Physical Cosmology Notes

Phil Bull, June 8, 2026

Contents

About this course

These notes are based on the Physical Cosmology course I taught at Queen Mary University of London from 2019/20 to 2021/22. They were written from the ground up, taking cues, inspiration, and notation from the precursor module by Karim Malik and *An Introduction to Modern Cosmology* by Andrew Liddle for the scope and depth of the course. I recommend using the latter as the main course textbook, and you will find suggested reading that references this and a few other texts in each section.

This course was taught for a mixed final-year BSc and MSc-level cohort, and is written with advanced undergraduates and starting graduate students in physics in mind. There were three 1-hour lectures per week, each covering one section of these notes. I understand that the course in its present form no longer uses these notes, and so I am releasing them publicly.

Please contact me at `phil.bull@manchester.ac.uk` with comments and corrections.

—- Phil Bull, Manchester, June 2026

License and copyright

1. Expanding universe

In this section you will learn some of the basic facts about cosmology, including a brief history of the Universe and the fact that space is expanding. As part of this, you will learn some of the history behind the discovery of the expansion of the Universe, and some of the key arguments in support of this observation.

Reading for this topic:

- *An Introduction to Modern Cosmology (A. Liddle), Chapter 1: Brief History of Cosmological Ideas.*
- *An Introduction to Modern Cosmology (A. Liddle), Chapter 2: Observational Overview*

1.1. What is cosmology?

Cosmology is the study of the whole Universe – how it formed, how it evolves with time, what its basic properties and constituents are, and how it is structured on the largest distance scales.

Cosmologists study the Universe through a combination of astronomical observations, mathematical theories, and computer simulations. The field has only been recognised as a scientific discipline for about 100 years – before then it was the domain of philosophers and theologians!

This module is about the fundamental facts and findings of cosmology, and the physical theories that we have developed to explain them.

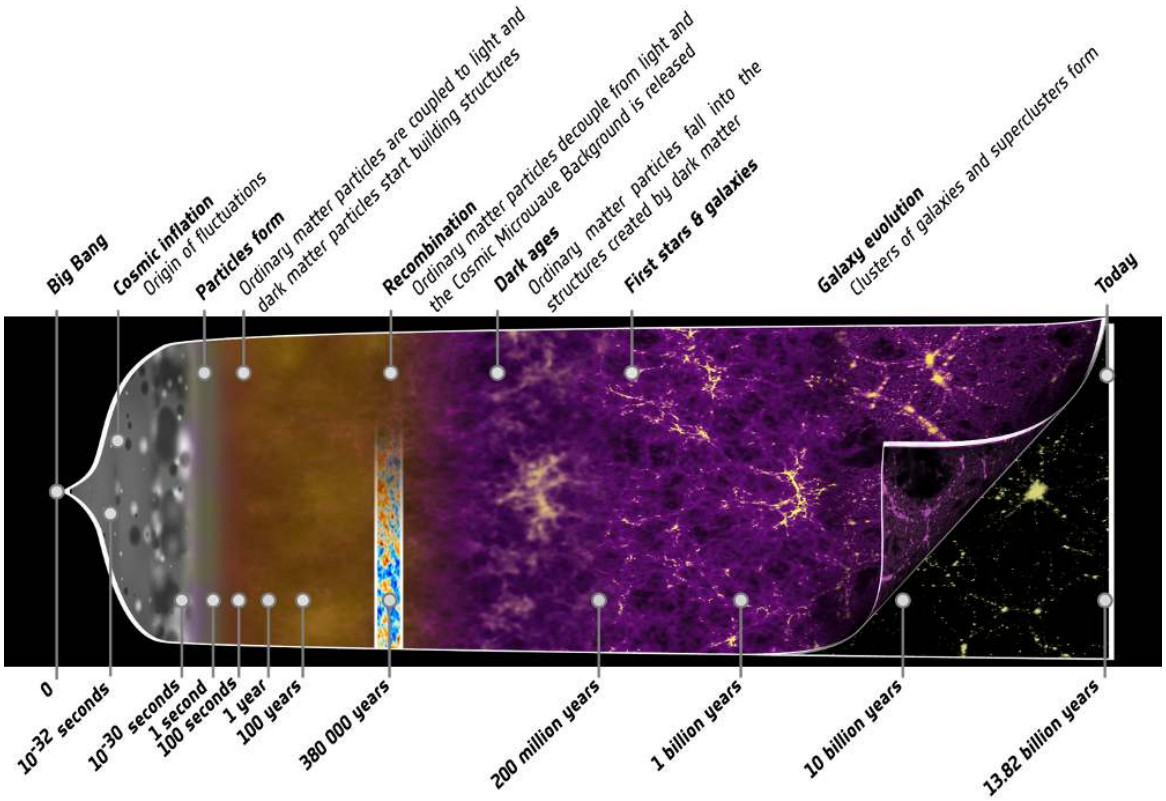

Figure 1: Illustration showing different periods of cosmic history, along with roughly what time they occurred since the Big Bang. A more detailed description is given here. (Adapted from figure by Planck / C. Carreau)

1.2. Brief history of the Universe

One of the most fundamental discoveries in cosmology is the fact that the Universe as a whole evolves with time. Cosmic history can be divided into several epochs, depending on what the dominant physical processes

were at the time. The transitions between these epoch are often very interesting physically, as they tend to imprint particular observable signatures.

The figure above illustrates the various phases of the Universe's history. A brief explanation of some of them is given below.

- **Big Bang** – The very beginning of the Universe ($t = 0$s).
- **Inflation epoch** – Brief period in the early Universe when space expanded exponentially ($t \approx 10^{-30}$ s).
- **Nucleosynthesis** – Protons and neutrons formed into the first nuclei ($t \approx 3$ min).
- **Recombination** – Nuclei and free electrons combined to form the first neutral atoms; the Universe became transparent to light ($t \approx 380,000$ years).
- **Dark ages** – The Universe was filled with neutral gas, and there were no luminous sources giving off light ($t \lesssim 200$ million years).
- **Cosmic Dawn** – The first stars and galaxies formed. Their light began to reionise the neutral gas in the Universe ($t \lesssim 1$ billion years).
- **Structure formation** – Galaxies began to form into large structures, forming a *cosmic web* of galaxy clusters, voids, and filaments. The galaxies themselves changed with time, as different stellar populations inside them evolved ($t \gtrsim 1$ billion years).

1.3. The state of the Universe today

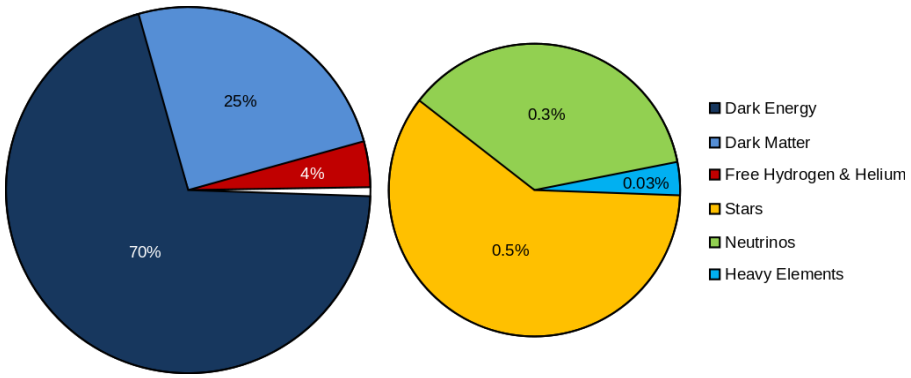

Figure 2: Chart showing the fraction of the cosmic energy density *today* taken up by different types of matter and radiation. These values are either *measured* from astronomical observations or *inferred* from a combination of observations and physical models. N.B. The chart on the right shows the fractions that belong in the very narrow white wedge in the plot on the left. (Wikipedia)

It is only in the last couple of decades that cosmologists have been able to measure the basic properties of the Universe with reasonable precision. We now know how much matter there is to better than 1% accuracy. We also know how fast the Universe is expanding with around percent-level accuracy too (the current best measurement is 67.36 ± 0.54 km/s/Mpc). As we'll see later, this lets us infer the age of Universe, which is currently best estimated to be 13.797 ± 0.023 Gyr.

The chart above shows our current best estimates of the composition of the Universe today (i.e. not its composition far into the past, which we'll see was quite different!). There are several different species of matter, radiation, and other types of energy, some of which are more abundant than others.

1.4. Olbers' paradox

Why is the night sky dark? We see plenty of stars, especially if we look through a telescope, but most of the sky is still dark. This should not be the case if the Universe is infinitely large and infinitely old though – every

direction you look in would, eventually, end up on a star. As a result, the whole sky would be bright. The statement of this problem is called *Olbers' paradox*.

Stars that are further away appear fainter though. How bright should we expect the sky to be if most of the stars are at great distances? Imagine that the Universe is filled with a uniform density of stars, n, on average. For simplicity, let's also assume that they all have identical luminosity, L. Each star will be observed on Earth with a flux $f = L/(4\pi r^2)$, where r is the distance of the star from Earth. If we draw a spherical shell of width dr at a distance r around Earth, we obtain a total number of stars $N = 4\pi r^2 n dr$ in that shell. The total flux of starlight from each shell is therefore

$$df_{\rm tot} = \frac{L}{4\pi r^2} 4\pi r^2 n dr = nLdr. \tag{1}$$

As you can see, it doesn't depend on distance! In fact, each shell contributes the same amount to the total flux received at Earth – while individual stars get fainter as the distance increases, the size of each shell (and therefore the number of stars per shell) increases by the same factor, and so the two effects cancel. If we integrate the flux over shells at all radii, we then get

$$f_{\rm tot} = \int df_{\rm tot} = \int_{r=0}^{\infty} n\,L\,dr = n\,L \int_{r=0}^{\infty} dr = \infty. \tag{2}$$

Needless to say, this is not what we observe, and so there must be something wrong with our assumptions. One thing we assumed is that all of the flux from all of the stars would reach us on Earth. This is not true if we consider that stars have a small, but non-zero, angular size, and are themselves opaque. Each star would block the light from any stars behind it, and so along each line of sight we would see only the light from the nearest star in that direction. This is an improvement – the brightness of the sky is no longer infinite – but it does not solve the paradox. Since every line of sight ends at a star, we would still have a uniformly bright night's sky about as bright as the surface of the Sun! *(Prove this in the first problem sheet!).*

Couldn't the darkness that we see be caused by intervening clouds of gas and dust that absorb the light? We certainly do see dust clouds in the sky, such as the Coalsack nebula (see the figure below), that block out the light from the stars behind them. But in an infinitely old Universe, the clouds would absorb so much energy over time that they would heat up and glow as brightly as the stars behind them!

The solution – obvious to us now – is that the Universe must either be finite in age, or expanding, or both, such that starlight has not had chance to reach Earth along every line of sight, even if it was filled uniformly with stars. Since the speed of light is finite, there is a maximum distance light could have travelled since the start of time. If the Universe is expanding, stars very far away would be expanding away from us faster than the light can travel towards us, and so their light would never reach us. These are both types of cosmic *horizon*, which we will learn more about in Section 4.

Figure 3: A picture of the Coalsack nebula. A more detailed description is given here. (Credit: ESO.)

1.5. Historical discovery of the expanding universe

Cosmology did not exist as a *scientific* field until after Einstein developed his theory of General Relativity. While some scientists did try to come up with 'world models' before then, they were often highly speculative.

Even the poet Edgar Allen Poe had a stab at it, and came surprisingly close to what we now know to be the right answer. Wikipedia has a timeline of cosmological theories that explains a lot of the historical ideas.

Few of the pre-20th Century thinkers had good scientific justifications or observational support for their theories however. Below is a brief timeline of the scientific discovery of the expansion of the Universe:

1908 Henrietta Swan Leavitt (astronomer and 'computer') discovered a relation between the pulsation period and absolute luminosity of Cepheid variable stars. This made it possible to measure the distance to very remote astronomical objects.

1913 Vesto Slipher (an astronomer) first measured the radial velocities of 'spiral nebulae'. It was not yet known that they were in fact galaxies separate from our own. By 1917 he had found that most spiral nebulae seemed to be moving away from Earth at quite high velocities.

1916 Albert Einstein published his General Theory of Relativity, including the Field Equations that connect the geometry of space-time to its matter/energy content.

1917 Willem de Sitter (an astronomer) found the first cosmological solution to the Field Equations – an empty 'vacuum' universe that seemed to expand.

1917 Einstein did not like this. He found another solution, called the Einstein static universe, that contained matter and had no expansion, but required a new 'cosmological constant' to be added into his equations to keep things from expanding. It was unstable, and Einstein himself thought it was ugly!

1922 Alexander Friedmann (a meteorologist) derived the Friedmann equations from Einstein's Field Equations. These showed how the Universe expands or contracts in the presence of a perfectly homogeneous fluid. Einstein himself reviewed Friedmann's calculations, but didn't realise their significance!

1924 Edwin Hubble (astronomer) observed Cepheid variables in spiral nebulae. He concluded that the nebulae were far too distant to be part of the Milky Way; they were galaxies separate from our own.

1927 Georges Lemaître (a Catholic priest and astronomer) derived a mathematical relation to explain how galaxies would seem to be travelling away from us in an expanding universe. His work was also not widely appreciated.

1929 Hubble plotted radial velocity data compiled by Milton Humason (his assistant) and Vesto Slipher against the distances to galaxies he had measured using Cepheid variables. He found that the further the galaxies were from us, the faster they seemed to be moving away. This was soon interpreted as evidence that the Universe is expanding. His estimate of the expansion rate was too large by a factor of seven however!

Further reading: Who discovered the expanding universe? (H. Kragh & R. W. Smith); Henrietta Leavitt – Celebrating the Forgotten Astronomer (AAVSO); The Contribution of V. M. Slipher to the Discovery of the Expanding Universe (C. O'Raifeartaigh); Edwin Hubble: Redshift increases with distance (Wikipedia)

1.6. Expansion and redshift

What does it mean for space to be expanding? We are used to thinking in a Newtonian way, where space itself is a fixed 'stage' on which particles and other objects move around. It seems much more natural to think of particles moving away from each other – expanding out from some location – than to think of the space itself changing. And yet that is what general relativity tells us is happening – and indeed *must* be happening. The expansion of the Universe is a relativistic effect, that we can only approximate using Newtonian physics.

Where is the centre of the Universe? The Universe does *not* have a centre! Every point is expanding away from every other point. Similarly, the Big Bang is *not* an explosion. Explosions travel out from a point, while the Big Bang happened *everywhere* at the same time.

One way to understand what's going on is by using the **raisin loaf model** as an analogy. This asks us to imagine a very large (actually, infinite!) lump of raisin loaf dough being baked in an oven. The raisins are embedded in the loaf, in the same way that galaxies are 'embedded' in space. As time goes by, the dough expands *homogeneously and isotropically*, i.e. at an equal rate everywhere, and in all directions. We can then

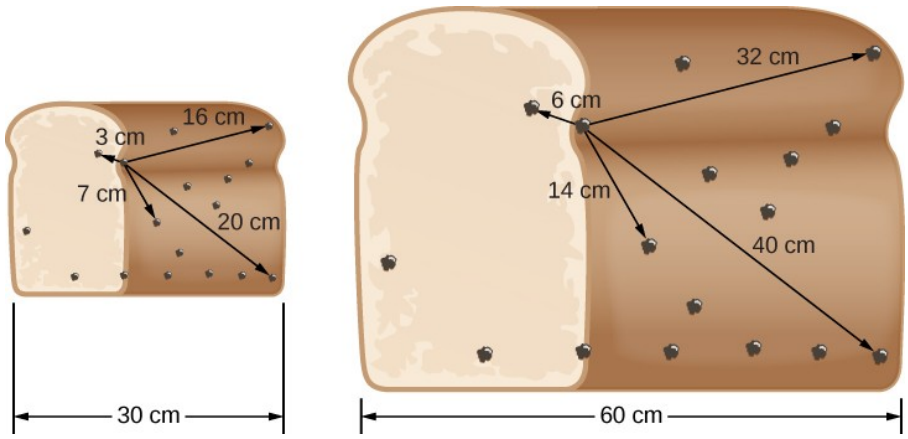

Figure 4: The 'raisin loaf' model provides an analogy for how the expansion of space is experienced by observers within the expanding space. A more detailed description is given here. (Credit: OpenStax Astronomy.)

ask: What happens to the distance between raisins (galaxies) as this happens? And how fast does each raisin appear to be moving away from the others?

For the first question, this analogy neatly shows that the distances should increase *proportionally* as the dough (space) expands. If raisins A and B are twice as far apart as raisins A and C at the start, they will *remain* twice as far apart as the dough expands.

For the second question, we can pick any raisin as a reference point and consider how the other raisins appear to be moving relative to it. We find that they are *all moving away*, and that the speed at which they are moving away (receding) is *proportional to their distance*. So, if we use raisin A as our reference, we will see raisin B moving away at twice the speed of raisin C (since B is twice the distance from A). This is because there is twice as much dough between A and B as between A and C, and so there is twice as much expansion (therefore the distance increased twice as much). It's exactly the same for galaxies embedded in an expanding space.

The raisin loaf model also gives some insight into why we don't need there to be a centre to the expansion. We could have chosen any raisin as our reference point, and would have found exactly the same behaviour. In the real Universe, all of space is expanding at the same rate in every direction, and so all galaxies are receding away from all other galaxies.

What expands when space expands? In principle, *everything* – the separation between every point in space and every other point increases with time. So certainly, the distance between two distant galaxies increases as space expands. But do the galaxies themselves expand? Or the stars within them? Do *we* expand?

First, we can work out how much an object of a certain length should be expanding. The current expansion rate of the Universe has been measured to be around 70 km/s/Mpc. That means every megaparsec of space expands by 70 km every second. For a 1.8m tall human, the corresponding expansion is $70\mathrm{km/s/Mpc} \times 1.8\mathrm{m} \times (3.086 \times 10^{22})^{-1}\mathrm{Mpc/m} \approx 4.1 \times 10^{-18}\mathrm{m/s}$. For comparison, a Hydrogen atom is around 10^{-10} m across, so clearly this is a tiny amount! Over a lifetime of 80 years, this would amount to a total change in height of only 10 nm.

A galaxy, however, is considerably larger than a human. If we take a typical spiral galaxy to be around 40 kpc in diameter, we get an expansion of 2.8 km/s. This seems small compared to the size of the galaxy, but would actually be observable with sensitive astronomical instrumentation! We don't see this expansion however.

The reason is that galaxies are *gravitationally bound*. Because they have so much mass concentrated in a relatively small volume, the force of their own gravity overcomes the cosmological expansion, and so they do not increase in size as the Universe expands. Similarly, humans are bound together by electromagnetic interactions between the atoms and molecules in our bodies that are vastly stronger than the cosmological expansion.

Further reading: The expanding universe (OpenStax Astronomy)

1.7. Spectra

One of the most important tools in astronomy is *spectroscopy*. Atoms and molecules emit and absorb electromagnetic radiation with distinctive lines in their spectra, corresponding to differences in energy levels of their electrons. By measuring spectra from distant sources, we can learn which atoms and molecules are present, and their abundance.

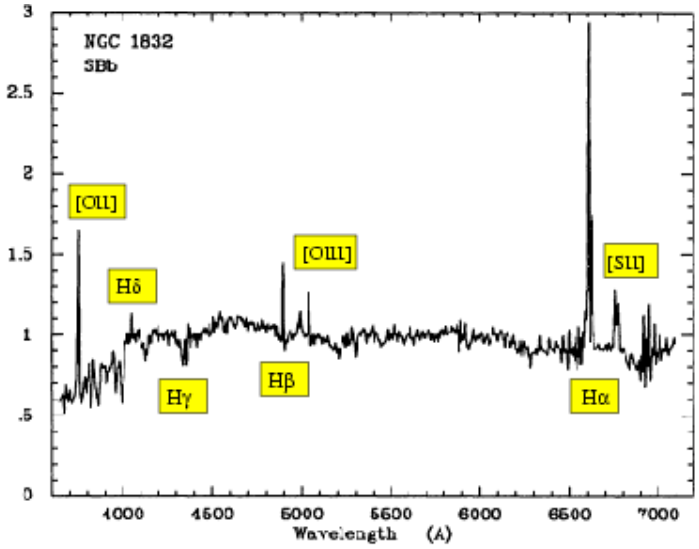

Figure 5: Spectrum of a spiral galaxy, with some spectral lines labelled. (Credit: Steward Observatory / R. Kennicutt)

As discussed above, the expansion of space causes the wavelength of light to increase. The greater the distance the light has travelled, the more stretching of space it will have experienced, and so the greater the change in wavelength. Emission and absorption lines will therefore be shifted to the redder (longer wavelength) end of the spectrum. By measuring the new wavelength of the lines, and comparing with what we know to be their 'rest-frame' wavelength (the wavelength when they were emitted), we can calculate how much space must have expanded since the light was emitted. This stretching factor is called the *redshift* (usually denoted by z), and can be calculated using

$$\lambda_{\rm obs}/\lambda_{\rm emit} = (1+z). \tag{3}$$

Spectra can be shifted for other reasons besides the stretching of space. Another is the *Doppler shift*, due to a relative velocity between the emitter and observer. The shift is given by

$$\lambda_{\rm obs}/\lambda_{\rm emit} = 1 + \frac{v}{c}, \tag{4}$$

where a positive velocity is defined as moving *away* from the observer. The spectral shifts we measure for astronomical objects is typically a combination of these two effects.

1.8. Expansion and the Hot Big Bang model

While the fact that the Universe is expanding was established in the late 1920's, there were several competing theories for how the expansion was happening and what this meant for the age of the Universe. One theory, the **hot Big Bang** model, stated that the Universe began in a hot and very dense state (mathematically, taking the form of a *singularity*), with its contents cooling rapidly as space expanded. This would mean that the Universe had finite age (a beginning, the Big Bang), and would look different (hotter, denser) in the past than the present (cooler, less dense).

Another theory, the **Steady State** model, said that the Universe was infinitely old, but was continually expanding at the same rate. As the expansion happened, new matter was being created throughout space at a slow but steady rate. Enough matter should be created to keep the density of the Universe constant in time,

cancelling out the diluting effect of the expansion. The properties of this Universe would not change with time; looking back into the past, the Universe would seem very similar to the Universe we see today.

Other models existed, such as **cyclic models** that allowed the Universe to expand from a Big Bang-like singularity, then collapse back in on themselves, and then 'bounce' back into a new expanding phase. This cycle would be repeated, possibly forever. Each phase of expansion would look like it came from a hot Big Bang until the Universe started to collapse again.

All of these theories were seriously entertained by scientists until around the late 1950's and early 1960's, when very distant objects called *quasars* were discovered. (We now know quasars to be the very bright, active nuclei of galaxies, associated with supermassive black holes.) The observations showed that the abundance of quasars changed significantly the further away (and therefore the further back in time) you look. This matched the predictions of the hot Big Bang model ("the Universe was denser in the past"), and did not support the Steady State theory ("the Universe has always looked very similar and doesn't change with time"). Further evidence came from the discovery in 1964 of the Cosmic Microwave Background – remnant radiation from a very hot previous phase of cosmic history. This was a key prediction of the hot Big Bang model, but not the Steady State model.

Further reading: Steady-state model (Wikipedia)

Learning outcomes:

- What is cosmology?
- What are the major epochs in the Universe's history?
- What physical processes are important during each of these epochs?
- What does it mean for space to be expanding?
- What does it mean for a universe to be homogeneous and isotropic?
- What is Olbers' paradox, and how is it resolved?
- How does the expansion of space affect the frequency of EM radiation?
- How are spectra used to measure redshift?
- What is the evidence for the hot Big Bang model?

2. Geometry and distance

In the last section we learned about the expansion of space, and how observers perceive that expansion. In this section, we will develop a mathematical understanding of the expansion, and how it affects how we define and measure distances. We will use the language of General Relativity to define some of the relevant quantities, for example by defining a mathematical object called the *spacetime metric* that tells us how to measure distances even when space is expanding. We will also learn about the possible geometries of the Universe – space itself can *curve*, as well as expand. We will examine the different ways it is allowed to curve.

Reading for this topic:

- *An Introduction to Modern Cosmology (A. Liddle), Chapters 2, 4, and 5*
- *An Introduction to Modern Cosmology (A. Liddle), Adv. Top. 1: General Relativistic Cosmology*

2.1. Recession velocity and Hubble's Law

Recall from the previous section that the expansion of space (if it is expanding homogeneously and isotropically) causes all distances to increase *proportionally*, i.e. by the same factor. As we saw in the raisin loaf model, this means that objects that are further away are expanding away from us faster than objects that are nearby – the increase in their distance per unit time is proportional to their current distance.

The rate of increase in distance due to the expansion can be interpreted as a velocity that we call the *recession velocity*. Objects at larger distances have larger recession velocities – an observation that we call *Hubble's Law*:

$$v = H_0 d. \tag{5}$$

Here, v is the recession velocity and d is the distance. *The recession velocity is proportional to the distance,* as we expected from thinking about the raisin loaf model. The constant of proportionality, H_0, is called the *Hubble parameter*.[1] It has units of inverse-time, but is most commonly quoted in a more convenient mixture of units: km/s/Mpc (speed per unit distance).

This relation was discovered *observationally* by Hubble – he plotted Slipher and Humason's measurements of the recession velocity of galaxies against his own measurements of distances from Cepheid variables (using the distance measurement technique of Swan Leavitt), and saw a linear trend between them. Prior to this, Lemaître had discovered the relation *theoretically*, by calculating how distances would change in an expanding space. The Hubble Law is therefore sometimes called the *Hubble-Lemaître Law* instead.

Note that if the Universe is expanding, the distance of an object from us is always increasing, and so the recession velocity is a positive quantity. The radial velocities we measure from spectra can sometimes be negative however, if the galaxies have a *peculiar velocity* towards us that counter-acts the expansion and causes a big enough Doppler shift. These galaxies are seen with a blueshift. (Of course the peculiar velocity could also be pointing in the direction away from us, in which case the galaxy would seem to have an even bigger redshift.)

Since the recession velocity increases with distance, it gets very large for very distant objects. First of all, this means that every distant galaxy has a redshift, and never a blueshift – at small distances, where the recession velocity is small, it is easy for galaxies to have a large enough peculiar velocity to cancel it out and appear to be moving towards us. At large distances, the recession velocity due to the expansion of space always wins over the Doppler shift.

We can see from the Hubble Law that as the distance, d, becomes larger and larger, the velocity also gets larger and larger – with no limit! In fact, when $d > c/H_0$, this implies that the recession velocity $v > c$! This is perfectly fine – plenty of galaxies have been found at such large distances. Remember that the recession velocity is not a *real* velocity – it's just a way of rewriting the redshift in a different way, by analogy with the Doppler shift. It is not really a Doppler shift however, so not really a velocity, and so doesn't need to follow the same rules as actual velocities (such as being less than the speed of light).

[1] It is not actually constant, as we will see in later sections.

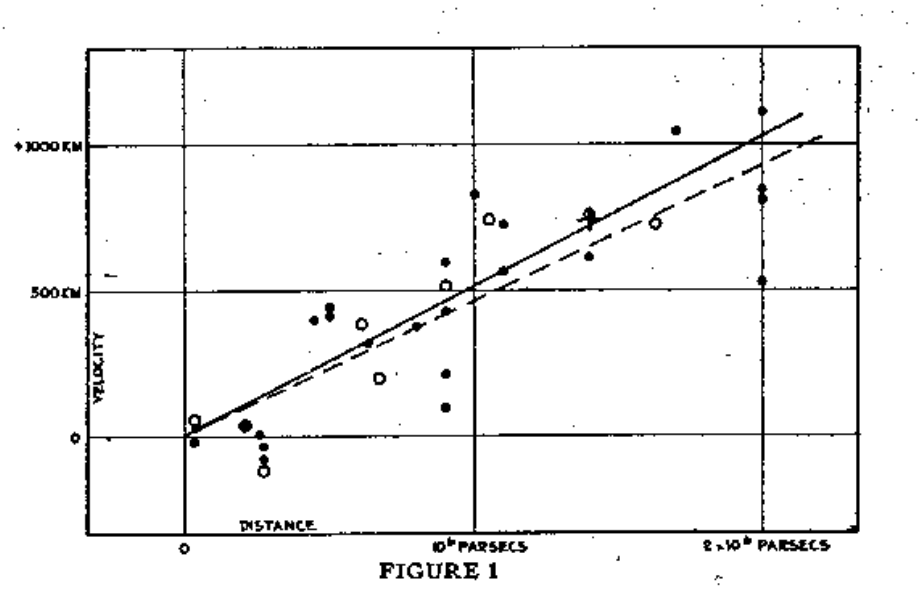

Figure 6: The 'discovery plot' for the expansion of the Universe. It shows distance (x axis) plotted against velocity (y axis) for a number of nearby galaxies. (Credit: E. Hubble)

Further reading: Hubble's Law (HyperPhysics); Hubble's Law (Khan Academy)

Taylor expansions

Almost any function can be written as a polynomial with an infinite number of terms,

$$f(x) = \sum_{n=0}^{\infty} b_n (x-a)^n, \tag{6}$$

where b_n are a set of coefficients that are to be found, and a is some reference point that we are free to choose. If the infinite series converges rapidly (i.e. the terms in the polynomial become smaller and smaller as n increases), we can accurately *approximate* the function using only the first few terms in the series.

This is the logic behind a *Taylor series expansion*. By taking the first few terms in an infinite series expansion around a suitable reference point (let's call it a), we can often get a very good and simple approximation to a function that is valid within some region around the reference point.

To calculate the value of each coefficient, we can take nth-order partial derivatives, i.e. $\partial^n f(x)/\partial x^n$, and then set the argument to the reference value, $x = a$. This isolates the coefficient for the nth term, up to a numerical factor: lower-order terms ($< n$) are removed by differentiation, while higher-order terms ($> n$) cancel when $x = a$. (You can prove this to yourself by repeatedly differentiating the expression above and evaluating at $x = a$ to see which terms remain.)

Using this insight, we can write a general expression for a Taylor series expansion around a point a as

$$f(x) \approx f(a) + \frac{1}{1!}\left.\frac{\partial f}{\partial x}\right|_a (x-a) + \frac{1}{2!}\left.\frac{\partial^2 f}{\partial x^2}\right|_a (x-a)^2 + \frac{1}{3!}\left.\frac{\partial^3 f}{\partial x^3}\right|_a (x-a)^3 + \ldots \tag{7}$$

Further reading: Taylor & Maclaurin polynomials (Khan Academy)

2.2. Parallax distance

The distance measurements used by Hubble were made by using Cepheid variable stars as "standard candles". Another way of measuring astronomical distances is by observing the *parallax*. This is the maximum change in the angular position of an object on the sky over the course of the year; as the Earth orbits the Sun, our viewing angle changes very slightly, leading to a tiny shift in the angular position of an object from one side of our orbit to the other. The further away the object is, the less its angle will change from one side of the Earth's orbit to another.

The parallax (change in angle) is most conveniently expressed in units of arcseconds. The measured parallax in arcseconds is related to the distance in parsecs by:

$$\frac{d_{\rm par}}{\rm pc} = \left(\frac{\Delta\theta}{\rm arcsec}\right)^{-1}. \tag{8}$$

So, an object that has a parallax of 0.5 arcsec is at a distance of 2 pc, while an object with a parallax of 1 μas (1 micro arcsec = 10^{-6} arcsec) is at a distance of 1 Mpc.

Parallax distances are only useful for relatively nearby objects, since the parallax angles that we would need to measure become really tiny for objects at cosmological distances. Measuring a parallax of 1 μas is very challenging, but can be done.

Further reading: Parallax (Wikipedia); Stellar distance using parallax (Khan Academy)

2.3. Proper vs comoving coordinates

Measuring distances in an expanding Universe is complicated, since space itself is constantly changing. We can choose a set of spatial coordinates to make this a bit more straightforward though.

First, imagine looking at a 'snapshot' of the Universe, frozen at a fixed time. What would the distances between the galaxies be? Those distances are called the *proper* distances, and increase as space expands. Importantly, *we can't directly measure proper distances!* How could we? Even if we did have a ruler big enough, we wouldn't be able to tell when one end had reached the galaxy we were trying to measure the distance to; we would have to wait for signals from the end of the ruler to travel back to us at the speed of light (or slower). By the time those signals made it back, the Universe would have expanded more, and the ruler wouldn't be touching the galaxy any more! (We will see how the proper distance is related to quantities we *can* measure later on.)

Proper coordinates are quite inconvenient, as the distances between galaxies and other distant objects are constantly changing, simply because the Universe is expanding. By defining coordinates that also expand, we can factor out the expansion. This is shown in the diagram below – in this coordinate system, the galaxies always stay in the same place on the coordinate grid, even though the (proper) distance between them is increasing.

These coordinates, with the expansion factored out, are called *comoving coordinates*. Proper coordinates are related to comoving coordinates by a factor of $a(t)$, like so:

$$\vec{x}_p = a(t)\vec{x}, \tag{9}$$

where $\vec{x}_p$ is a vector in proper coordinates and $\vec{x}$ is the same vector in comoving coordinates.

Galaxies and other objects *can* move within a comoving coordinate system, but this movement will not be due to the expansion (since that has been factored out). Instead, it will be due to their *peculiar velocity*, if they have one.

2.4. Measuring distances: the space-time metric

In a 3D Euclidean space, the *line element* – the infinitesimal unit along a straight line that connects two points – is given simply by $dr^2 = dx^2 + dy^2 + dz^2$, where dx, dy, and dz are the (infinitesimal) distance intervals in each dimension.

In 4 dimensions, we can define an analogous line element along a curve separating two points (or *events*) in space and time. There are some important differences when a time dimension is added though. In the

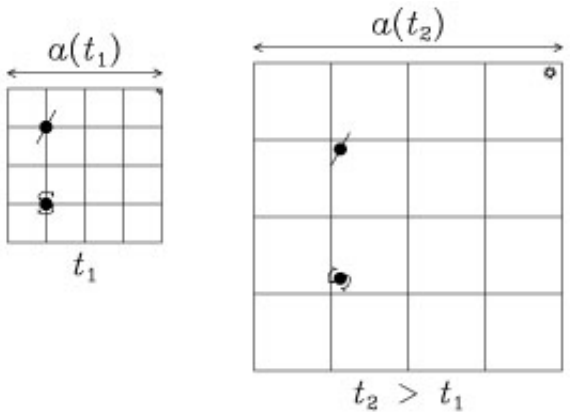

Figure 7: Relationship between comoving and proper coordinates. The galaxies stay at the same comoving coordinates as space expands (i.e. they stay fixed to the same points on the comoving grid). The proper distance between them increases however (the grid itself expands, so all proper distances increase). (Credit: E. Bertschinger)

Minkowski space-time of special relativity, the line element becomes

$$ds^2 = -c^2dt^2 + dx^2 + dy^2 + dz^2 = -c^2dt^2 + dr^2. \tag{10}$$

The time component is scaled by the speed of light, and contributes *negatively* to the space-time interval. A more detailed explanation is beyond the scope of this course, but light travels on paths where the space-time interval is always $ds^2 = 0$, while normal matter travels on paths where $ds^2 < 0$.

It is useful to define a quantity called the space-time *metric*. This describes how intervals in the individual dimensions contribute to the total space-time interval ds. The metric is a special type of 4×4 matrix called a *tensor*, and for a Minkowski space-time can be written as

$$g_{ab} = \begin{pmatrix} -c^2 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{pmatrix}. \tag{11}$$

This can be used to define the line element like so:

$$ds^2 = \sum_{a=0}^{3}\sum_{b=0}^{3} g_{ab}dx^a dx^b, \tag{12}$$

where indices a and b run over the 4 space-time dimensions, and we have defined a space-time vector (4-vector) as $x^a = (t, x, y, z)$.

Einstein summation convention

Lots of calculations in special and general relativity involve summing over one or more sets of indices on 4-vectors and tensors (as in the expression for ds^2 above). This gets very cumbersome for anything other than simple calculations.

To help with this, we can introduce some new notation called the *Einstein summation convention*. When using this convention, if you see a pair of repeated indices, the rule is that you're supposed to sum over them. For example, the example above can be rewritten as

$$ds^2 = g_{ab}x^a x^b. \tag{13}$$

Index a is repeated, which means we should do a sum over all values of a, from $0..3$. Likewise for b. Sometimes people use indices i, j, k to denote only the 3D (spatial) part of a 4-vector, i.e. components $1..3$. So, writing $x^i x^i$ would denote the sum $x^2 + y^2 + z^2$. Latin letters early in the alphabet (e.g. a, b, c) or Greek letters (e.g. μ, ν) normally denote 4D indices however.

There is also significance to whether the indices are *raised* (e.g. x^a) or *lowered* (e.g. x_a). This has to do with whether the 4-vector is a covariant or contravariant vector. You can convert between the two by multiplying by the metric, e.g. $x_a = g_{ab}x^b$.

Further reading: Einstein notation (Wikipedia); Covariant and contravariant vectors (Wikipedia); Covariant and contravariant vectors (animation).

2.5. Friedmann-Lemaître-Robertson-Walker (FLRW) metric

The metric for a homogeneous and isotropic expanding spacetime is

$$g_{ab} = \begin{pmatrix} -c^2 & & & \\ & a^2 & & \\ & & a^2 & \\ & & & a^2 \end{pmatrix}. \tag{14}$$

This is called the Friedmann-Lemaître-Robertson-Walker (FLRW) metric, sometimes just called the FRW metric (which is a bit mean to Lemaître). It is named after the people who came up with the corresponding solutions to Einstein's equations, and the people who wrote it down in this particular mathematical form.

Writing this as a line element, we obtain

$$ds^2 = -c^2dt^2 + a^2(t)dl^2, \tag{15}$$

where dl is the spatial part of the line element.

It is important to note that the scale factor, $a(t)$, is the same in *all* spatial directions. In other words, space stretches the same amount in every direction – the expansion is *isotropic*. What's more, note how the scale factor only depends on time, but not position. The stretching of space happens by the same amount in every location – the expansion is *homogeneous*.

So far we have been working in Cartesian coordinates (x, y, z) for the spatial part of the metric, but there is no reason why we need to. It is often more convenient to work in *spherical polar coordinates* (r, θ, ϕ) for example. The line element in spherical polars becomes

$$ds^2 = -c^2dt^2 + a^2(t)\left(dr^2 + r^2d\theta^2 + r^2\sin^2\theta d\phi^2\right). \tag{16}$$

As another point of notation, we will often drop the explicit dependence on time of quantities such as the scale factor, so we can write $a = a(t)$.

2.6. Geometry of space: open, closed, and flat universes

We are used to doing geometry in a Euclidean space; parallel lines never meet, and the angles of a triangle add up to $180°$. This is not true in the other types of geometry however! In a positively-curved space, the angles of

a triangle add up to greater than $180°$, and parallel lines always meet at some point! Conversely, in hyperbolic spaces, the angles of a triangle add up to less than $180°$ and parallel lines diverge.

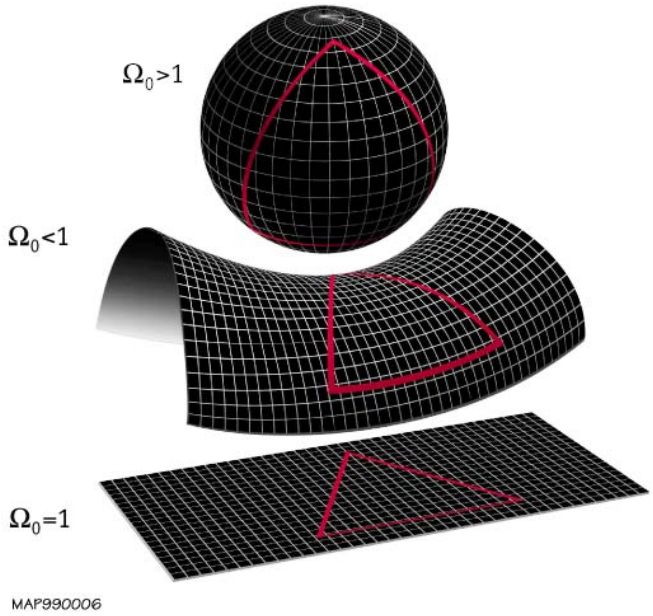

Figure 8: Illustrations of different type of curvature. Top is a closed surface, middle is an open surface, and bottom is a flat surface. These illustrations show curved 2D surfaces embedded in a 3D space; when we talk about the curvature of the Universe, we are talking about a 3D 'surface' embedded in a 4D space. (Credit: Wikipedia)

How should we visualise these different types of geometry? It's easiest to think about the 2D analogues of these spaces embedded in the 3D space that we're used to. The 2D analogue of 3D Euclidean space is a flat plane, extending to infinity in the x and y directions. You can just imagine an infinite sheet of paper. The 2D analogue of a hypersphere is the surface of a sphere (not the sphere itself!). You can imagine a globe for this one. The 2D analogue of a hyperbolic space is the surface of a saddle. This one is a bit more difficult, as you have to imagine the saddle extending off the infinity in all directions.

Importantly, all of these spaces are still *homogeneous and isotropic*. It doesn't matter where you are on a hypersphere for example; space still looks like same in every direction, and has the same properties at every point.

2.7. Space-time metric with curvature

With spatial curvature, the FLRW metric becomes

$$ds^2 = -c^2dt^2 + a^2(t)\left(\frac{dr^2}{1-kr^2} + r^2d\theta^2 + r^2\sin^2\theta d\phi^2\right). \tag{17}$$

Note how k has dimensions of $(\text{distance})^{-2}$ to make the units work out. Depending on the sign of k, we arrive at the three different possibilities for the curvature of space:

- $k = 0$: Flat, or Euclidean.
- $k > 0$: Positive curvature, i.e. a *closed* universe. The 3 space dimensions have the shape of a hypersphere.
- $k < 0$: Negative curvature, i.e. an *open* universe. The 3 space dimensions have a hyperbolic (saddle-like) shape.

2.8. Useful unit conversions

- 1 pc = 3.086×10^{16} m

- 1 yr = 3.154×10^7 s ($\approx \pi \times 10^7$ s)
- $c/(100\,\mathrm{km/s/Mpc}) = 2997.9$ Mpc (≈ 3000 Mpc)

Learning outcomes:

What are the definitions of 'recession velocity' and 'Hubble's Law'?

What are proper and comoving coordinates?

What is the FLRW metric and the corresponding line element for an expanding universe?

How can a Taylor series be used to approximate a function?

What does it mean for space to be curved?

What are the different types of curved spaces, and what are their properties?

3. Friedmann equation

In this section, you will learn about the most important equation in cosmology – the Friedmann equation, which describes how the expansion rate of the Universe depends on its matter content and geometry. You will learn two different ways of writing the Friedmann equation (one with units, one that is dimensionless), and how to solve the equation in a few different cases, so that you can calculate the age of the universe and its future fate. This section also includes an intuitive, non-relativistic derivation of the Friedmann equation based on Newtonian gravity.

Reading for this topic:

- *An Introduction to Modern Cosmology (A. Liddle), Chapters 3, 5, 6, and 8.*
- *An Introduction to Modern Cosmology (A. Liddle), Adv. Top. 1: General Relativistic Cosmology.*

3.1. The Friedmann equation

In previous sections we defined the scale factor, $a(t)$, which describes the factor by which the Universe has expanded at any given time. But how can we determine the scale factor and its time evolution in a given universe?

General Relativity gives us the tools to derive an *evolution equation* for the scale factor. GR links the geometry of the Universe and how it is changing to the matter/energy content of the Universe. Different types and configurations of matter and energy cause the geometry to respond and involve in different ways.

For a homogeneous and isotropic space-time, the equation that describes this link between geometry and energy content is the *Friedmann Equation*:

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G\rho}{3} - \frac{kc^2}{a^2} + \frac{\Lambda c^2}{3} \tag{18}$$

We will derive this later on, but for now let's define each of the terms. First, we see on the left a ratio of the time derivative of the scale factor, $\dot{a} = da/dt$, to the scale factor itself. On the right, the first term is the energy density, ρ, scaled by a factor; the second term depends on the curvature of space, k; and the third term involves a cosmological constant, Λ, which we will study in more detail later.

Note that only the factors of $a = a(t)$ and $\rho = \rho(t)$ depend on time in this equation. The curvature, k, and cosmological constant, Λ, are fixed, as are the physical constants G and c.

Importantly, the density ρ here is an *energy* density, not a mass density. It's very useful to use units where $c = 1$ though, so in practice we don't need to distinguish (e.g. since $E = mc^2$ can be written $E = m$ in these units). This density is the sum of the densities of all of the matter and radiation fields in the Universe (we will break this into its separate pieces later).

It is sometimes convenient to interpret the cosmological constant term as an energy density also, by defining

$$\frac{\Lambda c^2}{3} = \frac{8\pi G\rho_\Lambda}{3}. \tag{19}$$

We will discuss this interpretation in Section 4 – this simple rewriting of this term turns out to be related to one of the most fundamental problems in all of physics!

3.2. Hubble parameter and expansion rate

We previously met the Hubble parameter, H_0, in the context of the Hubble Law. It tells us the rate of expansion of the Universe today. We can relate this to the scale factor and its time derivative like so:

$$H_0 = (\dot{a}/a)_{t=t_0}. \tag{20}$$

That is, the Hubble parameter is the time derivative of the scale factor divided by the scale factor, all evaluated at t_0 (today). We can use a very similar definition to find the *expansion rate* at any given time:

$$H(t) \equiv \dot{a}/a. \tag{21}$$

As you can see, the left hand side of the Friedmann equation is nothing but the expansion rate squared! So the expansion rate depends on the matter/energy density of the Universe, its curvature, and the size of the cosmological constant.

3.3. Critical density and curvature

In the previous section, we discussed a way of classifying universes according to how space is curved. The sign of the curvature parameter, k, determines whether the universe is open, closed, or flat.

We can relate the curvature to the density and expansion rate, using a quantity called the *critical density*, $\rho_{\rm crit}$. For a given expansion rate, a universe that has a lower density than $\rho_{\rm crit}$ will be *open*, while one with a greater density than $\rho_{\rm crit}$ will be closed. A universe at *exactly* $\rho_{\rm crit}$ will be flat.

This can be understood as a prediction of General Relativity. In GR, as John Wheeler famously said, "spacetime tells matter how to move; matter tells spacetime how to curve". The greater the density of matter/energy in our spacetime, the more spacetime will curve around it. So, universes with a high density have positive curvature, $k > 0$, i.e. there is enough matter that they curve back in on themselves (closed universe). Universes with a low density have negative curvature, $k < 0$ (open universe). And universes where the density is 'just right', i.e. *exactly* balanced to equal $\rho_{\rm crit}$, are flat ($k = 0$).

The *critical density of the Universe today*, $\rho_{\rm crit,0} = \rho_{\rm crit}(t_0)$, is a useful reference value in cosmology. If we ignore the cosmological constant (or write it as a mass/energy density and add it into ρ), the Friedmann equation today in a flat (critical density) universe becomes

$$H_0^2 = \frac{8\pi G}{3}\rho_{\rm crit,0}. \tag{22}$$

We can rearrange this to get an expression that allows us to calculate the critical density for a given expansion rate,

$$\rho_{\rm crit,0} \equiv \frac{3H_0^2}{8\pi G}. \tag{23}$$

Let's see how we can use the critical density as a reference value. Instead of working with all of the different dimensionful prefactors for each term in the Friedmann equation, we can rewrite the whole equation in a (mostly) dimensionless way by dividing through by the critical density:

$$\frac{H^2(t)}{H_0^2} = \frac{8\pi G\rho/3}{8\pi G\rho_{\rm crit,0}/3} - \frac{kc^2/a^2}{H_0^2} \tag{24}$$

$$= \frac{\rho}{\rho_{\rm crit,0}} - \frac{kc^2}{a^2H_0^2} \tag{25}$$

Note how we chose to divide the curvature term by H_0^2 instead of $8\pi G\rho_{\rm crit,0}/3$ because it results in a neater expression. We have also continued to treat the Λ term as an energy density, ρ_Λ, so we can hide it away inside the ρ term. Next, let's evaluate this expression at $t = t_0$. We get:

$$\frac{H^2(t_0)}{H_0^2} = 1 = \frac{\rho(t_0)}{\rho_{\rm crit,0}} - \frac{kc^2}{H_0^2}. \tag{26}$$

Let's define the *fractional* energy density,

$$\Omega = \frac{\rho}{\rho_{\rm crit,0}}. \tag{27}$$

Using this definition, we can write the equation from above as

$$\Omega_0 - \frac{kc^2}{H_0^2} = 1. \tag{28}$$

It then makes sense to define an analogous parameter for the curvature,

$$\Omega_k \equiv 1 - \Omega_0 = -\frac{kc^2}{H_0^2}. \tag{29}$$

In other words, Ω_k is the fractional difference between the total energy density and the critical density. Note how Ω_k is positive when k is negative and vice versa.

3.4. Change in energy density as space expands

We saw in Section 1 that there are many different types of matter and radiation that contribute to the total energy density of the universe. As you might expect, their density typically *decreases* as the universe expands – because distances stretch according to the scale factor, $a(t)$, volumes increase by a factor a^3.

The way that the density changes depends on the type of energy we are dealing with. The simplest to understand is regular matter, like atoms, stars, galaxies and so on. In the later stages of cosmic history, matter is mostly conserved (neither created nor destroyed), and so the total mass of the different types of matter is constant. As the matter expands with the expansion of space, however, the volume that it takes up increases, like a^3. The density of matter therefore scales as

$$\rho_m \propto 1/a^3. \tag{30}$$

The density of cold dark matter is also expected to scale in the same way.

Radiation, like photons and highly relativistic particles, behaves differently. Its energy density is governed by its temperature, T, scaling as $\rho_r \propto T^4$. Taking a typical photon of temperature T, its energy is related to its wavelength by $k_B T \approx hc/\lambda$. We know how wavelength scales with the expansion of the universe ($\lambda \propto (1+z) \propto 1/a$), and so we can see that the energy density of radiation should scale as

$$\rho_r \propto 1/a^4. \tag{31}$$

There are other types of energy density that scale differently to matter and radiation, which we will discuss in later sections. But for now, matter and radiation, plus curvature and the cosmological constant, cover most of the possibilities that we will be interested in.

We can define a fractional energy density for any type of matter or energy that we like. For example, the fractional energy density of matter today is defined as $\Omega_m = \rho_m/\rho_{\mathrm{crit},0}$. Since we've normalised $a = 1$ at $t = t_0$, we can write

$$\rho_m \propto \Omega_m a^{-3}. \tag{32}$$

Similarly, for radiation,

$$\rho_r \propto \Omega_r a^{-4}. \tag{33}$$

Note how the matter and radiation energy density, ρ_m and ρ_r, depend on $a(t)$, which depends on t – their energy densities are time-dependent quantities. The fractional energy densities Ω_m and Ω_r are *not* time-dependent however – they are both evaluated at $t = t_0$, and so are constant.

Recall the definition of Ω_0 as being the total fraction of energy density in all types of matter and radiation at $t = t_0$. We can break this up into a sum of matter and radiation parts, plus a cosmological constant, plus the curvature term, to obtain

$$\Omega_m + \Omega_r + \Omega_k + \Omega_\Lambda = 1. \tag{34}$$

This lets us rewrite the Friedmann equation in a simpler form, with only a single dimensionful parameter, H_0:

$$H^2(a) = H_0^2 \left(\Omega_m a^{-3} + \Omega_r a^{-4} + \Omega_k a^{-2} + \Omega_\Lambda\right). \tag{35}$$

This is a very useful form for the Friedmann equation, as we only need to remember the fractional densities of matter, radiation etc. **You should remember this form of the equation!**

3.5. Matter-only solution to the Friedmann equation

Now we are in a position to solve the Friedmann equation for different compositions of the universe, to see how it affects the scale factor, $a(t)$. We'll start with a simple one, involving a flat universe that *only* contains matter.

First, we write down the Friedmann equation with only a matter component.

$$\left(\frac{\dot{a}}{a}\right)^2 = H_0^2 \Omega_m a^{-3}. \tag{36}$$

Since this universe only has matter, $\Omega_m = 1$. Taking the square root of both sides of the equation,

$$\frac{1}{a}\frac{da}{dt} = \pm H_0 a^{-\frac{3}{2}}, \tag{37}$$

we can then shuffle all terms that depend on a to one side to obtain

$$H_0 dt = \pm \frac{a^{\frac{3}{2}} da}{a} = \pm a^{\frac{1}{2}} da. \tag{38}$$

We now have everything we need to solve the equation – the left-hand side depends only on t, and the right-hand side only depends on a, so all we have to do is integrate both sides with the appropriate limits. Logically, we can start our integration at the Big Bang, where the scale factor is 0. What is the corresponding time coordinate of the Big Bang? Actually, we are free to choose – as long as we are consistent, we can choose whatever zero point for the time we like. We can even label the Big Bang as being at a negative time, so 'today' is at $t = 0$. For simplicity, we're going to choose the origin of the time coordinate, $t = 0$, to be at the Big Bang though. Our other integration limits will be the value of the scale factor a at some time t, so we can write

$$\int_0^t H_0 dt' = \pm \int_0^a (a')^{\frac{1}{2}} da', \tag{39}$$

where we have added primes to the integration variables to avoid confusion with the integration limits. Both integrals are straightforward; we obtain

$$H_0 \Big[t\Big]_0^t = \pm \left[\frac{2}{3} a^{\frac{3}{2}}\right]_0^a. \tag{40}$$

Inserting the integration limits, and taking the positive branch of the solution, we obtain

$$H_0 t = \frac{2}{3} a^{\frac{3}{2}} \tag{41}$$

$$\implies a = \left(\frac{3 H_0 t}{2}\right)^{\frac{2}{3}}. \tag{42}$$

This is the solution for the scale factor in a matter-dominated universe.

3.6. Interchangeability of time, redshift, and scale factor

In a universe that is expanding, the scale factor a always increases with time. Since $a(t)$ is monotonically increasing, the inverse relation $t(a)$ is also monotonically increasing, and so there is a unique value of a for every t and vice versa. This means that we can use the scale factor as an alternative time coordinate if we like.

Since the relation $a(z) = 1/(1+z)$ is also monotonic, this means that we can use the redshift as a time coordinate too (where increasing z implies that we are looking further back in time).

Not all cosmological models continue expanding forever, and so we can't always use a and z as alternative time coordinates. But since most of the universes we will be studying are continuously expanding, we will often find this property useful.

3.7. Age of the Universe

Given a cosmological model, we can use the Friedmann equation to calculate the age of the Universe (the time between the Big Bang and today). We start by taking the square root of the Friedmann equation to obtain

$$\frac{1}{a}\frac{da}{dt} = \pm H(a). \tag{43}$$

Rearranging and then integrating, we get

$$\frac{da}{aH} = dt \implies \int_0^1 \frac{da}{aH} = \int_0^{t_0} dt \tag{44}$$

$$\implies [t]_0^{t_0} = t_0 = \int_0^1 \frac{da}{aH}. \tag{45}$$

Note the integration limits. The scale factor integral runs from $a = 0$ (the Big Bang) to $a = 1$ (the scale factor today). The time integral runs from the corresponding times: $t = 0$ for the Big Bang, and $t = t_0$ for today. We have also chosen the positive sign from the square root. Now all we need is to plug in a model for the Hubble function, $H(a)$, and evaluate the final integral.

For a purely matter-dominated Universe, $H(a) = H_0 a^{-\frac{3}{2}}$. The integral becomes

$$\int_0^1 \frac{da}{a \times H_0 a^{-\frac{3}{2}}} = \frac{1}{H_0}\int_0^1 a^{\frac{1}{2}} da = \frac{2}{3H_0}[a^{\frac{3}{2}}]_0^1 = \frac{2}{3H_0} \tag{46}$$

3.8. Matter, curvature, and the fate of the Universe

We discussed above how universes with higher or lower energy densities than the critical density will be closed (positive curvature) or open (negative curvature). Let's now study a universe containing only matter and some amount of curvature to see how this affects the scale factor.

The Friedmann equation in this kind of universe can be written as

$$H^2(a) = H_0^2\left(\Omega_m a^{-3} + \Omega_k a^{-2}\right). \tag{47}$$

Taking the square root, cancelling some factors of a, and rearranging, we obtain

$$H_0 dt = \pm\frac{da}{a\sqrt{\Omega_m a^{-3} + \Omega_k a^{-2}}} \tag{48}$$

$$= \pm\frac{da}{\sqrt{\Omega_m a^{-1} + \Omega_k}}. \tag{49}$$

A solution to the integral of the right-hand side is not looking particularly obvious as it is currently written. But we can rearrange further to obtain

$$H_0 dt = \pm\frac{1}{\sqrt{\Omega_m}}\left(\frac{a}{1 + (\Omega_k/\Omega_m)a}\right)^{\frac{1}{2}} da. \tag{50}$$

Now we need to use some intuition! Integrals of this form often have trigonometric functions as their solutions. We must also be careful of the sign of Ω_k, since that can change the sign in the denominator and therefore the solution to the integral. While there is an analytic solution to this equation, it is very ugly! It is also difficult to invert, so we can't write down the function $a(t)$, which is what we really want.

A more elegant solution can be obtained by considering a *parametric solution*, where we rewrite the integral in terms of a proxy parameter that we can then separately relate to t. The *conformal time* happens to be a useful parameter for this purpose. Recall that the conformal time, τ, is defined through the relation $dt = ad\tau$. We now have two separate integrals to do:

$$H_0 d\tau = \pm\frac{1}{\sqrt{\Omega_m}}\left(\frac{1}{a + (\Omega_k/\Omega_m)a^2}\right)^{\frac{1}{2}} da \tag{51}$$

$$dt = ad\tau. \tag{52}$$

Here are the solutions. For a closed universe ($k > 0$):

$$a(\tau) \propto (1 - \cos\tau) \tag{53}$$
$$t(\tau) \propto (\tau - \sin\tau). \tag{54}$$

For an open universe ($k < 0$):

$$a(\tau) \propto (\cosh\tau - 1) \tag{55}$$
$$t(\tau) \propto (\sinh\tau - \tau). \tag{56}$$

You can verify that these are solutions to the Friedmann equation by differentiating them.

Let's inspect these solutions. First, let's note that conformal time is a parameter that we can continue to increase, and then see what happens to the scale factor and time coordinate as we do so. It is most informative to then plot a vs t, ignoring the fact that they are both really functions of τ.

The plot below shows the typical behaviour of the solutions $a(t)$ for open, closed, and flat universes. The open universe continues to expand, and never stops expanding. The flat universe also continues to expand, but eventually (asymptotically) stops expanding. The closed universe is the most interesting – it expands up to a point, and then starts to recollapse! This is called a 'Big Crunch' and is sometimes described as being like the Big Bang but in reverse. It would be very spectacular if it happened to our Universe!

Note that even a small amount of positive or negative curvature can completely change the future fate of the universe. A matter-dominated universe that is only slightly above the critical density would have a radically different outlook to one that is only slightly under. What is the fate of our Universe? Are we destined for a Big Crunch, or a lonely eternal expansion as all other galaxies progressively recede from view? Our current best measurement of the curvature is $\Omega_k \lesssim \pm 10^{-2}$. In other words, it's very close to flat, but might be non-zero. And if it is non-zero, we don't know if it's positive or negative!

All three types of universe are *decelerating* – the expansion is slowing down with time, as you can see from the changing gradients in the plot. We will see in the next section how adding a cosmological constant changes everything by causing the expansion to accelerate, changing the fate of the universe completely.

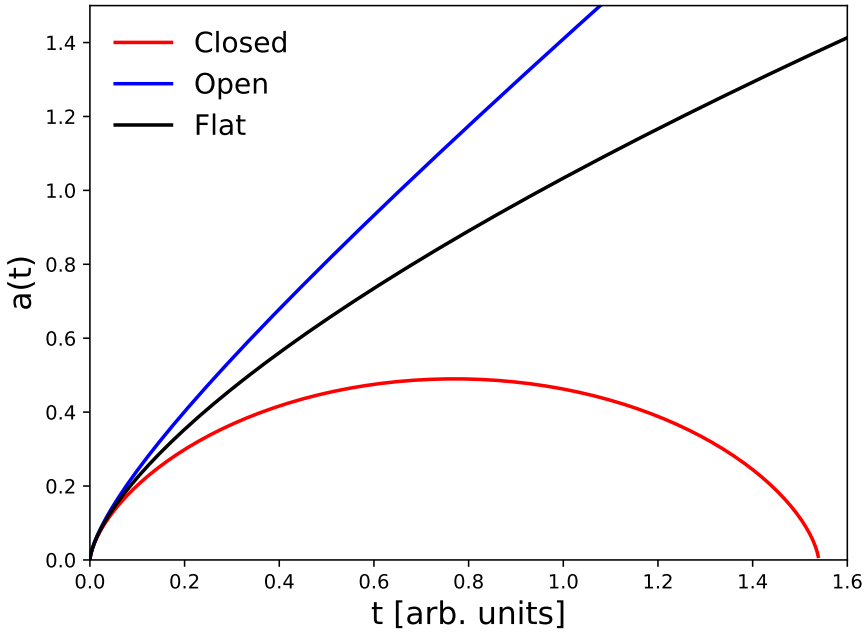

Figure 9: Solutions for $a(t)$ for universes with matter plus three different types of curvature.

3.9. Newtonian derivation of the Friedmann equation Optional

The Friedmann equation is a general relativistic result, but it turns out that we can get to the same expression by thinking about the homogeneous and isotropic expansion of matter with Newtonian equations of motion.

First, consider a thin shell around an arbitrary point (it can be *any* point) in a space filled with a homogeneous distribution of matter. Label this point with a radial coordinate $r = 0$ and place the shell at a radius of $r = R$. Now, what are the forces acting on this shell?

We know from Gauss' Law in Newtonian mechanics that the matter outside the shell has no effect on it – considering concentric shells of progressively larger radius cancel out, the forces from opposite points on those shells cancel each other out. Only matter inside the shell in question exerts a net force. This force is the same as if the matter inside the shell was concentrated into a point mass at the centre.

The total mass within the shell is

$$M(< R) = \frac{4}{3}\pi R^3 \rho, \tag{57}$$

for a density ρ. The force acting on a small segment of the shell (mass m) is then

$$F = -\frac{GmM(< R)}{R^2}. \tag{58}$$

We can calculate the acceleration of this small segment. To avoid confusion with the scale factor, let's write the acceleration as the second derivative of the radial position (R),

$$\frac{d^2R}{dt^2} = -\frac{GM(< R)}{R^2}. \tag{59}$$

So, each segment of the shell should be accelerating *towards* the centre.

Now, we can integrate this equation to find the velocity of the shell and so forth. The right-hand size depends only implicitly on t, since R is a function of t, which makes it a bit fiddly to do the integral. There is a neat trick we can use to help, though. First, multiply both sides by dR/dt:

$$\frac{dR}{dt}\frac{d^2R}{dt^2} = -\frac{GM(< R)}{R^2}\frac{dR}{dt}. \tag{60}$$

Then, integrate:

$$\int \frac{dR}{dt}\frac{d^2R}{dt^2}dt = -\int \frac{GM(< R)}{R^2}\frac{dR}{dt}dt. \tag{61}$$

The right-hand side can be integrated as follows:

$$-\int \frac{GM(< R)}{R^2}\frac{dR}{dt} = -\int \frac{GM(< R)}{R^2}dR = +\frac{GM(< R)}{R}. \tag{62}$$

There is a subtlety with this result – we have treated $M(< R)$ as if it is a constant. That's because it *is* constant if we think carefully about what it means. Before, when we calculated $M(< R)$, we just wanted to know how much mass was within a particular radius (at fixed time). Now, we are tracking the mass within a *particular shell* of matter. As the shell expands outwards, so does all of the material within it – the radius increases, the density decreases, but the mass enclosed stays the same. In other words, the quantity $R(t)$ that we are interested in is *the radius of a shell containing a constant mass*. Other definitions of the radius would involve also tracking a flux of matter into or out of the shell; our definition is unique in that it only tracks the same blob of matter at all times, no matter how much it collapses/expands.

The left-hand side can be integrated as follows. If we first write:

$$\int \frac{dR}{dt}\frac{d^2R}{dt^2}dt = \int \frac{dR}{dt}\frac{d}{dt}\left(\frac{dR}{dt}\right)dt, \tag{63}$$

we can make the replacement $u = dR/dt$, to obtain

$$\int u\frac{du}{dt}dt = \int u du = \frac{1}{2}u^2 + C. \tag{64}$$

Substituting the definition for u back in and equating to the result for the right-hand side, we obtain

$$\frac{1}{2}\left(\frac{dR}{dt}\right)^2 = \frac{GM(< R)}{R} + C. \tag{65}$$

This is an 'energy equation'; the LHS looks like a kinetic energy term (proportional to 'velocity' squared), while the RHS looks like the gravitational potential energy per unit mass. With this interpretation, we can identify the integration constant C as the total energy of the system.

Now, substituting in the expression for the mass enclosed within the shell, we obtain

$$\frac{1}{2}\left(\frac{dR}{dt}\right)^2 = G\frac{4\pi}{3}\frac{R^3\rho}{R} + C. \tag{66}$$

We can evaluate R with respect to some reference radius, R_0, at time t_0. Making the substitution $R(t) = a(t)R_0$ (where $a(t_0) = 1$) and rearranging, we obtain

$$\left(\frac{1}{a}\frac{da}{dt}\right)^2 = \frac{8\pi G}{3}\rho(t) + \frac{C}{a^2 r_0^2}. \tag{67}$$

We can now identify the total energy C with the curvature parameter multiplied by an overall distance/time-scale, $-kr_0^2/c^2$. The rest of the equation is already in the same form of the Friedmann equation, with $\rho(t) \propto R^{-3} \propto a^{-3}$ as expected.

Learning outcomes:

What is the Friedmann equation?

What does each term in the Friedmann equation mean, and what are the relevant units?

What is the critical density and how can it be used to rewrite the Friedmann equation?

How do the densities of matter and radiation depend on the scale factor?

How can you solve the Friedmann equation to find the scale factor as a function of time?

How can you solve the Friedmann equation to find the age of the Universe?

How does curvature affect the time evolution of the scale factor?

4. Distances and horizons

In this section, you will learn about the comoving distance that light travels from an object observed at some at redshift; how to calculate physically-meaningful distances in cosmology based on standard luminosities and sizes; how these distances are defined with respect to the comoving distance; and the concept of a cosmological *horizon*.

Reading for this topic:

- *An Introduction to Modern Cosmology (A. Liddle), Chapters 3, 6, and 7.*
- *An Introduction to Modern Cosmology (A. Liddle), Advanced Topic 2: Distances and Luminosities.*

4.1. Cosmological distances

We have previously discussed how defining distances in cosmology is a fraught process, because space itself is constantly changing (expanding), and because the distances are so large compared with how long it takes light to travel across them. We discussed in Section 2 how the FLRW metric can be used to convert from comoving distances to proper distances, and what the interpretation of those distances was. Neither was directly observable, however; they were both mathematical constructs that are useful for keeping track of things in our calculations.

In this section, we will learn about a few different types of distance that are much more closely related to things we can actually observe. All of them will turn out to be defined with reference to light rays that have travelled from distant objects.

4.2. Distance travelled by a light ray

All of our calculations in this section depend on the distance travelled by a light ray. Consider a photon emitted from a galaxy at some redshift, z. It is eventually observed by us, the observer, at $z = 0$. We know that the photon must have been emitted a long time ago; it takes light a long time to travel across the vast distances in our Universe, and that distance is also increasing, due to the cosmic expansion.

One way of measuring the distance travelled by the photon is to keep track of the *comoving distance* it has traversed between being emitted and being detected by us. We call this distance $r(z)$. It is uniquely defined for any given redshift; for a redshift z, we *only* see the photons that have travelled a comoving distance of $r(z)$, i.e. photons emitted by galaxies at a comoving coordinate of r away from us. Photons emitted at the same time by galaxies that are further away in comoving coordinates than r have not had chance to reach us yet, while those emitted at the same time by galaxies closer to us would have already reached Earth a long time ago (so we missed them). Hence, for a given redshift, we only see photons that were emitted from a comoving distance $r(z)$ away from us.

Recall that the *space-time* distance travelled by light rays is $ds = 0$. We can use the line element for an FLRW metric (assuming a flat universe), subject to this condition, to write down a relation between comoving coordinate of a photon and another variable (in this case, time):

$$ds^2 = -c^2dt^2 + a^2dr^2 = 0 \tag{68}$$

$$\implies cdt = \pm a dr. \tag{69}$$

Using the Friedmann equation, we can write $da/dt = \pm aH$ and substitute for dt, giving

$$dr = \pm c\frac{da}{a^2H} \tag{70}$$

$$\implies r(a) = \int_0^r dr' = -\int_1^a \frac{c\,da'}{(a')^2H(a')}, \tag{71}$$

where we have chosen the negative sign so that the distance comes out positive (this make sense; the light ray is travelling *from* a coordinate of r to reach us at a coordinate 0, so the order of the integration limits should be flipped unless a minus sign is used).

4.3. Luminosity distance and standard candles

The comoving distance travelled by a photon isn't something that we can measure directly – there are no properties of the photon that we can measure that would tell us how far it has moved since it was emitted. There are closely-related quantities that are measurable, though.

Consider light emitted from a distant object that has luminosity, L. Let's also assume that the light is emitted isotropically from the object. The light will therefore travel out from the object in a spherical shell. As the size of the shell increases, the light will be diluted, making the object appear fainter and fainter from greater and greater distances. What we can measure is the *flux* of light coming from the object, f, as we see it from a large distance away. The flux and luminosity are related to each other by the simple relation

$$f = \frac{L}{4\pi d_L^2}, \tag{72}$$

where $d_L$is the radius of the shell that corresponds to how much the light has been diluted.

We call $d_L(z)$ the *luminosity distance* to an object observed at redshift z. It is the distance we would infer if we knew the intrinsic luminosity of the source, measured its flux, and used the simple relation above. The derivation is beyond the scope of this module, but it can be shown that

$$d_L(z) = (1+z)\, r(z), \tag{73}$$

i.e. the luminosity distance is just the comoving distance travelled by photons, scaled by an additional factor of $(1+z)$ (this is related to the fact that the energy, and therefore flux, of the photons is also diluted by the expansion of space). Note that the redshift z here is the true cosmological redshift, and does not include peculiar velocities.

How can we measure the luminosity distance to an object at redshift z? Well, we can measure z and the flux directly using a telescope. The last remaining ingredient is the intrinsic luminosity, L. Certain types of astronomical objects have luminosities that we can figure out by inspecting some of their other properties. These are called *standard candles*. One such type of object are the Cepheid variable stars studied by Henrietta Swan-Leavitt; these exhibit a relationship between the period at which they pulsate vs their intrinsic luminosity that can be calibrated using measurements of very close-by Cepheids. If we then measure the period of pulsation of very far away Cepheids by making repeated measurements of the flux over time with a telescope, and then use the period-luminosity relation to derive their intrinsic luminosity L, we can then infer d_L.

4.4. Distance ladder

Other types of standard candle include Type Ia supernovae, which have very predictable luminosities when their progenitor stars explode. The luminosity can be measured from properties such as the colour of the supernova, and how long it takes for its light to fade. Type Ia supernovae have the advantage of being *extremely* bright, so we can see them from extremely large distances (even thousands of megaparsecs away). Cepheid stars are much fainter, however, and can only be detected in relatively nearby galaxies.

Supernovae are quite rare however, and so we don't have many examples from the local universe, where we would be able to calibrate the relations that tell us their luminosity. Instead, we have to construct a *distance ladder*. The idea is that we start with fainter, nearby types of standard candle (like Cepheids) that we can calibrate quite accurately. We then look for another type of standard candle that is brighter (but generally rarer) that we can see at larger distances. The game is to find galaxies that contain more than one type of standard candle, so we can use the already well-calibrated one to work out the calibration for the other. To calibrate the Type Ia supernovae, a distance ladder with several 'rungs' is needed!

Further reading: Cosmic distance ladder (Wikipedia).

4.5. Angular diameter distance

Instead of using standard candles, we can also use *standard rulers*, i.e. objects that we know the intrinsic size of. We can measure the angular sizes of objects like galaxies, again by taking images of them with telescopes. If we know their intrinsic (proper/physical) size d, we can then infer a distance, d_A, using the definition

$$\theta \approx \frac{d}{d_A}, \tag{74}$$

where we have used the small-angle approximation, $\tan\theta \approx \theta$, since most astronomical objects subtend a very small angle on the sky. The distance d_A is called the *angular diameter distance*.

The angular diameter distance is also related to the comoving distance, but in a different way to the luminosity distance:

$$d_A(z) = \frac{r(z)}{(1+z)}. \tag{75}$$

We can now relate the angular diameter distance to the luminosity distance, finding that

$$d_L(z) = (1+z)^2 d_A(z), \tag{76}$$

where the redshift z is the true cosmological redshift, and not the 'observed' redshift that includes peculiar velocities.

The angular diameter distance has a particularly counter-intuitive property. At small to intermediate redshifts, it increases with redshift, which means that objects of fixed physical size subtend smaller and smaller angles as they get further and further away. But for sufficiently large redshifts, the angular diameter distance begins to get smaller again! So, in an expanding universe, objects that are *very* far away (in terms of their comoving distance from us) can actually have a larger angular size than objects that are closer! This is just a projection effect caused by the expansion of the Universe, but it's still quite surprising.

4.6. Cosmological horizons

In the previous subsections, we learned about several different ways to define cosmological distances. Some, like the angular diameter distance, were defined according to how we make observations of distant objects. Others, like the comoving distance, were mathematical constructs that were useful for keeping track of coordinates.

Another useful type of distance is a *horizon*. It defines how far a particle or wave could have possibly travelled since it was emitted. Or, in other words, it is the maximum radius out to which some physical process could have influenced other events.

Horizons are very useful in cosmology. By measuring the size of a horizon, we can figure out how long a physical process must have been going on for the horizon to have reached that size. Different physical processes have different horizon sizes, and so by observing the Universe on distance scales longer and shorter than a given horizon, we can disentangle the different physical effects that must have been going on to cause the structures that we see.

A good example of a horizon is the *particle horizon*, r_H. This is a very fundamental distance in cosmology – it is the furthest (comoving) distance a particle (travelling at the speed of light or less) could have possibly travelled since the Big Bang. No physical process can have any effect over distances greater than the particle horizon; we say that regions of the Universe separated by a distance greater than the particle horizon are *causally disconnected*, since there is no way the two regions could ever have been in contact via a causal physical process.

To calculate the particle horizon, we need to calculate the comoving distance that could have (in principle) been covered by a light ray since the Big Bang. Using the definition of the comoving distance travelled by light, $r(z)$, we can write

$$r_H = r(z \to \infty) = \int_0^1 \frac{c\,da'}{(a')^2 H(a')}, \tag{77}$$

where we have integrated from our current value of the scale factor ($a = 1$) to the scale factor at the Big Bang ($a = 0$). This is the size of our particle horizon today.

How big was the particle horizon at some time in the past? If we look at an object at a redshift z from us, *their* particle horizon is

$$r_H(z) = \int_0^a \frac{c\,da'}{(a')^2 H(a')}, \tag{78}$$

where we have integrated back from *their* scale factor, $a = 1/(1+z)$, until the Big Bang at $a = 0$.

Another type of horizon is the *Hubble radius*, $r_{\rm HR}$. This is the approximate size of the observable Universe at a given time. We can define this using the Hubble Law, $v = H_0 d$ (which, remember, is only an approximation at low redshift, so this result will also be approximate). If we set the recession velocity $v_{\rm rec} = c$, we obtain $d = c/H_0$. Replacing d with a comoving distance, $d = a r_{\rm HR}$, and generalising to the expansion rate at any scale factor, $H(a)$,

$$r_{\rm HR} \approx \frac{c}{aH}. \tag{79}$$

This quantity is not strictly a horizon – we can see light that has travelled from further comoving distances than this – but it does correspond to the approximate distance over which causal processes can operate at any given time. In other words, only objects with a comoving separation $r < r_{\rm HR}$ can have any significant physical interaction between them.

Further reading: Cosmological horizon (Wikipedia).

Learning outcomes:

What is the comoving distance, $r(z)$?

How is the luminosity distance defined?

How is the angular diameter distance defined?

How are the luminosity distance and angular diameter distance related?

How can the different types of distance be measured?

What is the particle horizon? How can it be calculated?

What is the Hubble radius or Hubble horizon?

5. Cosmic acceleration

In this section, you will learn how the expansion of the Universe can be accelerating or decelerating, depending on what kind of matter it contains; how to measure acceleration using the deceleration parameter; how a universe dominated by the cosmological constant behaves; how the acceleration caused by a cosmological constant changes the size of the cosmological horizon; and what the likely fate of our own Universe is.

Reading for this topic:

- *An Introduction to Modern Cosmology (A. Liddle), Chapters 3, 6, and 7.*

5.1. Conservation equation

As we saw in previous sections, the density of different types of matter/energy scales with the scale factor in different ways; (non-relativistic) matter scales as $\rho_m \propto a^{-3}$ and (relativistic) radiation scales as $\rho_r \propto a^{-4}$ for example. We also saw that universe predominantly filled with different types of matter/energy expanded in different ways; the mix of matter, radiation, and curvature could quite radically change how the scale factor, $a(t)$, changes with time.

There is a useful equation that relates the time evolution of the density of a particular type of matter/energy to the time evolution of the scale factor. This is called the *conservation equation*, as it is derived from the stress-energy conservation equation of General Relativity. It can be written as:

$$\dot{\rho} = -3\frac{\dot{a}}{a}\left(\rho + \frac{p}{c^2}\right). \tag{80}$$

Note how this equation depends on the expansion rate ($\dot{a}/a = H$), as well as the density ρ and *relativistic pressure* p. Importantly, this equation holds *separately* for each type of matter/energy (as long as they are not interacting with each other). So, in a universe that contains both matter and radiation, we can write down a separate conservation equation for ρ_m and ρ_r, for example, with the same expansion rate, but different densities and pressures.

What do we mean by relativistic pressure? This is not quite the same as *thermal* pressure. First of all, we normally think of pressure as exerting an outward force (e.g. to keep stars from collapsing in on themselves), and so would expect it to counteract gravitational attraction somehow. This is what the thermal pressure does, but it's not really what the relativistic pressure in this equation describes. In fact, something with lots of relativistic pressure ($p > 0$) has a *stronger* gravitational pull! This is because pressure increases the amount of energy in a region, which causes a stronger space-time curvature (and so greater gravitational attraction).

Second, the thermal pressure of most regular (baryonic) matter and dark matter is very small compared to its rest mass, and so the density ρ dominates. The pressure can become larger in extreme systems such as the dense interiors of very massive stars, but this actually makes them more likely to collapse in on themselves, not less! Again, this is because relativistic pressure increases the gravitational attraction.

Instead, we can think of the relativistic pressure as a separate quantity that is characteristic of a particular type of matter/energy, and determined by its *equation of state*.

5.2. Equation of state

Different types of matter/energy have different characteristic amounts of relativistic pressure compared to their energy density. The relative amounts of pressure and density are encoded in the *equation of state* parameter w for that particular type of matter energy,

$$p = w\rho c^2. \tag{81}$$

We can work out the equation of state for matter and radiation by substituting this relation into the conservation equation and solving for w,

$$\dot{\rho} = -3\frac{\dot{a}}{a}\left(\rho + w\rho\right) = -3\frac{\dot{a}}{a}\rho\left(1 + w\right). \tag{82}$$

For matter, we can plug-in $\rho_m = \rho_{m,0} a^{-3}$ to obtain

$$-3\rho_{m,0}\, a^{-4}\, \dot{a} = -3\frac{\dot{a}}{a}\rho_{m,0}\, a^{-3}\,(1+w)\,, \tag{83}$$

where we have used $d\rho_m/dt = (d\rho_m/da)(da/dt) = \dot{a}(d\rho_m/da)$. Cancelling common factors from both sides, we get

$$1 = (1+w) \implies w = 0, \tag{84}$$

so the equation of state for matter is $w = 0$. This is what we expected – non-relativistic matter should have negligible relativistic pressure, $p \approx 0$, which implies $w = 0$. Repeating a similar exercise for radiation, we find $w = 1/3$.

More generally, we can rearrange the conservation equation to put all the factors of a on one side and ρ on the other:

$$\frac{\dot{\rho}}{\rho} = -3\frac{\dot{a}}{a}\,(1+w)\,. \tag{85}$$

Now, if we integrate with respect to time:

$$\int \frac{1}{\rho}\frac{d\rho}{dt}dt = -3\int \frac{1}{a}\frac{da}{dt}\,(1+w)\,dt \tag{86}$$

$$\implies \int \frac{d\rho}{\rho} = -3\int (1+w)\,\frac{da}{a}. \tag{87}$$

We can solve both sides of the equation once we know the equation of state. Note that the equation of state can, in general, depend on scale factor too, although for matter and radiation it is constant.

5.3. Cosmic acceleration and deceleration

The equation of state of different types of matter and energy is interesting because it determines how their densities scale with scale factor. It also helps determine whether the expansion of the universe is *accelerating* or *decelerating*.

We have used the Friedmann equation and the expansion rate $H = \dot{a}/a$ as a measure of whether the Universe is expanding. Very simply, if $\dot{a} > 0$, it's expanding. In the Big Crunch example, we saw that some types of universe can collapse too ($\dot{a} < 0$).

The next question to ask is whether the expansion is getting faster with time (accelerating), or slowing down (decelerating); in other words, whether the scale factor has a positive second derivative, $\ddot{a} > 0$ (accelerating) or a negative second derivative, $\ddot{a} < 0$ (decelerating). Note that $\ddot{a} > 0$ does *not* imply that the expansion rate H is getting bigger with time! Recall that $H = \dot{a}/a$. While it's true that $\dot{a}$ gets bigger with time in an accelerating Universe, a is also getting bigger!

General Relativity provides us with another equation that we can use to calculate the acceleration/deceleration of the cosmic expansion. It is called the Raychaudhuri equation,

$$2\frac{\ddot{a}}{a} + \left(\frac{\dot{a}}{a}\right)^2 + \frac{kc^2}{a^2} - \Lambda c^2 = -\frac{8\pi G}{c^2}p. \tag{88}$$

Note how it depends on the relativistic pressure, p, that we saw in the conservation equation, as well as some other terms (e.g. curvature, cosmological constant) that are familiar from the Friedmann equation. In fact, the Raychaudhuri equation can be derived by combining the Friedmann and conservation equations.

5.4. Deceleration parameter

Before anyone had measured the parameters of our cosmological model very precisely (in the 1960's, 70's, and 80's), astronomers tried to at least get some sense for how the scale factor was evolving with time given the imprecise data they had available. Instead of working with different FLRW models, they used a Taylor

expansion of the scale factor instead, which allowed them to fit a very simple model with only a couple of free parameters to data such as the luminosity distance (measured as a function of redshift).

Recall the definition of a Taylor expansion from Section 2. Expanding around the scale factor today, $a(t_0) = 1$, we can write a Taylor expansion of the scale factor in increasing powers of $t - t_0$. For some reason, astronomers at the time decided to add to this expansion by introducing factors of H_0, to obtain

$$a(t) \approx 1 + H_0(t - t_0) - \frac{1}{2} q_0 H_0^2 (t - t_0)^2 + \dots \tag{89}$$

Recall that the dots above the factors of a denote time derivatives, e.g. $\dot{a} = da/dt$. The coefficient of the linear term in this Taylor expansion is just the Hubble constant, H_0, which we have already seen is equal to the first derivative of the scale factor, divided by a, and evaluated at $t = t_0$. The next term involves H_0^2, and another coefficient, q_0. This is called the *deceleration parameter*, and is used to determine whether the expansion of the universe is slowing down or speeding up. Also for historical reasons, astronomers decided that the deceleration parameter would be defined with a minus sign in front of it, so $q_0 > 0$ means that the universe is decelerating, while $q_0 < 0$ means that it is accelerating.

The deceleration parameter can also be defined as a function of scale factor or redshift,

$$q(a) = -\left(\frac{a}{\dot{a}}\right)^2 \frac{\ddot{a}}{a} = -\left(1 + \frac{\dot{H}}{H^2}\right). \tag{90}$$

Matter- and radiation-dominated universes are *always decelerating*, regardless of whether they are open, closed, or flat, and so they have a positive deceleration parameter, $q > 0$. But other, more exotic, types of matter/energy can cause the expansion to accelerate, as we will see in the next subsection.

Historically, the deceleration parameter was important in the discovery that the expansion of our Universe is actually accelerating (a discovery that won the Nobel Prize in 2011). In the 1990s, it was becoming apparent that the Universe had much less mass than would be required for it to be flat ($\Omega_m \approx 0.3$, instead of $\Omega_m \approx 1$). This could have just meant that $\Omega_k \approx 0.7$, i.e. that we live in an open universe. This type of universe would still be decelerating, however. Measurements of the luminosity distance, made with distant Type Ia supernovae, were used to accurately measure the deceleration parameter for the first time, showing that $q_0 < 0$ in our universe – it's accelerating! Combined with the observation that the matter density is around $\Omega_m \approx 0.3$, this led to the conclusion that around 70% of the energy density of our Universe is made up of a cosmological constant, Λ – or perhaps something even more unusual.

5.5. Properties of the cosmological constant

The cosmological constant, Λ, is an unusual contribution to the total energy density of the Universe. First of all, it stays constant with time! It does not dilute as the Universe expands, as was the case with matter and radiation. If we use this information to solve the conservation equation (by setting $\rho = \text{const.}$, which implies $\dot{\rho} = 0$), we find that the equation of state for the cosmological constant is $w = -1$. Since $p = w\rho c^2$, this implies that the cosmological constant has a *negative* relativistic pressure! This is quite unusual, but the upshot is that the cosmological constant causes a kind of gravitational repulsion rather than attraction, causing the expansion of the universe to speed up rather than slow down.

The density of the cosmological constant has remained constant in time, while other components like matter and radiation have decreased in density with time. It therefore stands to reason that, back in the past, these other components must have had much higher densities than the cosmological constant. Since radiation- and matter-dominated universes are decelerating (see above), this implies that the deceleration parameter must have been positive in the past, but has become negative now that the cosmological constant is the dominant form of energy density. We say that the accelerating expansion of the Universe is a *late-time* phenomenon, that has only started occurring in the Universe 'recently' (that is, within the past few billion years...)

5.6. Cosmological constant solution

Now let's consider a flat universe that only has a cosmological constant, $\Omega_\Lambda = 1$. This is called a *de Sitter* space-time. The Friedmann equation reduces to

$$\left(\frac{\dot{a}}{a}\right)^2 = H_0^2 \Omega_\Lambda = H_0^2, \tag{91}$$

since $\Omega_\Lambda = 1$. Rearranging, we get

$$H_0 dt = \pm \frac{da}{a}. \tag{92}$$

Integrating from some reference value of the scale factor ($a = a_*$ at $t = t_*$), we find

$$\begin{aligned} \int_{t_*}^{t} H_0 \sqrt{\Omega_\Lambda} dt &= \pm \int_{a_*}^{a} \frac{da}{a} && (93) \\ \implies H_0(t - t_*) &= \pm (\log a - \log a_*) && (94) \\ &= \pm \log(a/a_*). && (95) \end{aligned}$$

Exponentiating both sides and rearranging once more, we get

$$a(t) = a_* e^{\pm H_0 (t - t_*)}. \tag{96}$$

The appropriate choice of sign in the exponent to get an expanding universe is the positive sign. Also, if we want to have $a = 1$ at $t = t_0$, we can set $a_* = 1$ and $t_* = t_0$, so we have the solution

$$a(t) = e^{H_0 (t - t_0)}. \tag{97}$$

This is an *exponential expansion* of space, and gives rise to some interesting properties. We have already seen from the Friedmann equation that the expansion rate, H, is constant in time. If we write the cosmological constant in the form of an energy density, ρ_Λ, we can see that this is constant too. So, as the universe expands, the energy density doesn't dilute (as with matter or radiation) – it stays the same! This seems quite unintuitive – as more space is created by the expansion, more energy is created too.

Doesn't it violate the law of energy conservation is the universe continually 'creating' energy? See the 'Energy conservation' box below for more discussion of energy conservation – it turns out that energy is *not* conserved in an expanding universe (but a more general quantity that involves energy *is* conserved). We will study this curious result in more detail in the next sub-section.

5.7. Age and Hubble radius in an exponentially-expanding space-time

Another curious feature arises when we try to work out the age of this universe. If we set $a = 0$ (denoting the Big Bang) and solve for $t = t_{\rm BB}$, we get $H_0(t_{\rm BB} - t_0) = \log 0 = -\infty$. So the age of the universe, $t_0 - t_{\rm BB} \to \infty$! A universe with *only* a cosmological constant (i.e. not even a tiny amount of any other kind of matter or energy) has no Big Bang – it is an eternal universe that has been expanding forever and will continue to expand forever.

For our next curious feature, let's work out the Hubble radius or Hubble horizon, $r_{\rm HR} = c(aH)^{-1}$. We obtain

$$r_{\rm HR}(t) = \frac{c}{e^{H_0(t-t_0)} H_0} = \frac{c}{H_0} e^{-H_0(t-t_0)}. \tag{98}$$

The Hubble radius is getting smaller with time! That is, as the universe expands, physical processes are only able to operate across smaller and smaller comoving distances. Distant parts of the universe are progressively falling out of contact with each other. This is the opposite of what happens in (e.g.) matter-only and radiation-only models, where the Hubble radius continues to expand.

5.8. The Cosmological Constant problem

Where does the energy for the cosmological constant come from? Strictly, the cosmological constant is just a term that crops up in the Friedmann equation when you derive it mathematically, sort of like an integration constant. You could think of it as an inherent property of space-time, in the same way that curvature is an inherent property of space. It is zero in some types of universe and non-zero in others.

Another common interpretation of the cosmological constant is that it represents the *vacuum energy* or zero-point energy of the universe. Recall the harmonic oscillator example from quantum mechanics, which has energy levels $E_n = \hbar\omega(n + \frac{1}{2})$. Its zero-point energy, at the lowest energy level ($n = 0$), is non-zero

($E_0 = \hbar\omega/2$). This is an example of a system that has a vacuum energy – even when it is 'empty' (i.e. in its ground state), there is still some energy there.

Our Universe isn't described by the harmonic oscillator, but it is described by the Standard Model of Particle Physics, which includes electromagnetism and the strong and weak nuclear forces. These forces are described within the framework of Quantum Field Theory (QFT), which allows every time of field to have a possibly non-zero vacuum energy. Each field has some kind of particle or particles associated with it, so there is an electron field, a photon field etc, and each of these can have some vacuum energy (which may be positive or even negative). If we add all of the vacuum energies from QFT together, we arrive at some number that we can interpret as the total vacuum energy of the universe, $\rho_{\rm vac}$.

What happens when we do this for our Universe? We get an absolutely huge number! The vacuum energy should be so large that atoms and nuclei could never have formed, as the expansion rate in the early Universe would have been too high. What's more, we have *measured* the apparent vacuum energy density in our Universe using distant supernovae and the luminosity distance-redshift relation. The measured value is around 60 orders of magnitude smaller than what we expect from QFT!

This discrepancy between the expected vacuum energy and the observed value of Λ is called the **Cosmological Constant problem**. It can be understood as a *fine-tuning* problem, where a parameter of our model must be tuned to a very, very specific number in order for things to work out. Theoretical physicists tend to hate fine tuning – it suggests that we don't understanding something about how the parameters that describe our model are actually being chosen (if they were being chosen randomly, how could they have possibly ended up with such specific values!). To see this fine-tuning problem for the cosmological constant, consider that the observed cosmological constant is actually the sum of the inherent value of Λ that space-time has (the 'integration constant' from GR, also called the 'bare' cosmological constant), and the QFT vacuum energy,

$$\Lambda_{\rm obs} = \Lambda_{\rm bare} + \frac{8\pi G}{c^2}\rho_{\rm vac} = \Lambda_{\rm bare} + \Lambda_{\rm vac}. \tag{99}$$

Both have the same effect on the Universe, and so we can only ever measure their sum, $\Lambda_{\rm obs}$, which we know to be small. We can also calculate $\rho_{\rm vac}$ from QFT, which we find to be very large. This implies that the fractional difference between the two terms is

$$\frac{|\Lambda_{\rm bare}| - |\Lambda_{\rm vac}|}{|\Lambda_{\rm vac}|} \approx 10^{-60}, \tag{100}$$

i.e. that they are almost exactly the same, except for a tiny, tiny difference. Why would the bare cosmological constant have anything to do with the QFT vacuum energy? Why would they be so similar? This puzzle is perhaps the biggest in all of fundamental physics.

5.9. The fate of our Universe

The shrinking-horizon property of exponentially expanding solutions will turn out to be important when we discuss cosmic inflation in later sections. It's also important for understanding the fate of our own Universe, which has a large fraction of its energy density in Λ ($\Omega_\Lambda \approx 0.7$). As our Universe continues to expand, the matter and radiation energy density will continue to dilute away, while the cosmological constant will remain the same. Eventually, the energy density of the cosmological constant will completely dominate over all other forms of matter and energy, and we will find a solution that is very, very close to the Λ-only (de Sitter) model that we have been studying. The Λ term even dominates over the curvature term, so even if our Universe did have a non-zero positive or negative curvature, it would still approach the de Sitter solution if we waited for a long enough time.

As the comoving Hubble radius continues to shrink in the Λ-only model, we will eventually find ourselves cut off from even the closest galaxies in the Hubble flow. Any light that could still reach us from distant galaxies would have been redshifted by an extreme amount. Eventually, our galaxy will have exhausted all of its raw materials for star formation, and the last stars will die. It is a very lonely, and seemingly inevitable fate! At least it will take tens of trillions of years to get to that point – our Sun only has about 4.5 billion years of fuel left, so would have died long before that point (and taken Earth with it). Even further into the future, it's possible that any remaining matter (e.g. neutron stars) could disintegrate due to proton decay, and black holes would eventually evaporate due to Hawking radiation (about 10^{100} years into the future). The Universe would

be left filled with a very low-density, low-energy, and homogeneous background of photons and other stable fundamental particles that would hardly ever interact with one another. Thinking about the thermodynamics of the Universe at this time, the entropy would be extremely high, and there would be no free energy for any physical or chemical processes to occur. Apart from the continuing expansion of space, the Universe would essentially be frozen. This future fate of the Universe is therefore often called the **Big Freeze**, or alternatively the **Heat Death of the Universe**, because of the absolute thermodynamic equilibrium that would be reached.

Can anything save our Universe from this chilly (and very boring) fate? Well, it could be worse! If our Universe is actually filled with a Dark Energy field, and not a cosmological constant, there could be other fates in store. One is called a **Big Rip**, where the expansion becomes faster than exponential (this happens for a cosmological fluid with an equation of state $w < -1$). This would eventually cause everything in the Universe to be ripped apart – even galaxies and atoms, which aren't in the Hubble flow. Depending on the value of w, this could even happen as soon as tens of billions of years in the future.

Another possibility is **false vacuum decay**. It could be that the apparent ground state of our Universe is not truly the lowest energy state – i.e. a *false vacuum* state – and that the matter/energy fields could spontaneously tunnel into a nearby, lower energy state after enough time. Measurements of the Higgs field from the LHC suggest that it could be in a metastable state, meaning that it could spontaneously find a lower-energy vacuum at some point in the future, but this is far from certain. The false vacuum decay would be a very violent event, with a wave of the new vacuum state ripping through our Universe at the speed of light and leaving completely reconfigured matter and energy fields in its wake. The effect on QFT fields could be such that chemistry, nuclear bonding etc. would become impossible and all of the matter would disintegrate instantly. Yikes.

Another, slightly more cheering, possibility is that the Big Freeze could eventually cycle through into the birth of a new universe. This idea, called **conformal cyclic cosmology**, has been championed by Roger Penrose. It relies on noticing that the metric of the Universe after the Big Freeze can be conformally transformed into the metric that we would expect to have at very early times in our own Universe (during cosmic inflation; see later sections). The old universe is in some sense mathematically equivalent a new universe, and so perhaps all of the same physical processes would kick in again and generate matter and radiation again? It is a speculative idea, but has testable predictions.

Further reading: Future of an expanding universe (Wikipedia); False vacuum decay (Wikipedia); Vacuum decay: the ultimate catastrophe (Cosmos Magazine); Conformal cylic cosmology (Wikipedia); Last Contact (short story about the Big Rip by Steven Baxter)

Energy conservation

The Cosmological Constant seems to violate energy conservation! As the Universe expands, the total volume gets bigger, but the energy density of the CC stays the same, implying that the total energy in the Universe continually increases. This happens even in a universe with only Λ (no matter, radiation, or curvature).

In fact, energy conservation *is* violated in many cosmological models, even ones without Λ. As first proven by Emmy Noether, conservation laws come from underlying symmetries in the laws of physics. Different symmetries give rise to different conservation laws; for example, rotational symmetry of a system typically gives rise to the conservation of angular momentum in that system.

The symmetry that gives rise to energy conservation is *time-translation invariance*. If the system follows laws that do not change with time, energy will be conserved. This is *not* the case for most cosmological models however, which are obviously changing with time due to the expansion (or contraction) of space.

Recall that, in relativity, we can transform between different space and time coordinate systems. There is nothing special about any given space or time coordinate. Space and time together *do* follow certain laws however (according to the *principle of general covariance*). We find that there are conservation laws associated with this generalised symmetry, such as the conservation of *stress-energy*. The stress-energy tensor, T_{ab} describes a generalised form of energy and momentum (and other things), and follows the conservation law $\nabla^a T_{ab} = 0$.

Energy density is the time-time component of the stress-energy tensor, T_{00}. Since this can be transformed to a different quantity under a coordinate transformation, it's clear that energy generally won't be conserved in this picture. The *covariant* quantity – the stress-energy tensor – *is* conserved though.

Learning outcomes:

- What is the conservation equation?
- How can it be used to work out how the density of a type of matter/energy depends on scale factor?
- What is the Raychaudhuri equation?
- How does the equation of state relate energy density and relativistic pressure?
- What is the equation of state for matter, radiation, and a cosmological constant?
- What does it mean to have accelerating expansion?
- How is the deceleration parameter defined?
- What is the Cosmological Constant?
- What is the Cosmological Constant problem?
- How does the Hubble radius behave during exponential expansion?

6. Big Bang Nucleosynthesis

In this section, we will learn how the lightest elements formed through a process called *Big Bang Nucleosynthesis* (BBN). We will study the production and decay of neutrons, and the nuclear reactions that produce light elements. As a result of this process, we can predict that around 75% of the normal matter in the Universe should be made from hydrogen, and just under 25% from Helium. The hot Big Bang model predicts the relative abundance of hydrogen, Helium, and other light elements very precisely, making this a very compelling line of evidence for the Big Bang theory.

Reading for this topic:

- *An Introduction to Modern Cosmology (A. Liddle), Chapters 11 and 12.*

6.1. Thermal history of the very early Universe

The most important feature of the hot Big Bang model is that the Universe was much hotter and denser in the past, and that it has been continually expanding and cooling down as time has gone by. As the temperature drops, different physical processes come into play. When the temperature is very high, there is sufficient energy available for quite extreme particle physics processes to occur. As the Universe cools, these stop being energetically favourable and so rapidly become rarer and rarer, leaving other processes to dominate.

In the very early Universe, the energy was high enough that it was filled with a quark-gluon plasma, and the electromagnetic and weak interactions were combined into a single *electroweak* force. As the universe cooled, the electroweak symmetry was broken, generating new particles and resulting in two separate forces. As the universe cooled even further, the first baryons (e.g. protons and neutrons) were able to form without instantly being blown apart into their quark/gluon constituents, in a process called *baryogenesis*. The first leptons were also created around this time in a process called *leptogenesis*.

We will not discuss these processes in detail here, but it's worth noting that there are some important puzzles associated with this epoch. Perhaps the biggest question is why matter and anti-matter were not created in equal abundances. If they had been, all of the particles would have rapidly annihilated with their antiparticles, leaving only photons! Some kind of asymmetry between matter and anti-matter must be present in the physical laws of our Universe that results in slightly more matter than anti-matter being produced at these early times.

Further reading: The very early universe (Wikipedia)

6.2. Neutron decay

Once the Universe has cooled down enough, it becomes possible to form bound nucleons like protons and neutrons. Free neutrons are unstable, and will decay into a proton and electron after a while (with a neutrino involved in order to conserve lepton flavour). This is because protons are more tightly bound, and so it is energetically favourable for the neutrons to decay according to one of the following processes:

$$n + \nu_e \quad \leftrightharpoons \quad p + e^- \tag{101}$$

$$n + e^+ \quad \leftrightharpoons \quad p + \bar{\nu}_e, \tag{102}$$

where n denotes a neutron, p a proton, $e^\pm$ a positron/electron, ν_e an electron neutrino, and $\bar{\nu}_e$ and electron anti-neutrino.

The density of the early Universe is high enough that collisions and other interactions are very frequent, and so particles tend to stay in thermal equilibrium. The reactions above can happen in both directions if enough energy is available, and so protons and neutrons are rapidly and frequently converted into one another in the early Universe.

Non-relativistic particles in thermal equilibrium follow the Maxwell-Boltzmann distribution, with number density

$$n \propto m^{\frac{3}{2}} \exp\left(-\frac{mc^2}{k_B T}\right), \tag{103}$$

where $E \approx mc^2$ because the particles are non-relativistic. We can use this simple equation to predict how the relative number density of protons and neutrons changes in the early Universe,[2] depending on whether it is energetically favourable for the neutrons to decay, and how often protons gain enough energy to be 'converted' into neutrons. Taking the ratio of the number densities of neutrons and protons, we obtain

$$\frac{n_n}{n_p} \propto \left(\frac{m_n}{m_p}\right)^{\frac{3}{2}} \exp\left(-\frac{(m_n - m_p)c^2}{k_B T}\right), \tag{104}$$

for neutron mass m_n and proton mass m_p. The difference in rest mass between a neutron and a proton is $\Delta m = m_n - m_p = 1.293$ MeV.

We can track the relative abundance of neutrons and protons as a function of redshift by using the formula $T \propto (1+z)$. The ratio n_n/n_p is plotted in the figure below. At high temperatures, neutrons and protons are almost equally abundant – there is enough energy around that creating a neutron is almost as easy as creating a proton. As the temperature decreases, however, it becomes less and less likely that a neutron will be created in a given reaction, and so the ratio decreases substantially.

At a temperature of $T \approx 0.7$ MeV (roughly 1 second after the Big Bang), the reactions between protons and neutrons became inefficient – the temperature and density of the Universe became too low, and so collisions and other interactions became rare. This is very similar to the decoupling process that happens to the photon-baryon fluid around the time of last scattering (although this time with protons and neutrons). Instead of decoupling, we call this the **freeze-out** of neutrons and protons, as they no longer interact with one another frequently.

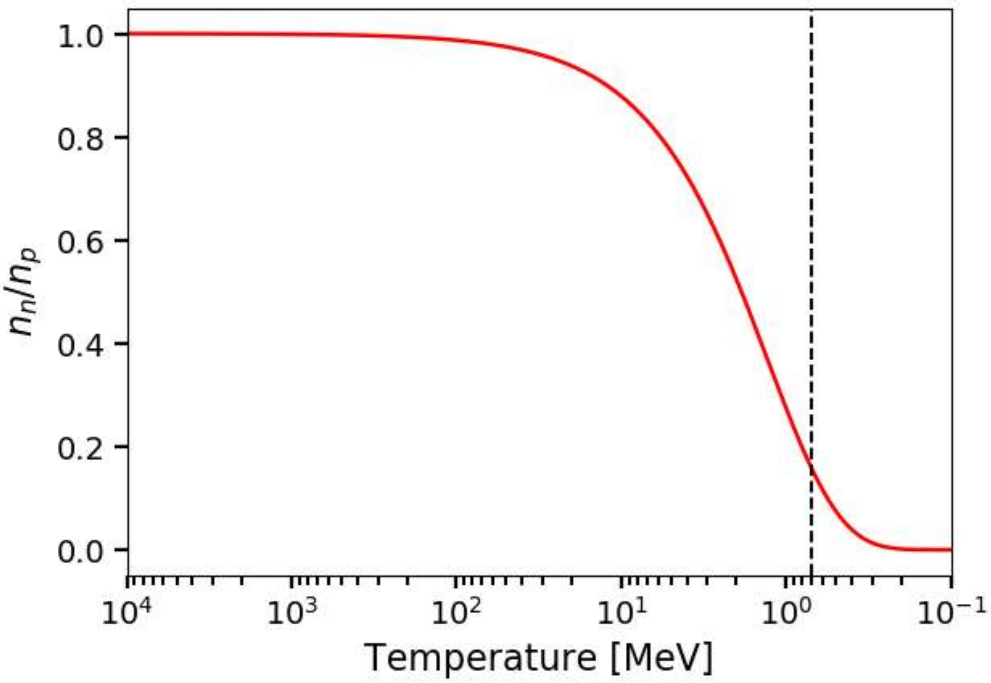

Figure 10: Ratio of the abundance of neutrons to protons as a function of temperature. The vertical dashed line shows the temperature when the protons and neutrons 'freeze out'.

If no other reactions occur, the neutron vs proton abundance would be fixed after freeze-out. Recall that free neutrons are unstable though. The half-life of a free neutron is about 600 sec. Protons, on the other hand, are thought to be stable (their half-life has been measured to be *at least* 10^{34} years, and may in fact be infinite). For the next few minutes after freeze-out, the temperature is still large enough that bound nuclei can't form efficiently, and so the protons and neutrons remain free. The abundance of protons stays constant, while the abundance of neutrons decays exponentially according to

$$n_n \propto n_n(t_{\rm freeze}) e^{-(t - t_{\rm freeze})/\tau_n}, \tag{105}$$

[2] Nucleons have a mass of around 1 GeV, so this non-relativistic equation will apply once the temperature has dropped below $T \lesssim 1$ GeV.

where τ_n is the neutron lifetime. By the time that bound nuclei form, the ratio of protons to neutrons is about 7:1.

Further reading: Free neutron decay (Wikipedia); Proton decay (Wikipedia)

6.3. Nucleosynthesis

The Universe becomes cool enough for bound nuclei to form at a temperature of around 80 keV (about 4 mins after the Big Bang). The process of forming nuclei is called **nucleosynthesis**. The simplest nucleus beyond hydrogen is **deuterium** – one proton and one neutron. A bound deuterium nucleus is a more energetically-favourable state than a free proton and neutron, with a binding energy of 2.2 MeV, and so deuterium nuclei formed very rapidly once the temperature became low enough, via the process

$$p + n \longrightarrow D + \gamma, \tag{106}$$

where D is a deuterium nucleus and γ is a photon that carries away any excess energy (many, many photons are produced during nucleosynthesis). Neutrons are stable once combined into a nucleus, and so this stopped the neutrons in the Universe decaying any further. Otherwise, we would be left only with protons, and so no heavier elements could ever have formed! Deuterium is therefore a critical step in the story of the formation of the elements.

Essentially all of the remaining neutrons rapidly combine into deuterium nuclei. Since the ratio of available protons to neutrons is about 7:1, this means that we get 1 deuterium nucleus (1p, 1n) for every 6 hydrogen nuclei (1p). To see this, consider 8 nucleons: 1 neutron and 7 protons. Of these, 1 neutron and 1 proton will be combined into a deuterium nucleus. The remaining 6 protons remain free as hydrogen atoms. As a fraction of the total *mass* of nuclei, deuterium makes up $2/8 = 25\%$ however.

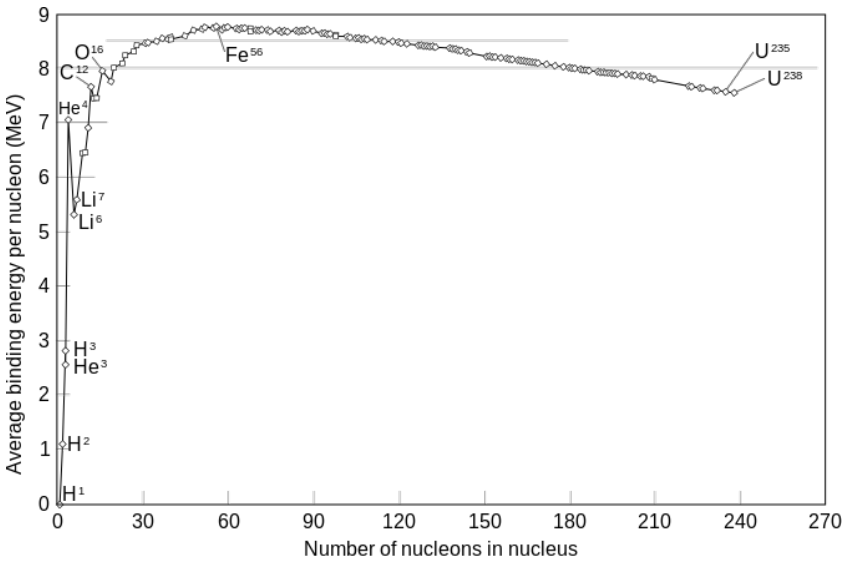

Figure 11: Average binding energy per nucleon in a nucleus, as a function of the total number of nucleons. The higher the average binding energy, the more stable an element will be. For more details see here. (Credit: Wikipedia.)

As long as it is energetically-favourable and the Universe is dense enough to support a high reaction rate, more massive nuclei will also form. The next step in the chain is **Helium**, which can be created by combining two deuterium nuclei or a deuterium nucleus and a proton to form Helium-3:

$$D + p \longrightarrow {}^3\mathrm{He} + \gamma \tag{107}$$

$$D + D \longrightarrow {}^3\mathrm{He} + n. \tag{108}$$

The Helium-3 can then combine with another deuterium atom or a proton to form Helium-4. This has a very high binding energy of around 28.3 MeV, and so is extremely energetically favourable, and so most of the deuterium in the Universe ends up combining into ^{4}He nuclei in one way or another. ^{4}He contains two neutrons

and 2 protons, so now consider a group of 16 nucleons (14p, 2n). Of these, 2n and 2p will be bound into ^{4}He nuclei (or slightly lighter nuclei), with the remaining 12 protons left as hydrogen nuclei. The mass fraction of Helium is therefore

$$Y_4 \approx \frac{2n+2p}{2n+14p} \approx \frac{4}{16} \approx 25\%. \tag{109}$$

(The notation Y_N is commonly used to denote the fraction of mass in nuclei with N nucleons.) So, *by mass*, about 25% of the normal baryonic matter in the Universe is Helium, and 75% is hydrogen. This is remarkably close to what we observe today, for example in the composition of stars and interstellar gas clouds! Practically all of the hydrogen and Helium in the Universe was therefore created in the first few minutes after the Big Bang!

What about other, heavier, nuclei? As shown in the figure above, the binding energy per nucleon of Helium-4 is remarkably large, and so it is very stable. There are also no stable isotopes with 5 or 8 nucleons, so these steps in the reaction chain are missing. The next available step in the chain would therefore be **Lithium**, ^{6}Li, but this is difficult to produce; we would need two Helium-3 nuclei (most of which have already been converted into Helium-4), or a Helium-4 and a deuterium nucleus (again, most of the deuterium has already been converted into Helium-4). If a stable nucleus with 5 nucleons existed, we could have efficiently combined Helium-4 with protons (of which there are many), and the chain could have continued, producing many more heavier elements. But as it is, only a small number of elements heavier than Helium-4 are produced by Big Bang Nucleosynthesis – the reactions needed to produce them occur much less often. If we do all the calculations properly and take into account all possible reaction chains, we find that only around 1 nucleus in a billion is Lithium!

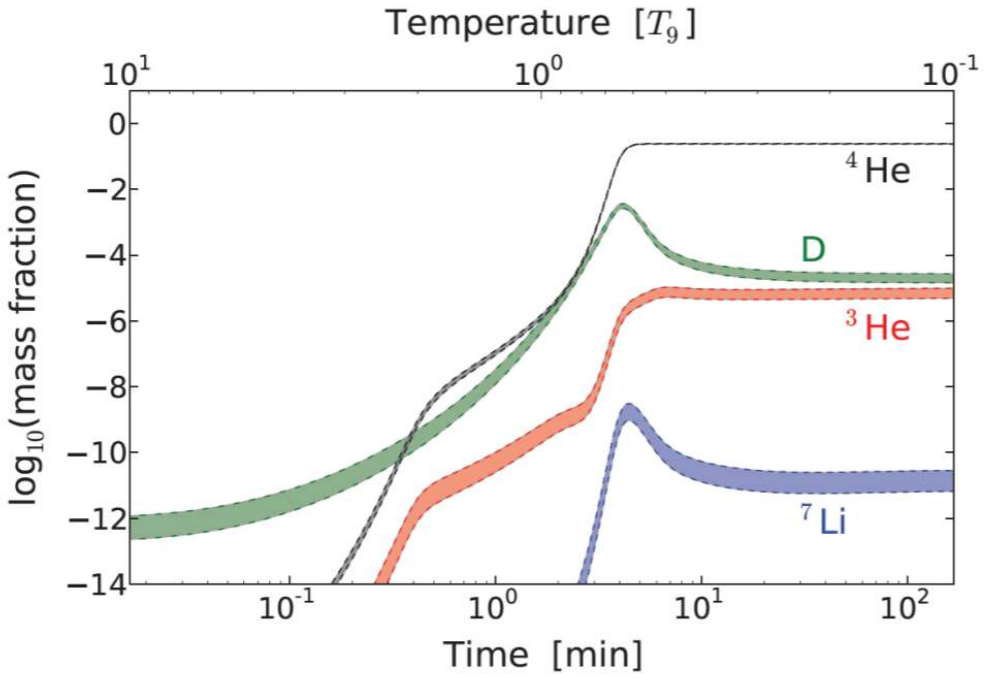

Figure 12: Fraction of the total mass of baryonic matter taken up by the most common types of nuclei (other than hydrogen) as BBN proceeds. The fraction is shown on a $\log_{10}$ scale, and the temperature is shown in units of 10^9 K. (Credit: Pizzone et al. 2014.)

In addition, not *all* of the lighter nuclei are combined into Helium-4. A small fraction of deuterium and Helium-3 nuclei remain after the density of the Universe has dropped sufficiently for nuclear reactions to stop happening. There is even a small amount of tritium (hydrogen-3). Again, the abundances of all of these intermediate reaction products is predicted very precisely by BBN (see the figure above).

The sequence of nuclear reactions that would have happened in the early Universe was first worked out by Ralph Alpher and George Gamow in the 1940s. They made remarkably detailed predictions for the abundance of the lightest elements, like hydrogen, Helium, Lithium, and Beryllium. These were beautifully confirmed

by observations (although a slight discrepancy in the Lithium abundance later showed up, and remains unexplained). These results were presented in a famous scientific paper, the Alpher-Bethe-Gamow paper. The name of Hans Bethe, a famous nuclear physicist, was added as a joke by Gamow!

What about the heavier elements? Their origins were explained over the course of about 20 years (including in another famous paper, the B^2FH paper by Margaret and Geoffrey Burbidge, Willie Fowler, and Fred Hoyle) as being from nucleosynthesis in the cores of stars of different masses, and in supernova explosions and neutron star mergers. See the figure below for a periodic table of the elements that explains where each element comes from.

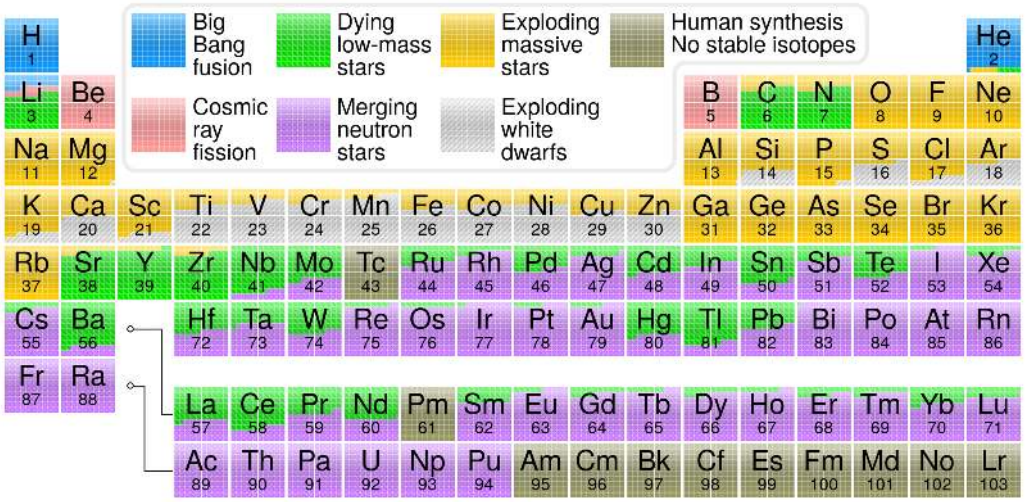

Figure 13: The periodic table of the elements, colour-coded according to how each element is formed. Elements formed by more than one process are shown in multiple colours, with the amount of colouring proportional to how much each process contributes. For more details see here. (Credit: Cmglee/Wikipedia.)

Further reading: The Deuteron (HyperPhysics); Big Bang Nucleosynthesis (HyperPhysics); The First Three Minutes (S. Weinberg/HyperPhysics); Origin of elements in the Solar System (J. Johnson); Cosmological Lithium problem (Wikipedia); Big Bang Nucelosynthesis (Particle Data Group) [pdf]

Binding energy

The binding energy of a nucleus is the amount of energy required to take its constituent nucleons (protons and neutrons) out of the nucleus and move them far away, where they would no longer feel any nuclear force from the other nucleons. Isotopes with a higher binding energy are more stable – it takes more energy to break them apart. Nuclei with negative binding energy, on the other hand, are *unstable*, and can decay spontaneously into lighter nuclei and other particles.

If free nucleons and/or lighter nuclei are in the vicinity of one another and could combine to form a heavier nucleus, the probability of them doing so will depend on the change in binding energy that would occur. If the binding energy would get larger if they combined, there will be a high probability that they form a heavier nucleus. If the binding energy is slightly lower, the chance is small (unless extra energy is available from somewhere; e.g. a photon, or collisional energy).

In general, free nucleons and nuclei that are able to interact with one another efficiently (e.g. in the dense early Universe) will rapidly combine to form the most stable nucleus possible – as long as there is a chain of reactions that can take them there. If there is a weak link in a chain of reactions – for example, an intermediate step that requires a decrease in binding energy – the reaction will be much less efficient, and so fewer nuclei will reach the higher-mass state.

Learning outcomes:

How does the relative abundance of neutrons and protons change with time?

What is freeze-out?

What is the lifetime/half-life of a free neutron?

What is Big Bang Nucleosynthesis (BBN)?

What are the basic reactions needed to form Helium-4 during BBN?

What is the mass fraction, Y_4, of Helium-4 left over from the Big Bang?

7. Cosmic Microwave Background Radiation

In this section you will learn about the Cosmic Microwave Background radiation, and how it formed when the hot, dense ionised gas that initially filled the Universe cooled down and became transparent and neutral. You will also learn about the frequency spectrum of the CMB, and the fact that it is very close to a perfect blackbody with a very low temperature.

Reading for this topic:

- *An Introduction to Modern Cosmology (A. Liddle), Chapters 10 and 12.*

7.1. What was the early universe like?

Recall that in the hot Big Bang model, the early universe is denser and hotter than the late universe. The further back in time we look (i.e. the higher the redshift), the higher the average temperature must have been. It stands to reason that, if we look back far enough in time, the Universe must have been hot enough to ionise neutral atoms. In fact, above a certain temperature, it must have been so hot on average that neutral atoms couldn't have existed – all of the normal (baryonic) matter in the Universe would have been ionised.

What are the implications of this? First of all, ionised gas is very good at scattering photons, especially if it is quite dense (as it would have been in the early Universe). The mean free path of photons (the distance a photon can travel, on average, before being scattered by an electron in an ionised gas), is given by

$$l_{\rm mfp} \approx \frac{1}{n_e \sigma_T}, \tag{110}$$

where n_e is the number density of free electrons, and σ_T is the scattering cross-section of electrons as seen by photons.

Since the energy density (and also number density) of regular matter scales like $\rho_m \propto a^{-3} \propto (1+z)^3$, the mean free path will get smaller at higher redshifts. When it becomes small enough, the probability of a photon travelling through the ionised gas without being scattered becomes very low – the gas becomes opaque to light. This is indeed what the early Universe was like – so densely filled with ionised gas that photons could not travel through it at all.

There are a couple of other important facts about the early universe that we need to understand:

1. Most of the normal baryonic matter in the early Universe was hydrogen produced soon after the Big Bang;
2. Processes soon after the Big Bang also produced large numbers of photons.[3]

So, the early Universe was hot, dense, and contained large numbers of photons. Neutral hydrogen has a binding energy of 13.6 eV, so photons with an energy greater than this can easily ionise it. This is indeed what happened during most of the early period of cosmic history; there were so many energetic photons around that if a neutral hydrogen atom managed to form, it would rapidly be ionised again by a photon. The early universe was filled with hydrogen, but it simply couldn't exist in a neutral form for more than a fraction of a second before being ionised again.

7.2. Formation of the Cosmic Microwave Background

Even further into the past, the Universe was of course even hotter and denser, and so it must have been opaque to light since the Big Bang. There must, therefore, have been a time when the Universe first became cool enough (and the density became low enough) that photons could finally travel freely – a time when the Universe first became transparent. The photons emitted or scattered at around this time would have been able to travel through

[3] In fact the energy density of the early universe was dominated by radiation more than matter, $\rho_r \gg \rho_m$, although by the time of recombination matter had become the dominant component.

the Universe virtually unimpeded since then; the mean free path would have become too large for there to be a significant chance of them scattering off any matter. Could we see those first unimpeded photons today?

The answer is yes! These photons give rise to a phenomenon called the **Cosmic Microwave Background** radiation, or CMB. The CMB has a number of important properties. One is that it is observed at microwave wavelengths! We will see shortly that the CMB photons were originally emitted primarily at visible/infrared wavelengths, but since then the Universe has expanded greatly and the photons have been significantly redshifted. The approximate redshifting factor happens to be ~ 1100 today, and so photons that were a few hundred nanometers in wavelength when emitted (i.e. visible light) would now be almost a millimetre in wavelength as seen today – in other words, at microwave frequencies.

Another important property of the CMB is that it was emitted from everywhere in the Universe at almost exactly the same time. Since our Universe is very close to being homogeneous and isotropic, the mean free path of photons will have dropped below the threshold required to make the Universe transparent at about the same time everywhere. The resulting photons will have travelled in all directions, from everywhere in space; the Universe is therefore filled with them! It is a *background* of photons that exists everywhere throughout space.

Further reading: Cosmic Microwave Background Explained [PBS/YouTube].

7.3. Recombination and decoupling

For the Universe to be opaque, the mean free path of photons has to be short enough that practically every photon will scatter off an electron/ion with a high probability after only a short time. As discussed above, the mean free path depends primarily on the number density of free electrons. For enough free electrons to be available, we need the Universe to be hot enough for gas to be ionised. We also need the Universe to be dense enough that there is a high probability of photons interacting with electrons after travelling only a short distance.

These two conditions are somewhat separate. We could imagine an early universe that was very dense but much colder, so that little of the gas was ionised. This would be transparent to some forms of EM radiation. We can also imagine a hot early universe that had a very low density of normal matter, so that ions/electrons were quite rare even though everything is ionised. In our Universe, the time when the Universe became cool enough for neutral atoms to form was also around the same time that its density dropped enough to allow photons to propagate freely though.

The time when the first neutral atoms formed is called **recombination**. It is poorly named, as this was the first time that ions and electrons had ever been combined – there is no "re" about it. Before recombination, a large fraction of the hydrogen in the Universe was ionised, and so there were many free electrons (and protons).

The time when photons were first able to travel freely, without being scattered by electrons, is called **decoupling**. Before this, we say that photons were 'coupled' to electrons by EM scattering processes.

7.4. Recombination

Forming neutral atoms was surprisingly hard in the early Universe! Consider the reaction

$$p + e^- \leftrightharpoons H + \gamma. \tag{111}$$

Ionised hydrogen is just a proton, p, and for every proton there will be an electron e^- if the Universe has no net charge. These can bind together to form a neutral hydrogen atom H and a photon γ. The opposite can also happen, where a photon come in and ionises a neutral hydrogen atom, leaving behind a proton and electron.

For recombination to happen, this process has to proceed very efficiently from left to right much more frequently than the opposite ionisation process from right to left. As the Universe expands and cools, there is less energy available to ionise any neutral hydrogen that forms, and so the recombination process does begin to happen more efficiently. Unfortunately, this process also emits a photon with energy equal to the binding energy, so every time we create a neutral hydrogen atom, we release a photon that is capable of ionising another neutral hydrogen atom! In the dense early Universe, these photons soon find another hydrogen atom to ionise. The result is that almost all of the hydrogen stays ionised, even when the temperature drops well below the ionisation energy.

The relative abundance of protons, electrons, and neutral hydrogen atoms can be calculated using the *Saha equation*. This is an approximate equation that relates the occupancy of two states to the energy difference

between them. It assumes that a thermal equilibrium exists (essentially, that switching between states can happen very rapidly and efficiently), which is a good approximation in the early Universe.

The two states that we are interested in as far as recombination is concerned are the left- and right-hand sides of the reaction above. One is the 'ionised' state, with separate protons and electrons, while the other is the 'neutral' state, consisting only of neutral hydrogen (and photons). Working in terms of the number densities of each type of particle, we can write

$$\frac{n_p\, n_e}{n_H} \approx \left(\frac{m_e k_\mathrm{B} T}{2\pi\hbar^2}\right)^{\frac{3}{2}} \exp\left(-\frac{E_\infty}{k_\mathrm{B} T}\right), \tag{112}$$

where n_p, n_e, n_H are the number densities of the particles, m_e is the electron mass, T is the temperature, and E_∞ is the energy required to ionise hydrogen from its ground state, $E_\infty = 13.6$ eV.

The right-hand side of this equation might remind you of the Maxwell-Boltzmann distribution, which describes the distribution of velocities of a thermal gas of massive particles. It is closely related. If a gas of particles (in this case, free electrons) is in thermal equilibrium at temperature T, the particles will tend to have a distribution of energies given by this form of equation. The threshold energy E_∞ determines whether the electrons have enough energy to remain free, or whether there is a good chance of them being absorbed to form a neutral atom.

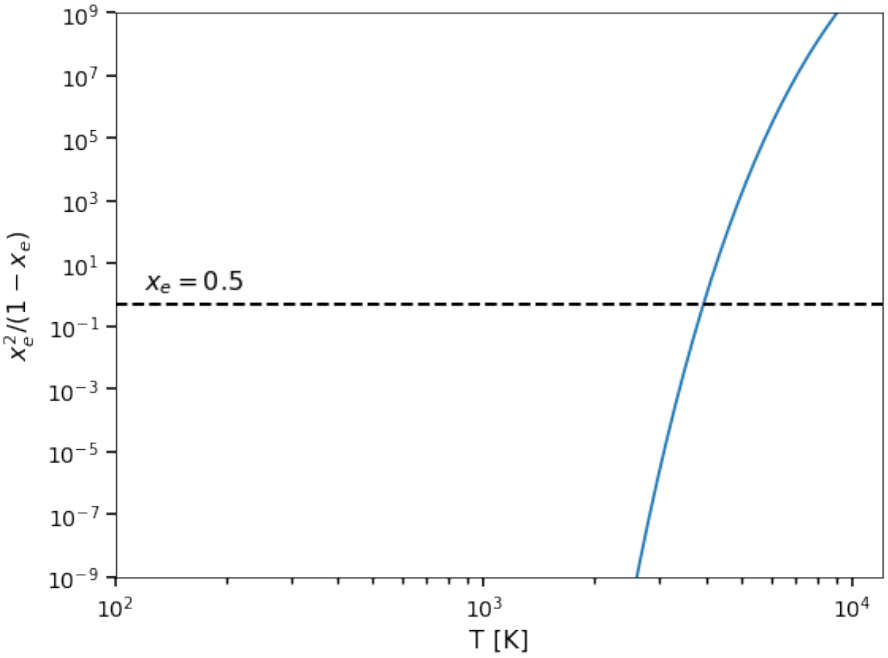

Figure 14: Plot showing the left-hand side of the Saha equation (y axis) as a function of temperature, T (blue line). The dashed black line shows the y value where $x_e = 0.5$. This intercepts the Saha equation where $T_\mathrm{rec} \approx 4000$ K. This is a good rough estimate of the temperature of recombination.

We can solve this equation to get a rough estimate of when recombination must have happened. First, we note that the Universe seems to be neutral on average, so there must be roughly one electron for every proton. This lets us set $n_e \approx n_p$. We also need to know the total number density of hydrogen (neutral or ionised) in the Universe, which is measured to be $n_\mathrm{H} + n_p \approx 1.6\,(1+z)^3$ m^{-3} (hydrogen is a type of matter, so its density scales like $a^{-3} = (1+z)^3$). Recall that the temperature itself scales with redshift like $T \propto (1+z)$.

If we now define the ratio of the electron number density to the total hydrogen number density as

$$x_e = \frac{n_e}{n_p + n_H}, \tag{113}$$

we can see that $x_e \to 0$ when the hydrogen is completely neutral, while $x_e \to 1$ when it is completely ionised. If we rewrite the Saha equation in terms of this variable, we get

$$\frac{x_e^2}{1-x_e} \approx \frac{1}{n_p + n_H}\left(\frac{m_e k_\mathrm{B} T}{2\pi\hbar^2}\right)^{\frac{3}{2}} \exp\left(-\frac{E_\infty}{k_\mathrm{B} T}\right), \tag{114}$$

A reasonable guess for when recombination happens is when around half the hydrogen is ionised and half is neutral, so $x_e \approx 0.5$. We can solve the Saha equation graphically (i.e. by plotting it) to find the temperature at which the RHS is equal to the LHS with $x_e = 0.5$. The result is shown in Fig. 14: $x_e = 0.5$ when $T \approx 4000$ K. This corresponds to a recombination redshift of $z_{\rm rec} \approx 1450$, which is not too far away from the value of $z_{\rm rec} \approx 1090$ obtained from a more precise calculation.

A temperature of 4000 K corresponds to an energy of approximately 0.35 eV – *much* lower than the ionisation energy of hydrogen (13.6 eV). This is how much extra the Universe had to cool in order for neutral hydrogen to form without energetic photons rapidly reionising it again.

7.5. Decoupling

We have seen how the Universe needs to cool down significantly before neutral atoms can exist. Now let's figure out when the Universe would have become transparent to light.

The condition for light to be able to travel freely in the Universe is that its mean free path must be large. Large compared to what? Well, ideally, each photon would interact with matter at most once as it travelled across the Universe. Otherwise, it would potentially be scattered multiple times, and the Universe would again be opaque out to some distance.

As a rough guide, the mean free path of photons should therefore exceed the Hubble radius, $r_{\rm HR} \propto (aH)^{-1}$, at a particular time if we want the Universe to be transparent. Recall that this is the approximate maximum distance over which physical processes can operate at a given time in cosmic history. If the mean free path is larger than $r_{\rm HR}$, then a typical photon will never scatter off an electron as it travels through space.

We can find when this should happen by simply equating the expression for the mean free path (see above) with the expression for the Hubble radius (see previous sections). If we're careful about remembering whether these quantities are in proper or comoving coordinates, we arrive at the following expression that defines when decoupling happens:

$$\frac{l_{\rm mfp}}{a} \gtrsim \frac{c}{aH(a)}, \tag{115}$$

where we have written the *comoving* mean free path on the left-hand side. Substituting the definition of the mean free path from above, we obtain

$$\frac{1}{n_e \sigma_T \, a} \gtrsim \frac{c}{aH(a)} \implies H(a) \gtrsim n_e \sigma_T c, \tag{116}$$

which implies that decoupling happens at a scale factor $a = a_{\rm dec}$ twhen

$$H(a_{\rm dec}) \simeq n_e(a_{\rm dec}) \sigma_T c. \tag{117}$$

We know that decoupling happens in the matter-dominated era of cosmic history, when $\rho_m \gg \rho_r$, so $H(a) \propto a^{-\frac{3}{2}}$ at this time. We also know from above that the number density of electrons should scale like matter, $n_e \propto a^{-3}$, while the Universe remains strongly ionised. We could therefore solve this equation for $a_{\rm dec}$ to get an estimate of the redshift at which decoupling happened.

A complication is that n_e will drop even more rapidly when recombination starts to happen, as electrons become bound into neutral hydrogen atoms and $x_e \to 0$. In turn, this will cause the mean free path to rapidly increase. As a result, decoupling must therefore happen at a very similar time to recombination, as recombination will rapidly drive $n_e \to 0$, therefore satisfying the condition for decoupling to happen from above for practically any value of the Hubble radius.

7.6. The surface of last scattering

Recall the definition of the comoving distance travelled by light from Section 4. The CMB photons we see on Earth today are the ones that have travelled from regions of the Universe at a comoving distance of $r(z \approx 1100)$ away from us. Photons emitted from regions that are closer to us already travelled past us a long time ago, while photons emitted from further away are yet to arrive. They were all emitted at the same *time* though.

If we were to draw the regions that we are currently receiving CMB photons from, we would trace out a spherical surface around us at a constant comoving distance $r = r(z \approx 1100)$. This surface is called the *last-scattering surface*. We can draw a *different* last-scattering surface around every point in the Universe; different observers in the Universe will all see a slightly different CMB, emitted from different regions in space. It always corresponds to photons that were emitted at the same time though – when the Universe first became transparent.

7.7. Blackbody spectrum of the CMB

The ionised hydrogen gas and radiation were very close to being in thermal equilibrium before recombination and decoupling happened. We saw that this gave a particular distribution of electron energies in the Saha equation. It also gives the radiation a very particular energy distribution, known as a *blackbody* or Planck distribution, which has the form

$$B_\nu(T) = \frac{2h\nu^3}{c^2} \frac{1}{e^{\frac{h\nu}{k_B T}} - 1}. \tag{118}$$

When last scattering happened, the Universe was so close to being in thermal equilibrium that the resulting radiation – the CMB – was emitted with an almost perfect blackbody spectrum. Measurements of this spectrum made by the COBE space mission are shown in the plot below. The measurement is so precise that the error bars have been blown up by a factor of 400 just so that you can see them, but even so the CMB fits the expected blackbody curve almost exactly.

The measured temperature of the CMB today is $T = 2.725$ K. By scaling this temperature with redshift, we can figure out what the temperature of the Universe must have been at any redshift, $T(z) = T_0(1 + z)$, where $T_0 = 2.725$ K. We can also use this to work out the energy density of radiation today, since $\rho_r \propto T^4$.

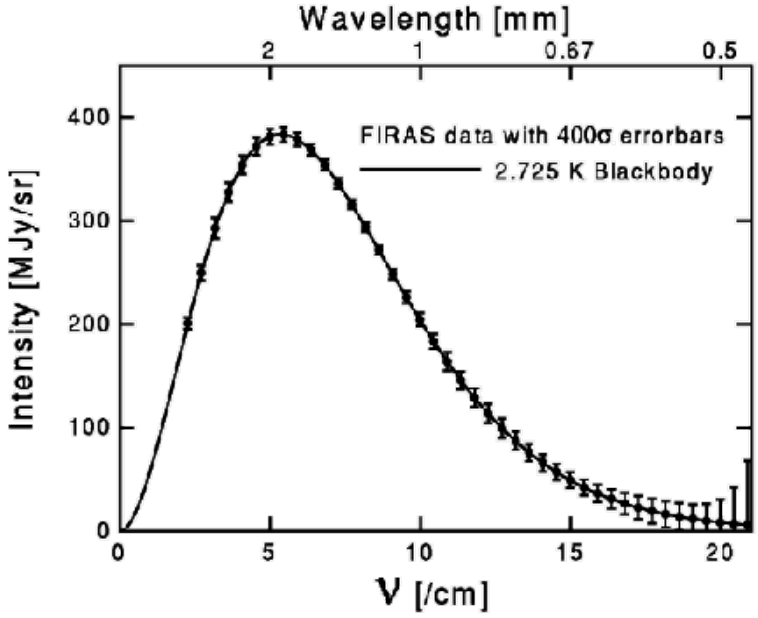

Figure 15: Frequency spectrum of the cosmic microwave background, measured by the COBE mission. It is extremely close to a blackbody spectrum with temperature $T = 2.725$ K; the error bars on the measurement have been drawn 400× bigger than they really are just to be visible! (Credit: COBE FIRAS)

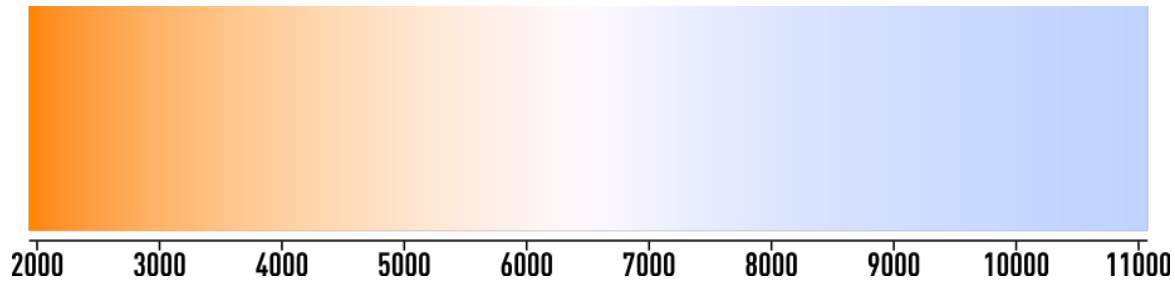

Figure 16: Colours corresponding to blackbody radiation of different temperatures. (Credit: Wikimedia)

Learning outcomes:

What is the Cosmic Microwave Background (CMB)?
What was the Universe like before the CMB formed?
What is the mean free path of a photon in an ionised medium?
What are decoupling and recombination?
What roles did decoupling and recombination play in the formation of the CMB?
What is the last scattering surface of the CMB? How is it defined for different observers?
What is a blackbody frequency spectrum?
How does the temperature of a blackbody spectrum vary with redshift?
At approximately what redshift did the CMB form?

8. Cosmic Microwave Background Anisotropies

In this section you will learn about the small fluctuations (temperature anisotropies) that we observe in the Cosmic Microwave Background radiation. You will see how the CMB anisotropies can be analysed using the *spherical harmonic* expansion and a statistical tool called the power spectrum, and how different features in the power spectrum are attributable to different physical effects.

Reading for this topic:

- *An Introduction to Modern Cosmology (A. Liddle), Chapters 10 and 12.*
- *An Introduction to Modern Cosmology (A. Liddle), Advanced Topic 5.4 (CMB Anisotropies).*

8.1. CMB anisotropies

The early Universe was extremely homogeneous – typical fluctuations in the radiation and matter energy density were around 1 part in 10^5 or so. This homogeneity was caused by two main processes. The first was *inflation*, a process that occurred in the first fractions of a second of cosmic history, and which we will learn about in the following section. Inflation set the 'initial conditions' of our Universe, smoothing out any inhomogeneities that may have been there before, and leaving behind only tiny fluctuations by the time it finished (again, in the first fractions of a second).

The second was *thermalisation*, the tendency for any temperature and pressure differences to average out because the early Universe was in a state very close to thermal equilibrium. As we learned in the previous section, the mean free path of photons was very low before last-scattering; photons could typically travel only a very short distance before being scattered off an electron. On average, this tends to make larger fluctuations smaller – slightly hotter regions of the Universe would tend to lose energy, while slightly colder regions would gain energy. This process resulted in the almost-perfect blackbody radiation that we now see as the CMB.

The fact is that there *were* small fluctuations in the early Universe however. Different physical processes shaped these fluctuations; some processes worked to erase them (like thermalisation); some generated them (like inflation); and some allowed existing fluctuations to grow (like gravitational attraction).

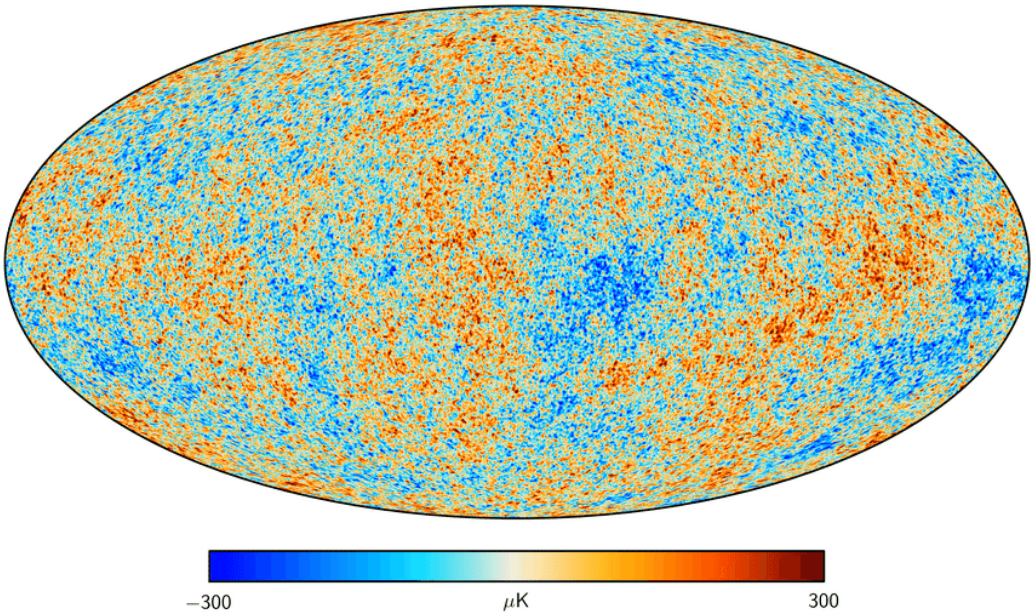

Figure 17: Temperature anisotropies of the CMB, shown in a Mollweide projection. (Credit: Planck Collaboration)

In fact, there were many physical processes at work in the early Universe. Each of them had a distinctive effect on the fluctuations in the energy density of radiation, baryons, and dark matter, which we are now able to

observe because they caused fluctuations in the CMB temperature. These fluctuations are called **anisotropies**, because we see them as deviations from the average CMB temperature that are different in different directions. Without these fluctuations, the CMB would be perfectly isotropic, i.e. perfectly uniform across the whole sky. By observing these fluctuations, we are in some sense getting a 'snapshot' of the Universe as it was very early in its history, around 380,000 years after the Big Bang (when last-scattering occurred).

A map of these fluctuations is shown in the figure above. This map was observed by the Planck satellite, an ESA mission that mapped the CMB temperature very precisely at microwave frequencies. The figure you see is a projection of the whole sky (a sphere) onto an egg-like shape. This type of projection is called a Mollweide projection, and is also used for making maps of the Earth for example, as in the figure below. Note how it distorts the shapes of objects (like continents). It preserves the *area* of the objects though, so the continents in this map are at least the correct size. We use the Mollweide projection when we make maps of the CMB precisely because of this area-preserving property – the size of the fluctuations is an important signature of the physical processes that caused the anisotropies, as we shall soon see.

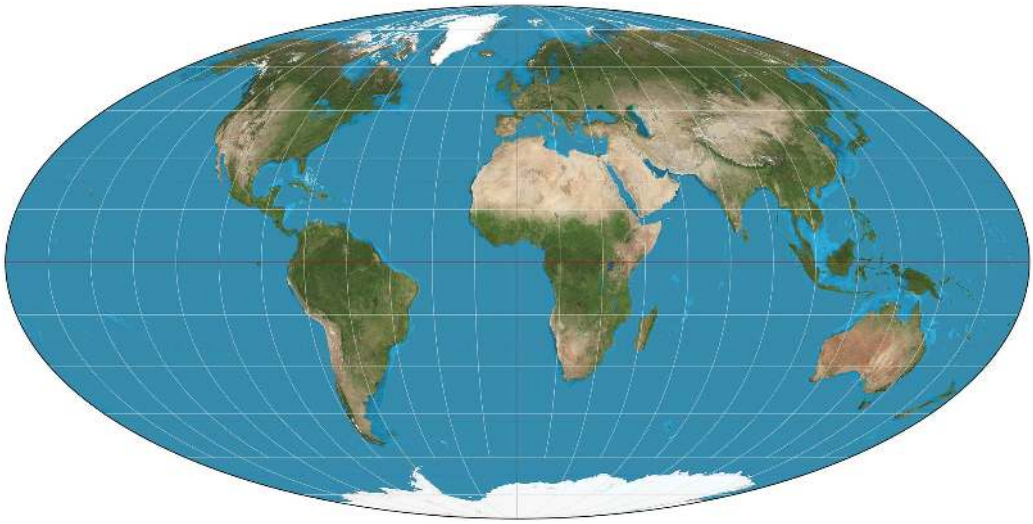

Figure 18: A Mollweide projection of the surface of the Earth. (Credit: Strube/Wikipedia)

Also note the temperature scale of the CMB map. The mean temperature ($T = 2.725$ K) has been subtracted from this map, leaving behind fluctuations that are at most a few hundred μK (micro-Kelvin) in size. A quick calculation gives us the typical size of the fluctuations as a fraction of the mean temperature:

$$\frac{\Delta T}{T} \sim \frac{100\,\mu\mathrm{K}}{2.725\,\mathrm{K}} \approx 4 \times 10^{-5}. \tag{119}$$

This is pretty tiny! To measure the fluctuations, we need a device that can tell apart the temperatures in a radiation field to 1 part in 10^5 or better. A lot of the technology to make these measurements possible was developed right here in the Physics Department at Queen Mary in the 1960's and 70's, back when it was called Queen Mary College. The research group was eventually spun out into a company called QMC Instruments, which is now based in Cardiff and still plays an important role in developing sensitive EM radiation detectors for use in astronomy.

8.2. Physical processes that cause anisotropies

As mentioned above, many different physical processes contribute to cause the temperature anisotropies that we see in the CMB. What are these processes, and how do they cause the anisotropies?

Any process that is capable of changing the temperature of the CMB radiation field *locally* will cause anisotropies. Processes such as cosmological redshift are global, in the sense that they affect everywhere in the Universe in the same way at the same time. As such, while the cosmological redshift does change the temperature of the CMB, it cannot cause anisotropies because it doesn't change the temperature by different amounts in different places/directions.

What do we mean by 'local'? We mean that the process can occur more or less strongly in some regions than in others. The 'strength' of the effect will depend on some underlying physics, to be discussed shortly. The *size* of the region affected will also depend on the underlying physical process. This is a really important point – different physical processes act over different distances. So, if we map out the CMB anisotropies as a function of their size, we can potentially identify which processes were at work, and how strong those processes were. We can build a picture of what physics was like in the early Universe!

What can locally change the temperature of the CMB radiation field? Recall that before last scattering, photons and baryons were tightly coupled together by Thomson scattering (scattering of photons off electrons in the highly-ionised gas). As a result, we would expect any changes in the baryon distribution to very rapidly get picked up by the photon distribution and vice versa. The coupling between the two is so strong that we refer to the photons and baryons as a 'coupled photon-baryon fluid' before last scattering. If we can cause fluctuations in this fluid somehow, they will be converted into temperature anisotropies very efficiently.

The easiest way to do this is to change the energy density of the baryons locally. If a region has baryons with a slightly larger energy density, some of that excess energy will be transmitted into the photons, hence increasing their temperature (recall that $\rho_r \propto T^4$). So, by increasing the density of baryons in a region, we can increase the CMB temperature. This can be achieved through **gravitational collapse** – if a region starts off with a slightly higher matter/energy density than average, gravity will tend to attract more matter/energy towards it, thus increasing its density further. Likewise, regions that are slightly lower density than the average will tend to decrease further in density as matter is attracted away from them by their surroundings.

This works just as well for the dark matter too; fluctuations in the dark matter density are also enhanced by gravitational collapse. Since there is about 5 times as much dark matter as baryonic matter in our Universe, the dark matter fluctuations can be quite important – baryons tend to fall into the potential wells caused by the collapsing dark matter, thus enhancing the fluctuations in the baryon distribution too. This is of course then transmitted to the photon distribution, and so fluctuations in both dark matter and baryons cause temperature anisotropies.

The process of gravitational collapse requires matter to move into some regions and out from others. There are therefore *flows* of matter in the early Universe, with a range of velocities. CMB photons that are scattered off baryons that are moving with respect to everything else receive a small **Doppler shift**, changing their frequency (and therefore energy) by a small amount depending on the direction of their relative motion. Region of the last-scattering surface that were moving towards us when last scattering happened will have their photons slightly blueshifted, while regions moving away from us at that time will be slightly redshifted. If we recall that temperature depends on redshift as $T = T_0(1+z)$, we can see that a similar relation must also hold for the Doppler shift. Writing this as a fractional change in the temperature, we obtain

$$\frac{\Delta T}{T} = -\frac{v}{c}. \tag{120}$$

Regions with a velocity away from us ($v > 0$) are redshifted, which reduces the temperature – hence the minus sign.

Another type of redshift is **gravitational redshift**. Photons travelling out from a gravitational potential well lose energy in the process, and so are redshifted slightly, reducing their temperature. Likewise, photons falling into a potential well gain energy and so are blueshifted. This effect is called the **Sachs-Wolfe effect**, and affects photons that were inside a potential well when last scattering occurred.

To recap:

- Fluctuations in the energy density of photons, baryons, and dark matter are enhanced by gravitational attraction. Increased density 'heats up' the photon temperature.
- The motion of matter flowing into/out of over-dense/under-dense regions causes a Doppler shift. This redshifts/ blueshifts the photons, also affecting their temperature.
- Photons travelling into/out of gravitational wells are blue/redshifted, which also changes their temperature.

8.3. Baryon acoustic oscillations

The fact that the photons and baryons are so tightly coupled, while also collapsing under gravity, gives rise to an interesting phenomenon.

We saw above that gravitational collapse makes over-dense regions even denser. The compression of the baryons heats them up, which also heats up the photons within that region. Both photons and baryons exert an outward pressure – radiation pressure and thermal pressure respectively – that counteracts the gravitational collapse.

A similar situation can be found in stars that are forming. The stars collapse under gravity, causing the proto-stellar material to heat up. In this case, nuclear fusion starts, and generates a very large outward pressure that prevents the star from collapsing further. The star eventually enters *hydrostatic equilibrium* – the outward pressure exactly balances the gravitational collapse and the size of the star stabilises.

This does *not* happen in the early Universe though. First of all, the densities are not high enough for nuclear fusion to start. The energy density of the photon-baryon fluid is only increased by adiabatic compression, and there is no additional source of energy to increase the pressure as the fluid is compressed. In fact, the photon-baryon fluid responds to the compression with a significant increase in outward pressure that overcomes the gravitational attraction and causes the region to expand outwards. This expansion cannot continue for long due to the inward pressure of the surrounding photon-baryon fluid though, and so gravitational collapse soon takes over again. The compression and expansion cycles continue, setting up **oscillations** of the photon-baryon fluid. The fluid does not stabilise into hydrostatic equilibrium – it continues to oscillate until decoupling suddenly prevents the photons and baryons from strongly interacting with each other any more.

These oscillations propagate as sound waves in the photon-baryon fluid, hence the name 'baryon acoustic oscillations' (BAO). They are *almost* like standing waves, in that they have wavelengths that are close to multiples of the **sound horizon** – the maximum distance a sound wave could have travelled since the Big Bang. The sound horizon is smaller than the Hubble radius and particle horizon because the speed of sound in the photon-baryon fluid is smaller than the speed of light in vacuum. The sound speed (squared) can be calculated as

$$c_s^2 = \frac{\dot{p}_\gamma c^2}{\dot{\rho}_\gamma}, \tag{121}$$

where p_γ and ρ_γ are the pressure and density of the photons respectively. Since we know how the pressure and density of photons vary with time (they are types of radiation), we can work out that $c_s^2 \simeq c^2/3$. The sound horizon can be worked out by integrating the sound speed over time, and for our Universe is found to have a value of $r_s \simeq 150$ Mpc (comoving) at $z \approx 1090$.

If the BAO were a standing wave, the sound horizon would give us the wavelength of the fundamental mode. There are also harmonic modes, with wavelengths that are integer fractions of the sound horizon. The Universe is constantly expanding though, so the size of the fundamental mode should be constantly increasing – the actual acoustic waves therefore never quite manage to resonate across the sound horizon. The photons also cause a small amount of dissipation of the waves (more on this later), which also damps the resonance. As a result, the waves do not have a single, well-defined wavelength (plus harmonics) – they are smeared out a bit, and so take a broader range of wavelengths around the acoustic horizon size.

Nevertheless, this leaves a **preferred scale** in the photon-baryon fluid. The typical distance between two over-dense clumps of baryons will tend to be some multiple of the wavelength of the sound waves – there are more baryons around the peak of a wave, and fewer around its trough. Recall that higher density regions of baryons tend to be hotter and therefore give rise to positive temperature fluctuations. The BAOs therefore cause temperature anisotropies across the last scattering, with sizes of order the sound horizon (or harmonics of this). As we will see shortly, these sound waves leave a *very* distinctive statistical pattern in the CMB anisotropies.

What happens when decoupling occurs, and the photons and baryons are no longer strongly coupled together? The photons immediately stream away, reducing the outward pressure and preventing any further oscillations from occurring. The acoustic waves stall, left in whatever configuration that they were in immediately before decoupling. The pattern of peaks and troughs is **frozen in** to the baryon distribution. Regions with a high density of baryons can subsequently collapse gravitationally, but the distance between peaks in the baryon density left by the sound waves will not change. This preferred distance scale is called the BAO scale, and is left imprinted in the distribution of baryons (and dark matter), unchanged even today.

This brings us to an important subtlety that we haven't mentioned yet, which is the role of dark matter. While dark matter is not strongly coupled to the photons during the pre-decoupling era, it does still interact with the photons and baryons gravitationally. In our Universe, the dark matter density is around 5 times larger than the baryon density, so if there are over-densities in the dark matter, the baryons will tend to fall towards them. Over-densities in the baryons can also drag the dark matter towards them a bit, although this is a less prominent effect. The dark matter therefore responds to the waves in the baryon distribution more slowly, acting as a sort of counterweight to the oscillations.

When decoupling happens, the dark matter continues to be gravitationally attracted towards the baryon over-densities, and eventually catches up with them. This is called the **drag epoch**, as the baryons are dragging the dark matter towards them (and vice versa; the baryons will be attracted back towards the dark matter too). The sound horizon at decoupling is therefore not *quite* the final size of the BAO waves that we observe today; instead we see a slightly modified wavelength, due to the brief period of baryon/dark matter dragging. This doesn't last for very long however, and only changes the wavelength by a couple of Mpc. The redshift at which the dragging completes is only slightly smaller than decoupling, $z_{\rm drag} \approx 1060$, and the final BAO feature has a comoving size of $r_{\rm BAO} \approx 147$ Mpc.

A final subtlety: the acoustic oscillations are happening everywhere in the Universe at the same time. As a result, there are many, many overlapping sound waves being generated across the Universe at the same time, with similar wavelengths. So, we don't see individual waves in the distribution of baryons when we observe the CMB anisotropies; we see a superposition of many such waves, all jumbled up on top of each other. A similar effect can be obtained by dropping many stones of the same size in a pond. The waves they generate will all be broadly similar in wavelength, but the fact that many of them have been dropped at different positions results in a complicated superposition of many waves on the surface of the pond. If we look at the CMB anisotropies, there is a similar effect; we don't see the BAO as individual waves, but when we analyse the anisotropies statistically, we see that there is a preferred separation of higher-temperature (over-dense) regions that is around the BAO scale.

Does this sound complicated? It is! The theory behind baryon acoustic oscillations was only worked out in the late nineties, and the conclusive measurements of the phenomenon were made in the early 2000s. The fact that we see them is a beautiful confirmation that we understand a lot of what's going on in the early Universe though.

To recap:

- Baryons and photons are tightly coupled before last scattering.
- Gravitational attraction compresses the baryon-photon fluid, heating it up. This generates outward thermal/radiation pressure, which pushes it out again.
- This sets up oscillations, with a typical wavelength given by the sound horizon (which can be calculated from the sound speed, $c_s \approx c/\sqrt{3}$).
- After decoupling, the sound waves in the baryons are frozen. They leave an imprint in the CMB temperature anisotropies, as well as a preferred distance scale in the matter distribution that we can observe even today.
- The dark matter is also affected by these sound waves, and is dragged along with the baryons by gravitational attraction.

8.4. Diffusion damping

While the mean free path of the photons is small before decoupling, it is not completely negligible – they do travel a short distance before scattering. This allows the photons to diffuse out from hotter, higher-density regions into surrounding cooler, lower-density regions, as long as those regions are sufficiently nearby. A combination of the photon pressure and gravitational attraction drags the baryons along with the photons too, causing the baryons to diffuse also. Dark matter is pulled along with the photons and baryons too. This effect smooths out the temperature anisotropies over distances smaller than the photon mean free path, erasing the anisotropies generated by other processes. It is called *diffusion damping* or *Silk damping*.

The mean free path slowly increases as the Universe expands, so the diffusion affects larger and larger regions as time goes by. The diffusion process takes time, however; regions that have been smaller than the photon mean free path for only a short time will not have suffered from as much diffusion, so the anisotropies will only be mildly damped, while those that have been smaller than the photon mean free path for a long time will have been almost entirely smoothed out. This causes a scale-dependent damping of the CMB anisotropies – the smaller distance/angular scales you look at, the stronger the damping effect. At sufficiently small angular scales, there are no primary CMB anisotropies at all, although secondary anisotropies (see the next section) can be generated on small scales long after last scattering. On sufficiently large angular scales, larger than the mean free path of photons immediately before decoupling, there is no diffusion damping effect.

8.5. Secondary anisotropies

All of the anisotropies that we've discussed so far are caused by physical processes happening at the time of last scattering, on the surface of last scattering. They are known as *primary anisotropies*.

The temperature of the CMB photons can also be changed as they travel through space after last-scattering, causing *secondary anisotropies*. These tend to be much weaker than the primary anisotropies, as after last-scattering, CMB photons are not confined to one local region, but travel rapidly through the Universe. As such, they typically traverse a number of regions with different local physical conditions before reaching the observer. Since some regions will increase the temperature and others will decrease it, travelling through many regions means that the net effect will be the sum of many positive and negative contributions, which tend to cancel each other out on average. Still, some small secondary effects are certainly observable.

The figure below shows a map of **gravitational lensing** of the CMB, as observed by the Planck satellite. As CMB photons travel through the Universe, their paths are bent, very slightly, by regions of higher mass density that they pass close to. The fact that light is bent by mass is a key prediction of general relativity, and this map is a wonderful confirmation of this prediction on the scale of the entire Universe! What you are seeing is essentially a map of all of the mass between our surface of last scattering and Earth, averaged into a single number from each line of sight (i.e. in each direction on the sky). Several CMB experiments have made this measurement by making a very high-resolution map of the CMB, and then figuring out which direction the CMB photons *would* have come from if they hadn't been lensed. It's an incredible measurement that has only been made for the first time within the past 10 years.

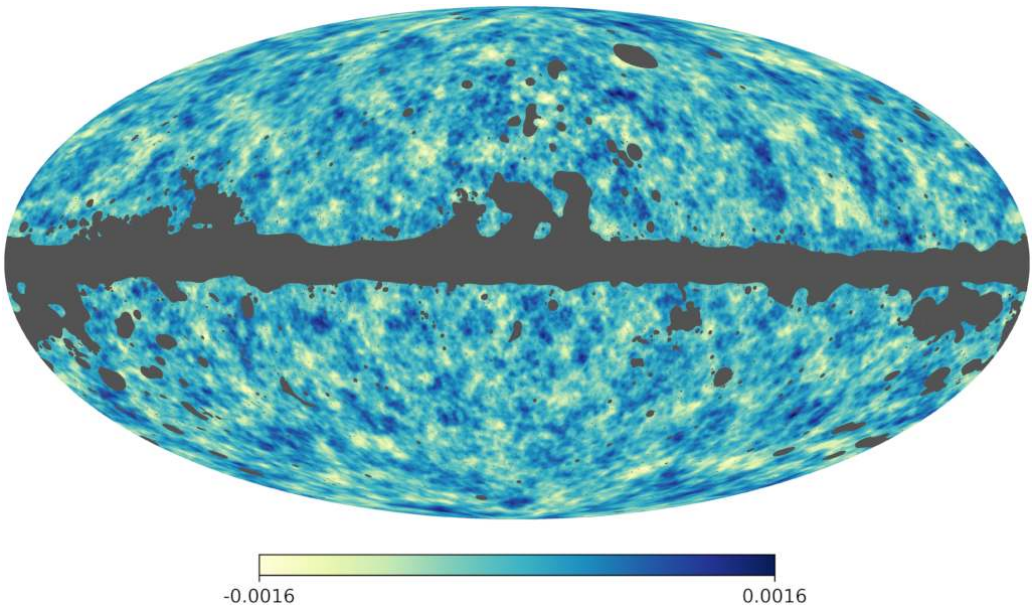

Figure 19: A Mollweide projection of the CMB lensing potential as measured by the Planck satellite. This is essentially a map of the projected mass density throughout the entire Universe, from last-scattering until today. The grey regions are regions that couldn't be measured because our own galaxy is in the way. (Credit: ESA/Planck)

Another type of secondary anisotropy is the **integrated Sachs-Wolfe** (ISW) effect. This is similar to the

Sachs-Wolfe effect, in that it is caused by the gravitational redshifting of CMB photons. This time, the photons are passing through gravitational potential wells on their way to the observer, instead of at the surface of last scattering. As they fall into potential wells they gain energy and are blueshifted, but this is exactly cancelled by the redshift as they travel back out of the well. On average, then, there is no net blue- or redshifting caused by potentials that the photons travel through after being emitted.

This is only true if the potentials do not change with time however. If a potential gets deeper in the time it takes a photon to travel through it, the photon will experience a steeper slope on the way out than it did on the way in. The blueshift experienced when travelling into the potential (when it was shallower) will therefore be less than the redshift experienced when leaving the potential (now it has grown deeper). In our Universe, potential wells remain almost constant until quite late times, and so the ISW effect doesn't happen until the photons have travelled a long distance. When dark energy starts to dominate the cosmic energy budget at late times, however (when $\rho_\Lambda > \rho_m$), it causes the potential wells to decay (get shallower)! This adds a small but non-negligible ISW contribution to the CMB anisotropies that we see. The fact that the ISW effect can be measured from the CMB is strong evidence that dark energy is real – without it, there would be no ISW secondary anisotropy to observe.

A final type of secondary anisotropy is called a **spectral distortion**. All of the anisotropies that we have discussed so far simply change the temperature of the CMB photons by a small amount, but still leave their energy distribution as an almost perfect blackbody (Planck) distribution. Spectral distortions, on the other hand, also induce deviations in the photon energy distribution away from a perfect blackbody. This can be achieved by giving energy to the photons from matter that is not in thermal equilibrium with them.

The biggest source of such energy is the extremely hot but low-density gas that exists inside galaxy clusters. This gas has been heated up to millions of Kelvin during the formation of the galaxy clusters, and is so hot that it glows at X-ray wavelengths! As CMB photons travel through this gas, they occasionally scatter off a high-energy electron, gaining some of its energy and thus departing from the blackbody energy distribution that they had previously. The physical process at work is called *Compton scattering*, and gives rise to different types of spectral distortion called the **thermal and kinetic Sunyaev Zel'dovich effects**. By measuring the size of these effects, we can learn about the distribution of galaxy clusters between us and the last-scattering surface, and how hot gas comes to be distributed throughout the Universe long after recombination.

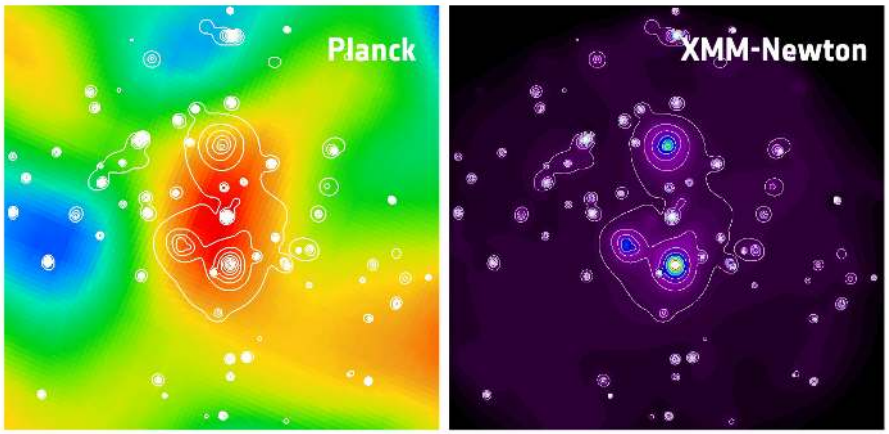

Figure 20: A massive galaxy cluster, seen via the CMB temperature anisotropy caused by the Sunyaev-Zel'dovich effect (left) and in X-ray emission (right). Particular hot-spots in X-ray emission are shown using white contours in both panels. (Credit: ESA/Planck)

8.6. Spherical harmonics

We can't get much information by simply staring at a map of the CMB. As discussed above, the really important information comes from figuring out how the temperature anisotropies depend on the size of the region being observed, as this tells us which physical processes were at work at the time the anisotropies were generated. Some physical processes operate across larger distances than others, and some have a bigger effect than others.

To see this, what we would really like is some measure of the amount of anisotropy as a function of *angular*

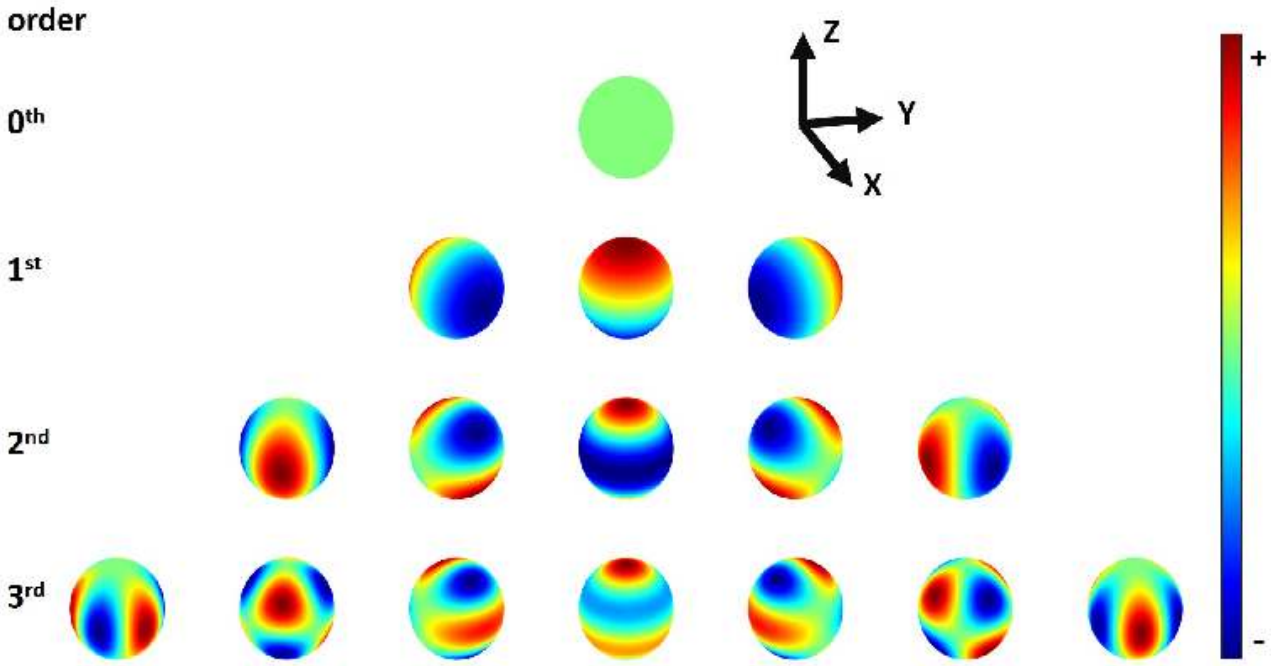

Figure 21: Spherical harmonic modes. Each row ('order') corresponds to an ℓ mode, with $2\ell + 1$ m-modes per row (Credit: M. Adjeiwaah et al.)

scale, i.e. the size of the observed region. The mathematical tool that allows us to do this is called the *spherical harmonic transform*, and it is very similar in principle to a Fourier transform. Instead of viewing a map of the CMB, we can use the spherical harmonic transform to break up the map into a sum of many waves on the sky, with a range of different 'wavelengths'. The amplitude of each wave component tells us how much of the anisotropy comes from angular scales corresponding to the wavelength of that wave. Long waves (on the sky) correspond to large angular scales, while short waves correspond to small angular scales.

Mathematically, we can write the spherical harmonic expansion of the temperature anisotropies as a sum,

$$\frac{\Delta T}{T}(\hat{n}) = \sum_{l} \sum_{m} a_{\ell m} Y_{\ell m}(\hat{n}). \tag{122}$$

This says that we can work out the size of the fractional temperature anisotropy, $\Delta T/T$, in any direction on the sky, $\hat{n}$, by summing over all of the functions $Y_{\ell m}$, where each function has some amplitude $a_{\ell m}$. The ℓ and m label different modes, or wavenumbers. There are two wavenumbers because we are performing this expansion on the sky, which is a 2D surface (so we need 2D coordinates).

Recall that Fourier transforms work in a similar way. A function $f(x)$ can be written as a sum over Fourier modes,

$$f(x) = \sum_{n} f_n e^{2\pi i n x/L} \tag{123}$$

where $k_n = 2\pi n/L$ is the Fourier wavenumber, and f_n is the amplitude of each Fourier mode. Larger values of n correspond to larger Fourier wavenumbers, which have *smaller* wavelengths. We can write any reasonable continuous 1D function as a sum over Fourier modes. Similarly, we can write any continuous 2D field on a sphere (like the sky) as a sum over spherical harmonic modes. Just as a function is uniquely specified by a set of Fourier coefficients f_n, a 2D field on a sphere is uniquely specified by a set of spherical harmonic coefficients, $a_{\ell m}$.

The figure above shows the spherical harmonic functions $Y_{\ell m}$ for the first few values of ℓ and m. For a given value of ℓ, only integer values of m between $-\ell$ and $+\ell$ are allowed. For each ℓ mode, there are $2\ell + 1$ different m modes.

Let's pay particular attention to the first two rows of the figure. The first row is $\ell = 0$. This is called the **monopole** mode, and there is only one m-mode ($m = 0$). This mode is perfectly uniform over the entire sky.

Its interpretation is as the average over the whole sky – if you take any map of the sky and average over all angles, you will get the value a_{00}, corresponding to $(\ell, m) = (0, 0)$.

The next row corresponds to the **dipole**. This corresponds to an anisotropy pattern on the sky that is positive in one direction and negative in the opposite direction. If this were a Fourier mode, we would have fit a single wavelength across the whole sky – one peak and one trough. In general, this wave could be 'pointing' in any direction, correspond to the three Cartesian coordinates, x, y, z, or some combination of them. This is why there are three m-modes for $\ell = 1$; each of $a_{1,-1}$, $a_{1,0}$, $a_{1,+1}$ tells us how much of this mode is pointing in each of the directions x, y, and z.

It gets harder to visualise, but we can extend this mathematical pattern to higher values of ℓ. You can see that the $m = 0$ mode always 'points' in the Cartesian z direction for any value of ℓ. You can also see that higher values of ℓ mean that more 'wavelengths' of the wave can be fit across the sky. Larger ℓ implies shorter wavelengths, and therefore smaller angular scales on the sky. The approximate relation between wavenumber ℓ and angular scale $\Delta\theta$ is

$$\Delta\theta \simeq \frac{\pi}{\ell}, \tag{124}$$

where angles are measured in radians. An angle of 1 degree corresponds to $l \approx 180$, while 10 degrees is $\ell \approx 18$.

To calculate the values of the spherical harmonic coefficients, we can take a map of the sky (i.e. a value of T in every direction $\hat{n}$), multiply it by the complex conjugate of the spherical harmonic function, and then integrate over the whole sky:

$$a_{\ell m} = \int T(\hat{n}) Y^*_{\ell m}(\hat{n}) d\Omega. \tag{125}$$

This result follows from the fact that the spherical harmonics form an *orthonormal basis*.

8.7. Power spectrum of the CMB

To see how big the anisotropies are as a function of angular scale, we can calculate a quantity called the *power spectrum*. The power spectrum is a statistical quantity related to the *variance* of the temperature anisotropies.

Why do we need to take the variance? It's because the anisotropies themselves are *random but correlated*. The initial conditions of the Universe are random (which we'll discuss when we learn about inflation), which leads to the anisotropies having a high level of randomness too – we can't predict where a particular temperature fluctuation will occur, for example, or exactly how big it will be. The physical processes at work in the early Universe *correlate* the anisotropies though, as they interact and so can come to share some of the same properties. Because they are correlated, if we know the size of one anisotropy, we can make a reasonable estimate of the size of another, neighbouring, one.

The upshot of all this is that the anisotropies are random, but we can predict some things about them – particularly their typical size as a function of angular scale. This is exactly what the variance (and therefore the power spectrum) tells us. Note that the mean of the anisotropies is zero, by construction – we are measuring them as fluctuations around the average CMB temperature, which we measure separately.

We normally measure the power spectrum as an average of the variance of the m-modes,

$$C_\ell = \frac{1}{2\ell + 1} \sum_{m=-\ell}^{+\ell} |a_{\ell m}|^2. \tag{126}$$

The square of the coefficients, $|a_{\ell m}|^2$, gives the variance of each spherical harmonic mode. The factor $2\ell + 1$ is the number of m-modes per ℓ-mode, and so dividing by this factor gives the average. The reason we average over the m-modes is because the CMB is expected to be *statistically isotropic*. Since we believe our Universe to be almost homogeneous, this means that the temperature fluctuations on the surface of last scattering should have been very similar everywhere. As we now observe them in the CMB radiation, the fluctuations should therefore have no preferred direction; they should not be larger on one side of the Universe compared to the other, for example. As a result of this statistical isotropy, we expect all of the m-modes to have the same statistical distribution (i.e. no dependence on direction), and therefore the same variance. This is why we can average them together, which we do in order to reduce the statistical uncertainty on the measurement.

The figure below shows the power spectrum of CMB temperature fluctuations measured by the Planck satellite. The x-axis shows the spherical harmonic wavenumber, ℓ. Recall that larger values of ℓ mean smaller angular scales.

8.8. Features in the CMB power spectrum

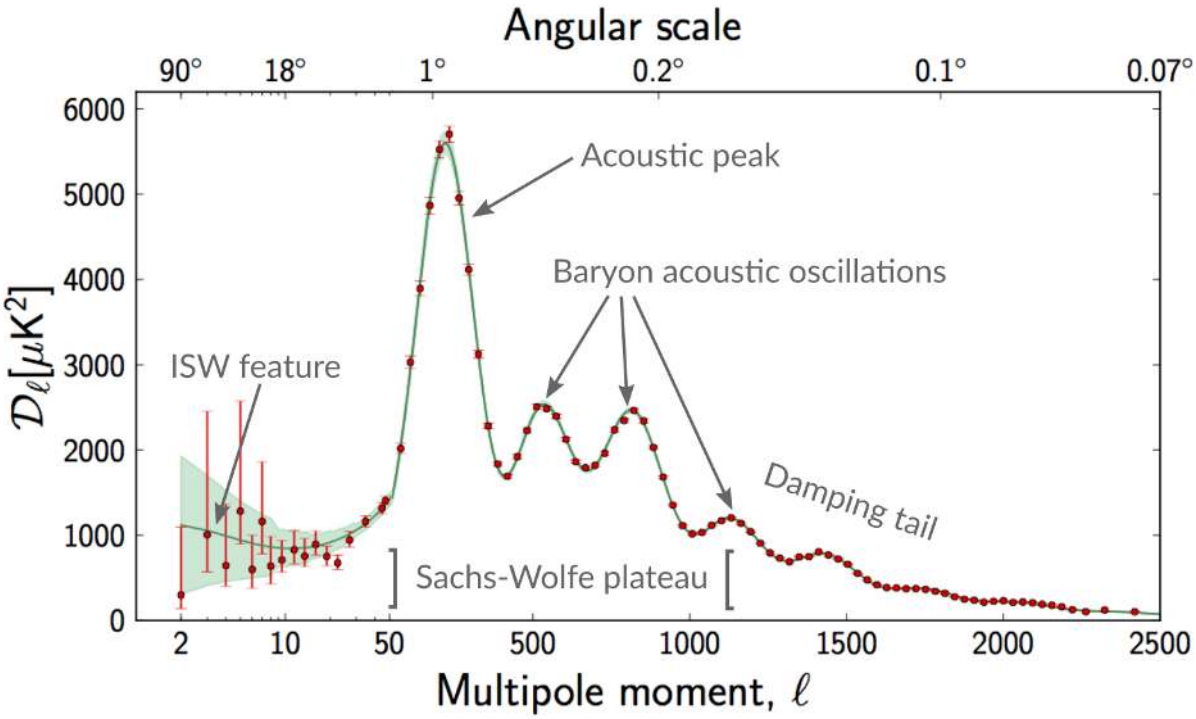

Figure 22: CMB power spectrum measured by the Planck satellite. The spherical harmonic multipole ℓ is shown along the bottom x-axis, with corresponding angular scale shown along the top x-axis. (Credit: Planck Collaboration)

- **Acoustic peak** – There are actually several acoustic peaks (see *baryon acoustic oscillations* below), but the first one is the biggest. It is observed at an angular scale of about 1°, and corresponds to the *sound horizon* of the photon-baryon fluid at the time of last scattering. This is essentially the fundamental mode of the sound waves in the photon-baryon fluid. It's important because it is the biggest, and therefore sets the typical angular size of the temperature fluctuations – if you look at the CMB map above, the typical size of the fluctuations that are most easily seen by eye is about 1°.
  The first acoustic peak is a very useful *standard ruler*. We can work out the size of the sound horizon at last scattering if we know some basic properties of the photon-baryon fluid, such as the relative densities of photons and baryons. The observed angular size of the acoustic peak can then be used to infer the angular diameter distance to last scattering, i.e. $d_A(z \approx 1090) = r_s/\Delta\theta_{\rm peak}$, where r_s is the size of the sound horizon.
  This measurement is very important in establishing the geometry of the Universe; if it was open or closed, the typical size of the anisotropies would be different, and so the acoustic peak would be at a different angular scale. The fact that it's observed at around a degree, when combined with other observations, tells us that our Universe must be very close to flat.

- **Baryon acoustic oscillations** – The sound waves in the photon-baryon fluid are also present on scales smaller than the sound horizon, and so we see a series of other peaks in the power spectrum extending to smaller scales. These are preferentially seen at distance scales that are a multiple of the sound horizon – i.e. harmonics of the fundamental acoustic mode. You can think of them as the modes that exist as standing waves in the photon-baryon fluid.
  The peaks are not narrow, and the power spectrum does not go to zero in between them, as we might

expect for harmonic modes of a musical instrument, for example. The peaks are broader because of damping, and because the size of the Universe is changing as the photon-baryon fluid oscillates (due to cosmic expansion). This makes the harmonics less well defined.
The power spectrum does not go to zero in between the peaks because of the Doppler shift. The peaks themselves are caused by the acoustic waves in the *density* of the photon-baryon fluid (higher density means higher CMB photon temperature and so on). Recall that the *velocity* of the fluid also causes a temperature change however. If you think back to the simple harmonic oscillator, you may remember that the velocity of the oscillator is highest when the displacement of the oscillator is lowest (e.g. a pendulum travels fastest when it is perfectly vertical). The same is true here – the velocity is largest in between the density fluctuations of the photon-baryon fluid. This means that there are temperature fluctuations caused by the Doppler shift with wavelengths exactly in between the wavelengths of the baryon density fluctuations. This tends to fill in the power spectrum between the acoustic oscillation peaks, although the effect is not as large because the Doppler shift produces a smaller temperature anisotropy than the density fluctuations.
As you can see, a lot of different physics is at work in producing the BAO peaks, and so it stands to reason that we can learn about a lot of physics by measuring them! These peaks do indeed tell us a lot about the early Universe, including the expansion rate at last scattering, and the relative amounts of dark matter, photons, and baryons. In particular, if matter was made purely of baryons (i.e. if there was no cold dark matter), these peaks would be much bigger.

- **Damping tail** – The amplitude of the CMB power spectrum decays away as we look on smaller and smaller scales (larger values of ℓ). This is primarily due to the diffusion damping effect, which is stronger for smaller angular scales. Experiments have actually been able to measure the power spectrum out to around the ninth BAO peak, at an ℓ of around 3000 or so. Beyond this, the primary CMB anisotropies are obscured by secondary anisotropies like the Sunyaev-Zel'dovich effect, which preferentially show up on smaller scales.

- **Sachs-Wolfe plateau** – The flat, low part of the power spectrum, visible at $\ell \lesssim 50$, is called the Sachs-Wolfe plateau. This is caused by gravitational redshifts at the surface of last scattering. Inflation left fluctuations in the gravitational potential on all distance scales – there are about as many small potential wells as large ones for example – so the Sachs-Wolfe plateau therefore extends across most of the range in ℓ in the figure. It's just that it's easiest to see at low values of ℓ, where other effects don't matter as much, or the anisotropies haven't been damped away.

- **Integrated Sachs-Wolfe feature** – The slight uptick in the power spectrum as we go to very low ℓ is caused by the *integrated Sachs-Wolfe (ISW) effect*. This is similar to the Sachs-Wolfe effect in that it is caused by gravitational redshifts – photons gaining or losing energy (and therefore changing temperature) as they pass through potential wells. Unlike the Sachs-Wolfe effect, however, this is not due to the potential wells that the photons were emitted from, but the ones that they pass through as they travel from the last scattering surface to the observer. The ISW effect only occurs in the recent, dark energy-dominated phase of cosmic history, as before then the potential wells were almost constant (neither growing nor decaying). The fact that we see an ISW feature at all is therefore evidence for the existence of dark energy, or something like it.

- **Amplitude** – The overall amplitude of the power spectrum describes how big the CMB fluctuations are in general. Increasing the amplitude would make all of the values on the y-axis bigger for example. It is set by the factor $A_s e^{-\tau}$, where A_s is the amplitude of the initial 'primordial' power spectrum of fluctuations left by inflation, and τ the optical depth to last scattering.
The primordial amplitude A_s gives us a starting point for the size of the fluctuations – the bigger the initial amplitude (just after inflation at $t \approx 0$), the bigger the amplitude at last scattering ($t \approx 380,000$ yr).
The optical depth determines how many CMB photons have been scattered (e.g. by intervening free electrons) in the ~ 13.8 Gyr between last scattering and the time we observe them, today. The more free electrons there are along the line of sight to the CMB, the larger the optical depth, and so the larger the probability of CMB photons being scattered. Scattering smooths out the CMB fluctuations, reducing the

observed amplitude of the power spectrum. It does this uniformly for all angular scales however, so this effect doesn't change the shape of the power spectrum.

- **Tilt** – Inflation left behind a distribution of shallow gravitational potential wells on all distance scales (i.e. of all widths). These were the seeds of fluctuations in the photon-baryon fluid, which eventually became the CMB temperature anisotropies. The initial distribution of potential wells is described by the *primordial power spectrum*, which we will learn about in the section on Inflation. This power spectrum tells us about the size distribution of the potential wells – how many large vs. small wells were left after inflation ended for example.
  In our Universe, the distribution is slightly 'tilted', in that there are slightly more wide potential wells than narrow ones. Since these potential wells cause the Sachs-Wolfe effect, and also affected how the photon-baryon fluctuations formed, the CMB power spectrum also gains a slight tilt in it (so the power spectrum is slightly larger at low ℓ than high ℓ, all other things being equal). It's hard to see by eye, but it is necessary to take the tilt into account when we make detailed predictions of the power spectrum.

Further reading: How the CMB shows us that dark matter exists (Ethan Siegel).

8.9. Dependence on cosmological parameters

The CMB power spectrum is one of our most precise observational probes. CMB measurements are now routinely used to constrain the cosmological parameters that describe our Universe with accuracies of 1% or even better.

The way this works is to make the most precise measurements of the power spectrum as possible using a CMB experiment. We then compare the measurements to theoretical calculations of the power spectrum for a range of different values of the main cosmological parameters. The figure below shows how the CMB power spectrum varies when several cosmological parameters are varied. By finding the theoretical prediction that best matches the observed power spectrum, we can figure out what the cosmological parameters corresponding to the real Universe are. The most accurate measurements of these parameters so far come from the Planck CMB satellite. For the parameters that we have come across previously in this course, Planck has found the following values:

$$H_0 = 67.27 \pm 0.60\,\mathrm{km/s/Mpc} \quad (127)$$

$$\Omega_m = 0.3166 \pm 0.0084 \quad (128)$$

$$\Omega_\Lambda = 0.6834 \pm 0.0084 \quad (129)$$

$$r_s = 144.39 \pm 0.30\,\mathrm{Mpc} \quad (130)$$

$$z_{\mathrm{eq}} = 3407 \pm 31. \quad (131)$$

These values were derived under the assumption that the Universe is spatially flat. If we allow the curvature to vary too, Planck finds a value of $\Omega_k = 0.001 \pm 0.002$. This is consistent with the Universe being flat.

Further reading: CMB map simulator (interactive tool)

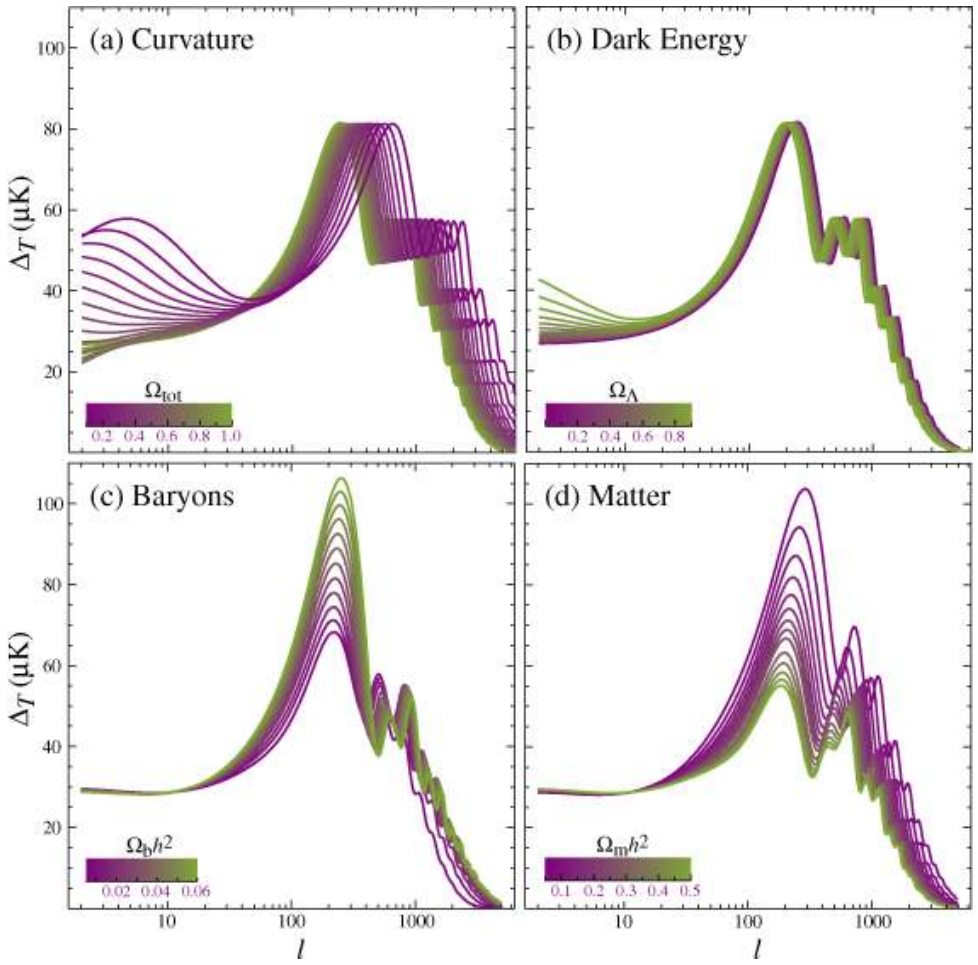

Figure 23: The CMB power spectrum for a range of different cosmological parameter values (only one parameter is changed at a time; the other parameters are fixed). This figure shows how sensitive the power spectrum is to the values of the cosmological parameters. Taken from CMB Parameter Sensitivity (Credit: W. Hu).

Learning outcomes:

What are CMB anisotropies?

What physical effects cause CMB anisotropies?

What are the baryon acoustic oscillations, how do they form, and what do they look like in the CMB power spectrum?

What is the approximate comoving size of the BAO feature?

What is diffusion damping?

What are the Sachs-Wolfe and integrated Sachs-Wolfe effects?

What are spherical harmonics, and why are they used to analyse the CMB?

What patterns do spherical harmonics make on the sky?

What is the relationship between spherical harmonic mode, ℓ, and angle on the sky, θ?

What is a power spectrum? Why do we use it to study the CMB anisotropies?

What does the CMB power spectrum look like, and where do its various features come from?

What is the acoustic peak and how can it be used to measure the distance to the CMB?

9. Inflation

In this section you will learn about inflation, a period very early in the Universe's history where space is thought to have expanded very rapidly. Through the mechanism of accelerating exponential expansion, inflation can explain a number of puzzling properties of the Universe, including why it is observed to be so close to flat and so close to homogeneous. We will study simple models for the inflationary mechanism, based on a dynamical scalar field called the *inflaton*, and will develop some useful mathematical results that allow us to predict how inflation affects observable properties of our Universe. We'll also see how inflation generates the 'seeds' of galaxies and other large-scale structure that we see today.

Reading for this topic:

- *An Introduction to Modern Cosmology (A. Liddle), Chapter 13.*

9.1. How special is our Universe?

The theory of inflation is a fundamental ingredient of the standard model of cosmology. While it describes a very brief and very remote period of cosmic history, tiny fractions of a second after the 'Big Bang', it is essential to make scientific sense of the state that we have found the Universe to be in for much of its subsequent history. In particular, inflation provides us with a physical mechanism to solve a handful of problems that would otherwise make our Universe seem extremely, almost impossibly, improbable.

We will learn about these problems in the following three sections, but all of them ask a similar kind of question – why does our Universe seem to be in such a special configuration, when so many other possible universes seem much more likely to have occurred? There are actually quite a few ways in which our Universe is special:

- It is very close to being spatially flat (and was even flatter in the past);
- It is quite old, even though it could have easily started out so dense that it would recollapse immediately;
- It has a cosmological constant that is just the right size to be observable today, without having blown the whole Universe apart a long time ago;
- It is filled with normal matter, and doesn't have an equal amount of anti-matter that would have annihilated it all;
- It is very close to being homogeneous and isotropic everywhere, even when comparing regions that have never been in causal contact;
- It isn't filled with weird high-energy particle relics that would frequently rip through planets, stars, galaxies etc. and destroy them;
- It is hospitable enough that at least one part of it has hosted complex living things for several billion years.

Why does our Universe have these properties? In other words, why does it appear to be so **finely-tuned** to be flat/smooth/old/hospitable to life? This is an absolutely **huge** question that straddles cosmology, fundamental physics, and even philosophy.

It turns out that inflation can give us compelling answers to at least a few of these questions without needing the Universe to have randomly begun in a very, very, very particular and special initial state. In other words, it provides a physical mechanism for some important properties of the Universe to generically end up in the 'special' configurations we see them in, without needing the initial conditions of the Universe themselves to be special.

9.2. The horizon problem

The CMB shows that the Universe is very smooth on very large distance scales. This is surprising because parts of our last-scattering surface that are widely enough separated should never have been in causal contact! Signals from one side of our last scattering surface haven't had time to reach the other side, even at the speed of light, so how could the two regions have possibly come so very close to being in thermal equilibrium with one another? What are the odds that two completely independent regions of the Universe would randomly happen to have almost the exact same temperature, density, and expansion rate? What are the odds that *every* region of the Universe that we can see on our last scattering surface had randomly ended up to be almost the same, to within one part in about 100,000? This sounds fishy – as if the initial conditions of our Universe have been fine-tuned so that every region has very similar properties from the start (subsequently evolving independently, but producing almost the same results since the initial conditions were so similar).

This strange occurrence is called the **horizon problem**, and can be put on a firmer mathematical footing by calculating which regions could and couldn't have been in causal contact since the Big Bang.

Since the CMB is our 'smoking gun' for this problem, let's start by calculating the Hubble radius at the time of last-scattering:

$$r_{\rm HR}(a_{\rm LS}) = \frac{c}{a_{\rm LS} H(a_{\rm LS})} \approx 234\,{\rm Mpc}, \tag{132}$$

where we have used the expansion rate at last-scattering inferred from CMB observations ($H_{\rm LS} \approx 1.4 \times 10^6$ km/s/Mpc) and $z_{\rm LS} \approx 1090$. This comoving distance is larger than the comoving sound horizon, $r_s \approx 150$ Mpc, as expected – the acoustic peaks in the CMB were formed due to physical processes before last-scattering, and so *should* be within the Hubble radius.

Using the angular diameter distance to the CMB, we can also calculate the angular scale corresponding to the Hubble radius at this redshift. The angular diameter distance[4] can be calculated by observing that the first acoustic peak arises at $\ell_* \approx 220$ (which actually corresponds to a comoving distance scale of $r_* \approx 0.75\, r_s$), so

$$\Delta\theta_s = \frac{r_*}{(1+z)\, d_A} = \frac{\pi}{\ell_*} \implies (1+z)\, d_A = 13.88\,{\rm Gpc}. \tag{133}$$

The corresponding angular scale for the Hubble radius at last-scattering is therefore

$$\ell_{\rm HR} \approx (1+z)\frac{\pi\, d_A}{r_{\rm HR}} \approx 186. \tag{134}$$

If regions of our last scattering surface separated by distances greater than $r_{\rm HR}$ were truly independent, we would not expect the CMB anisotropies to be correlated on angular scales larger than this ($\ell \lesssim 186$). The CMB power spectrum could look like anything here, with potentially very large, uncorrelated variations. The map of CMB anisotropies itself would not have any coherent structures on larger angular scales than this (except for secondary anisotropies). In other words, if we filtered out all of the structures on smaller angular scales, we would expect to be left with a random, uncorrelated distribution of potentially quite large anisotropies.

Our measurements of the CMB map and power spectrum show that this isn't the case however; the power spectrum still has features at $\ell \lesssim 186$ (see above), and the filtered CMB map does not look like random noise or have very large temperature differences on opposite sides of the sky.

Note: The calculation of $\ell_{\rm HR}$ in the video is slightly incorrect; see above for a corrected version.

9.3. The flatness problem

Another curious observation is that we find space to be so close to flat today ($|\Omega_k| \lesssim 5 \times 10^{-3}$ according to Planck). It could have started off with any sort of curvature, depending on how much mass/energy the Universe started with. Why should we find the total density of mass/energy to be so close to the critical density today? Again, it seems as if the Universe has been fine-tuned to have just enough mass/energy to keep it almost spatially flat.

[4] Since r_s and $r_{\rm HR}$ are in comoving units, they must be multiplied by $a = 1/(1+z)$ so we can use the usual angular diameter distance formula, which is in proper units.

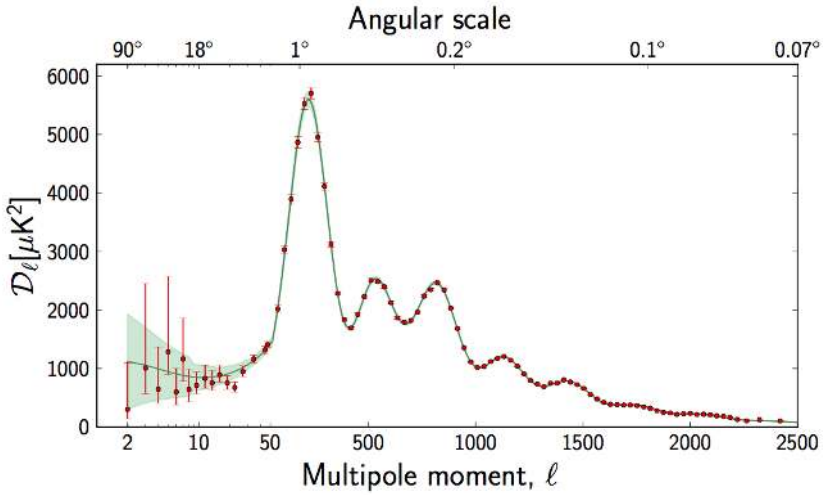

Figure 24: The CMB power spectrum as measured by the Planck satellite. (Credit: ESA/Planck)

This observation gets even more curious if we consider how close to flat the Universe must have been in the past. Recall that the curvature term in the Friedmann scales like $\Omega_k a^{-2}$, and so grows with time. For the curvature to be small today, it must have been even smaller in the past. In fact, it turns out that the initial spatial curvature of the Universe, fractions of a second after the Big Bang, must have been absolutely tiny in order to produce the observational bounds of $\lesssim 1\%$ on Ω_k that we see today.

To see this, let's rewrite the curvature term as the fractional difference between the total energy density of the Universe and the critical density at a given time. Recall that $\Omega_{\rm tot} = \rho_{\rm tot}(t_0)/\rho_{\rm cr,0} = 1 - \Omega_k$, all evaluated at t_0 (today). Using the Friedmann equation, we can write the total energy density as a fraction of the critical density *at any time* t. First, divide both sides of the Friedmann equation by H^2:

$$\frac{H^2}{H^2} = \frac{8\pi G\rho_{\rm tot}(t)}{3H^2} - \frac{kc^2}{a^2H^2}. \tag{135}$$

The left-hand side is just 1, while the first term on the right-hand side can be simplified by noticing that the critical density as a function of time is

$$\rho_{\rm cr}(t) = \frac{3H^2}{8\pi G}. \tag{136}$$

If we borrow the fractional density ('Ω') notation to define $\Omega_{\rm tot}(t) = \rho_{\rm tot}/\rho_{\rm cr}(t)$, we can write the Friedmann equation as

$$1 = \Omega_{\rm tot}(t) - \frac{kc^2}{a^2H^2} = \Omega_{\rm tot}(t) + \frac{H_0^2\Omega_k}{(aH)^2} \quad \Longrightarrow \quad 1 - \Omega_{\rm tot}(t) = \frac{H_0^2\Omega_k}{(aH)^2}. \tag{137}$$

(Note that we have folded the dark energy density into the total energy density as usual.)

Consider a Universe with a small deviation from flatness today, $\Omega_k = 10^{-2}$. We can use the expression above to calculate the fractional deviation from flatness at any time in the past. For example, at decoupling, $a_{\rm dec} = 1/(1+1090) = 9.17 \times 10^{-4}$, and the Universe was matter-dominated, so $H^2 \approx H_0^2\Omega_m a^{-3}$. Plugging this in, we obtain

$$1 - \Omega_{\rm tot}(a_{\rm dec}) = \frac{\Omega_k H_0^2}{a_{\rm dec}^2 H_0^2 \Omega_m a_{\rm dec}^{-3}} = \frac{\Omega_k}{\Omega_m} a_{\rm dec} \approx 3 \times 10^{-5}, \tag{138}$$

where we took $\Omega_m \approx 0.3$. So, a 1% deviation from flatness today requires that the Universe was only 0.003% away from being perfectly flat at decoupling! Let's look even further back, at the time when the Universe had cooled enough for the first bound hadrons to form ($T_{\rm had} \approx 10^{10}$ K, corresponding to energies of around 1 MeV). This happened at a redshift[5] of $z \approx 3.7 \times 10^9$, when the Universe was strongly radiation-dominated, so

[5] You can work out the redshift corresponding to this temperature from $T = T_0(1+z)$, where $T_0 \approx 2.725$ K for our Universe.

$H^2 \approx H_0^2 \Omega_r a^{-4}$. Plugging in $a_{\rm had} = 1/(1+z) \approx 3 \times 10^{-10}$ and $\Omega_r \approx 10^{-5}$, we obtain

$$1 - \Omega_{\rm tot}(a_{\rm had}) = \frac{\Omega_k}{\Omega_r} a_{\rm had}^2 \approx 10^{-16}. \tag{139}$$

This is a *tiny* number, and we haven't even gone back as far in time as physics reliably allows us to – the further back we look, the worse the problem gets! For the Universe to be only 1% flat today, it must have been almost *perfectly* flat at very early times. What could have caused the early Universe to be so almost perfectly flat so that it still looks close to flat today? Even if the number above had been 10^{-15} instead of 10^{-16}, the value of Ω_k would be ten times larger and space would be very noticeably non-Euclidean. It seems that the (almost-)flatness of space also requires a very large fine-tuning of the initial state of the Universe.

9.4. The (magnetic) monopole problem

In the early 1980s, a lot of theoretical work revolved around trying to find a theory that unified all of the fundamental forces of nature. At sufficiently high energies, the EM and weak nuclear forces merge into one 'electroweak' force. This was a very pleasing feature to physicists at the time, as it suggested that there is a deeper symmetry to the Universe than the various symmetries suggested by the four fundamental forces that we experience at lower energies. A lot of serious effort went into trying to find a *Grand Unified Theory* (GUT), which would also unify the electroweak force with the strong nuclear force.

A common prediction of theories that were put forward as candidates for the GUT is that *magnetic monopoles* should be generated in large quantities in the early Universe. Normal magnets are dipoles, with a north and south pole, and no net magnetic charge. Monopoles are only either a north or a south pole however, and so can have a net positive or negative magnetic charge. Magnetic monopoles have never been observed in nature.

Why would magnetic monopoles form? They are a class of object called a **topological defect**. As the Universe cooled from its initially high-energy state, where the three forces were unified, it would eventually undergo *spontaneous symmetry breaking* (SSB), where the forces became separate again. Spontaneous symmetry breaking results in a phase transition in the quantum fields that pervade the Universe, and topological defects are associated with these phase transitions. You can think of them as boundaries between regions of the Universe that froze into different vacuum states after the phase transition (see the discussion of vacuum energy in Section 5). A common analogy is to think about the boundary between two parts of a crystal that have begun forming separately and eventually grow together; the crystals on each side of the boundary are unlikely to be aligned, and so a distinct boundary between the two different crystal phases gets locked in between them and cannot easily be changed. In fact, the mathematical description of phase transitions and topological defects in cosmology has a lot of similarities to the way similar phenomena are described in solid state physics! This field – applying solid-state theory to cosmology – was pioneered by Tom Kibble at Imperial College London, amongst others.

Topological defects have very strange properties, and are completely unlike any normal form of matter or energy that we are used to. They can come in a small number of possible shapes: monopoles are point-like (0-dimensional) objects, *cosmic strings* are like lines (1-D objects), and *domain walls* are like surfaces (2-D objects). Each type of object describes a boundary between different types of regions, and each type of object has a different equation of state; $w = -1/3$ for cosmic strings and $w = -2/3$ for domain walls, for example.

They would have a number of interesting observational consequences if we could observe them, including causing very strong lensing artefacts in the CMB. These have been searched for extensively, and *we don't see them*. According to the GUT theories, we would expect at least a handful of topological defects within every Hubble radius. Our observations are consistent with there being *none*. This is the monopole problem – despite topological defects being a clear, and indeed almost inevitable, consequence of higher-energy theories like GUTs, why do we not see any in nature? They should be very noticeable!

Further reading: Magnetic monopoles in Grand Unified Theories (Wikipedia); Cosmic strings (Wikipedia).

9.5. The inflationary mechanism

Cue the invention of *cosmic inflation*, a theory of the early universe that solves these problems in a compelling way. So compelling, in fact, that since it was first proposed by Alan Guth and others in the early 1980s,

many thousands of scientific papers have been written about it. It even has the philosophers arguing about its implications for the nature of reality.

What is inflation, and how does it solve these problems? The idea is reasonably simple: instead of the Universe being filled with radiation at very early times (around 10^{-36} sec after the Big Bang), what if there was a new force of nature that took over instead? This is similar to what you might expect from a Grand Unified Theory, where the EM, weak, and strong nuclear forces are all combined into a single force. There are various options for how such a force might behave, but a simple and theoretically-compelling option is to model it as a **scalar field**.

Recall that the electric and magnetic fields are vector fields, which have a direction and magnitude at every point in space. A scalar field is more like the electric potential, in the sense that it only has a magnitude at every point in space. This is about as simple as it gets for a force of nature. Fundamental force fields are associated with force-carrying particles, and the hypothetical early universe scalar field is no exception. Just as the Higgs boson (force-carrying particle) is associated with the Higgs field (a scalar field), the *inflaton* is the particle associated with the early universe scalar field.

The inflaton has two important properties that allow it to solve the horizon, flatness, and monopole problems. First, it has an equation of state of $w \approx -1$, which means that it causes accelerated cosmic expansion, much like the cosmological constant. Recall from Section 5 that accelerating expansion causes the Hubble radius to shrink instead of grow. This will turn out to be the most important property of the inflaton. This kind of expansion also causes the scale factor to increase almost exponentially, rapidly 'inflating' the size of the Universe (hence the name).

The second property of inflation is that, unlike the cosmological constant, the inflaton stops causing acceleration after a while. If it simply had $w = -1$ forever, it would dominate the energy density of the Universe forever, and no radiation- or matter-dominated periods would ever arise. The inflaton eventually decays in a process called *reheating* however, converting its energy into matter and radiation.

How does a shrinking Hubble radius and exponential expansion help to solve the various problems?

- **Horizon problem**: In the inflationary paradigm, everything that we see today (including every region of our last scattering surface) was once contained in a very small patch of the Universe that was in thermal equilibrium before inflation began. Inflation then caused a brief but intense period of exponential expansion that rapidly increased the size of this patch by many orders of magnitude (typically a factor of e^{50} or more). This expansion was 'faster than light', in the sense that the space between objects expanded faster than light could travel between them, effectively removing them from being in causal contact with one another. (There is no speed limit on how fast space can expand, so this doesn't violate special relativity.)

  This solves the horizon problem. The reason that different regions of our last-scattering surface look so similar despite being outside one another's Hubble radius is that they were actually previously in causal contact. Inflation then rapidly expanded them apart from one another, placing them outside the Hubble radius (i.e. the Hubble radius effectively shrank during inflation).

- **Flatness problem**: During exponential expansion, $H \approx \text{const.}$, so $|1 - \Omega_{\text{tot}}(a)| \propto a^{-2}$. During inflation, the fractional difference between the total matter density and the critical density therefore goes down, causing the Universe to get flatter. Since the scale factor increases very rapidly during exponential expansion, a very large flattening effect can build up in just a short time. For a factor of e^{50} increase in the scale factor during inflation, $|1 - \Omega_{\text{tot}}|$ can be reduced by a factor of $e^{100} \approx 10^{43}$. Inflation therefore almost inevitably sets the Universe to be very close to flat, as we observe today.

- **Monopole problem**: The solution to the monopole problem is more or less the same as the solution of the horizon problem – while many monopoles and other topological defects may have been produced before inflation, they are expanded outside our Hubble radius during inflation, and so we can't see them any more. Inflation predicts that at most a handful would be observable within our Hubble radius today, and likely fewer than that. Our observations are consistent with there being none within our Hubble radius.

Note that after inflation, when normal radiation- and then matter-dominated expansion happened, the Hubble radius started to grow again. Slowly but surely, regions of the Universe that had been pushed outside our Hubble

radius by inflation re-entered it, and so came back into causal contact with us. The structures and correlations that we see on very large angular scales in the CMB today are relics of cosmic history that have lain untouched since the very first moments after the Big Bang, when the Universe was less than 10^{-30} seconds old.

Further reading: Inflationary universe: A possible solution to the horizon and flatness problems (A. Guth)

What is an e-fold?
Inflation causes the scale factor to increase by a very large amount in a very short time. To keep track of this when discussing possible theoretical models of inflation, cosmologists often measure the expansion factor in *e-folds*.

An e-fold is just an increase in the scale factor of one power of e. Ten e-folds would therefore be a factor of $e^{10} \approx 2.2 \times 10^4$. Inflation must typically last for at least 50 e-folds to solve the horizon and flatness problems. It could have gone on for much longer than this however!

If the number of e-folds was very large, our initial patch of the Universe could have been blown up to a size many, many times larger than the Hubble radius today. This is one reason why we think that the actual Universe is much larger than the *observable* Universe.

9.6. Cosmological Klein-Gordon equation

For most models of inflation, the evolution of the inflaton field can be described using the cosmological Klein-Gordon (KG) equation,

$$\ddot{\phi} + 3H\dot{\phi} + \frac{dV}{d\phi} = 0. \tag{140}$$

This is the equation of motion for a homogeneous scalar field, ϕ, in an expanding universe. Recall that dots are our shorthand notation for derivatives with respect to cosmic time, t. The first term therefore describes the 'acceleration' of the scalar field, i.e. the rate at which its 'speed' $\dot{\phi}$ is changing. The second term is the damping term, and depends on the expansion rate, H. The third term depends on the gradient of the potential that the scalar field inhabits, $V(\phi)$. Different models of inflation have different potentials, that will give rise to different behaviours of $\phi(t)$ when the scalar field is allowed to evolve according to the KG equation.

Note that the KG and Friedmann equations are coupled to one another, as the scalar field appears as a source of energy density in the Friedmann equation. The scalar field alters the solution for $a(t)$, which also alters the solution for $\phi(t)$ (through the factor of H in the second term).

The KG equation is a second-order ODE, and can be solved once initial conditions for $\phi(t_i) = \phi_i$ and $\dot{\phi}(t_i) = \dot{\phi}_i$ are given, and a form for the potential $V(\phi)$ has been chosen.

In the early Universe, H was very large, and so the second term tends to take over initially, allowing us to approximate the KG equation as $3H\dot{\phi} \approx 0$. The effect of this term is to drive the 'speed' $\dot{\phi}$ to zero, hence why it is called the damping term. The initial speed $\dot{\phi}_i$ therefore doesn't matter so much. This kind of behaviour is called an *attractor* solution, as it doesn't matter where you start – after a while, the equations always draw you into the same place (i.e. where $\dot{\phi} \approx 0$).

Once $\dot{\phi} \approx 0$, the other terms can become important however. We will study them in more detail when we discuss *slow-roll inflation*. For now, the important thing to realise is that the shape of the potential, $V(\phi)$, is very important in determining what happens to the scalar field next. The typical analogy is to think of the scalar field as a ball rolling up and down hills (the potential). The value of the scalar field, ϕ, is analogous to the position of the ball in the potential.

Steeper potentials will allow the inflaton to lose or gain more energy as it rolls up or down them respectively. Shallower potentials will result in a more gentle evolution however. The question is what kind of behaviour is needed for the inflaton to behave in a way that can solve the various problems discussed above.

9.7. Scalar field dynamics

Continuing with the analogy of the inflaton as a ball rolling around in a landscape described by the potential, $V(\phi)$, we can define the kinetic energy of the inflaton to be $\dot{\phi}^2/2$. This should look familiar from Newtonian

dynamics – it's simply the speed of the field squared, divided by two (we don't need to worry about a 'mass' of the field, as we can always redefine some units to make the mass equal to 1). The potential energy of the field is just given by the value of the potential, $V(\phi)$.

We can use these definitions of kinetic and potential energy to understand the expressions for the energy density and relativistic pressure of the scalar field,

$$\rho_\phi = \frac{\dot{\phi}^2}{2} + V(\phi) \tag{141}$$

$$p_\phi = \frac{\dot{\phi}^2}{2} - V(\phi), \tag{142}$$

where we have used units where $c = 1$. The energy density is just the sum of the two different types of energy, while the pressure is their difference (recall that relativistic pressure is different from thermal pressure). We can then write the equation of state for the scalar field as

$$w_\phi = \frac{p_\phi}{\rho_\phi} = \frac{\frac{\dot{\phi}^2}{2} - V(\phi)}{\frac{\dot{\phi}^2}{2} + V(\phi)}. \tag{143}$$

Let's consider two limits of the equation of state – one where the scalar field is dominated by its potential energy ($|V| \gg \dot{\phi}^2$), and one where it is dominated by its kinetic energy ($|V| \ll \dot{\phi}^2$). In the first case, we see that $w_\phi \approx -V/V = -1$, which you may remember as the equation of state of a cosmological constant. This means that a scalar field that is potential-dominated will behave in the same way as a cosmological constant and therefore cause accelerating expansion! This is exactly what we need to solve our horizon, flatness, and monopole problems – by having the Universe expand exponentially, we can shrink the Hubble radius, therefore removing the pesky monopoles and any inhomogeneous regions from view. Accelerating expansion will also cause gravitational potentials to decay (recall the ISW effect from the last section), smoothing out inhomogeneities in the Universe in the process.

The second case leads to $w_\phi \approx \dot{\phi}^2/\dot{\phi}^2 = +1$, which is quite an unusual equation of state. In this case, the scalar field does not cause accelerated expansion, but it doesn't behave like matter ($w = 0$) or radiation ($w = 1/3$) either. In fact, it will tend to make the Universe collapse in on itself, which is obviously a situation that should be avoided if we're trying to model the early Universe.

9.8. Slow-roll approximation

As shown above, we want to avoid the inflaton getting too much kinetic energy, and would quite like to end up with an accelerating expansion with $w_\phi \approx -1$ instead. We are therefore interested in building inflationary models that can produce this behaviour, especially if it can be achieved without too much fine-tuning. How can we do this?

The key is to construct a potential that keeps the inflaton from gaining too much kinetic energy, while still allowing it to evolve. The former condition is trivially satisfied if we trap the inflaton at the bottom of a steep potential that keeps the field stuck with exactly $\dot{\phi} = 0$. In this case there is no way for inflation to end, however – the field would stay stuck in the same configuration forever, without being able to move, and so the Universe would continue to expand at an accelerating rate forever. This does not fit in with what we observe in our Universe. A better option is to put the scalar field in a shallow potential and then allow it to slowly roll down the slope.

This leads to an important approximation called the *slow-roll approximation*. When the field is slowly-rolling, we can neglect a handful of terms in the Friedmann and KG equations that makes them easier to solve.

Under the slow-roll approximation, we require that the kinetic energy of the scalar field is significantly smaller than its potential energy, $\dot{\phi}^2/2 \ll |V(\phi)|$. If we apply this approximation to the energy density of the field, we obtain $\rho_\phi \approx V(\phi)$. Plugging this into the Friedmann equation (for a flat Universe containing only a scalar field), we obtain

$$H^2 = \frac{8\pi G}{3}\rho_\phi \approx \frac{8\pi G}{3}V. \tag{144}$$

In natural units, the prefactor can be rewritten in terms of the Planck mass, $8\pi G = M_{\rm pl}^{-2}$, if needed. Continuing with the slow-roll approximation, the KG equation becomes

$$3H\dot{\phi} + \frac{dV}{d\phi} \approx 0. \tag{145}$$

Note that we have neglected $\ddot{\phi}$; the slow-roll approximation says that $\dot{\phi}$ is small compared to V, which means that the time derivative of $\dot{\phi}$ must be small also. The term involving $3H\dot{\phi}$ is *not* neglected however, as while $\dot{\phi}$ may be small, the prefactor $3H$ is actually very large (the expansion rate H is very large in the early universe, and is actually proportional to $\sqrt{V}$ as we can see from the Friedmann equation).

With a bit of rearranging, we can now turn the Friedmann and KG equations into a single integral equation. First, we can use the Friedmann equation to write $da/dt = aH$. Rewriting $d\phi/dt = (d\phi/da)(da/dt) = aHd\phi/da$, the KG equation becomes

$$3aH^2\frac{d\phi}{da} = -\frac{dV}{d\phi}. \tag{146}$$

Substituting the Friedmann equation for H^2, and moving all of the factors of a to the right-hand side, we obtain

$$-\frac{8\pi G\,V}{(dV/d\phi)}d\phi = \frac{da}{a}. \tag{147}$$

If we know what $V(\phi)$ is, we can evaluate the left-hand side, perform an integral of both sides, and thus find an expression for $a(\phi)$. This can be inverted to find $\phi(a)$.

Note that we could also solve the Friedmann equation under this approximation. But, since we know that $w_\phi \approx -1$ in the slow-roll approximation, we already know that the solution we obtain should be very close to $a(t) \propto e^{Ht}$, where $H \approx \text{const}$.

9.9. Quantum fluctuations and the primordial power spectrum

We have seen that inflation flattens out and homogenises the Universe, leaving it very close to a perfect FLRW universe (i.e. one that is *perfectly* homogeneous and isotropic). If this was the end of the story, our Universe would be even smoother than we see it today, with far fewer structures and even smaller anisotropies in the CMB.

It turns out that inflation also *generates* small fluctuations in the cosmic energy density though. These are seen as small fluctuations in the gravitational potential that give rise to the Sachs-Wolfe effect in the CMB. They also act as the seeds of the fluctuations in the baryon/photon/dark matter density that become acoustic oscillations and other inhomogeneities. Ultimately, these fluctuations grow via gravitational collapse to become galaxies and galaxy clusters. These original seed fluctuations are called *primordial energy density perturbations*.

The small fluctuations from inflation are generated by **quantum fluctuations** in the inflaton field itself. At the very high energies present in the early Universe, it is no surprise that some quantum phenomena might arise, and so we often describe the inflaton as a 'semi-classical field', that has some non-quantum and some quantum properties. The mildly quantum nature of the inflaton means that there is some uncertainty in when inflation ends, for example – the end of inflation corresponds to when the temperature of the Universe dips below a certain value, but the Heisenberg Uncertainty Principle tells us that we can't know both the energy and time of an event perfectly precisely. The slight variations in when inflation ended therefore led to some parts of the Universe inflating slightly more than others, leading to small fluctuations in the energy density.

These quantum fluctuations are generated for the entire duration of inflation. As they are generated, they too are expanded outside the Hubble radius, where they cannot evolve in time (gravitational collapse is a causal process, so cannot affect fluctuations larger than the horizon). The fluctuations are therefore frozen in to the energy density distribution on very large scales, and do not change or evolve. After inflation ends, when the Hubble radius starts to grow again, the fluctuations **re-enter the horizon**, and can start growing due to gravitational collapse etc. This means that the fluctuations we see on the very large angular scales in the CMB are the ones that have most recently re-entered the horizon. They are frozen relics from the very first moments of cosmic history!

Inflation makes quite specific predictions for the size of the primordial perturbations. Since they are being generated throughout inflation, they respond to any changes in the physical behaviour of the inflaton field. Most inflationary models predict that the inflaton slowly rolls at a more or less constant rate however, with $H \approx \mathrm{const.}$, so all of the fluctuations that are generated have more or less the same general statistical properties. The time at which the fluctuation is generated determines the eventual distance scale on which it will be observed; fluctuations generated early on during inflation will be expanded the most, and so will appear on very large scales, while fluctuations generated towards the end of inflation will be expanded less and so will be seen on smaller scales. The typical size of the fluctuations is almost constant however – fluctuations on large scales will have almost the same variance as fluctuations on small scales.

Putting this all together, we find that most inflationary models predict an **almost scale-invariant power spectrum** for the primordial perturbations. Recall that a power spectrum is a measure of the variance of a fluctuation field as a function of distance scale. Scale-invariant means that the fluctuations have the same variance on all distance scales, so their typical size on small scales is the same as on large scales. The inflaton is not perfectly slowly rolling (i.e. $\dot{\phi} \neq 0$), so the inflaton does change its kinetic energy slightly with time, and so the fluctuations generated around the start of inflation are slightly different to the ones generated at the end. The power spectrum is therefore slightly tilted, with slightly lower variance on small scales than on large scales. We normally write the primordial power spectrum as $P(k)$, the variance as a function of Fourier mode k. Most inflationary models predict

$$P(k) \propto k^{n_s - 1}, \tag{148}$$

where n_s is called the *tilt* or *spectral index* of the primordial power spectrum. We can measure n_s from the CMB power spectrum, and find a value of around $n_s \approx 0.96$ – very close to scale invariant, but not exactly. This is in important prediction of inflation.

Another prediction of inflation is that the fluctuations should follow an almost Gaussian statistical distribution. We can measure the statistical distribution of the CMB temperature fluctuations and use this to figure out if the primordial fluctuations were Gaussian or not. So far, all of our observations suggest that they are – another small hint that inflation really happened in the early Universe.

9.10. Reheating

An important feature of inflation is that after a short time (typically 10^{-32} seconds or so) it ends. This can happen when the inflaton gains enough kinetic energy, for example, which makes the equation of state less negative ($w_\phi > -1$) so the expansion of the Universe stops accelerating. When building theoretical models of inflation, and common way to get it to end is to add a sudden dip in the potential. After slowly rolling fro a while, the inflaton falls into the dip and rolls up and down inside it, sort of like a harmonic oscillator trapped in a potential well. These oscillations cause the inflaton to decay, converting the energy density locked in the inflaton field into high-energy particles and radiation. The process of converting the inflaton into normal radiation/matter is called **reheating**. After a short while, the inflaton completely decays away and essentially vanishes, never affecting the Universe again – it has done its job of setting up the initial conditions of the Universe in exactly the right way.

What *really* happened at the Big Bang?

What we call the Big Bang – the point at time $t = 0$, when the Universe was infinitely dense, and all of space was crushed into a singularity – is just an extrapolation of our solutions to the Friedmann equation as far back in time as we can go. Most cosmologists don't think there was actually a real, physical singularity like this! Instead, we think that there was possibly a phase of the Universe's history immediately before inflation started.

We don't know how long this pre-inflationary phase lasted, or even if it makes sense to talk about time passing during this phase! The Universe would have been so hot and dense at this time that energies, lengths, and times would be around the Planck scale, where our theories of gravity and quantum mechanics break down. Perhaps theories of quantum gravity, like string theory, will give us sensible solutions to what could have happened during this time. It seems unlikely that we can ever access this epoch observationally however, since inflation does such a good job of smoothing out any structure that existed before. Perhaps part of the Universe's history is forever hidden from our view.

Talk about a pre-inflationary phase does lead us to the idea of a multiverse however. Perhaps there is a much larger reality out there that contains many Universes like our own, some of which have not started expanding yet, and some of which have been expanding for a long time or have even started to contract and have crunched back in on themselves. Perhaps this multiverse contains more than the 3+1 dimensions of space and time that we observe in our Universe. Perhaps, sometimes, universes within this multiverse collide with one another and leave faint marks or scars on each other that can be observed? It's impossible to say much about any of these scenarios yet, as we don't have any observational evidence for any of them, or even a fully consistent theory that we can do calculations with.

Regardless, we have abundant evidence that something very much like a Big Bang happened – the Universe began to rapidly grow from an extremely hot, dense state, and then continued expanding. It just probably didn't come from a mathematical singularity.

Learning outcomes:

- What are the horizon, flatness, and monopole problems?
- What is the inflationary mechanism?
- How does inflation solve these problems?
- What is the inflaton and what are its properties?
- What is an e-fold?
- What does the cosmological Klein-Gordon equation describe?
- How are the density and pressure of the inflaton related to its kinetic energy and potential?
- What is the slow-roll approximation?
- How does inflation generate primordial fluctuations?

10. Dark matter

In this section we will learn about the observational evidence for dark matter, and the many different explanations that have been proposed to explain its existence. We will also study some of the observed properties of dark matter, particularly how it clusters together and forms gravitationally bound *dark matter halos*. These properties imply that it is *cold*, and also explain how the large-scale structure of the Universe formed *hierarchically*.

Reading for this topic:

- *An Introduction to Modern Cosmology (A. Liddle), Chapter 9 and Advanced Topics 3 and 5.*

10.1. Observational evidence for dark matter

Virial velocities of galaxies – The first hints of the existence of dark matter were found by Fritz Zwicky and others in the 1930s. Zwicky measured the peculiar velocities of individual galaxies in the nearby *Coma galaxy cluster*, adding them all up to find the total kinetic energy within the cluster. He also estimated the total mass of all the galaxies in the cluster, and used that to estimate its gravitational potential energy. The **Virial Theorem** tells us that if the cluster was in equilibrium (i.e. neither collapsing in on itself or flying apart), the kinetic and potential energy should be in balance. What he found was that the galaxies had *much* more kinetic energy (i.e. were travelling faster) than expected however, to the point that many of them would be travelling faster than the escape velocity of the cluster and should by flying away from it. Yet, they were not! Zwicky used this observation to infer the existence of what he called 'dunkel materie' (dark matter). If there was a large amount of unseen dark matter within the cluster, in addition to the matter that could be seen, the potential energy would be larger than observed and the Virial Theorem would be satisfied, explaining why the cluster hadn't flown apart.

Galaxy rotation curves – In the early 1970s, Vera Rubin and Kent Ford used a new, sensitive spectrograph to measure the rotational velocity of spiral galaxies as a function of the distance from the centre of the galaxy. This is known as a **rotation curve**. They found that the rotation grew rapidly with radius close to the centre (in a region called the *galactic bulge*), then flattened out to a roughly constant velocity at larger distances (see figure below). This was completely unexpected! By measuring the brightness of the galaxy as a function of distance from the centre, it seemed that most of the stars (and so, presumably, most of the mass) was concentrated in the centre, with the outskirts of the galaxy containing progressively less material. In this case, we would expect the galaxy to follow a **Keplerian rotation curve**, similar to the way the planets orbit the Sun. The rotation would increase rapidly inside the bulge, but should then decrease outside the bulge, as if everything in the outer reaches of the galaxy was orbiting a single massive object in the centre. The fact that a Keplerian rotation curve is not observed suggests that the distribution of matter is quite different from expected. In fact, the observed rotation curves could be explained if the galaxy was embedded in a much larger *halo* of matter with a significant amount of mass distributed beyond the outer reaches of the galaxy. This matter is not seen, even though there must be a very large amount of it.

Acoustic peaks in the CMB – The most precise and convincing evidence for the existence of dark matter now comes from the CMB temperature anisotropies. The size of the baryon acoustic oscillation features in the CMB power spectrum depends on the abundance of baryons. We know that the total fraction of matter in the Universe today is $\Omega_m \sim 0.3$. If this consisted entirely of normal baryonic matter (hydrogen, electrons etc.), the baryon acoustic oscillations would be much stronger than we observe; there would be a lot more baryons to interact with photons in the photon-baryon fluid, and so the acoustic oscillations would be more intense. If a large fraction of the matter does *not* couple to the photons in the period before decoupling, however, it will act as a sort of gravitational counterweight, damping the oscillations by making it harder for them to propagate (due to its gravitational attraction). This would make the oscillations smaller. Our observations of the CMB power spectrum are consistent with there being around 4-5 times more non-interacting matter than baryonic matter; there *must* be a large amount of matter out there that does not interact appreciably through the EM force, as otherwise it would have contributed to making the BAOs much larger.

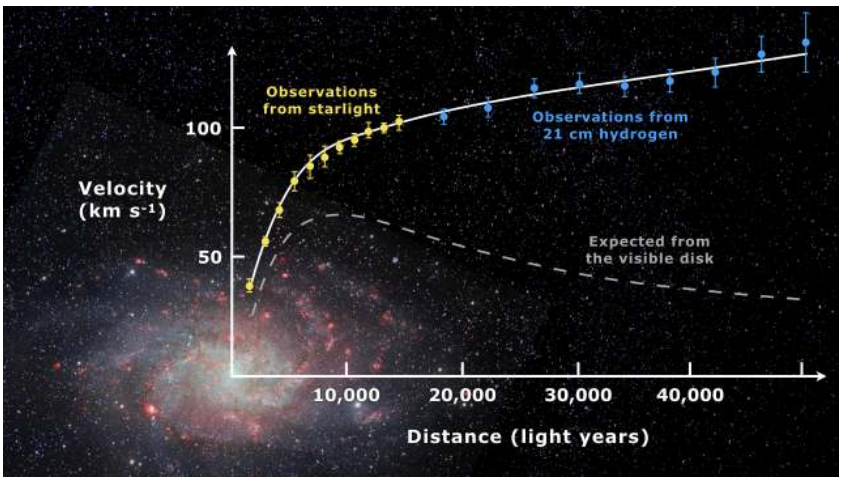

Figure 25: Rotation curve (rotation velocity vs radius) of the galaxy M33. (Credit: M. De Leo / CC-BY-SA 4.0)

10.2. Properties of dark matter

As dark matter became established as a real phenomenon, with a significant weight of observational evidence behind it, cosmologists started seeking explanations for what it might be made of. While its existence was clear, there was considerably less information about what its properties might be. All we really knew was that:

- There is more dark matter than normal (baryonic/luminous) matter
- It is almost completely **transparent or invisible** (doesn't interact strongly via the electromagnetic force)
- It can **cluster together**, i.e. it is distributed inhomogeneously throughout space (unlike a cosmological constant, which has the same energy density everywhere).

So, gravitationally, dark matter seems to behave in a similar way to normal matter, while electromagnetically it is different.

At first, people wondered whether the dark matter could be made up of neutrinos. These would certainly fit the bill – they interact only very weakly with normal matter[6], do not emit, absorb, or scatter light, and are known to be produced in abundance by nuclear reactions (e.g. like those in the early Universe, and inside stars). The problem is, neutrinos have a very low mass.[7] Neutrinos produced shortly after the Big Bang would be highly energetic, and would therefore be moving at highly relativistic speeds – in other words, they would behave like radiation rather than as matter. This would make it hard for them to cluster together, as like photons around the time of decoupling, they would 'free stream' away, smoothing out fluctuations in the mass distribution as they went. Still, if neutrinos had a small but non-zero mass, they could still do a small amount of clustering, although this would happen on very large scales – only inhomogeneities that were a similar size to the neutrino mean-free path or larger would remain intact. This model of dark matter, where it is made up mostly of neutrinos, is called the **hot dark matter (HDM)** model, as the dark matter in this case would have a high energy and travel semi-relativistically.

The HDM model quite quickly runs into difficulties – it's hard to produce enough neutrinos to make it work, and it's hard to get large structures (galaxy clusters etc.) to form in the way that we observe. Instead, people started to consider a couple of more exotic alternatives to explain dark matter: perhaps it is a new form of massive particle, or highly compact, macroscopic blobs of normal matter?

10.3. Particle dark matter

The first option is called **particle dark matter**, and looks to particle physics for an explanation. Perhaps there is a new type of force or field that we haven't yet seen in particle colliders or other experiments? This was seen

[6]There are hundreds of billions of neutrinos passing through each square centimetre of your body each second, with precious few ever interacting with any part of you.

[7]The fact that neutrinos are not massless was discovered around the same time as the rotation curve evidence for dark matter.

as quite a natural explanation, as it came at a time when particle theorists were finding new models that extend the Standard Model of Particle Physics to explain various other puzzles. Many new types of particles were expected to come out of these theories, especially *supersymmetric partners* to known Standard Model partners. If new particles exist that have a mass of around 100 GeV are are able to interact only via the weak nuclear force, they would have all of the properties required to explain dark matter, and would be produced in almost exactly the right abundance in the high temperatures of the early Universe. This type of particle is called a *WIMP (weakly-interacting massive particle)*, and the fact that it would be produced in the right abundance is called the *WIMP miracle*.

This was seen as an extremely compelling explanation, and many different groups have spent the last 4 decades building sensitive experiments to **directly detect** these particles. The detectors often rely on large quantities of some special gas or crystal (e.g. liquid Xenon, Germanium crystals), and are placed underground to shield them from common forms of particle such as cosmic rays. While the hypothesised WIMP particles only rarely interact with normal matter, the occasional ones that *do* interact with the detector cause the nuclei or electrons in the detector material to recoil, in a way that can be seen and amplified (e.g. by detecting photons that are produced). Different detectors are sensitive to different types of particles with different properties and masses, while larger detectors provide more material for the WIMPs to interact with, thus making it more likely to see one of the rare interactions. So far (with the exception of some claims that are widely believed to be flawed), no WIMPs have been seen! This has now reached the point that the WIMPs, if they do exist, would have to have a very different mass than expected, or be much more weakly interacting than hypothesised. This spoils the 'WIMP miracle', and so these models now look much less compelling as an explanation of dark matter.

Figure 26: A major component of the XENON-1T direct dark matter detection experiment. A more detailed description is given here. (Credit: E. Sacchetti)

There are many other types of particle dark matter however, each with their pros and cons. A popular kind are called **axions**. These come out quite naturally from some high-energy particle theories, and have different properties from WIMPs. They tend to be much lighter, form in a different way, and exist in a 'fuzzy' quantum state called a Bose-Einstein condensate, which prevents structures from forming on small scales. These properties produce observational consequences that can be probed by astronomers. They also interact with the EM force occasionally; if axions pass through a very strong magnetic field, they can cause disturbances in the field that can be detected by a sufficiently sensitive detector. No evidence of axions has been found yet, but several different experimental searches are ongoing.

Further reading: Ghostlike neutrinos (NASA); Top Dark Matter Candidate Loses Ground to Tiniest Com-

petitor (Quanta Magazine)

10.4. Baryonic dark matter and compact objects

Instead of a new kind of particle, what if dark matter was made out of regular matter that existed in a state that prevents it from interacting with EM radiation? A good way to achieve something like this is to combine as much of the matter as possible into a relatively small number of compact but massive objects, such as asteroids, planets, or brown dwarf stars. While these objects most certainly do interact with EM radiation, this happens at a relatively low level considering the amount of mass that they contain. Most of the material inside a planet is shielded from interacting with outside photons for example, while typical planets and asteroids are quite cold, and so only give off a small amount of infrared radiation. If the relative sizes and abundances of the compact objects were in the right range, it would be difficult to detect them, even though a large amount of mass could be stored away in this form.

Recall that the rotation curve observations require that the dark matter is spread across large halos that galaxies are embedded in. The compact objects would also have to be distributed in this way if they are to explain dark matter. Hence the generic name for them – **massive compact halo objects (MACHOs)**. What would cause them to have such a different distribution to the rest of the matter in the galaxy, which is mostly found in the bulge and disk? They would also have to form very early in the Universe's history to explain the CMB observations – but how would they have had time to collapse (and where would the large amounts of elements heavier than hydrogen have come from)? As you can see, MACHOs raise a lot of questions, even if they can solve the dark matter problem.

In fact, many types of MACHO are ruled-out observationally, through several different avenues. One important set of constraints comes from **microlensing** – MACHOs would occasionally pass almost exactly between Earth and a star in our galaxy. While the MACHO would rarely be so well-aligned that it would block-out the light from the star completely, it could cause the light to be *gravitationally lensed*, making the star temporarily appear brighter. By monitoring many stars for spikes in brightness of this kind, astronomers have been able to put tight constraints on the number of unseen compact objects in our galaxy, effectively ruling out this explanation for dark matter. There are simply not enough of them!

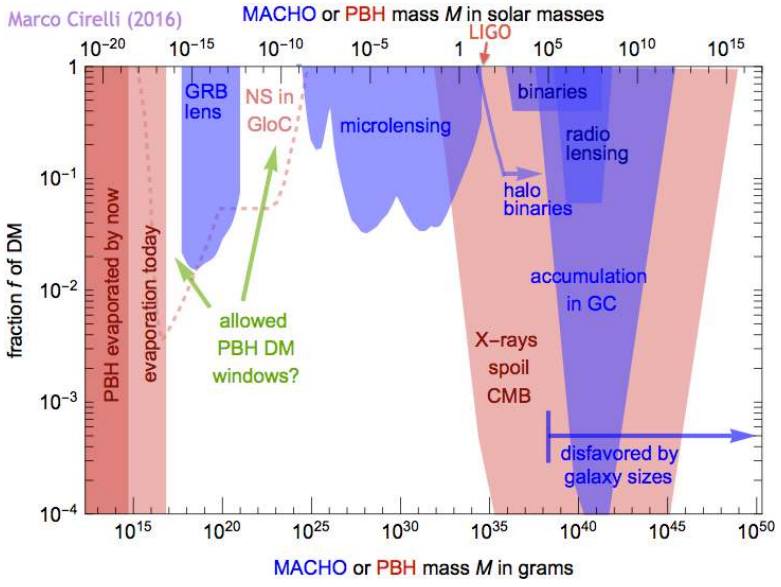

Figure 27: An *exclusion plot* showing which values of MACHO mass and fraction of dark matter are allowed by observations. The white regions are allowed, while the coloured regions have been excluded by different types of observations. For example, observations allow MACHO dark matter to have a mass of $10^{-5} M_\odot$ (upper x-axis) and make up a fraction of $\leq 10^{-2}$ of the total dark matter density (y axis). They cannot have this mass and make up 100% of the dark matter though, as this would be in the blue region *excluded* by microlensing observations. A more detailed description is given here. (Credit: M. Cirelli)

Another possibility is that a large number of black holes (**primordial black holes, or PBHs**) formed in the early Universe. This is an intriguing idea that was pioneered by Bernard Carr (in the Astronomy Unit

at QMUL) and Stephen Hawking, amongst others. Black holes certainly don't emit light, and so would be invisible. They can also be formed in the very early Universe if the conditions are right (although this requires inflation to proceed in a more complicated way in order to create enough PBHs). The fact that the black holes were created primordially would also help to explain why the distribution of MACHOs is different to everything else in the galaxy. This idea was disfavoured until recently however, as the PBHs would have to be created in a narrow mass range in order to stay within observationally-allowed bounds, which seemed unlikely (see the figure above for a compilation of observational constraints). The recent detection of gravitational waves from colliding black holes by the LIGO and VIRGO experiments just so happened to uncover a population of black holes with masses in this range, however – several tens of Solar masses, or thereabouts. This was somewhat surprising (we didn't expect to see quite so many black holes in this mass range), and led to a suggestion that they could actually be the PBHs needed to explain dark matter. Further work is fast closing this window however, and it seems unlikely that MACHOs could make up more than a small fraction (perhaps a few percent) of the dark matter.

Further reading: MACHOs (Wikipedia); Black holes in the early Universe (B. Carr and S. Hawking (1974)

10.5. Warm vs cold dark matter

If the dark matter is made of some as-yet undiscovered particle, we can obtain some basic clues about how and when it was produced by figuring out its mean energy or temperature. This is hard to do precisely, but there are sufficiently clear differences in how it behaves depending on whether the particles are moving at relativistic speeds or not. Since highly-relativistic hot dark matter (HDM) is off the table, let's instead consider two options: warm and cold dark matter.

Warm dark matter (WDM) corresponds to a type of particle that is moving at mildly relativistic speeds. This type of particle may start off with higher energy, behaving like a highly relativistic particle (i.e. radiation), but then cool down as the Universe expands, becoming slower and moving more like regular non-relativistic matter. Its high speed means that its mean free path is relatively large. Recall how the changing mean free path of photons affected the formation of the CMB temperature anisotropies – as the photon-baryon fluid became less opaque, photons were able to travel larger distances before being scattered. As they diffused out of high-density regions, they removed energy and dragged the baryons with them, smoothing out fluctuations that had formed on smaller scales. WDM does something similar; because of its large mean free path, it can diffuse outwards quite far, dragging other matter along with it due to its gravitational attraction. This tends to smooth out fluctuations in the matter distribution on small distance scales. The warmer the dark matter, the further it can diffuse, and so the more fluctuations on small scales will be damped away.

Cold dark matter (CDM), on the other hand, begins non-relativistic and stays non-relativistic. Its mean free path is small (but not zero!), and so fluctuations on small scales are not diffused away. This leads to a firm prediction – if dark matter is warm, the fluctuations on small scales should be significantly weaker than on large scales, at least up to a distance scale corresponding to the mean free path of the particles. If dark matter is cold, however, then there will be appreciable fluctuations both large and small scales.

10.6. Hierarchical structure formation

Fluctuations in the dark matter distribution tend to grow bigger with time. This is due to *gravitational collapse* – regions with a higher density than the average tend to attract matter from their surroundings and grown even denser, while lower-density regions more easily lose material to their surroundings and become even emptier. This process, beginning from the seed fluctuations left over from inflation and the various processes that occurred before decoupling, and allowed to continue for many billions of years, is called **structure formation**, and results in the complex distribution of matter that we see in the Universe today, otherwise know as the **large-scale structure of the Universe**.

There are several different types of structures in the Universe. First, we can divide structures into two types: gravitationally **bound** and **unbound** structures. Those that are bound have grown to a high enough density to resist the cosmic expansion. Instead of expanding with the Universe, the dominant force is their own self-gravity. We say that these structures are 'outside the Hubble flow'. Unbound structures are stretched as the Universe expands however, and so are 'in the Hubble flow'.

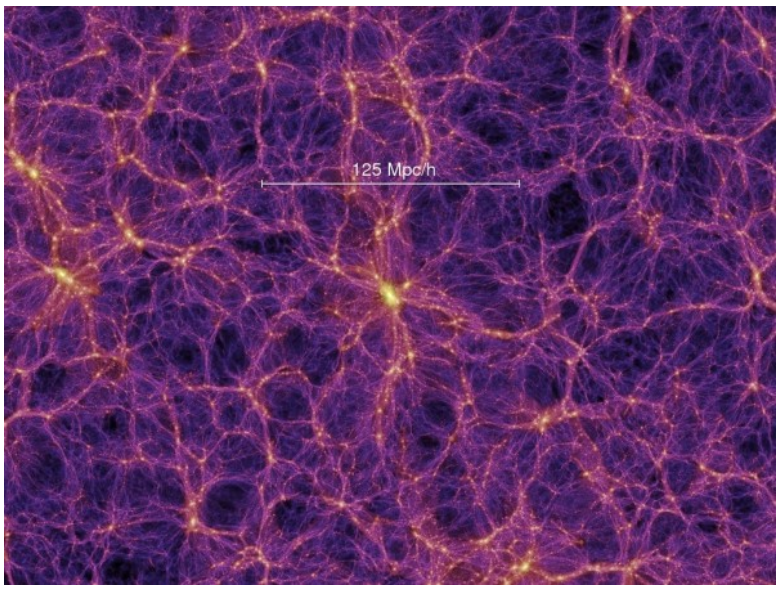

Figure 28: A slice through a simulation of the large-scale structure of the Universe. The brighter the colour, the higher the density of cold dark matter. Note the small, high-density regions in yellow, which are galaxy clusters, and how they are connected by a 'web' of narrow, lower-density filaments. (Credit: V. Springel/Virgo Consortium)

The largest gravitationally-bound structures in the Universe are **galaxy clusters** – large agglomerations of hundreds to thousands of galaxies that are sufficiently dense that their constituent galaxies resist the cosmic expansion. Next are smaller **galaxy groups**, made up of tens to perhaps a hundred galaxies. On smaller scales are galaxies themselves, which contain even smaller bound objects such as gas clouds, stars, and planets.

On larger scales than a galaxy cluster, we find two types of unbound structure. **Filaments** are regions of higher than average matter density that run between clusters, while **voids** are regions of lower than average density that may contain only a small number of galaxies. Voids tend to be very large, as there has been enough time for clusters and filaments to attract matter towards themselves from a large surrounding region. Clusters tend to show up at the intersection of filaments.

As you can see, there is a *hierarchy* of structures. The largest structures are formed from a collection of smaller ones. But how did this hierarchy form?

There were originally two competing models. **Top-down** structure formation is when the largest structures formed first. The smaller structures then formed within them. For example, a cluster or filaments might form, and then subsequently fragment into smaller structures as it continued to collapse gravitationally. These smaller structures would then form into galaxies. Similarly, as the galaxies formed, they would fragment into smaller structures that then became stars. Conversely, **bottom-up** structure formation is when the smallest objects formed first. As time went by, they would then fall into the potential wells created by seed fluctuations on larger scales, enhancing those fluctuations and therefore causing the growth of the larger structures.

Which process happens depends on the initial seed fluctuations. If the seed fluctuations are most prominent on large scales, but washed out/damped on small scales, top-down structure formation will occur. Structures on smaller scales can only form from the fragmentation of the larger structures. If the seed fluctuations are prominent on all scales, however, the ones on smaller scales will have had more time to grow gravitationally (smaller scales have been inside the Hubble radius for longer), and so bottom-up structure formation will occur.

In our Universe, we now know that structure formed bottom-up. The reason is that it seems to mostly contain *cold dark matter*. As discussed in the previous section, CDM does not wash out fluctuations on small scales, while WDM does. If dark matter was warm or hot, the small-scale fluctuations would have been washed out and structure formation would have proceeded top-down instead. As explained in the next section, this would result in far fewer low-mass objects, such as *dwarf galaxies*.

Further reading: Origin of Structure (R. Guzmán)

10.7. Dark matter halos

Galaxy clusters are the visible signs of the largest collapsed structures, but in reality it is the invisible dark matter that makes up most of the large-scale structure and that dominates the process of structure formation. Where the dark matter goes, the galaxies tend to go, and so what we see as a cluster of galaxies is actually only the 'tip' of a massive 'iceberg' of dark matter called a **dark matter halo**. It is the dark matter halos that are the main collapsed structures; the baryonic matter in galaxies (and intergalactic gas called the intergalactic medium) contributes only about a fifth of the total mass of the collapsed structure.

We can predict quite early on whether a particular region of the Universe will collapse to form a dark matter halo. If the total amount of matter within some radius is above a particular threshold, we can calculate its subsequent evolution using Newtonian gravity to find that it will resist cosmic expansion and collapse in on itself to form a bound object (a dark matter halo). The halo will not continue to collapse into a point; once it reaches a relatively compact size, it will stop collapsing as the dark matter particles will have gained kinetic energy through the collapse and reached thermal equilibrium, with their thermal energy providing an outward pressure to support the halo. Halos that have reached this stage are said to be **virialised**, as they have reached an equilibrium between gravitational potential energy and thermal/kinetic energy.

The boundary of a halo is defined by the **virial radius**. Inside this radius matter is gravitationally bound to the halo, while outside this radius it can escape (and so 'feels' the cosmic expansion). In our Universe, the virial radius of a halo occurs when the density of dark matter is around 200 times the critical density. Inside the halo, the density increases significantly towards the centre, following a characteristic density profile called the **Navarro-Frenk-White (NFW) profile**. This profile can be used to calculate the mass and rotational velocity of the matter in the halo, amongst other things.

Halos form with a range of different masses, depending on the size of the original over-dense region that they formed from. More massive halos require a larger over-dense region and so on. Halos can also grow by merging together to form even more massive halos. The distribution of halo masses can be predicted both analytically and by using cosmological simulations. The number density (number per unit volume) of halos per unit mass is called the **halo mass function**. However it is calculated, it tends to predict an increasing number density of halos towards lower masses, with a sharp exponential cut-off at high masses. This is because small halos are easier to form, so there will be more of them. Large halos are difficult to form however, and above a certain mass can't form at all due to the cosmic expansion that counteracts gravitational collapse. Some example halo mass functions are shown in the figure below.

Thinking back to the warm vs cold dark matter question, it turns out that the abundance of dark matter halos (as described by the halo mass function) is strongly affected by the temperature of the dark matter. In a warm dark matter scenario, there won't be as many fluctuations on small scales, and so not as many small (low-mass) DM halos will be able to form. Larger (high-mass) halos will be unaffected however, as they form from the gravitational collapse of larger regions, corresponding to fluctuations that are too big to have been damped away. Smaller objects will only be able to form later on, as the larger objects fragment into smaller pieces that can then collapse gravitationally.

In our Universe, the halo mass function, simulations, and other observations point towards CDM – the fluctuations on small scales do not seem to have been suppressed or damped in any way beyond what is expected from the diffusion damping of the CMB. Since galaxies follow the dark matter, we would be able to see this effect in the distribution of galaxies of different masses. For example, if dark matter was warm instead of cold, we would see fewer low-mass galaxies called **dwarf galaxies**. Whatever dark matter is actually made from, it seems to be *cold* (non-relativistic).

Further reading: Formation and structure of dark matter halos (Wikipedia); Spherical collapse model of halo formation (X. Shi) [PDF]; Virial mass (Wikipedia); Navarro-Frenk-White profile (Wikipedia)

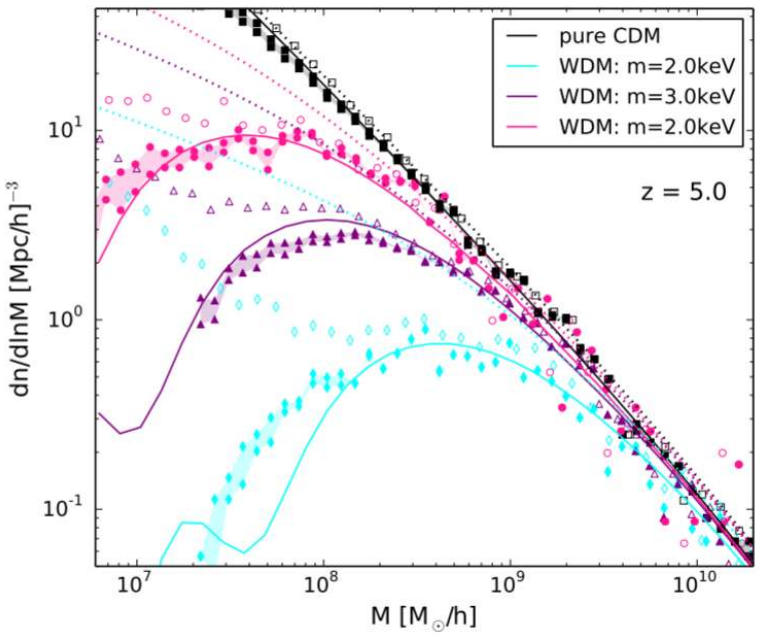

Figure 29: The halo mass function (relative abundance of dark matter halos as a function of mass) for several different types of dark matter. The lower the mass/energy of the dark matter, the 'warmer' (more relativistic) it is, and so the more fluctuations on small scales are damped away. This causes the number of low-mass halos to be suppressed compared to CDM. This plot is from a simulation; see here for more details (Credit: A. Schneider).

Learning outcomes:

What is dark matter?

What are the three main pieces of observational evidence for dark matter?

What are the known properties of dark matter?

What are some of the proposals for what dark matter might be made of?

What are WIMPs and how can they be detected?

What are MACHOs and how can they be detected?

What are the differences between hot, warm, and cold dark matter?

What is the difference between a bound and unbound structure?

What is the difference between top-down and bottom-up structure formation?

What is a dark matter halo?

What is one piece of evidence that disfavours WDM compared to CDM?

11. Structure formation

In this section, we will learn how to describe the evolution of the large-scale structure in terms of perturbations to a perfectly homogeneous and isotropic FLRW universe. The the mathematical formalism used to describe this is called *Cosmological Perturbation Theory*, and while the full details are quite involved, it is possible to arrive at a relatively straightforward equation that describes the *growth rate of structure* as long as the deviations from homogeneity and isotropy are small. We will see some applications of this result, and learn how it can be used to predict statistical quantities such as the *matter power spectrum* and *galaxy correlation function*. In support of these results, we will also need to use the 3D Fourier transform.

Reading for this topic:

- *An Introduction to Modern Cosmology (A. Liddle), Chapter 9 and Advanced Topic 5.*

11.1. Perturbation theory

The fact that the Universe is close to homogeneous and isotropic allows us to use an extremely useful mathematical tool to study the large-scale structure. As long as the deviations from perfect homogeneity are small (which is true at least on large distance scales), we can use *perturbation theory* to describe how structures in the Universe grow and evolve with time.

Perturbation theory is similar in spirit to performing a Taylor expansion. We pick something to expand around (in this case, the solution for a perfectly homogeneous FLRW cosmological model, which we call the *background*), and then simplify the resulting equations by neglecting higher powers of a small variable (the perturbations).

Which equation, and what quantities, do we perturb (expand in a small parameter)? The answer is – all of them! For example, we must certainly perturb the Friedmann equation – if the Universe is slightly homogeneous, the expansion rate will now need to depend on position as well as time (we have already met the relevant perturbed quantity – the peculiar velocity, which would be zero in a perfectly homogeneous and isotropic Universe). So too will the density. There will also be new equations that determine how the perturbations themselves change with time and location. The important thing is to perturb all of the relevant equations in a consistent manner, so that (e.g.) the perturbed matter density that appears in one equation is the same as the perturbed density in any other equation.

There is a well-defined procedure to derive all of the perturbed equations that we need for cosmology, but unfortunately it's too involved to go into here. Instead, we will take a look at some of the underlying principles of the method, and study a couple of the end results.

A useful way of making sure that all of the perturbed equations are consistent is to start from the most 'basic' object possible and then derive all of the other quantities that we need from a perturbed version of that. For the geometry of space-time, a good place to start is with the *metric*. We can split the metric into a perfectly homogeneous and isotropic background part ($\bar{g}_{ab}$), and a small fluctuating part (h_{ab}),

$$g_{ab}(\vec{x},t) = \bar{g}_{ab}(t) + h_{ab}(\vec{x},t). \tag{149}$$

This is essentially a Taylor expansion around the FLRW metric, but the small parameter is a tensor, h_{ab}, rather than a scalar quantity like we're used to seeing.

Remember that, in general, the metric in a 4D spacetime looks like a 4x4 matrix. In principle, each element of the metric could be perturbed differently. Using some simple symmetry properties and a bit of physical intuition, it turns out that only a couple of perturbed quantities are needed to get a pretty good description of the large-scale structure however. Written as a line element, we have

$$ds^2 = -c^2(1+2\Psi(\vec{x},t))dt^2 + (1-2\Phi(\vec{x},t))a^2 d\vec{x}^2. \tag{150}$$

The small quantities Φ and Ψ are called the *metric potentials*, and are closely related to the *gravitational potential* that we use in Newtonian gravity (also recall our discussion of gravitational potentials when we

studied the Sachs-Wolfe effect in the CMB; these are the same potentials[8]). If they are zero everywhere, we recover the perfectly homogeneous and isotropic FLRW metric. The line element as written above is from the **perturbed FLRW metric**, and includes only **scalar perturbations** (we could have included some perturbations that look like vectors and tensors too, but they are small so we can neglect them). It is perhaps slightly clearer to see this written in tensor (matrix) form,

$$g_{ab} = \begin{pmatrix} -c^2(1+2\Psi) & & & \\ & a^2(1-2\Phi) & & \\ & & a^2(1-2\Phi) & \\ & & & a^2(1-2\Phi) \end{pmatrix}. \tag{151}$$

Recall that the metric tells us how to measure distances in space and time. If our spacetime is slightly inhomogeneous, the potentials Φ and Ψ will tell us all about the deviations from homogeneity. Whenever we calculate distances, redshifts, times etc, we must now take these potentials into account.

Since "spacetime tells matter how to move; matter tells spacetime how to curve", we must now consider how the perturbations to the metric (i.e. to the spacetime geometry) are related to perturbations in the matter/energy density. To start off with, we can write the energy density as a perturbation to the background energy density. If we only consider the matter density for now, we can write

$$\rho_m(\vec{x},t) = \bar{\rho}_m(t) + \delta\rho_m(\vec{x},t) = \bar{\rho}_m(t)(1+\delta(\vec{x},t)), \tag{152}$$

where δ is the **density contrast**. If the Universe was perfectly homogeneous and isotropic, we would find $\delta = 0$, and so $\rho_m = \bar{\rho}_m$, i.e. only the background energy would exist, with no perturbations. From the equation above, we can figure out that δ is dimensionless – it is a *fractional* fluctuation in the matter density. We can also see that if $\delta = -1$ in some region, the matter density there will be zero. The matter density can't be negative, so we must always have $\delta \geq -1$.

As long as δ stays close to 0, we can use it as a small perturbative parameter. Note that we can calculate δ for *any* density fluctuation however – a galaxy, a planet, a person, or even a black hole. These have much higher densities than the cosmic average, so $\delta \gg 0$. For these objects, δ cannot be used as a perturbative parameter. This is generally true of most objects found on small distance scales. On large scales however (larger than a galaxy cluster), typical density contrasts are in the range $-1 \lesssim \delta \lesssim +1$, and so our perturbative expansion can be used.

11.2. Growth of matter fluctuations

By following the perturbation theory method through, we can derive equations that tell us how the perturbed quantities should evolve in time and space. For the matter fluctuations, a remarkably straightforward equation pops out called the *growth equation*. As long as the perturbations in the matter density stay small, they can be separated into a time-dependent factor and a space-dependent factor:

$$\delta(\vec{x},t) = D(t)\delta(\vec{x},t_i). \tag{153}$$

The term $D(t)$ is the **growth factor**, and depends only on time. The term $\delta(\vec{x},t_i)$ is the *initial condition* of the matter fluctuations, evaluated at some time t_i. These initial conditions are given by inflation, and are random (see the discussion of the *primordial power spectrum* for more details on this). They are modified slightly by the physics that occurs between the end of inflation and decoupling/last scattering of the CMB too. After that time, however, the spatial distribution of the fluctuations remains the same – an over-dense region will stay over-dense, an under-dense region will stay under-dense. The matter fluctuations simply get bigger or smaller, in a way that is proportional to $D(t)$

The growth equation is

$$\frac{d}{da}\left(a^3 H(a)\frac{dD}{da}\right) = \frac{3}{2}\frac{H_0^2\Omega_m a^{-3}}{H^2(a)} aH(a)D(a). \tag{154}$$

We can solve this with appropriate initial conditions to find a solution for how the growth factor, D, evolves with time/scale factor. By inspecting this equation, we can see that it depends on the expansion rate, H, in a

[8] But *not* the same potentials as the scalar field potentials that we needed to explain inflation.

couple of places. The solution for the growth factor (and hence how the matter fluctuations grow with time) depends on what types of matter/energy are present in the Universe.

Note that we have a choice of which time to define the 'initial' conditions of δ at. A common choice is to defined them at $t_i = t_0$, today. In this case, they are really 'final conditions' rather than initial conditions, but the maths works out the same; we can just solve the growth equation by integrating it backwards in time instead of forwards to get the same result. This choice also allows us to set $D = 1$ at $a = 1$ ($t = t_0$), so D can be interpreted as a sort of scale factor for the size of matter fluctuations.

It's often useful to study the rate of change of the growth factor, rather than the growth factor itself. It is most convenient to define the **linear growth rate** as a logarithmic derivative with respect to the scale factor,

$$f(a) \equiv \frac{d\ln D}{d\ln a} = \frac{a}{D}\frac{dD}{da}. \tag{155}$$

The value of f tells us how rapidly matter fluctuations grow with time. If the energy density of the Universe is dominated by matter, we find $f = 1$, while if there is also a cosmological constant, $f < 1$. We can see that dark energy therefore *suppresses* the growth of structure; it makes matter fluctuations grow more slowly than they would otherwise. (We previously saw another manifestation of this when we discussed the *integrated Sachs-Wolfe effect* in the CMB.)

11.3. Poisson equation

We can also derive equations that describe how the metric potentials (the perturbation to the geometry) are related to the density fluctuations. These are reasonably complicated, but on moderately small scales (a few tens to hundreds of Mpc) can be reduced to a simple form called the **Poisson equation**:

$$\nabla^2\Phi = -4\pi G a^2 \bar{\rho}_m \delta. \tag{156}$$

This looks very similar to the Poisson equation from Newtonian gravity, except for the factors of a. This equation tells us that the density fluctuations are proportional to the second spatial derivative of the potential, Φ. So, wherever there are fluctuations in the metric potential, we will also expect to see fluctuations in the matter density (their distribution following the second derivative of the potential). We can also see that the prefactors of δ in this equation are time-dependent, so δ will also change with time, even if the potential Φ does not. This is described by the growth equation in the last section.

11.4. Fourier transforms

In general, the spatial distribution of fluctuations in the potential and the density contrast will be quite complicated, and so it's not clear how we should model them mathematically. Because the time- and space-dependence of the matter fluctuations can be factorised, however, we can use a neat trick to simplify things. That trick is to take a **3D Fourier transform** of the fluctuations.

The Fourier transform allows us to separate the fluctuations according to their *wavenumber*, k. This is approximately related to the comoving size, r, of a fluctuation by $k = 2\pi/r$. So, bigger wavenumbers correspond to fluctuations on smaller scales, and smaller wavenumbers correspond to fluctuations on larger scales. The 3D Fourier transform and its inverse for any scalar quantity f can be calculated using

$$\begin{aligned} f(\vec{x}) &= \int \tilde{f}(\vec{k}) e^{i\vec{k}\cdot\vec{x}} \frac{d^3k}{(2\pi)^3} \\ \tilde{f}(\vec{k}) &= \int f(\vec{x}) e^{-i\vec{k}\cdot\vec{x}} d^3x. \end{aligned} \tag{157}$$

The Fourier transform is a *linear operation*. Mathematically, we can always swap the order of linear operations. The nice thing about cosmological perturbation theory, at least on the distances scales we are interested in here, is that it is linear. So, we can use Fourier transforms to simplify some of our equations by Fourier-transforming a quantity before we apply another linear operator (like the ∇^2 operator in the Poisson equation).

Another nice result of taking the Fourier transform is that *each mode evolves independently*. So, to figure out how a density fluctuation $\delta(k)$ depends on the potential, we only need to know $\Phi(k)$ (and not $\Phi(k')$, for

any $k \neq k'$. This is very useful, as it allows us to analyse the distribution of matter fluctuations mode by mode, completely ignoring whatever any other modes are doing.

To see an example of why 3D Fourier transforms are useful, let's apply them to the Poisson equation. First, let's substitute the definition of the Fourier transform into this equation for $\Phi(\vec{x}$ and $\delta(\vec{x})$:

$$\nabla^2 \left(\int \Phi(\vec{k}) e^{i\vec{k}\cdot\vec{x}} \frac{d^3k}{(2\pi)^3} \right) = -4\pi G a^2 \bar{\rho}_m \int \delta(\vec{k}) e^{i\vec{k}\cdot\vec{x}} \frac{d^3k}{(2\pi)^3}. \tag{158}$$

The Fourier transform is a linear operation, and so is ∇^2, so we are allowed to bring ∇^2 inside the integral. Recall that ∇^2 is the (second) spatial derivative, so only depends on $\vec{x}$, and not $\vec{k}$. So, we can swap its order with $\Phi(\vec{k})$ too! On the right-hand side, you can also see that we've been able to bring purely time-dependent quantities outside the integral too, since the Fourier transform doesn't depend on time. Concentrating on the left-hand side, we now have

$$\nabla^2 \left(\int \Phi(\vec{k}) e^{i\vec{k}\cdot\vec{x}} \frac{d^3k}{(2\pi)^3} \right) = \int \Phi(\vec{k}) \left(\nabla^2 e^{i\vec{k}\cdot\vec{x}} \right) \frac{d^3k}{(2\pi)^3}. \tag{159}$$

Recalling a bit of vector calculus, we find that

$$\nabla^2 e^{i\vec{k}\cdot\vec{x}} = (i\vec{k}) \cdot (i\vec{k}) e^{i\vec{k}\cdot\vec{x}} = -k^2 e^{i\vec{k}\cdot\vec{x}}. \tag{160}$$

(We have used the definition $k^2 = \vec{k} \cdot \vec{k}$, so k is the magnitude of the wavevector $\vec{k}$.) Plugging this back into our transformed Poisson equation, we obtain

$$-\int \Phi(\vec{k}) k^2 e^{i\vec{k}\cdot\vec{x}} \frac{d^3k}{(2\pi)^3} = -4\pi G a^2 \bar{\rho}_m \int \delta(\vec{k}) e^{i\vec{k}\cdot\vec{x}} \frac{d^3k}{(2\pi)^3}. \tag{161}$$

If we put everything back inside the integrals, we get

$$\int \left(\Phi(\vec{k}) k^2 \right) e^{i\vec{k}\cdot\vec{x}} \frac{d^3k}{(2\pi)^3} = \int \left(4\pi G a^2 \bar{\rho}_m \delta(\vec{k}) \right) e^{i\vec{k}\cdot\vec{x}} \frac{d^3k}{(2\pi)^3}. \tag{162}$$

This implies that the terms in brackets are equal, so we can write

$$\Phi(\vec{k}) k^2 = 4\pi G a^2 \bar{\rho}_m \delta(\vec{k}). \tag{163}$$

This shows one of the most useful features of the 3D Fourier transform – it changes differential equations (e.g. with a ∇^2 in them) into algebraic ones (no spatial derivatives)! We can now see that $\delta \propto k^2 \Phi$, so for a given potential, the density fluctuations will be larger on smaller scales (higher values of k).

11.5. Matter power spectrum

Remember that the initial conditions of the universe (set by inflation) are *random but correlated*. This means that we can't predict exactly what the value of the potential or density contrast will be at any given point, but we *can* predict (with some degree of uncertainty) how similar two nearby points will be. For example, galaxies and dark matter halos tend to cluster together due to their mutual gravitational attraction, so if we observe one galaxy in a particular location, we also expect to see a few other galaxies nearby.

This is OK as a qualitative picture, but can we *quantify* how clustered the distribution of matter is? We would like to be able to say if it is more or less clumpy, and on which distance scales it is most clumped together (i.e. are most structures large or small?). We can use a familiar statistical tool to answer these questions – the **power spectrum**. We have already studied the power spectrum of the CMB temperature fluctuations, and briefly discussed the primordial power spectrum too. In the context of large-scale structure, we usually talk about the *matter power spectrum*, i.e. the power spectrum of the density fluctuations, δ.

Recall that the power spectrum is really measuring the *variance* of the fluctuations – what their typical size is – as a function of distance scale. In this case, the distance scales are represented by Fourier wavenumbers, k. So, the power spectrum $P(k)$ tells us the variance of the density fluctuations on distance scales corresponding to comoving distances of around $r \approx 2\pi/k$. A large value of $P(k)$ means that the variance is large for that

wavenumber, which means that there are significant fluctuations (i.e. more structure) on that scale. Conversely, a small $P(k)$ means that there is not much structure on that particular scale.

As a slightly tortured analogy, consider a beach made up of sand and pebbles. There are lots of small sand grains packed close together, so the 'power spectrum' of the beach would be quite large on small scales (high k; in this case, a millimetre or smaller). There are fewer pebbles, scattered more sparsely across the beach, so the power spectrum would be lower on intermediate scales (perhaps a few centimetres or so). On larger scales, we wouldn't pick up individual objects like the sand or pebbles any more – we would be looking at the structure of the entire beach. If the beach had large sand dunes, the power spectrum would be large on large scales (small k), whereas if it was almost perfectly flat, the power spectrum would be small.

The precise definition of the matter power spectrum is as follows:

$$\langle \delta(\vec{k}) \delta^*(\vec{k}') \rangle = (2\pi)^3 \delta^{(3)}(\vec{k} - \vec{k}') P(k). \tag{164}$$

The left-hand side is just the notation for the *variance* of $\delta(\vec{k})$ (the variance is the square of the standard deviation). The angle brackets just mean 'an average over all fluctuations with these Fourier modes', and you can see that were are essentially 'squaring' δ, as is needed when we calculate a variance.

On the right-hand side, we have a factor of $(2\pi)^3$ that just comes from our Fourier transform convention. The next factor is a 3D *Dirac delta function*, $\delta^{(3)}$. Note that this has nothing to do with the density contrast, δ; it just happens to use the same Greek symbol. The Dirac delta function is zero everywhere, except at the exact point where $\vec{k} = \vec{k}'$. This delta function just tells us that all of the Fourier modes are uncorrelated with one another (i.e. are independent from one another). Physically, this means that a fluctuation mode with wavevector $\vec{k}$ doesn't know anything about any other wavevector $\vec{k}'$ – it evolves in a way that is completely independent, no matter how big or small the fluctuations on other distances scales are. This is quite a useful property – it means that we can study each Fourier mode in isolation, without having to worry about what any of the others are doing.

The final factor on the right-hand side is the *matter power spectrum*, $P(k)$. This is similar to the CMB power spectrum, in that it tells us how many fluctuations we have as a function of distance scale. Bigger values of $P(k)$ mean that there are more fluctuations on distance scales $r = 2\pi/k$, while smaller values would mean that there are fewer fluctuations on those scales (i.e. the matter distribution would be smoother). Note how $P(k)$ only depends on k, which is just the magnitude of the wavevector, $k = |\vec{k}|$. The matter distribution is statistically *homogeneous and isotropic*, and so has no dependence on position or direction. As a result, the power spectrum in this case only depends on k and not $\vec{k}$.

11.6. Correlation function

The correlation function is the equivalent of the power spectrum in real space (as opposed to Fourier space). It describes the *covariance* of the density contrast between two points, separated by a vector $\vec{r}$.

$$\xi(\vec{r}) = \langle \delta(\vec{x}) \delta^*(\vec{x} + \vec{r}) \rangle. \tag{165}$$

The angle brackets denote an average[9] over all possible positions in space, $\vec{x}$. Since we expect the matter distribution in our Universe to be statistically homogeneous and isotropic, we expect the statistical properties of the distribution to be the same in every region, and so the correlation function should *not* depend on position, $\vec{x}$. Instead, it only depends on the *separation* between two points, $\vec{r}$; in other words it is *translation invariant*.

The correlation function can be interpreted as the probability that we will find a pair of galaxies or dark matter halos with some separation (compared with how often you'd expect to find that separation for a completely random distribution of galaxies, with no large-scale structure). If $\xi(r)$ is large for some value of r, then many galaxies/halos will be found separated by r. A measurement of the correlation function for galaxies is shown below. You can see that it is large for small separations ($r \lesssim 50$ Mpc). Galaxies are often found clustered together, especially within the same dark matter halo. So, many pairs of galaxies/halos can be found close together. At very large separations, the correlation function goes to zero. This does not mean that *no*

[9] There is a subtlety here that I am glossing over. The angle brackets are actually an *ensemble average*, i.e. an average over all possible statistical realisations of the density contrast. Due to something called the *ergodic theorem*, we can reinterpret the angle brackets as an average over all points in space in our Universe however.

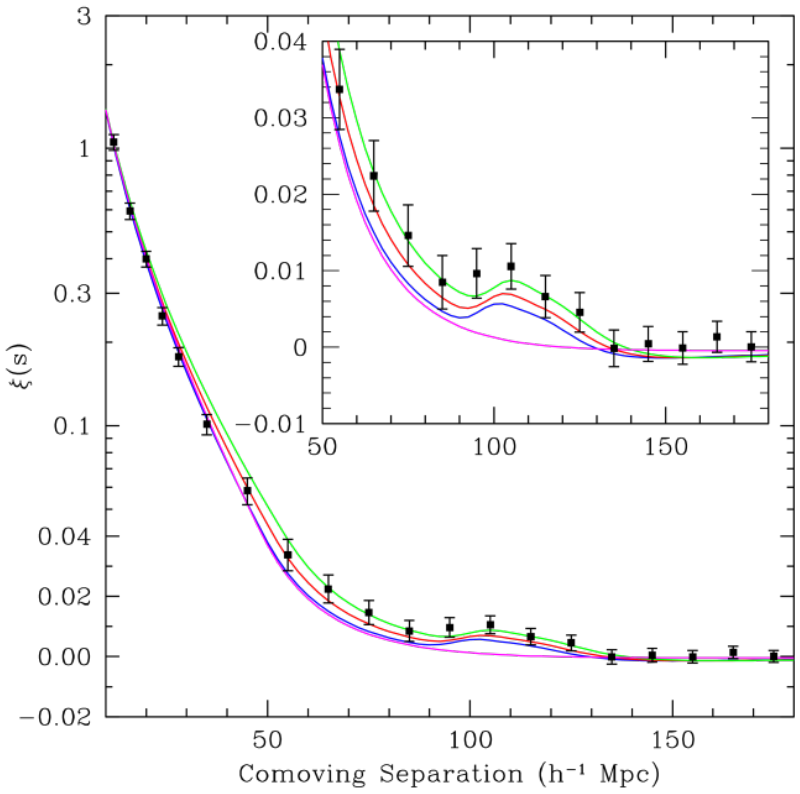

Figure 30: The galaxy correlation function, $\xi(r)$, measured by the SDSS/BOSS galaxy survey. The BAO bump is shown in more detail in the inset. It occurs at a separation of **about 150 Mpc** (which is approximately 100 Mpc/h if you multiply by the dimensionless Hubble parameter, as has been done in this plot). See the original paper for more details (Credit: D. Eisenstein et al.).

galaxies/halos are found with large separations ($r \gtrsim 150$ Mpc); it just means that the number of galaxies found with that separation is very similar to what you'd expect to find for a completely random distribution.

If the correlation function is larger than zero, it means that there are structures on those distance scales that some kind of physical process must have created. This is similar to seeing peaks in the CMB power spectrum, like the acoustic peak. We know from the CMB that the *baryon acoustic oscillations* imprinted some structures in the matter distribution around the time of last scattering, at a distance scale of around 150 Mpc (corresponding to the sound horizon at around that time). That same distance scale is seen in the galaxy correlation function too! The figure below shows that a small bump appears exactly where predicted – at around 150 Mpc (which is $\sim 100h^{-1}$ Mpc in the units used in that figure). This bump is closely related to the acoustic scale just after last scattering, caused by the BAO phenomenon!

Different types of galaxies form depending on where they are in a dark matter halo, and depending on the mass of the halo itself. For example, large red elliptical galaxies tend to be found in the centre of very massive halos, whereas blue spiral galaxies are more often found in the outskirts. Small, irregularly-shaped dwarf galaxies are found in smaller, low-mass halos. If we look at the distribution of galaxies over very large scales, it doesn't particularly matter exactly where the galaxies are within a halo; they are still tracing the positions of themselves to a good degree of accuracy. So, any type of galaxy can be used to trace the underlying dark matter distribution. The main difference is that the galaxy correlation function will have a *different amplitude* depending on the type of galaxy that is used. The amplitude is described by a parameter called the **galaxy bias**, b, which is different for each kind of galaxy. It depends on the typical mass of the dark matter halos that those galaxies are found in. Galaxies found in very massive halos have a large bias ($b \gg 1$), for example.

The bias relates the galaxy correlation function to the matter correlation function:

$$\xi_g(r) = b^2 \xi_m(r), \tag{166}$$

i.e. the galaxy correlation function has the same shape as the matter correlation function, but has an amplitude

proportional to b^2. This relation is true on reasonably large scales, above a few tens of Mpc, but breaks down on smaller distance scales, where the exact position of the galaxies within the halos starts to matter more.

What does the galaxy bias *mean*? You can think about it as measuring the likelihood that a particular kind of galaxy will be found near a peak (over-density) in the dark matter distribution. Red elliptical galaxies have a high bias because they are almost always found in the biggest peaks of the dark matter distribution (where the most massive halos are). Blue spiral galaxies are often, but not always, found near slightly smaller peaks (lower-mass halos), so have a lower bias.

11.7. Peculiar velocities

We have already discussed how the density contrast δ and gravitational potential Φ are related, but what about the peculiar velocity? It turns out that this is related to the time derivative of the density contrast,

$$\frac{d\delta(k,a)}{dt} = -ik\frac{v(k,a)}{a},$$

where δ and v are both evaluated in Fourier space, and $\delta(k,a) = D(a)\delta(k,a=1)$. Note the factor of i! This doesn't mean that the velocity is an imaginary number. It's actually a phase factor; recall that $i = e^{i\pi/2}$. So, it just means that the Fourier modes of the velocity are $\pi/2$ out of phase with the density contrast – wherever the density contrast Fourier mode is at a maximum, the velocity Fourier mode will be zero and so on. If we do the inverse Fourier transform, the quantity that we get back for the velocity is real, so there is nothing fishy going on here.

Now let's look at the time derivative of the density contrast. That can be simplified by remembering that the density contrast factorises into a time-dependent and space-dependent part, $\delta(t) = D(t)\delta(t_i)$. It doesn't matter that the δ in the equation above has been Fourier transformed; the growth factor D doesn't depend on $\vec{x}$ or $\vec{k}$, and so is unaffected. We can then write

$$\frac{d\delta(k,a)}{dt} = \frac{dD}{dt}\delta(\vec{k},a=1) = \dot{D}\delta(\vec{k},a=1). \tag{167}$$

If we want to keep things in terms of $\delta(k,a)$, we can write

$$\delta(\vec{k},a=1) = \frac{\delta(k,a)}{D(a)} \implies \frac{d\delta(k,a)}{dt} = \frac{\dot{D}}{D}\delta(k,a). \tag{168}$$

So, the peculiar velocity is proportional to the density contrast, up to a factor that depends on time derivatives of the growth factor, and the inverse of the wavenumber, k^{-1}.

Learning outcomes:

What does it mean to perturb the metric?

What is the definition of the density contrast?

What is the growth factor, D, and growth rate, f?

How does dark energy affect the growth rate?

What is the matter power spectrum and how is it related to the density contrast?

What is galaxy bias?

What is the galaxy correlation function and how is it related to the matter correlation function?

What does the Baryon Acoustic Oscillation feature look like in the galaxy correlation function?

How can Fourier transforms be used to simplify perturbation equations?

How is the velocity related to the density contrast in Fourier space?

12. Observational cosmology

In this section, we will learn about some of the observational techniques that are used to measure the expansion rate, geometry, and large-scale structure of the Universe.

Reading for this topic:

- *An Introduction to Modern Cosmology (A. Liddle), Chapter 9.*

12.1. Type Ia supernovae

We can use Type Ia supernovae as standard candles, to measure the luminosity distance $d_L(z)$ as a function of redshift. Since $d_L(z)$ depends on an integral that involves the expansion rate, $H(z)$, different values of the cosmological parameters (like H_0 and Ω_m) will produce different $d_L(z)$ curves. By comparing our measurements of d_L from many different supernovae at many different redshifts which can figure out which curve (and therefore which set of cosmological parameters) best fits the real Universe. This method was used to provide the first compelling evidence that the Universe was expanding, by measuring $\Omega_\Lambda > 0$.

How do Type Ia supernovae work as standard candles? These supernovae occur in a binary star system, where there is a white dwarf and a companion star. The white dwarf slowly accretes material from the other star onto its surface. White dwarfs are supported by electron degeneracy pressure, and will collapse if they grow beyond a mass of around $1.4M_\odot$ (the Chandrasekhar mass), as the degenerate electrons will no longer be able to counteract the gravitational attraction of the matter in the star. This means that all Type Ia supernovae occur when the white dwarf hits the same mass, and so a very similar amount of energy is available in each Type Ia explosion. As a result, all Type Ia explosions have a similar luminosity, and so can be used as standard candles.

The way we measure the Type Ia luminosity is by observing them multiple times over several night after the explosion. The brightness of the supernova actually goes up for a few days, since radioactive material in the supernova debris decays and produces a lot of energy over this timeframe. A plot of the brightness of the supernova as a function of time is called a **lightcurve**, and it is these that can be used to give a standardised measurement of the luminosity.

Further reading: Type Ia supernovae

12.2. Galaxy surveys

The large-scale structure of the Universe is dominated by fluctuations in the dark matter distribution, such as the halos, filaments, and voids that were discussed in a previous section. We can't directly observe dark matter however, and so we need some other way to observe the large-scale structures. Galaxies are an excellent candidate as **tracers** of large-scale structure; they are bright (and so quite easy to observe), there are many of them, and the distribution of galaxies traces the distribution of dark matter (to to some bias parameter; see the previous section).

If we can measure the angular positions and redshifts of many galaxies, we can build up a 3D map that will presumably trace the underlying dark matter distribution also. From this we can measure the matter power spectrum and correlation function, for example, which can also be used to measure the baryon acoustic oscillation (BAO) feature. Recall that the BAO feature occurs at a particular comoving distance, and so by observing its location in the correlation function, we can use it as a 'standard ruler' to infer the angular diameter distance, d_A.

There are two main ways of doing a galaxy survey. The first is a **spectroscopic survey**. For this kind of survey, we first take images of large patches of the sky to identify suitable target galaxies. Then, a spectrograph is used to take a frequency spectrum of each galaxy, one by one. Taking spectra takes quite a long time, as most target galaxies are faint, and further splitting their light into many frequency channels makes them effectively even fainter. It is therefore important to observe for a long enough time, so that enough photons can reach the detector to build up the spectrum, without being swamped by random noise. Modern instruments can actually take spectra of many galaxies at once, using *multi-object spectrographs*. These may use fibre optic cables

plugged into specially-machined plates with holes at the location of each target galaxy, or even fibres that are positioned by small robot arms called fibre positioners.

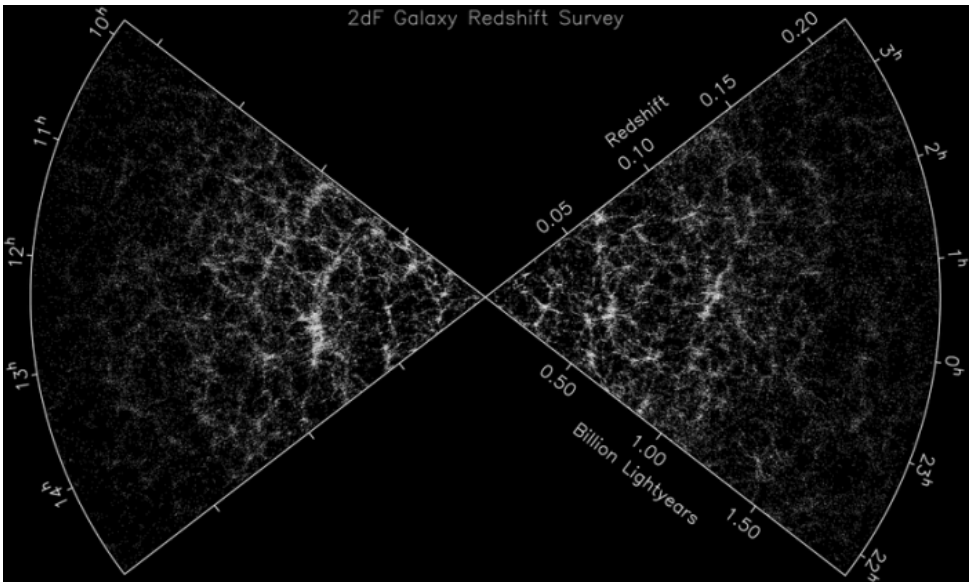

Figure 31: Positions of galaxies detected in the 2dF galaxy survey. The 'cosmic web' structure is very visible: voids, filaments, and clusters of galaxies. (Credit: 2dF survey).

With spectra in hand for each galaxy, we can then identify prominent emission or absorption lines from various elements and use these to infer the redshift of the galaxy very precisely. This gives us the 3D position of the galaxy in spherical polar coordinates – the redshift corresponds to some comoving distance, $r(z)$, from Earth, while the position of the galaxy on the sky corresponds to the angular coordinates θ and ϕ. Modern spectroscopic galaxy surveys have measured the positions for several million galaxies in this way (see figure above).

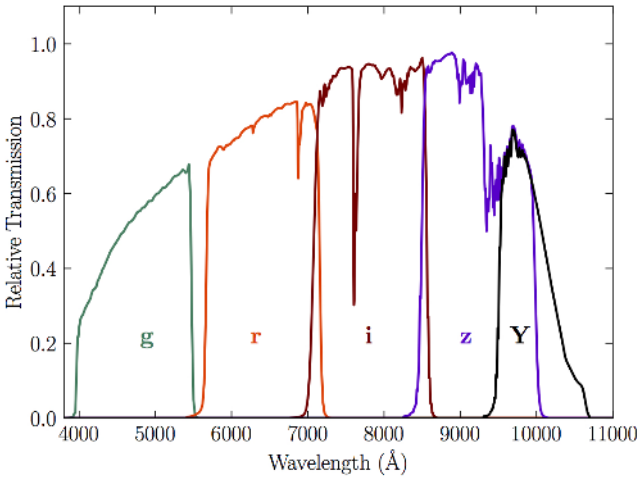

Figure 32: The sensitivity of the colour filters used by the DES galaxy survey. Each filter is sensitive to a broad range of wavelengths, roughly corresponding to a visible/near-IR colour. See here for more details (Credit: CTIO).

The second type of survey is a **photometric survey**. As explained above, it takes a long time to get a good spectrum of a galaxy, and so making 3D maps of the Universe is very time consuming. In practice, the size or depth (maximum redshift) of the maps is limited if we can only use the spectroscopic technique. Photometric surveys solve this problem by obtaining a rough estimate of the redshift by using simpler colour filters instead of a spectrograph (see the figure above). Each filter takes in light across a broad range of wavelengths, roughly

corresponding to a colour of visible light (e.g. green or red). Because the spectrum of each galaxy is only being split into a small number of broad bands, more photons arrive at the detector per band each second, and so measurements can be made much more rapidly.

This technique is like obtaining a very low-resolution spectrum. The amount of light from each band can then be compared to estimate how much redder a galaxy is than if we had observed it at redshift $z = 0$. This gives an estimate of the redshift; more emission in the red filter means that the galaxy is likely to be at a higher redshift and so on. This technique is quite inaccurate however, and so the redshift can typically only be measured to a precision of a few percent. The uncertainty on the redshift of the galaxy caused by using this method is called the **photometric redshift (photo-z) error**. There is also a risk of 'catastrophic outliers', where some galaxies have unusual spectra that may make them seem redder or bluer (and therefore at higher or lower reshifts) than they actually are. There are many trade-offs involved in performing photometric surveys compared to spectroscopic surveys, and so in general we try to do both and see how well they match up.

12.3. Gravitational lensing

As well as making maps of the positions of galaxies, we can also measure their shapes and orientations. While galaxies are mostly randomly oriented, we do see very slight correlations in their shapes. This is mostly caused by **weak gravitational lensing**. As General Relativity predicts, light rays can be deflected by large concentrations of mass. The masses effectively act like a lens, slightly focusing the light and therefore distorting the observed shapes and positions of the galaxies behind them. Galaxies that are lensed by the same massive object will have slightly correlated shapes, and so if we can measure this effect for many galaxies, we can figure out how strong the lens effect is, and therefore how much mass it must contain.

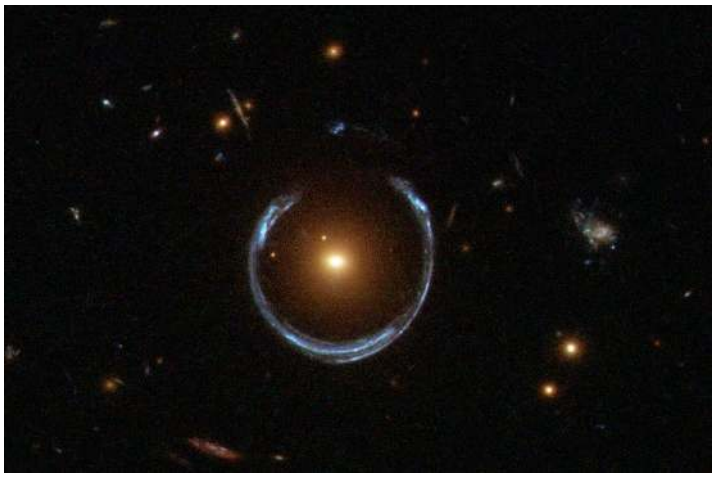

Figure 33: A strongly lensed galaxy. The bright yellow galaxy in the middle is a massive galaxy that acts as the lens. The stretched and distorted blue galaxy is an 'image' galaxy in the background that has been lensed. (Credit: NASA).

All of the matter in the Universe contributes to this effect; a photon travelling from a distant galaxy will be lensed slightly by all of the matter it encounters as it travels towards the observer. By measuring the deflections of light caused by lensing for galaxies at different distances, we can therefore build up a map of all of the mass in the Universe; regardless of whether it is dark matter or baryonic matter. A similar effect (CMB lensing) is seen in the temperature anisotropies of the CMB. The difference with using galaxies as the lensed objects is that they lie at a range of different distances from us, and so we can use a technique called **tomography** to figure out how much lensing is caused by matter lying at a particular range of distances from us. This allows us to create a 3D map of all the matter in the Universe, rather than just a 2D projection as was the case with CMB lensing.

Most of the lensing effect is very small, changing the shape of galaxies by only 1% or less. This is called **weak gravitational lensing**. Occasionally photons pass close to dense objects, such as the inner regions of galaxy clusters, however, and the lensing effect is much stronger. When the effect is large enough to significantly deflect the photons from their original path, the effect is called **strong gravitational lensing**. Strong

gravitational lenses are much rarer, but cause more extreme observational effects – for example, the lens can be strong enough to create multiple images of a galaxy, or even spread the light into a 'ring' around the lens (see figure above). This is similar to the effect you get from looking at a light through the base of a wine glass.

Further reading: Strong gravitational lensing (Wikipedia).

Learning outcomes:

How can Type Ia supernovae be used as standard candles?

What is a galaxy survey?

What is the difference between a spectroscopic and photometric galaxy survey?

What are the relative advantages and disadvantages of spectroscopic and photometric surveys?

What is the difference between weak and strong gravitational lensing?