-dense region acts like a universe with positive curvature (i.e. $k = +1$). The inner sphere will then behave like the closed universe we met in Section 1.3.2: it first continues to expand, before slowing and subsequently collapsing back in on itself.

We presented the solution for a closed universe in parametric form in (1.57) and (1.58); you can check that the following expressions satisfy (3.63)

$$\begin{aligned} r(\tilde{\tau}) &= A\,(1 - \cos\tilde{\tau}) \\ t(\tilde{\tau}) &= B\,(\tilde{\tau} - \sin\tilde{\tau}) \end{aligned} \tag{3.66}$$

where the constants are

$$A = \frac{GM}{2|E|} \quad \text{and} \quad B = \frac{GM}{(2|E|)^{3/2}} \quad \Rightarrow \quad A^3 = GMB^2 \tag{3.67}$$

We can apply the solution (3.66) to the edge of the over-dense region, i.e. the point with $r(t_i) = r_i$. We see that the spatial extent of the perturbation continues to grow for some time, swept along by the expansion of the universe. At early times $\tilde{\tau} \ll 1$, we can linearise the solution to find

$$r(\tilde{\tau}) \approx \frac{1}{2}A\tilde{\tau}^2 \quad \text{and} \quad t(\tilde{\tau}) \approx \frac{1}{6}B\tilde{\tau}^3 \quad \Rightarrow \quad r(t) \approx \frac{A}{2}\left(\frac{6}{B}\right)^{2/3} t^{2/3} \tag{3.68}$$

Thus, initially, the growth of the over-dense region has the same time dependence as the region outside the shell (3.64).

However, the excess mass in the over-dense region causes the expansion to slow. From (3.66), we see that the expansion halts and then starts to collapse again at time $\tilde{\tau}_{\rm turn} = \pi$. This is the turn-around time.

Taken at face value, the solution (3.66) then collapses back to a point at the time $\tilde{\tau}_{\rm col} = 2\pi$. We will discuss what really happens here in Section 3.3.2.

The Density in Spherical Collapse

From the solution (3.66), it is straightforward to figure out how the density evolves. At a given time, the density of the over-dense region is

$$\rho(\tilde{\tau}) = \frac{M_i c^2}{(4\pi/3)r^3} = \frac{3M_i c^2}{4\pi A^3}\frac{1}{(1-\cos\tilde{\tau})^3}$$

Meanwhile, the critical density evolves as (3.65)

$$\bar{\rho}(\tilde{\tau}) = \frac{c^2}{6\pi G}\frac{1}{t^2} = \frac{c^2}{6\pi GB^2}\frac{1}{(\tilde{\tau}-\sin\tilde{\tau})^2}$$

The density contrast $\delta = \delta\rho/\bar{\rho}$ can be computed from the ratio of the two,

$$(1+\delta) = \frac{\rho}{\bar{\rho}} = \frac{9}{2}\frac{(\tilde{\tau}-\sin\tilde{\tau})^2}{(1-\cos\tilde{\tau})^3} \tag{3.69}$$

where we've used the fact that $A^3 = GMB^2$.

Again, we can see what happens at early times. We Taylor expand each of the terms, but this time we need to go to second order: $\tilde{\tau} - \sin\tau \approx \frac{1}{3!}\tilde{\tau}^3 - \frac{1}{5!}\tilde{\tau}^5$ and $1 - \cos\tilde{\tau} \approx \frac{1}{2}\tilde{\tau}^2 - \frac{1}{4!}\tilde{\tau}^4$. This gives

$$1 + \delta_{\rm lin}(\tilde{\tau}) \approx \frac{(1-\frac{1}{20}\tilde{\tau}^2)^2}{(1-\frac{1}{12}\tilde{\tau}^2)^3} \approx 1 + \frac{3}{20}\tilde{\tau}^2 \tag{3.70}$$

But, from (3.68), we can write this as

$$\delta_{\rm lin}(t) = \frac{3}{20}\left(\frac{6}{B}\right)^{2/3} t^{2/3} \tag{3.71}$$

Happily, this coincides with the $t^{2/3}$ time dependence that we found in (3.31) when discussing linear perturbation theory.

When we reach turn-around, at $\tilde{\tau} = \pi$, the density is

$$\delta(\tilde{\tau}_{\rm turn}) = \frac{9\pi^2}{16} - 1 \approx 4.55$$

For what follows, it will prove useful to ask the following, slightly artificial question: what would the density contrast be at turn-around if we were to extrapolate the linear solution? From (3.66), we have $t_{\rm turn} = B\pi$, so we can write the linear solution (3.71) as

$$\delta_{\rm lin}(t) = \frac{3}{20}(6\pi)^{2/3}\left(\frac{t}{t_{\rm turn}}\right)^{2/3} \quad \Rightarrow \quad \delta_{\rm lin}(t_{\rm turn}) = \frac{3}{20}(6\pi)^{2/3} \approx 1.06$$

Meanwhile, when the perturbation has completely collapsed at $\tilde{\tau}_{\rm col} = 2\pi$, the true density is

$$\delta(\tilde{\tau}_{\rm col}) = \infty$$

and we'll see how to interpret this shortly. We can again ask the artificial question: what would the density contrast be at collapse if we were to extrapolate the linear solution. This time, from (3.66), we have $t_{\rm col} = 2B\pi = 2t_{\rm turn}$, so

$$\delta_{\rm lin}(t_{\rm col}) = \frac{3}{20}(12\pi)^{2/3} \approx 1.69$$

A simplistic interpretation of this result is as follows: if we work within linear perturbation theory, and the density contrast reaches $\delta_{\rm lin} \approx 1.69$, then we should interpret this as a complete collapse.

3.3.2 Virialisation and Dark Matter Halos

As we have seen, the simple spherical collapse model predicts that an initial over-density will ultimately collapse down to a point with infinite density. The interpretation of such a singularity is a black hole.

Yet our universe is not dominated by black holes. This is because the assumption of spherical collapse is not particularly realistic, and while this is not too much of a problem for much of the discussion, it becomes important as the end point nears. Here, the random motion of the matter, together with interactions, means that the matter will ultimately settle down into an equilibrium configuration with the kinetic energy balanced by the potential energy. The end result is a *dark matter halo*, an extended region of dark matter in which galaxies are embedded.

This process in which equilibrium is reached is known, rather wonderfully, as *violent relaxation*. Or, less evocatively, as *virialisation*. This latter name reflects the fact that by the time the system has settled down, it obeys the virial theorem, with the average kinetic energy T related to the average potential energy V by

$$\bar{T} = -\frac{1}{2}\bar{V}$$

We proved this theorem in Section 1.4.3.

Let's now apply this to our collapse model. Our original formula (3.63) is conveniently written in terms of the kinetic energy $T = \frac{1}{2}\dot{r}^2$ and the potential energy $V = -GM/r$. We can start by considering the turn-around point, where the kinetic energy vanishes, $T = 0$, and

$$V_{\rm turn} = -\frac{GM}{r_{\rm turn}}$$

The total energy $E = T + V$ is conserved. This means that after virialisation, when $T = -\frac{1}{2}V$, we must have

$$T_{\rm vir} + V_{\rm vir} = \frac{1}{2}V_{\rm vir} = V_{\rm turn} \quad \Rightarrow \quad \begin{cases} r_{\rm vir} = \frac{1}{2}r_{\rm turn} \\ \rho_{\rm vir} = 8\rho_{\rm turn} \end{cases}$$

Our real interest is in the density contrast, $1 + \delta_{\rm vir} = \rho_{\rm vir}/\bar{\rho}_{\rm vir}$. We take the virialisation time to coincide with the collapse time, $t_{\rm vir} = t_{\rm col} = 2t_{\rm turn}$. Since the universe scales as $a \sim t^{2/3}$, the critical energy has diluted by a factor of 4 between turn-around and virialisation, so $\bar{\rho}_{\rm vir} = \bar{\rho}_{\rm turn}/4$. Putting this together, we have

$$\delta_{\rm vir} = \frac{\rho_{\rm vir}}{\bar{\rho}_{\rm vir}} - 1 = 32\,\frac{\rho_{\rm turn}}{\bar{\rho}_{\rm turn}} - 1$$

But from (3.69), using $\tau_{\rm turn} = \pi$, we have $\rho_{\rm turn}/\bar{\rho}_{\rm turn} = 9\pi^2/16$. The upshot is that the density contrast in a dark matter halo is expected to be

$$\delta_{\rm vir} = 18\pi^2 - 1 \approx 177$$

Once again referring to our linear model, we learn that whenever $\delta_{\rm lin} \gtrsim 1.69$, we may expect to form a dark matter halo whose density ρ is roughly 200 times greater than the background density $\bar{\rho}$.

3.3.3 Why the Universe Wouldn't be Home Without Dark Matter

We can try to put together some of the statements that we have seen so far to get a sense for when structures form.

The right way to do this is to use the window function that we introduced in Section 3.2.5, to define spatial variations smoothed on different scales R. The spatial variations are computed by integrating the power spectrum against the window function, as in (3.61). We can then trace the evolution of these spatial perturbations to see how they evolve.

Here, instead, we're going to do a quick and dirty calculation to get some sense of the time scale. Indeed, taken at face value, there seems to be a problem. The CMB tells us that $\delta T/T \sim 10^{-5}$ at redshift $z \approx 1000$. Yet we know that, in the matter dominated era, perturbations grow linearly with scale (3.31). This would naively suggest that, even today, we have only $\delta \sim 10^{-2}$ which, given our discussion above, is not enough for structures to form. What's going on?

In large part, this issue arises because we need to do a better job of defining the spatial variations. But there is also some important physics buried in this simple observation which we mentioned briefly before, but is worth highlighting. The CMB figure of $\delta T/T \sim 10^{-5}$ is telling us about the fluctuations in radiation and, through this, fluctuations in baryonic matter at recombination. This is not sufficient for galaxies to form. To get the universe we see today, it's necessary to have dark matter. Between $z \approx 3000$ and $z \approx 1000$, when the universe was matter dominated, perturbations in dark matter were growing while the baryon-photon fluid was sloshing back and forth. This can be further enhanced by the logarithmic growth (3.36) of dark matter perturbations during the radiation dominated era.

Even accounting for dark matter, it's not obvious, using our results above, that there is enough time for structures to form. Fortunately, there are a bunch of scrappy factors floating around which get us close to the right ballpark. For example, the fluctuations in matter density are related to those in temperature by $\delta_m \approx 3 \times \delta T/T$. (We will see this in (3.73).) Furthermore, we should focus on the peaks of the fluctuations rather than the average: these come in around $\delta T/T \approx 6 \times 10^{-5}$. The Sachs-Wolfe effect (which we will describe in Section 3.4 provides another small boost. All told, these

factors conspire to give $\delta_m \approx 10^{-3}$ at $z \approx 1000$. This tells us that we expect dark matter halos to form at redshift $z \approx 1$ which is roughly right.

However, an important take-home message is that the existence of dark matter, which is decoupled from the photon fluid and so starts to grow as soon as the universe is matter dominated, is crucial for structure to form on a viable time scale.

3.3.4 The Cosmological Constant Revisited

We can repeat the argument above in the presence of a cosmological constant. We saw in (1.60) that the cosmological constant changes the equation (3.63), describing the radial motion of a particle, to include a term that looks like an inverted harmonic oscillator

$$\frac{1}{2}\dot{r}^2 - \frac{GM(r)}{r} - \frac{1}{6}\Lambda r^2 = E \tag{3.72}$$

Let's now play our earlier game. We start with a universe comprising of both matter and a cosmological constant with critical density, so that $E = 0$.

Now we create an over-density by squeezing the sphere at $r = R_i$ to a smaller radius, $r = r_i$. Particles with $r(t_i) < r_i$ have negative energy $E < 0$. If, as previously, this over-dense region is to turn around and subsequently collapse then there must be a time when $\dot{r} = 0$ and $r(t)$ solves the cubic equation

$$\frac{1}{6}\Lambda r^3(t) - |E|r(t) + GM = 0$$

with M the constant mass contained in the over-dense region. We want to know if this equation has a solution with $r(t) > 0$?

To answer this, first note that the cubic has stationary points at $r = \pm\sqrt{2|E|/\Lambda}$. The cubic only has a root with $r > 0$ if the positive stationary point lies below the real axis, or

$$\frac{1}{6}\Lambda\left(\frac{2|E|}{\Lambda}\right)^{3/2} - |E|\left(\frac{2|E|}{\Lambda}\right)^{1/2} + GM < 0 \quad \Rightarrow \quad \Lambda^{1/2} < \frac{(2|E|)^{3/2}}{3GM}$$

We write this upper bound on Λ as

$$\Lambda^{1/2} < \frac{1}{3B}$$

where $B = GM/(2|E|)^{3/2}$ was defined previously in (3.67). We need to relate this constant B to the initial density perturbation. For this, note that if we make the

density perturbation at early times, then the cosmological constant is negligible and the universe evolves as if it is matter dominated. In this case, we can use our earlier result (3.71)

$$\delta(t) = \frac{3}{20}\left(\frac{6}{B}\right)^{2/3} t^{2/3}$$

Using this to eliminate B, and evaluating the various constants, we have an upper bound on Λ

$$\Lambda^{1/2} \lesssim 0.1\,\frac{\delta^{3/2}}{t}$$

Note that $\delta^{3/2}/t$ is the combination which, in linear perturbation theory, stays constant in the matter dominated era as seen in (3.31). We see that if we want gravitational collapse to occur and galaxies to form (which, let's face it, would be nice) then there is an upper bound on the cosmological constant Λ, which depends on the strength of the initial perturbations.

What is this bound for our universe? It's a bit tricky to get an accurate statement using the information that we have gathered so far in this course, but we can get a ball-park figure. We argued in Section 3.3.3 that it is sensible to take $\delta_m \sim 10^{-3}$ at $z \approx 1000$, which is roughly the time of last scattering $t_{\rm last} \approx 350,000$ years $\approx 10^{13}$ s. This gives an upper bound on the cosmological constant of

$$\Lambda \lesssim 10^{-37}\ {\rm s}^{-2}$$

and a corresponding bound on the vacuum energy of

$$\rho_\Lambda = \frac{\Lambda c^2}{8\pi G} = \frac{M_{\rm pl}^2 c^4 \Lambda}{\hbar c^3} \approx (10^{47}\Lambda)\ {\rm eV\,m^{-3}s^2} \lesssim 10^{10}\ {\rm eV\,m^{-3}}$$

This is only a factor of 10 higher than the observed value of $\rho_\Lambda \approx 10^9\ {\rm eV\,m^{-3}}$! Although the calculation above involved quite a lot of hand-waving and order-of-magnitude estimates, the conclusion is the right one[14]: if the cosmological constant were much larger than we observe today, then galaxies would not have formed. We are, it appears, living on the edge.

[14] A better version of this calculation models the size of density perturbations using the σ_8 variable defined in (3.62), and takes into account the non-vanishing radiation contribution to the energy density in the early universe. Some of this discussion can be found in the original paper by Weinberg "*Anthropic Bound on the Cosmological Constant*" in Physical Review Letters vol 59 (1987).

3.4 The Cosmic Microwave Background

The cosmic microwave background (CMB) provides the snapshot of the early universe. In section 2.2, we described the how the CMB is an almost perfect blackbody. At temperature $T \approx 2.73$ K. However, there are small fluctuations in the CMB, with magnitude

$$\frac{\delta T}{T} \approx 10^{-5}$$

We already mentioned this at the very start of these lectures as evidence that the early universe was homogeneous and isotropic. As we now explain, these temperature fluctuations contain a near-perfect imprint of the anisotropies at the time of recombination. Moreover, we can trace the fate of these perturbations back in time to get another handle on the primordial power spectrum.

In Section 3.2.1, we stated that the perturbations in the early universe were adiabatic, meaning that perturbations in all fluids are proportional. In particular, the density perturbations in matter and radiation are related by

$$\delta_r = \frac{4}{3}\delta_m$$

It is more convenient to express this in terms of the temperature of the CMB. From our discussion of blackbody radiation, we know that $\rho_r \sim T^4$, so

$$\delta_r = \frac{\delta\rho_r}{\rho_r} = 4\frac{\delta T}{T} \quad \Rightarrow \quad \frac{\delta T}{T} = \frac{1}{3}\delta_m \tag{3.73}$$

We might, therefore expect that temperature fluctuations of the CMB contain a direct imprint of the matter fluctuations in the early universe. In fact, there is a subtlety which means that this is not quite true.

3.4.1 Gravitational Red-Shift

The new physics is gravitational redshift. This is an effect that arises from general relativity. Here we just give a heuristic sketch of the basic idea.

As a warm-up, first consider throwing a particle from the Earth upwards into space. We know that it must lose kinetic energy to escape the Earth's gravitational potential $\Phi = -GM/R$.

What happens if we do the same for light? Clearly light can't slow down, but it does lose energy. This manifests itself in a reduction in the frequency of the light, or a stretching of the wavelength. In other words, the light is redshifted. In the Newtonian limit, this redshift is

$$\frac{\delta\lambda}{\lambda} = -\frac{\Phi}{c^2} \tag{3.74}$$

Now consider a spatially varying gravitational potential $\delta\Phi(\mathbf{x})$ of the kind that permeates the early universe. To reach us, the photons from any point in space $\mathbf{x}$ will have to climb out of the gravitational potential and will be redshifted. This, in turn, shifts the temperature of the CMB. A straightforward generalisation of (3.74) suggests

$$\frac{\delta T(\hat{\mathbf{n}})}{T} = \frac{\delta\Phi(\mathbf{x}_{\rm last})}{c^2}$$

where $\mathbf{x}_{\rm last} = |\mathbf{x}_{\rm last}|\hat{\mathbf{n}}$ sits on the surface of last scattering, where the CMB was formed. In fact, this too misses an important piece of physics. The slight increase in $\delta\Phi$ results in a slight change in the local expansion rate of the universe which, since the CMB forms in the matter dominated era, scales as $a(t) \sim t^{2/3}$. This is known as the *Sachs-Wolfe effect*. It turns out that this gives an extra contribution of $-\frac{2}{3}\Phi/c^2$. This means that the temperature fluctuation in the CMB is related to the gravitational perturbation by

$$\frac{\delta T(\hat{\mathbf{n}})}{T} = \frac{\delta\Phi(\mathbf{x}_{\rm last})}{3c^2} \tag{3.75}$$

We learn that there are two, competing contributions to the temperature fluctuations in the CMB: the initial adiabatic perturbation (3.73) and the gravitational perturbation leading to the redshift (3.75). The question is: which is bigger?

The two contributions are not independent. They are related by the Poisson equation (3.51),

$$\delta\Phi(\mathbf{k}) = -\frac{4\pi G}{c^2k^2}\bar{\rho}a^2\delta_m(\mathbf{k}) \tag{3.76}$$

We see that the redshift contribution dominates for large wavelengths (k small) while the adiabatic contribution dominates for small wavelengths (k large). The cross-over happens at the critical value of k

$$k_{\rm crit}^2 \sim \frac{4\pi G}{c^4}\bar{\rho}a^2 \quad \Rightarrow \quad k_{\rm crit} \sim \frac{aH}{c}$$

But we recognise this as the size of the co-moving horizon. This means that modes that are were outside the horizon at last scattering will be dominated by the redshift and the Sachs-Wolfe effect; those which were inside the horizon at last scattering will exhibit the matter power spectrum.

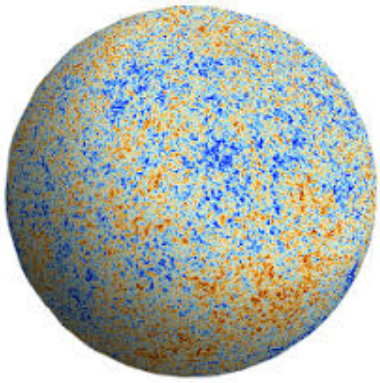

Figure 38: The CMB in its natural setting.

3.4.2 The CMB Power Spectrum

We don't have a three-dimensional map of the microwave background. Instead, the famous picture of the CMB lives on a sphere which surrounds us, as shown in the figure. This is clear in (3.75), where the temperature fluctuations depends only on the direction $\hat{\mathbf{n}}$.

We introduce spherical polar coordinates, and label the direction $\hat{\mathbf{n}}$ by the usual angles θ and ϕ. We then expand the temperature fluctuation in spherical polar coordinates as

$$\frac{\delta T(\hat{\mathbf{n}})}{T} = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} a_{l,m}\, Y_{l,m}(\theta, \phi)$$

Here $Y_{l,m}(\theta, \phi)$ are spherical harmonics, given by

$$Y_{l,m}(\theta, \phi) = N_{l,m} e^{im\phi} P_l^m(\cos\theta)$$

with $P_l^m(\cos\theta)$ the associated Legendre polynomial and $N_{l,m}$ an appropriate normalisation. Shortly, we will need $N_{l,0} = (2l+1)/4\pi$.

The measured coefficients $a_{l,m}$ the temperature anisotropies at different angular separation. Small l corresponds to large angles on the sky. We will now relate these to the primordial power spectrum $P(k)$.

As in the previous section, we are interested in correlations in the temperature fluctuations. The temperature two-point correlation function boils down to understanding the spatial average of

$$\langle a_{l,m}\, a^{\star}_{l',m'} \rangle = C_l\, \delta_{l,l'} \delta_{m,m'}$$

where statistical rotational invariance ensures that the average depends only on the angular momentum label l, and not on m. The coefficients C_l are called *multipole moments*.

The temperature correlation function can be written in terms of C_l. We pick spherical polar coordinates such that $\hat{\mathbf{n}} \cdot \hat{\mathbf{n}}' = \cos\theta$. Using θ and ϕ. Using $P_l^0(1) = 1$ and $P_l^m(1) = 0$ for $m \neq 0$, we then have

$$\begin{aligned}\frac{\langle \delta T(\hat{\mathbf{n}}) \delta T(\hat{\mathbf{n}}')\rangle}{T^2} &= \sum_{l,m}\sum_{l'm'} \langle a_{l,m} a_{l'm'}\rangle \; Y_{l,0}(\theta,\phi) \\ &= \sum_l \frac{2l+1}{4\pi} C_l \, P_l(\cos\theta)\end{aligned}$$

We would like to relate these coefficients C_l to the power spectrum. We will focus on large scales, with small l, where, as discussed above, we expect the temperature fluctuations to be dominated by the Sachs-Wolfe effect (3.75). In practice, this holds for $l \lesssim 50$.

It is a straightforward, if somewhat fiddly, exercise to write C_l in terms of the gravitational power spectrum (3.52).

$$\langle \delta\Phi(\mathbf{k}) \, \delta\Phi(\mathbf{k}')\rangle = (2\pi)^3 \, \delta_D^3(\mathbf{k}+\mathbf{k}') \, P_\Phi(k)$$

We do not give all the details here. (See, for example, the book by Weinberg.) After decomposing the Fourier mode $\delta\Phi(\mathbf{k})$ in spherical harmonics, one finds that the coefficients of the two-point function can be written as

$$C_l = \frac{16\pi T^2}{9} \int dk \; k^2 P_\Phi(k) j_l^2(kr)$$

with $j_l(kr)$ a spherical Bessel function. The primordial gravitational power spectrum takes the form (3.53)

$$P_\Phi(k) \sim k^{n-4}$$

which differs by a power of k^{-4} compared to the matter power spectrum, a fact which follows from the relation (3.76). For the Harrison-Zel'dovich spectrum, $n = 1$, one then finds

$$C_l \sim \frac{1}{l(l+1)}$$

It remains to compare this to the observed CMB power spectrum.

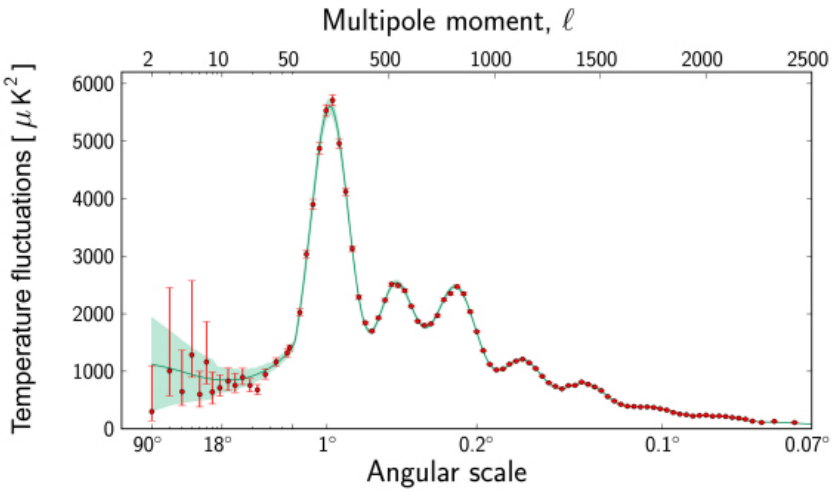

Figure 39: The CMB power spectrum measured by Planck. The combination $l(l+1)C_l$ is plotted on the vertical axis.

3.4.3 A Very Brief Introduction to CMB Physics

There has been an enormous effort, over many decades, to accurately measure the fluctuation coefficients C_l. The results from the Planck satellite are shown in Figure 39, with the combination $l(l+1)C_l$ plotted on the vertical axis; the red dots are data, shown with error bars, while the green line is the best theoretical fit.

The power spectrum exhibits a distinctive pattern of peaks and troughs. These are again a remnant of the acoustic oscillations in the early universe. A quantitative understanding of how these arise is somewhat beyond what this course. (You can learn more next year in Part III.) Here we give just a taster:

- At low l, the temperature fluctuations have the advertised scale $\delta T/T \approx 10^{-5}$. Here the plot is roughly constant. This confirms that the CMB is close to the Harrison-Zel'dovich spectrum, with $C_l \sim 1/l(l+1)$, as expected. In fact, a detailed analysis gives

  $$n \approx 0.97$$

  in good agreement with the measurements from galaxy distributions.

- The first peak sits at $l \approx 200$ and sets the characteristic angular scales of fluctuations that one can see by eye in the CMB maps. At this point, the fluctuations have risen to $\delta T/T \approx 6 \times 10^{-5}$.

This peak arises from an acoustic wave that had time to undergo just a single compression before decoupling. This is the same physics that led to the baryon acoustic peak shown in Figure 37. The angular size in the sky is determined both by the horizon at decoupling (usually referred to as the *sound horizon*) and the subsequent expansion history of the universe. In particular, its angular value is very sensitive to the curvature of the universe. The location of this first peak is our best evidence that the universe is very close to flat (or $k = 0$ in the language of Section 1.)

Given the observed fact that the matter and radiation in the universe sits well below the critical value, the position of the first peak also provides corroborating evidence for dark energy.

- The second and third peaks contain information about the amount of baryonic and dark matter in the early universe. This is because the amplitudes of successive oscillations depends on both the baryon-to-photon ratio in the plasma, and the gravitational potentials created by dark matter.

- The microwave background doesn't just contain information from the temperature anisotropies. One can also extract information from the polarisation of the photons. These are two kinds of polarisation pattern, known as *E-modes* and *B-modes*.

  The E-mode polarisation has been measured and is found to be correlated with the temperature anisotropies. Interestingly, these correlations (really anti-correlations) extend down below $l < 200$. This is important because modes of this size were outside the horizon at the time the CMB was formed. Such correlations could only arise if there was some causal interaction between the modes, pointing clearly to the need for a period of inflation in the very early universe.

  B-modes in the CMB have been found but, somewhat disappointingly, arise because of contamination due to interstellar dust. A discovery of primordial B-modes would be *extremely* exciting since they are thought to be generated by gravitational waves, created by quantum effects at play during in inflation. The observation of primordial B-modes imprinted in the CMB would provide our first experimental window into quantum gravity!

- For very low $l \lesssim 10$, there are both large error bars and poor agreement with the theoretical expectations. The large error bars arise because we only have one sky to observe and only a handful of independent observables, with $-l \leq m \leq l$. This

issue is known as *cosmic variance*. It makes it difficult to know if the disagreement with theory is telling us something deep, or is just random chance.

3.5 Inflation Revisited

> "With the new cosmology the universe must have started off in some very simple way. What, then, becomes of the initial conditions required by dynamical theory? Plainly there cannot be any, or they must be trivial. We are left in a situation which would be untenable with the old mechanics. If the universe were simply the motion which follow from a given scheme of equations of motion with trivial initial conditions, it could not contain the complexity we observe. Quantum mechanics provides an escape from the difficulty. It enables us to ascribe the complexity to the quantum jumps, lying outside the scheme of equations of motion."
>
> A very prescient Paul Dirac, in 1939

Until now, we have only focussed only on the evolution of some initial density perturbations that were mysteriously laid down in the very early universe. The obvious question is: where did these perturbations come from in the first place?

There is an astonishing answer to this question. The density perturbations are quantum fluctuations from the very first moment after the Big Bang, fluctuations which were caught in the act and subsequently stretched to cosmological scales by the rapid expansion of the universe during inflation, where they laid the seeds for the formation of galaxies and other structures that we see around us.

This idea that the origin of the largest objects in the universe can be traced back to quantum fluctuations taking place at the very earliest times is nothing short of awe-inspiring. Yet, as we will see, the process of inflation generates perturbations on a super-horizon scale. These perturbations are adiabatic, Gaussian and with a power spectrum $P(k) \sim k^n$ with $n \approx 1$. In other words, the perturbations are exactly of the form required to describe our universe.

3.5.1 Superhorizon Perturbations

Before we get to the nitty gritty, let's first understand why inflation provides a very natural environment in which to create perturbations which, subsequently, have wavelength greater than the apparent horizon. During inflation, the universe undergoes an accelerated expansion (1.90) which, for simplicity, we approximate as an exponential de Sitter phase,

$$a(t) = a(0) \exp\left(H_{\rm inf} t\right)$$

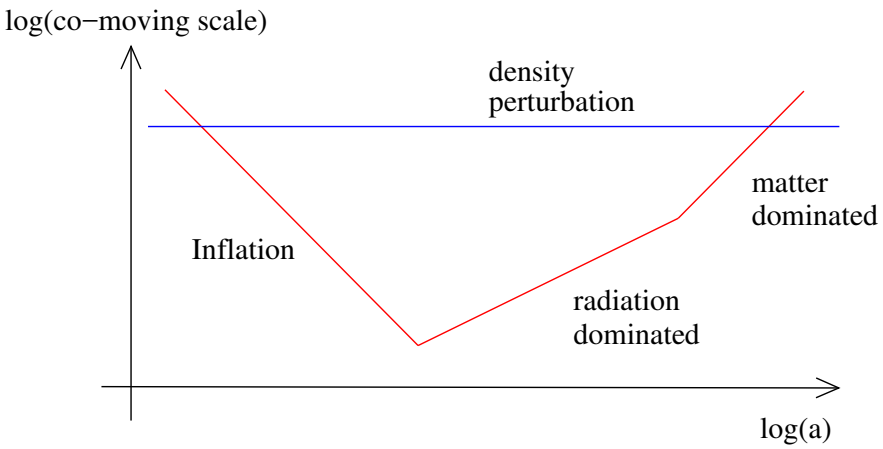

Figure 40: The density perturbations are created during inflation and exit the co-moving horizon, shown in red. Then they wait. Later, during the hot Big Bang phase of radiation or matter domination, the co-moving horizon expands and the density perturbations re-enter where we see them today.

The key observation is that, in an accelerating phase of this type, the co-moving horizon is shrinking,

$$\chi_H = \frac{c}{aH_{\rm inf}} \tag{3.77}$$

Focussing on the co-moving horizon (rather than the physical horizon) gives us a view of inflation in which we zoom into some small patch of space, which subsequently becomes our entire universe.

Any perturbation created during inflation with co-moving wavevector $\mathbf{k}$ will rapidly move outside the horizon, where they linger until the expansion of the universe slows to a more sedentary pace, after which the co-moving horizon expands, as in (3.38), and the perturbations created during inflation can now re-enter. This is shown in Figure 40. In this way, inflation can naturally generate superhorizon perturbations that seem to be needed to explain the universe we see around us. This picture also makes it clear that the longer wavelength perturbations must have been created earlier in the universe's past.

3.5.2 Classical Inflationary Perturbations

It remains for us to explain how these density perturbations arose in the first place. A full discussion requires both quantum field theory and general relativity. Here we give the essence of the idea.

Recall that inflation requires the introduction of a new degree of freedom, the *inflaton* scalar field with action (1.82),

$$S = \int d^3x\, dt\; a^3(t) \left[\frac{1}{2}\dot{\phi}^2 - \frac{c^2}{2a^2(t)} \nabla\phi \cdot \nabla\phi - V(\phi) \right]$$

The scalar field ϕ rolls from some initial starting point, high up on the potential, and in doing so, drives inflation. In this process, ϕ also undergoes quantum fluctuations; these will be the seeds for density perturbations.

We start by looking at a stripped down version of this story. We will take the potential $V(\phi) =$ constant, which is the same thing as a cosmological constant. This ensures that the universe sits in a de Sitter phase with $a(t) \sim e^{H_{\rm inf} t}$. We then look at the dynamics of ϕ in this background. The classical equation of motion is

$$\frac{d^2\phi}{dt^2} + 3H_{\rm inf}\frac{d\phi}{dt} - \frac{c^2}{a^2}\nabla^2\phi = 0 \tag{3.78}$$

Ultimately, we want to treat $\phi(\mathbf{x}, t)$ as a quantum variable. To do this, we will massage the equation of motion in various ways until it looks like something more familiar. First, we decompose the spatial variation of $\phi(\mathbf{x}, t)$ in Fourier modes,

$$\phi(\mathbf{x}, t) = \int \frac{d^3k}{(2\pi)^3}\; e^{-i\mathbf{k}\cdot\mathbf{x}}\, \phi_{\mathbf{k}}(t)$$

The reality of $\phi(\mathbf{x}, t)$ means that we must have $\phi^\star_{\mathbf{k}} = \phi_{-\mathbf{k}}$. The equation of motion (3.78) then becomes decoupled equations for each $\phi_{\mathbf{k}}$,

$$\frac{d^2\phi_{\mathbf{k}}}{dt^2} + 3H_{\rm inf}\frac{d\phi_{\mathbf{k}}}{dt} + \frac{c^2k^2}{a^2}\phi_{\mathbf{k}} = 0 \tag{3.79}$$

This equation takes the form of a damped harmonic oscillator, with some time dependence hiding in the $1/a^2$ part of the final term. A time dependent frequency is something we can deal with in quantum mechanics, but friction is not. For this reason, we want to make a further change of variables that gets rid of the damping term proportional to $\dot{\phi}_{\mathbf{k}}$. To achieve this, we work in conformal time (1.26)

$$\tau = \int^t \frac{dt'}{a(t')} = -\frac{1}{aH_{\rm inf}}$$

Note that, for a de Sitter universe, conformal time sits in the range $\tau \in (-\infty, 0)$ so $\tau \to 0^-$ is the far future. We then have

$$\frac{d^2\phi}{dt^2} = \frac{1}{a^2}\frac{d^2\phi}{d\tau^2} - \frac{H}{a}\frac{d\phi}{d\tau} \quad \text{and} \quad \frac{d\phi}{dt} = \frac{1}{a}\frac{d\phi}{d\tau}$$

and the equation of motion (3.79) becomes an equation for $\phi_{\mathbf{k}}(\tau)$,

$$\frac{d^2\phi_{\mathbf{k}}}{d\tau^2} - \frac{2}{\tau}\frac{d\phi_{\mathbf{k}}}{d\tau} + c^2k^2\phi_{\mathbf{k}} = 0$$

This doesn't seem to have done much good, simply changing the coefficient of the damping term. But things start looking rosier if we define

$$\tilde{\phi}_{\mathbf{k}} = -\frac{1}{H_{\rm inf}\tau}\,\phi_{\mathbf{k}} \tag{3.80}$$

Using $\dot{a} = H_{\rm inf}a$, the equation becomes

$$\frac{d^2\tilde{\phi}_{\mathbf{k}}}{d\tau^2} + \left(c^2k^2 - \frac{2}{\tau^2}\right)\tilde{\phi}_{\mathbf{k}} = 0 \tag{3.81}$$

This is the final form that we want. Each $\tilde{\phi}_{\mathbf{k}}$ obeys the equation of a harmonic oscillator, with a frequency

$$\omega_k^2 = c^2k^2 - \frac{2}{\tau^2} \tag{3.82}$$

that depends on both k and on conformal time τ. In the far past, $\tau \to -\infty$, the time-dependent $1/\tau^2$ term is negligible. However, as we move forward in time, ω^2 first goes to zero and then becomes negative, corresponding to a harmonic oscillator with an upside-down potential. The co-moving horizon (3.77) is $\chi_H = c/aH_{\rm inf} = -c\tau$. This means that, for a given perturbation $\mathbf{k}$, the wavelength $\lambda = 2\pi/k$ exits the horizon at more or less the time that the frequency of the associated harmonic oscillator is $\omega_k^2 = 0$.

It is not too difficult to write down a solution to the time-dependent harmonic oscillator (3.81). It is a second order differential equation, so we expect two linearly independent solutions. You can check that the general form is given by

$$\tilde{\phi}_{\mathbf{k}} = \alpha e^{-ick\tau}\left(1 - \frac{i}{ck\tau}\right) + \beta e^{+ick\tau}\left(1 + \frac{i}{ck\tau}\right) \tag{3.83}$$

where α and β are integration constants. In the far past, $ck\tau \to -\infty$, these modes oscillate just like a normal harmonic oscillator. But as inflation proceeds, and $ck\tau \to 0^-$, the oscillations stop. Expanding out the $e^{\pm ick\tau}$ in this limit, we find that the modes grow as $\tilde{\phi}_{\mathbf{k}} \approx (\beta - \alpha)/ck\tau$. If we then translate back to the original field $\phi_{\mathbf{k}}$ using (3.80), we find that the Fourier modes obey

$$\phi_{\mathbf{k}} = -\frac{\alpha H_{\rm inf}}{ck}e^{-ick\tau}\,(ck\tau - i) - \frac{\beta H_{\rm inf}}{ck}e^{+ick\tau}\,(ck\tau + i)$$

These modes now oscillate wildly at the beginning of inflation, $ck\tau \to -\infty$, but settle down to become constant after the mode has exited the horizon and $ck\tau \to 0^-$.

3.5.3 The Quantum Harmonic Oscillator

Our ultimate goal is to understand the quantum fluctuations of the inflaton field $\phi(\mathbf{x}, t)$. At first glance, this sounds like a daunting problem. But the analysis above shows the way forward, because each (rescaled) Fourier mode $\tilde{\phi}_{\mathbf{k}}$ obeys the equation for a simple harmonic oscillator (3.81). And we know how to quantise the harmonic oscillator. The only subtlety is that the frequency ω_k is time dependent. But this too is a problem that we can address purely within quantum mechanics.

A Review of the Harmonic Oscillator

Let's first review the solution to the familiar harmonic oscillator in which the frequency ω does not vary with time. The Hamiltonian is

$$\hat{H} = \frac{1}{2}\hat{p}^2 + \frac{1}{2}\omega^2\hat{q}^2$$

where we've set the usual mass $m = 1$. The position and momentum obey the canonical commutation relation

$$[\hat{q}, \hat{p}] = i\hbar$$

The slick way to solve this is to introduce annihilation and creation operators. These are defined by

$$\hat{a} = \sqrt{\frac{\omega}{2\hbar}}\hat{q} + i\sqrt{\frac{1}{2\hbar\omega}}\hat{p} \quad \text{and} \quad \hat{a}^\dagger = \sqrt{\frac{\omega}{2\hbar}}\hat{q} - i\sqrt{\frac{1}{2\hbar\omega}}\hat{p}$$

and the inverse is

$$\hat{q} = \sqrt{\frac{\hbar}{2\omega}}(\hat{a} + \hat{a}^\dagger) \quad \text{and} \quad \hat{p} = -i\sqrt{\frac{\hbar\omega}{2}}(\hat{a} - \hat{a}^\dagger) \tag{3.84}$$

You can check that these obey the commutation relations

$$[\hat{a}, \hat{a}^\dagger] = 1$$

When written in terms of annihilation and creation operators, the Hamiltonian takes the simple form

$$\hat{H} = \frac{1}{2}\hbar\omega(\hat{a}\hat{a}^\dagger + \hat{a}^\dagger\hat{a}) = \hbar\omega\left(\hat{a}^\dagger\hat{a} + \frac{1}{2}\right)$$

Now it is straightforward to build the energy eigenstates of the system. The ground state is written as $|0\rangle$ and obeys

$$\hat{a}|0\rangle = 0$$

Excited states then constructed by acting with $\hat{a}^\dagger$, giving

$$|n\rangle = \frac{1}{\sqrt{n!}}\hat{a}^{\dagger n}|0\rangle \quad \Rightarrow \quad \hat{H}|n\rangle = \hbar\omega\left(n+\frac{1}{2}\right)|n\rangle$$

In what follows, we will be particularly interested in the variance in the ground state $|0\rangle$. First, recall that the expectation value of $\hat{q}$ vanishes in the ground state (or, indeed, in any energy eigenstate),

$$\langle 0|\hat{q}|0\rangle = \sqrt{\frac{\hbar}{2\omega}}\langle 0|(\hat{a}+\hat{a}^\dagger)|0\rangle = 0$$

where we use the property of the ground state $\hat{a}|0\rangle = 0$ or, equivalently, $\langle 0|\hat{a}^\dagger = 0$. However, the variance is non-vanishing, and given by

$$\langle 0|\hat{q}^2|0\rangle = \frac{\hbar}{2\omega}\langle 0|(\hat{a}+\hat{a}^\dagger)^2|0\rangle = \frac{\hbar}{2\omega}\langle 0|\hat{a}^\dagger\hat{a}|0\rangle = \frac{\hbar}{2\omega}$$

We write this as

$$\langle \hat{q}^2\rangle = \frac{\hbar}{2\omega} \tag{3.85}$$

These will be the fluctuations which we will apply to the inflaton field. But first we need to see the effects of a time dependent frequency.

A Review of the Heisenberg Picture

There are two ways to think about time evolution in quantum mechanics. In the first, known as the *Schrödinger picture*, the states evolve in time while the operators are fixed. In the second, known as the *Heisenberg picture*, the states are fixed while the operators evolve in time. Both give the same answers for any physical observable (i.e. expectation functions) but one approach may be more convenient for any given problem. It will turn out that the Heisenberg picture is best suited for cosmological purposes, so we pause to review it here.

The Schrödinger picture is perhaps the most intuitive. Here the evolution of states is determined by the time-dependent Schrödinger equation

$$i\hbar\frac{d|\psi\rangle}{dt} = \hat{H}|\psi\rangle$$

Alternatively, we can introduce a unitary evolution operator $U(t)$ which dictates how the states evolve,

$$|\psi(t)\rangle = \hat{U}(t)|\psi(0)\rangle$$

The Schrödinger equation tells us that this operator must obey

$$i\hbar\frac{d\hat{U}}{dt} = \hat{H}\hat{U} \tag{3.86}$$

If $\hat{H}$ is time-independent then this is solved by $\hat{U} = \exp\left(-i\hat{H}t/\hbar\right)$. However, if $\hat{H}$ is time-dependent (as it will be for us) we must be more careful.

In the Heisenberg picture, this time dependence is moved onto the operators. We consider the state to be fixed, while operators evolve as

$$\hat{\mathcal{O}}(t) = U^\dagger(t)\,\hat{\mathcal{O}}\,\hat{U}(t)$$

From (3.86), we find that these time-dependent operators obey

$$\frac{d\hat{\mathcal{O}}}{dt} = \frac{i}{\hbar}[\hat{H}, \hat{\mathcal{O}}] \tag{3.87}$$

We can look at how this works for the harmonic oscillator with a fixed frequency ω. The creation and annihilation operators $\hat{a}$ and $\hat{a}^\dagger$ have a particularly simple time evolution,

$$\begin{aligned} [\hat{H}, \hat{a}] = -\hbar\omega\hat{a} \quad &\Rightarrow \quad \hat{a}(t) = e^{-i\omega(t-t_0)}\,\hat{a}(t_0) \\ [\hat{H}, \hat{a}^\dagger] = +\hbar\omega\hat{a}^\dagger \quad &\Rightarrow \quad \hat{a}(t) = e^{+i\omega(t-t_0)}\,\hat{a}^\dagger(t_0) \end{aligned}$$

We can then simply substitute this into (3.84) to see how $\hat{q}(t)$ and $\hat{p}(t)$ evolve in time. We have

$$\begin{aligned} \hat{q}(t) &= \sqrt{\frac{\hbar}{2\omega}}\left(e^{-i\omega(t-t_0)}\,\hat{a}(t_0) + e^{+i\omega(t-t-0)}\,\hat{a}^\dagger(t_0)\right) \\ \hat{p}(t) &= -i\sqrt{\frac{\hbar\omega}{2}}\left(e^{-i\omega(t-t_0)}\,\hat{a}(t_0) - e^{+i\omega(t-t-0)}\,\hat{a}^\dagger(t_0)\right) \end{aligned} \tag{3.88}$$

Note that these obey the operator equation of motion (3.87), with

$$\frac{d\hat{q}}{dt} = \frac{i}{\hbar}[\hat{H}, \hat{q}] = \hat{p} \quad \text{and} \quad \frac{d\hat{p}}{dt} = \frac{i}{\hbar}[\hat{H}, \hat{p}] = -\omega^2\hat{q}$$

The Time-Dependent Harmonic Oscillator

For our cosmological application, we need to understand the physics of a harmonic oscillator with a time-dependent frequency,

$$\hat{H}(t) = \frac{1}{2}\hat{p}^2 + \frac{1}{2}\omega^2(t)\hat{q}^2$$

Our real interest is in the specific time-dependence (3.82) but, for now, we will keep $\omega(t)$ arbitrary.

A time-dependent Hamiltonian opens up different kinds of questions. We could, for example, pick some fixed moment in time t_0 at which we diagonalise the Hamiltonian. We do this by introducing the usual annihilation and creation operators, and place the system in the instantaneous ground state

$$\hat{a}(t_0)|0\rangle = 0$$

Now the system subsequently evolves. But, with a time-dependent Hamiltonian it will no longer sit in the ground state (in the Schrödinger picture). This is related to the fact that energy is no longer conserved when the Hamiltonian is time-dependent. We want to understand how the variance (3.85) evolves in this situation.

We will work in the Heisenberg picture. In analogy with (3.88), we expand the position operator in terms of $\hat{a}(t_0)$ and $\hat{a}^\dagger(t_0)$, with some time-dependent coefficients

$$\hat{q}(t) = v(t)\,\hat{a}(t_0) + v^\star(t)\,\hat{a}^\dagger(t_0) \tag{3.89}$$

The momentum is then

$$\hat{p}(t) = \frac{d\hat{q}}{dt} = \dot{v}(t)\,\hat{a}(t_0) + \dot{v}^\star(t)\,\hat{a}^\dagger(t_0)$$

Taking a second time derivative, we have

$$\frac{d\hat{p}}{dt} = \ddot{v}(t)\,\hat{a}(t_0) + \ddot{v}^\star(t)\,\hat{a}^\dagger(t_0) = -\omega^2(t)\hat{q}(t)$$

where the second equality comes from the operator equation of motion (3.87). Comparing coefficients of $\hat{a}(t_0)$ and $\hat{a}^\dagger(t_0)$, we see that the coefficient $v(t)$ must obey the original equation of motion

$$\ddot{v} + \omega^2(t)v = 0 \tag{3.90}$$

Meanwhile, we can normalise $v(t)$ by insisting that $[\hat{q}(t), \hat{p}(t)] = i\hbar$ and $[\hat{a}(t_0), \hat{a}^\dagger(t_0)] = 1$. These are compatible provided

$$v\dot{v}^\star - v^\star\dot{v} = i\hbar \tag{3.91}$$

When ω is constant, this agrees with what we saw before: we had $v = \sqrt{\hbar/2\omega}e^{-i\omega(t-t_0)}$, which is a solution to the harmonic oscillator (3.90), with the normalisation fixed by (3.91).

Finally, we can answer the main question: if we place the time-dependent harmonic oscillator in the ground state $|0\rangle$ at some time t_0, how does the variance of $\hat{q}(t)$ subsequently evolve? Using (3.89), we have

$$\langle \hat{q}^2(t) \rangle = |v(t)|^2 \tag{3.92}$$

This is the result we need to evaluate the size of quantum fluctuations during inflation.

3.5.4 Quantum Inflationary Perturbations

We can now import the quantum mechanical story above directly to the inflaton field. Recall that each (rescaled) Fourier mode of the inflaton acts like a harmonic oscillator with a time-dependent frequency,

$$\frac{d^2\tilde{\phi}_{\mathbf{k}}}{d\tau^2} + \omega_{\mathbf{k}}^2(\tau)\tilde{\phi}_{\mathbf{k}} = 0 \quad \text{with} \quad \omega_k^2(\tau) = c^2k^2 - \frac{2}{\tau^2}$$

We treat each Fourier component as an independent quantum operator which, piling hat on hat, we write as $\hat{\tilde{\phi}}_{\mathbf{k}}$. This is analogous to $\hat{q}$ in the harmonic oscillator that we described above. Following (3.89), we write

$$\hat{\tilde{\phi}}_{\mathbf{k}}(\tau) = v_{\mathbf{k}}(\tau)\,\hat{a}_{\mathbf{k}}(\tau_0) + v_{\mathbf{k}}^\star(\tau)\,\hat{a_{\mathbf{k}}}^\dagger(\tau_0) \tag{3.93}$$

where, as we've seen, $v(\tau)$ must obey the original harmonic oscillator equation (3.90), together with the normalisation condition (3.91) (with $\dot{v} = dv/d\tau$ in these equations).

First, we must decide when we're going to place the system in its ground state. The only sensible option is to do this right at the beginning of inflation, with $\tau_0 \to -\infty$. At this point, the frequency is simply $\omega_k^2 = c^2k^2$ and we get the normal harmonic oscillator. In the context of inflation, this choice is referred to as the *Bunch-Davies vacuum*. As we will see, this simple choice for the initial conditions at the very beginning of the universe is the one that ultimately agrees with what we see around us today.

Next, we must determine the coefficient $v_{\mathbf{k}}(\tau)$. We know that the general solution to (3.90) is (3.83)

$$v_{\mathbf{k}}(\tau) = \alpha e^{-ick\tau}\left(1 - \frac{i}{ck\tau}\right) + \beta e^{+ick\tau}\left(1 + \frac{i}{ck\tau}\right)$$

We need only to fix the integration constants α and β. We set $\beta = 0$ to ensure that, as $\tau \to -\infty$, the operator expansion (3.93) agrees with that of the normal harmonic oscillator. The normalisation of α is then fixed by (3.91)

$$v_{\mathbf{k}}\dot{v}_{\mathbf{k}}^\star - v_{\mathbf{k}}^\star\dot{v}_{\mathbf{k}} = 2\alpha^2 ick = i\hbar \quad \Rightarrow \quad \alpha^2 = \frac{\hbar}{2ck}$$

Now we're home and dry. The time-dependent coefficient in the expansion of the Fourier mode $\hat{\tilde{\phi}}_{\mathbf{k}}$ is

$$v_{\mathbf{k}}(\tau) = \sqrt{\frac{\hbar}{2ck}} e^{-ick\tau}\left(1 - \frac{i}{ck\tau}\right)$$

So the quantum fluctuations in the field $\tilde{\phi}_{\mathbf{k}}$ can be read off from (3.92),

$$\langle \hat{\tilde{\phi}}_{\mathbf{k}} \hat{\tilde{\phi}}^{\dagger}_{\mathbf{k}} \rangle = \frac{\hbar}{2ck}\left(1 + \frac{1}{c^2k^2\tau^2}\right)$$

where we have to take $\hat{\phi}\hat{\phi}^{\dagger}$ because, in contrast to $\hat{q}$, the Fourier mode $\hat{\phi}_{\mathbf{k}}$ is complex. Our interest is in the original field $\phi_{\mathbf{k}} = -H_{\rm inf}\tau\tilde{\phi}_{\mathbf{k}}$. (This rescaling was introduced back in (3.80).) The fluctuations of this field are given by

$$\langle \hat{\phi}_{\mathbf{k}} \hat{\phi}^{\dagger}_{\mathbf{k}} \rangle = \frac{\hbar H^2_{\rm inf}}{2ck}\left(\frac{1}{c^2k^2} + \tau^2\right)$$

At early times, the fluctuations are large. However, at late times, $ck\tau \to 0^-$, the fluctuations become constant in time. The cross-over happens at $ck\tau \approx 1$, which is when the fluctuations exit the horizon. At later times, the k dependence of the fluctuations is given by

$$\lim_{ck\tau\to 0^-} \langle \hat{\phi}_{\mathbf{k}} \hat{\phi}^{\dagger}_{\mathbf{k}} \rangle = \frac{\hbar H^2_{\rm inf}}{2c^3k^3} \tag{3.94}$$

This is the famous inflationary power spectrum. It takes the Harrison-Zel'dovich "scale invariant" form, a statement which, as we explained in Section 3.2.1, is manifest only when written in terms of the power spectrum introduced in (3.50),

$$\Delta_{\phi}(k) = \frac{4\pi k^3 \langle \hat{\phi}_{\mathbf{k}} \hat{\phi}^{\dagger}_{\mathbf{k}} \rangle}{(2\pi)^3} = \frac{\hbar H^2_{\rm inf}}{4\pi^2 c^3}$$

This is indeed independent of k. These fluctuations remain frozen outside the horizon, until they subsequently re-enter during the radiation dominated era or, for very long wavelength, matter dominated era.

The fact that the power spectrum $\Delta(k)$ does not depend on the wavelength can be traced to an underlying, scale invariance symmetry of de Sitter space.

A Rolling Inflation

The calculation above holds for a scalar field ϕ with $V(\phi) = \text{constant}$. This, of course, is not the realistic situation for inflation, but it's a good approximation when the scalar field rolls down a rather flat potential. In this case, the shorter wavelength modes (larger k) which exit the horizon later will have a slightly smaller H and, correspondingly, slightly smaller fluctuations. This means that the power spectrum is almost, but not quite, scale invariant.

We will not present this longer calculation here; we quote only the answer which we write as

$$\Delta_\phi(k) \sim k^{n_S - 1}$$

Here *scalar spectral index* n_s is close to 1. It turns out that, to leading order,

$$n_S = 1 - 2\epsilon \tag{3.95}$$

where ϵ is a dimensionless number known as a *slow-roll parameter*. It is one of two such parameters which are commonly used to characterise the shape of the inflaton potential,

$$\epsilon = \frac{M_{\rm pl}^2}{2}\left(\frac{V'}{V}\right)^2 \quad \text{and} \quad \eta = M_{\rm pl}^2 \frac{V''}{V}$$

with the Planck mass given by $M_{\rm pl}^2 = \hbar c/8\pi G$.

The Gravitational Power Spectrum

To compare to observations, we must turn the fluctuations of the inflaton field ϕ into fluctuations in the energy density or, as explained in (3.2.1), the gravitational potential Φ. As with many details, a full treatment needs a relativistic analysis. It turns out that the inflationary perturbations imprint themselves directly as fluctuations of the gravitational potential,

$$\Delta_\phi(k) \mapsto \Delta_\Phi(k)$$

But this is exactly what we need! The almost scale-invariant power spectrum of the inflaton gives rise to the almost scale-invariant power spectrum needed to explain the structure of galaxies in our universe. Moreover, the observed spectral index $n \approx 0.97$ can be used to infer something about the dynamics of the inflaton in the early universe.

There are many remarkable things about the inflationary origin of density perturbations. Here is another: the fluctuations that we computed in (3.94) are quantum. They measure the spread in the wavefunction. Yet these must turn into classical probabilities which, subsequently, correspond to the random distribution of galaxies in the universe. This is, at heart, no different from the quantum measurement problem in any other setting, now writ large across the sky. But one may worry that, in the absence of any observers, the problem is more acute. Closer analysis suggests that the modes decohere, and evolve from quantum to classical, as they exit the horizon.

3.5.5 Things We Haven't (Yet?) Seen

There is much more to tell about inflation, both things that work and things that don't. Here, as a taster, is a brief description of two putative features of inflation which might, with some luck, be detected in the future.

Gravitational Waves

It's not just the inflaton that suffers quantum fluctuations during inflation. There are also quantum fluctuations of spacetime itself.

It's a common misconception that we don't understand quantum gravity. There is, of course, some truth to this: there are lots of things that we don't understand about quantum gravity, such as what happens inside the singularity of a black hole. But provided that the curvature of spacetime is not too large, we can do trustworthy quantum gravity calculations, and inflation provides just such an opportunity.

These quantum gravity fluctuations leave an imprint on spacetime and, subsequently, on the CMB. This can be traced back to the fact that the graviton is a particle with spin 2. Correspondingly, these fluctuations have a distinctive swirly pattern, known as B-mode polarisation.

We have not yet observed such B-modes in the CMB, although it's not for the want of trying. Finding them would be a very big deal: not only would it be our first observational evidence of quantum gravity, but they would tell us directly the scale at which inflation occurs, meaning that we can determine $H_{\rm inf}$, or equivalently, the magnitude of the potential $V(\phi)$. (In contrast, the density perturbations that we have observed depend on both $V(\phi)$ and the slow-roll parameter ϵ as we can see in (3.95).)

The power spectrum of tensor modes is denoted Δ_T (with T for tensor). It also predicted to take (almost) Harrison-Zel'dovich form, but with a slightly different spectral index from the scalar modes. Cosmologists place limits on the strength of these tensor

perturbations relative to the scalar modes Δ_ϕ formed by the inflaton. The ratio is defined to be

$$r = \frac{\Delta_T}{\Delta_\phi}$$

Currently, the lack of observation only allows us to place an upper limit of $r \leq 0.07$, although it's possible to relax this if we allow some flexibility with other parameters. Roughly speaking, if inflation is driven by physics close to the Planck scale or GUT scale then we have a hope of detecting $r \neq 0$. If, however, the scale of inflation is closer to the TeV scale (the current limit of our knowledge in particle physics) then it seems unlikely we will find tensor modes in our lifetime.

Non-Gaussianity

We saw in Section 3.2.1 that the observed spectrum of density perturbations is well described by a Gaussian probability distribution. This too is a success of inflation: one can show that in slow-roll inflation three point functions $\langle \hat{\phi}_{\mathbf{k}_1} \hat{\phi}_{\mathbf{k}_2} \hat{\phi}_{\mathbf{k}_3} \rangle$ are suppressed by the slow-roll parameters ϵ^2 and η^2.

Nonetheless, this hasn't stopped people hoping. The discovery of non-Gaussian primordial density fluctuations would provide us with a wealth of precious information about the detailed dynamics of the inflation in the early universe. While the two-point function tells us just two numbers — n_S and the overall scale of the power spectrum — the three-point correlator $\langle \hat{\phi}_{\mathbf{k}_1} \hat{\phi}_{\mathbf{k}_2} \hat{\phi}_{\mathbf{k}_3} \rangle \sim f_{NL}\, \delta_D^3(\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3)$ is a function of every triangle you can draw on the (Fourier transformed) sky. For this reason, there has been a big push to try to detect a primordial non-Gaussian signal in the CMB or large scale structure. Alas, so far, to no avail. Meanwhile, ever optimistic theorists have proposed more creative versions of inflation which give rise to non-Gaussianity at a detectable level[15]. Sadly, there is little evidence that these theorists are going to be validated any time soon.

[15] See, for example, M. Alishahiha, E. Silverstein and D. Tong, "*DBI in the sky*", Phys. Rev. D **70**, 123505 (2004) [hep-th/0404084].