Combined Reference Notes

Lesson 1: Space and Time

TWO FUNDAMENTAL PRINCIPLES

All of modern cosmology is based around two fundamental principles/observations:

1. Uniformity: The universe is the same everywhere.
2. The Hubble Law: Everything is moving away from us, with a speed that is proportional to its distance.

UNIFORMITY

Is the universe really homogeneous, the same everywhere? It is certainly not uniform on scales of planetary systems - the difference in density between the middle of the Earth and outer space is extreme.

On scales of thousands of light-years, it is also very non-uniform: you have dense galaxies in some regions and empty space in others. Galaxies themselves are gathered into groups and clusters, which are themselves gathered into superclusters, separated by voids. Still not uniform here!

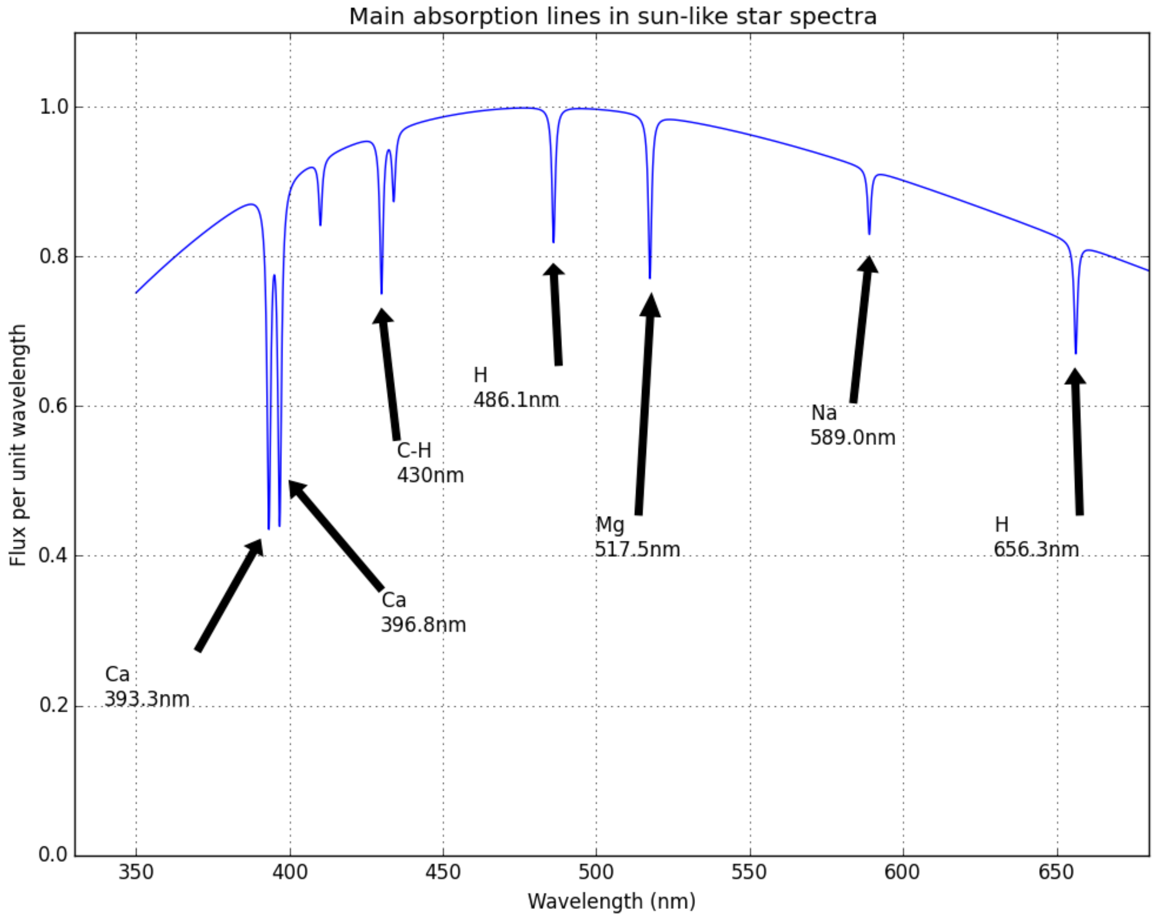
But on larger scales still, things do appear to smooth out. Superclusters do not themselves seem to cluster into anything bigger. Surveys for things like quasars that probe distances of billions of light years do find a very uniform distribution.

So - on really big scales, the universe does appear to be homogeneous.

THE HUBBLE LAW

For a more detailed explanation of the Hubble Law and expanding universe, see the first course in this series, ANU-ASTRO1x “Greatest Unsolved Mysteries of the Universe”.

If you take a spectrum of anything out in space, you will usually see spectral lines, caused by electrons jumping up and down energy levels and either emitting or absorbing light. Each element has characteristic wavelengths at which it produces lines: for example, the very famous H-alpha line of hydrogen occurs at 656.3nm wavelength. Here is a plot of a star spectrum showing some of the strongest and most common spectral lines.



Spectral Lines

If you look at distant galaxies, however, all these spectral lines appear shifted to longer wavelengths. This effect is called redshift, and the redshift of a given object is defined as:

$$z = \frac{\Delta\lambda}{\lambda_0}$$

where z is the redshift and $\Delta\lambda$ is the shift in wavelength (i.e. observed wavelength minus the wavelength λ_0 you would get for this spectral line in a laboratory on Earth.

This shift is caused by the doppler effect: all these distant galaxies are moving away from us at a speed v, which can be found from the equation $v = cz$ where c is the speed of light.

Edwin Hubble showed that the recession velocity is proportional to distance: i.e. that $\vec{v} = H_0\vec{r}$, where H_0 is Hubble's constant and has a value of around $70\text{km s}^{-1}\text{Mpc}^{-1}$ (i.e. a galaxy that is one mega-parsec from us is moving away at a speed of around 70 km/s).

Does this mean we are in a special place and everything is moving away from us? No - using vectors you can show that you would see exactly the same thing wherever you are.

So - we live in a uniform and seemingly endless universe, and wherever you are in this universe, everything seems to be moving away from you with a velocity proportional to its distance.

TWO MASSES

Einstein came up with this theory by pondering the puzzle of why mass crops up in two quite different contexts in physics.

The first context is gravity: mass is something like a gravitational "charge" that indicates how strongly an object attracts another object or is attracted by it. This "gravitational mass" is the mass that you find in the equation

$$F = \frac{GMm}{r^2}$$

The second context is inertia - how much something resists being pushed around. This “inertial mass” is the mass you find in the equation

$$F = ma$$

The puzzle is why these two should be the same. Why should something that resists being pushed around also attract other objects? There is nothing like this for the other forces of nature. Take electromagnetism, for example. How much something attracts or is attracted by this force is determined by its charge. But charge has nothing to do with inertia. There are objects with the same charge but very different inertia (like a positron and a proton).

One concrete manifestation of this is that all objects fall at the same rate - the mass cancels when gravity accelerates something. But an electric field will not attract everything equally - indeed this is used in a mass spectrometer to separate out different materials.

Einstein was wondering - why are these two masses the same? To answer this, he had to think very deeply about the nature of space and time.

THE METRIC

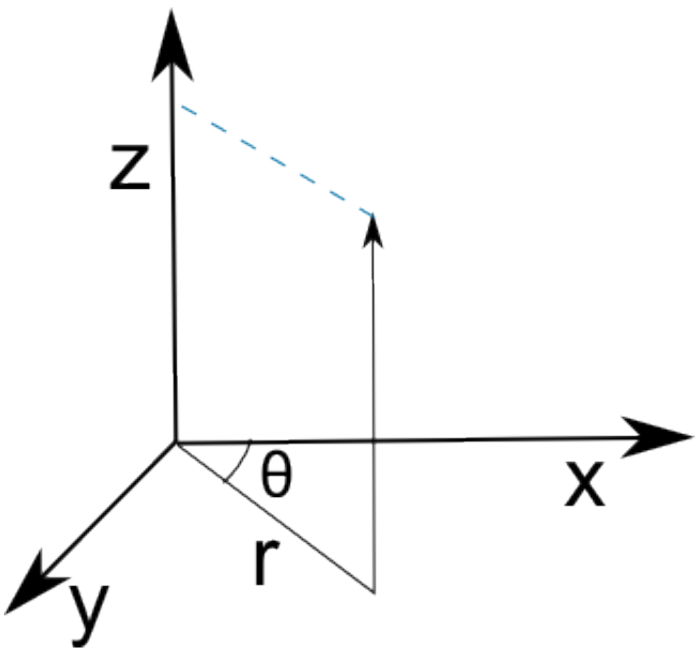
How can you define space, without using self-referential terms? The only real way is using mathematics. Space is defined by three numbers, possessed by every particle. These can be their x, y and z coordinates, or something different in a different coordinate frame, but there have to be three of them because we live in a three dimensional universe.

In addition to the coordinates, we need some way to measure how much one particle affects another. This is called distance (written as s). In Cartesian (x, y z) coordinates, the distance between two objects is given by Pythagorus' Theorem:

$$\delta s^2 = \delta x^2 + \delta y^2 + \delta z^2$$

where δs is a small distance element coming from small changes δx, δy and δz in the three coordinates.

We could also use cylindrical polar coordinates r, θ and z, as shown below.



In this case, the distance element is given by

$$\delta s^2 = \delta r^2 + r^2 \delta\theta^2 + \delta z^2$$

.

This equation relating distance to the three coordinates is called the metric. Einstein's brilliant idea was to modify it. This will make objects move in strange ways. If, for example, you leave out the r^2 in the cylindrical polar version of the metric, objects will move in circles rather than straight lines. The way you work out motion in a strange metric is to imagine a wavefront and see how far each edge of it moves. Where these waves add up in phase is where something will go.

GENERAL RELATIVITY

This is Einstein's theory of general relativity. There is no such thing as gravity. Instead, matter changes the metric of space around it. This change in the metric causes spacecraft to go in circles (orbits) and it causes objects to fall. When you drop something, it doesn't fall because a force is applied to it. It falls because accelerating downwards is its natural motion in the curved space-time of the Earth.

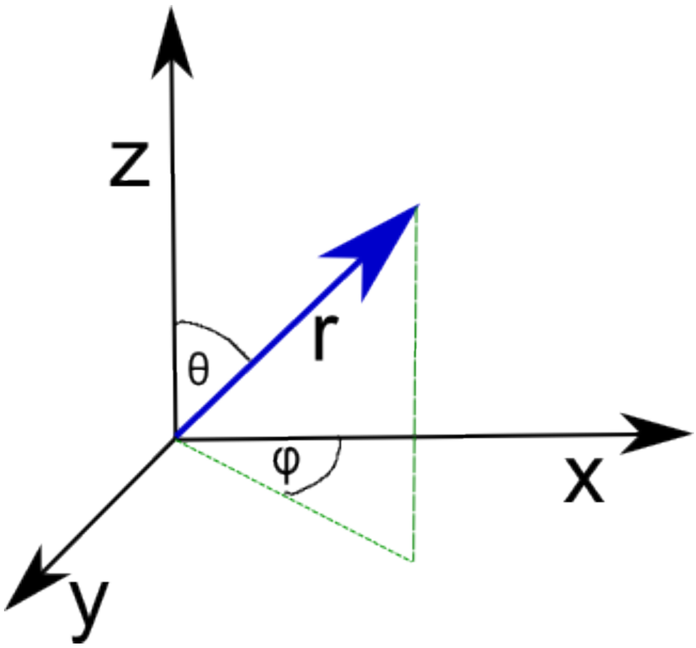
If something doesn't fall (for example if it is resting on a table), it is being accelerated away from its natural motion by the compression of the tabletop beneath it. This force from the table is the only real force present. There is no such thing as a gravitational force. Massive things are heavier because the table needs to apply a bigger force to accelerate them away from their natural motion.

Thus gravity turns out to be a ficticious force like centrifugal force. We feel a centrifugal force when a car goes around a corner, but in reality our body is just trying to follow its natural motion (a straight line) and the car has to push us sideways. Similarly we feel gravity pulling us downwards, but all that's really happening is the ground pushing us upwards against our natural motion.

This natural motion is called the Geodesic.

Lesson 2: Dynamics and Geometry of our Universe

THE ROBERTSON-WALKER METRIC



In spherical polar coordinates, the metric of normal space is

$$\delta s^2 = \delta r^2 + r^2 \delta\theta^2 + r^2 \sin^2(\theta)\, \delta\phi^2$$

It turns out that if we make the assumption (once again) that the universe is uniform and isotropic, there is only one possible metric, the Robertson-Walker metric, which is a slightly modified version of the conventional spherical polars metric:

$$\delta s^2 = a(t)\left(\frac{\delta r^2}{1 - kr^2} + r^2 \delta\theta^2 + r^2 \sin^2(\theta)\, \delta\phi^2\right)$$

There are two differences here - a(t) which can make the whole universe expand or shrink, and k, which changes the geometry of space.

We can work out how k changes the value of π in our universe. π is the circumference of a circle divided by its diameter. We can work out both the radius and the diameter in a universe with a Robertson-Walker metric by adding up all the little length elements δs around the circumference and across the diameter of a circle. If you do this, you find that the measured value of π is given by:

$$\pi = \pi_0 \frac{r\sqrt{k}}{\arcsin{(r\sqrt{k})}}$$

if $k > 0$ and

$$\pi = \pi_0 \frac{r\sqrt{k}}{\text{arcsinh}\ (r\sqrt{k})}$$

if $k < 0$. π_0 is the value of π in a geometrically flat universe, a universe with $k = 0$ (i.e. 3.141592…).

How do these solutions behave? Regardless of the value of k, both give the normal value of $(pi$ when the radius of a circle is much smaller than the radius of curvature of the universe. But as the circles become larger, if $k > 0$, the value of π drops, while if $k < 0$, the value of π rises. So π is not a constant, on large enough scales, unless k=0.

These different geometries can perhaps be best understood through a two-dimensional analogy. A universe with k>0 is analogous to a two-dimensional life-form on the surface of a sphere. If you go far enough in any direction, you will end up back where you started. The universe is finite, but has no edge. π is small on large scales, parallel lines will ultimately meet, and the interior angles of a triangle add up to more than 180 degrees.

A universe with k<0, on the other hand, is analogous to a saddle-shape. The universe is infinite, π is large on large scales, parallel lines diverge, and the interior angles of a triangle add up to less than 180 degrees.

THE FRIEDMANN EQUATION

How about a(t)? To work out how that behaves, we use the conservation of energy. We balance the change in potential and kinetic energy to derive the Friedmann equation:

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{kc^2}{a_0^2}$$

where ρ is the average density of the universe at that moment, and $\dot{a}$ is the rate of change of a. The current rate of the expansion H_0 (Hubble's constant) is given by

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

THE CRITICAL DENSITY

Using the Friedmann equation, and substituting in H_0, we see that to have a geometrically flat universe, we need the density today to be

$$\rho_c = \frac{3H_0^2}{8\pi G} \sim 9 \times 10^{-27}\,\mathrm{kg\,m^{-3}}$$

This is known as the critical density. If the density is larger than this, $k > 0$ and we live in a "spherical", finite universe, while if the density is smaller than this critical value, $k < 0$ and we live in a saddle-shaped open universe. Because this critical density is so important, the average density of various components of our universe is often quoted as the ratio Ω to this critical density:

$$\Omega = \frac{\rho}{\rho_c}$$

DENSITY EVOLUTION

To solve the Friedmann equation, we need to know how the density of the universe changes as the universe expands or contracts. For matter, the density is proportional to $1/a^3$, because the amount of matter is constant but the volume increases. For radiation (or matter moving very close to the speed of light), the energy of each wave/particle drops due to the expansion of space (and hence increase in wavelength), which means that density is instead proportional to $1/a^4$.

This means that the density of radiation drops faster than that of matter. Our universe is currently matter dominated (i.e. the density of matter is much greater than that of radiation), but as we go further back in time, radiation becomes more and more important, and indeed dominated the very early universe.

SOLVING THE FRIEDMANN EQUATION

The Friedmann equation is a differential equation (i.e. one including derivatives). In general it cannot be analytically solved, but has to be numerically approximated. It can however, be solved exactly for the case where k=0. This solution can be found in the appendix below. For a purely matter dominated universe,

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

where t_0 is the time today. For a radiation-dominated universe,

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

In general, however, you will need to solve the equation numerically. You can rearrange the Friedmann equation to give the rate of change of a at any given moment:

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

Starting from today (when a=1 by definition), you can then step forwards or backwards in tiny time steps (Δt) and estimate the value of a a short time in the future or past, using the equation

$$a(t + \Delta t) \simeq a(t) + \dot{a}\Delta t$$

(which just comes from the definition of differentiation - this is known as Euler's method). A python program that does this is available here.

We find that if k=0, the universe always expands but the rate of expansion always decreases. If k>0, the expansion of the universe will eventually stop, and the universe will start shrinking, until everything comes together in an almighty "Big Crunch" or "Gnab Gib" (the latter being "Big Bang" spelled backwards). If k<0, the universe will expand forever, with its expansion rate asymptoting to a positive constant.

CONCLUSIONS

So - we have three classes of cosmological model.

k>0, $\Omega > 1$, a finite, spherical universe that will eventually collapse. In this universe, π is small on large scales and parallel lines will eventually meet.

k=0, $\Omega = 1$, an infinite flat universe with normal geometry, that will expand forever but at a perpetually decreasing rate.

k<0, $\Omega < 1$, an infinite saddle-shaped universe that will expand forever. In this universe, π is larger on large scales and parallel lines diverge.

Which (if any) of these models is true? That is a question to be answered experimentally.

APPENDIX: EXACT SOLUTIONS TO THE FRIEDMANN EQUATION

The Friedmann eqution can be solved exactly if k=0 and we assume a purely radiation or matter dominated universe. Consider the case of a matter dominated universe. The density is thus

$$\rho = \frac{\rho_0}{a^3}$$

where ρ_0 is the density now and we've chosen to define a(t) = 1 right now ($t = t_0$). If k=0, the Friedmann equation is thus:

$$\dot{a}^2 = \frac{8\pi G\rho_0}{3}\frac{1}{a}$$

The easiest way to solve this is to guess that the answer might be in the form of a power-law: i.e. that $a \propto t^q$, where q is some unknown index. Substitute this into the above equation, and you will find that the left-hand side depends on t^{2q-2} while the right-hand side depends on t^{-q}. For a valid solution, these two must behave the same way, so $2q - 2 = -q$, and hence $q = 2/3$.

Using exactly the same trick, you can find that for a radiation-dominated universe, $q = 1/2$.

Lesson 3: Inflation

THREE PUZZLES

The theory of cosmic inflation gives possible answers to three puzzles:

Why is the universe so uniform on large scales? Normally different objects are the same if they've been in contact at some time. But the most distant regions we can see in opposite directions across the universe have never been in contact with each other in a standard cosmology.
Why is the geometry of the universe so close to being flat? If the universe was even slightly curved one way or another early on, this initial curvature would have grown exponentially. Thus to be within even an order of magnitude of having a flat universe now implies that early on the universe must have been flat to spectacular accuracy.
While the universe is pretty uniform on large scales, there are lumps at a whole range of scales which have turned over time into stars, galaxies, clusters and super-clusters of galaxies. Where did these fluctuations come from?

SIMPLICITY AND SYMMETRY

The guiding principles of much of particle physics are simplicity and symmetry.

Simplicity

We currently know of four forces of nature:

1. Electromagnetism
2. The Strong Force
3. The Weak Force
4. Gravity

This is not simple - so particle physicists have been trying to find a way to unify these four forces into one "unified" force.

Symmetry

In physics, symmetry has a slightly more general meaning than in every-day life. If you have a transformation you can apply to something, and this transformation produces an unchanged result, then we say there is a symmetry there.

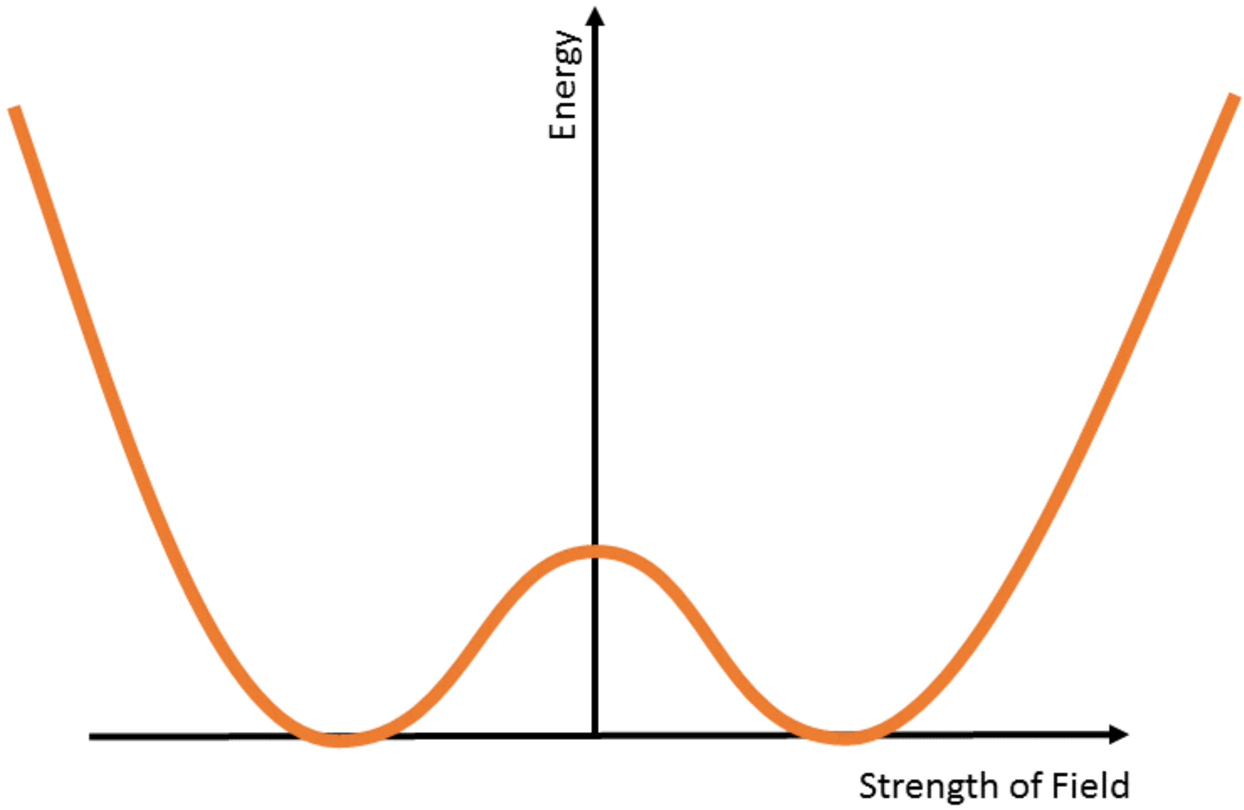
For example, if you flip something over and the end result looks the same, then there is a flipping symmetry. If you rotate something and it looks the same, there is a rotational symmetry. If you move something in space and it still acts the same, there is a translational space symmetry - it turns out that you can deduce Newton's laws of motion from this symmetry. And if something does not change at different times, there is a time symmetry, from which you can deduce the law of conservation of energy.

Spontaneous Symmetry Breaking

There is a contradiction between the principles of symmetry and simplicity. If all the forces of nature are really one, why do they act differently? A symmetrical principle of nature should not make a force behave in different ways, surely?

One possible resolution is the idea of spontaneous symmetry breaking. This allows symmetrical laws of physics to produce non-symmetrical results. It was first developed to explain how liquids can cool to form crystals and in the process pick definite directions in which to align their atoms.

The basic idea is that space is filled with a field (a Gauge field). This field tells particles how to behave. Depending on the value of this field at a given location in space, the different forces of nature will behave in different ways. Crucially, if you plot the energy of this field against its value, you get a “Mexican Hat” diagram:



When the universe is young (and hence energy levels are high) the field will jitter over a whole range of values, but with an average of zero. But as the energy level drops, and touches the bump in the middle of the hat, the field can no longer keep a value of zero but must chose to go either positive or negative. The “hat” is symmetrical, but the outcome is not.

FALSE VACUUM AND INFLATION

If the field just rapidly picked a positive or negative value, there would be no inflation. But if the top of the hat is flat enough, the universe can get stuck on top of the hat, even as the overall energy level of the universe drops down to the lowest points of the hat. This is known as a false vacuum state - a state with dramatically more energy than the energy as a whole - one cubic millimetre of false vacuum has a million times more energy that the entire observable universe!

So while it is in this state, every part of the universe has a constant and very large density. From the Friedmann equation, we can see that in this case the rate of expansion of the universe $\dot{a}$ is proportional to the scale factor a. Thus the universe will grow very rapidly, which will make the expansion rate even greater, so it will expand even more, making the size and hence the expansion rate greater still. Exponential growth.

This era of exponential growth means that regions on the far side of the universe today can actually have been in contact before the era of inflation, which naturally explains how their properties can be so similar. By stretching the whole universe enormously, the original curvature would be almost perfectly flattened out, explaining why the geometry of space is so close to flat. And the expansion amplified quantum-mechanical vacuum fluctuations, turning them into the primordial seeds of stars, galaxies and clusters today. This theory correctly predicts that there should be roughly the same amplitude of fluctuations on all scales, though it cannot predict the amplitude of the fluctuations.

THE END OF INFLATION

How does inflation end? Well, eventually, the universe will roll one way or another down the peak in the mexican hat, causing the forces of nature to split from each other, and the exponential growth to end.

This may not happen everywhere at once - the nudge to start rolling down the hat may come from random quantum mechanical fluctuations which may happen in some places earlier than in others. Once some part of the universe has undergone this phase transformation, split the laws of physics and stopped growing exponentally, it will look much like our own universe did when it was very young. This region of “normal” space-time will expand at the speed of light. Other regions of “normal” space-time may also exist, also expanding at the speed of light, and perhaps with the forces of nature split in a different way (depending on which side of the mexican hat the part of the universe rolled down).

It is unlikely that these regions of “normal” space-time will meet, because the parts of the universe between them are still growing exponentially. Thus inflation may never stop - while parts of the universe are constantly stopping inflating, this loss is easily made up by the continued exponential growth of the remainder.

INTERVIEW WITH LAWRENCE KRAUSS

How secure is the evidence for inflation? It predicts that the fluctuations in the universe (seen most easily in the microwave background) are adiabatic and gaussian, which is observed. Inflation theories, however, have many free parameters, so the theory could probably have been fit to different fluctuation patterns, so this is perhaps not a very strong prediction. A more compelling one is that inflation should leave a very particular pattern in the polarisation of the microwave background. In 2014, the BICEP2 experiment claimed to have detected this, though at the time of writing, it appears that they may have been fooled by foreground radiation from dust grains in our own galaxy.

The basic idea of gauge fields and spontaneous symmetry breaking is very well motivated in particle physics, and indeed explains the unification of electromagnetism and the weak force (carried by the W and Z bosons, which have been well studied). It also predicted the recently discovered Higgs boson. The predicted energy at which inflation should take place is a good match to the observations.

On the other hand, the Higgs field unexpectedly seems to carry no energy, and hence couldn’t drive exponential growth. The splitting of the electroweak force happened too rapidly to drive anything like inflation (i.e. the universe did not get stuck in a false vacuum state).

If you extrapolate the forces of nature as they are today, there is no energy at which they all unify, unless you make modifications. These modifications (super-symmetry) make predictions which are being tested at the Large Hadron Collider (LHC). So far, no signs have been seen, and indeed most (but not all) possibilities have been ruled out. When the LHC re-opens, if it still fails to find any signs of super-symmetry, the theory will be in trouble.

Another problem is that to get inflation of the correct duration you need to fine-tune the shape of your mexican hat. There is no good theoretical reason for this - you just have to fudge it to fit the data.

Lesson 4: Observational Cosmology

MEASURING THE AVERAGE DENSITY OF THE UNIVERSE

How can we work out observationally whether the theories we've discussed are correct, and if so what type of universe we live in?

One possible approach would be to measure the average density of the universe. As we saw earlier, there is a critical density: if the universe is more dense than this it has a closed geometry and will re-collapse, while if the universe is less dense, it will have an open geometry and will expand forever.

Unfortunately, measuring the density is hard because most of the mass in the universe is in the form of dark matter, which by definition is invisible. We know that galaxies contain large amounts of dark matter because they spin too fast. In galaxy clusters, there is yet more dark matter, as we can see from gravitational lensing and from the pressure balance of their x-ray emitting gas.

On larger scales yet, we can use "peculiar motions" to measure the amount of mass.

PECULIAR MOTIONS

The observed motion of a galaxy is the sum of two components - the motion due to the expansion of space ("Hubble Flow") plus the "peculiar motion", which is due to the gravitational pull of nearby galaxies and clusters.

Within our local group of galaxies, peculiar motions are larger than the effect of the expansion of space - indeed, the Andromeda galaxy will eventually collide with the Milky Way. On scales of ten mega-parsecs, the expansion of space is larger, but peculiar motions are still a significant contributer. For example, the Virgo Cluster would be moving away from us at 1400km/s if there were no peculiar motions. Its gravity is, however, pulling us in at around 300 km/s, which reduces its apparent redshift to 1100 km/s. To see the uncontaminated expansion of space, you need to go out to scales of around a hundred mega-parsecs or larger.

These peculiar motions can be used to measure the density of the universe on really large scales. Giant redshift surveys, such as the 2dF Galaxy Redshift Survey, look at the velocity and position differences between galaxies. If galaxies are falling in towards other galaxies, the redshift extent of structure is compressed compared to the positional extent (because galaxies on the near side of a structure are being pulled away from us, while those on the far side are being pulled towards us).

Using observations of this infall, we find that the density of the universe is around 30% of the critical value.

MEASURING THE GEOMETRY OF THE UNIVERSE

In principle we could measure the geometry of the universe by mapping out a really big triangle and measuring the internal angles. Unfortunately this is not possible in practice - the triangle would need to be billions of light-years across.

There is, however, another possibility. If galaxies are uniformly distributed in space, then the number you see out to a given distance D will be proportional to the volume within a sphere of radius D. But if space is curved, that will affect the value of π and hence the volume of this sphere. As we look further out, we would thus see the number of galaxies increase faster or slower than one would expect in a flat universe.

It is hard to measure the distance to vast numbers of galaxies, but it is relatively easy to measure their flux. If space is flat and they are uniformly distributed, then we showed in course 1 that the number n of galaxies brighter than some flux limit f is

$$n \propto f^{-3/2}$$

If space has a different geometry, the number would increase more or less steeply with decreasing flux. This could potentially be observable.

One complication is that distance is not well defined in a curved expanding universe. Do we mean the distance now, or the distance when the light set out? Or the distance the light actually travelled? We can't assume that flux follows the inverse square law, because the inverse square law was derived by assuming the photons were spread uniformly over a spherical surface of area $4\pi D^2$ which will not be true if space is curved. We've also ignored the redshifting of the photon energies and arrival rates. Luckily all these effects can be computed for a given cosmology. Traditionally, we retain the inverse square law, but define a "luminosity distance" d_L which factors in these effects and allows the inverse square law to still work - i.e.

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

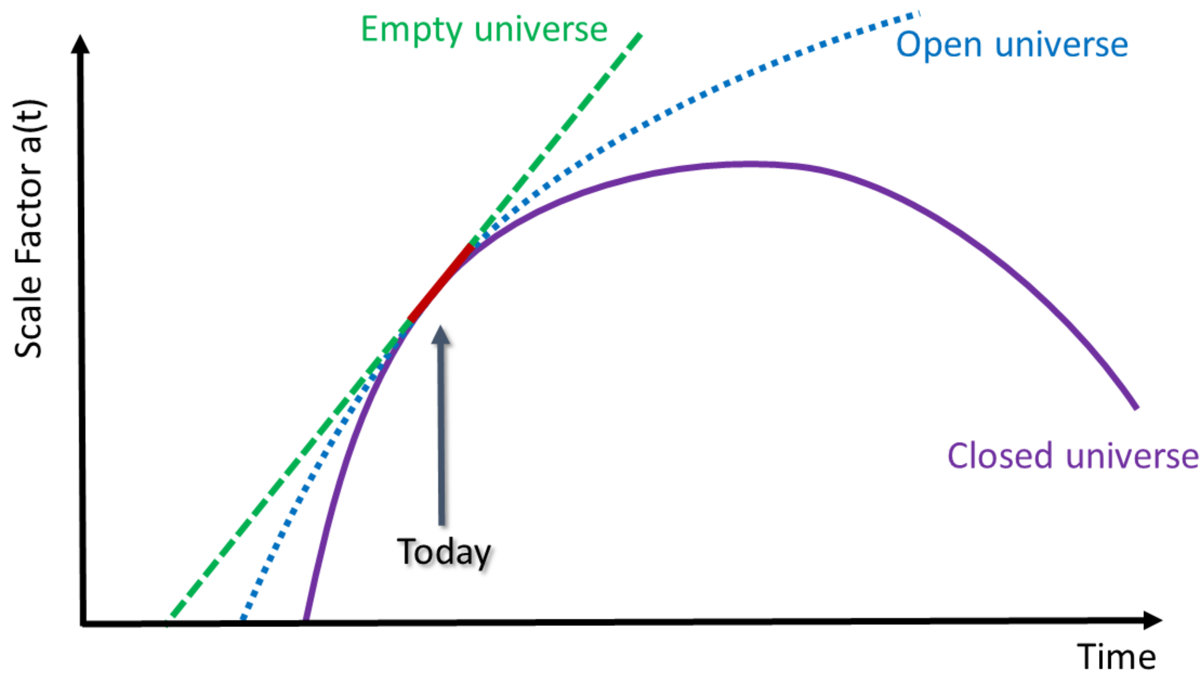
Calculating luminosity distances is beyond the scope of this course, but there are online calculators that will work it out for you for any given assumed cosmology.

When you do this, you find that there are relatively far too many faint galaxies. Unfortunately, this is probably caused by evolution in galaxy properties rather than the geometry of the universe. When the universe was young, galaxies were forming lots of stars, which meant there were many massive bright young stars, which made the galaxies more luminous than they are now. And more luminous galaxies are easier to see.

So while in principle, we could count galaxies and work out the geometry of the universe, in practice we could only do this if we correct for the evolution of galaxies. And we do not understand galaxy evolution anything like well enough to make this correction.

SCALE FACTOR EVOLUTION

Another possible way to determine what sort of universe we live in would be to measure how the scale factor of the universe a(t) varies.



We can only observe scale factors in the past, but if we could measure those, they would tell us which of these curves is the best fit.

Redshift tells us how much the universe has expanded while the light has been flying - for example a redshift of 0.1 tells us that space has expanded by 10% while the light has been travelling to us. In general, the scale factor when the light was emitted is given by

$$a(t) = \frac{1}{1+z}$$

So we can measure the vertical (y) axis of the above graph easily by measuring redshifts. Unfortunately, the horizontal (x) axis, time, is much harder. To get this, we need to measure the distance to a distant object. This then tells us how long the light has been flying before reaching us, and hence the time of emission. But measuring distances is hard.

THE DISTANCE LADDER

There are lots of ways to measure distances in astronomy, but most only work over a limited range of distances. So typically we have to use a whole series of different measures, starting with close-in measures and using them to calibrate the further measures, which in turn calibrate measures that work at even greater distances.

Rung 1: Parallax

The simplest method is parallax. As the Earth moves around the Sun, it causes our viewing angle to slightly shift, making nearby stars appear to move relative to far-distant background stars. The parallax angle is defined as the difference in apparent position you would get if the Earth moved by one astronomical unit (in reality the Earth moves by two astronomical units as it goes around the Sun, so the observed angular shift will be twice the parallax angle). If the parallax angle of a star is one arcsecond, it lies at a distance of one parsec (by definition).

Unfortunately space is very big - there are no stars within a parsec of the Earth. So parallax angles are typically very small and hard to measure. Even the best ground-based telescopes can only measure parallaxes to a handful of nearby stars. Space missions like Hipparchos extended it out a little further, crucially including the nearest star cluster, the Hyades.

Rung 2: Main Sequence Fitting

The second rung of the distance ladder is main-sequence fitting. Most stars are burning hydrogen to form helium, and if you plot their luminosity against their colour (a so-called Hertsprung-Russell diagram), you find that they lie in a line called the “Main Sequence”. More massive stars are luminous and hot, and hence appear blue. Less massive stars are faint and red. Using parallax distances to the Hyades Cluster we can convert the measured fluxes into luminosities and hence fit the main sequence and thus determine the luminosity of a star of a main-sequence star of any particular colour.

We then go out to the nearest easily observable galaxy, the Large Magellanic Cloud (LMC). We measure the fluxes and colours of lots of stars in the LMC, and identify the main sequence. We then look at how much fainter stars of a given colour are compared to similar stars in the Hyades.

Consider two stars of the same luminosity L. Their observed fluxes are given by the inverse square law

$$f = \frac{L}{4\pi D^2}$$

Take the ratio of the fluxes of the two stars. Rearranging, you find that

$$\frac{D_2}{D_1} = \sqrt{\frac{f_1}{f_2}}$$

Using this, you can estimate the distance to the LMC. Unfortunately, you cannot get distances to any more distant galaxies, and main-sequence stars are too faint and too crowded to measure individual brightnesses clearly even in the Andromeda galaxy.

One problem - the stars in the Hyades have a different chemical composition to those in the LMC. In particular, they have more “metals”. To an astronomer, a metal is any element heavier than hydrogen and helium! It could well be that this composition difference affects the luminosities and colours of main-sequence stars. In addition, the luminosity of a star on the main sequence depends somewhat on how old it is - the Hyades is much younger than most LMC stars. It is possible to correct for these effects, but it is unclear how accurate these corrections are.

Rung 3: Cepheids

To go beyond the LMC, we need something brighter than main sequence stars, so they can be seen out to greater distances. Cepheid variables are pulsing giant stars (and hence nice and luminous). They have a layer of doubly ionised Helium within them that traps heat inside. This causes the stars to expand as the heat builds up. As they expand, they cool down, until eventually the Helium ceases to be doubly ionised. At this point, the built up heat can escape and the star shrinks. Eventually it gets hot enough for the helium to become ionised once more, and the whole process repeats.

Crucially, the luminosity of cepheid variables correlates with their pulsation period (the Leavitt law). This calibration is measured in the Large Magellanic Cloud (LMC). You can then look for pulsing stars in more distant galaxies, and use the ratio of fluxes to work out the ratio of distances. With ground-based telescopes this method easily gets you out to the Andromeda galaxy, and perhaps ten times further. But that’s still only far enough to get a handful of galaxies (as we will see later, space-based measurements extend this considerably further).

Unfortunately, many Cepheid variables are partially obscured by dust clouds, which can make them appear abnormally faint. Infra-red observations can reduce this effect, but are hard. It may also be that the Leavitt law depends on the chemical composition (amount of “metals”), which would be a problem as the law is calibrated in the LMC (low metals) but mostly applied to galaxies like Andromeda (high metals). The metal dependence is controversial.

Rung 4: Tully-Fisher

Using galaxies in clusters, it has been found that for certain types of spiral galaxy, their luminosity correlates with their rotation speed, in the sense that brighter galaxies spin faster. This tells us that the amount of dark matter in a galaxy (which determines the rotation speed) correlates with the number of stars (which determines the luminosity).

We can use the doppler effect to measure the rotation speed, and hence the luminosity. The flux then gives us the distance, via the inverse square law. Unfortunately, there are very few spiral galaxies close enough to measure their distances using Cepheid variables, so the Tully-Fisher relation is not well calibrated. It is also not very accurate - there is an 18% scatter in distances measured using it.

Rung 5: Type Ia Supernovae

Type Ia supernovae are carbon-oxygen white dwarf stars that become dense enough to ignite fusion and hence explode. They were covered in considerable detail in Course 3 of this series. They are all of roughly the same luminosity. It turns out that their exact peak luminosity correlates with how long they stay bright, and using this relationship they can be used to produce quite accurate distances. And they are so bright that they can be seen out to enormous distances.

Unfortunately, few have occurred close enough that we have some other way to measure their distance, so while we know that they all have the same peak luminosity, we have trouble knowing what this peak luminosity is.

CONTROVERSY

So measuring distances is hard. We have five rungs in our distance ladder, and each of them has problems. The problems all multiply together as you move up the ladder, so that the end result is extremely uncertain. For decades, cosmologists fought over this, often getting widely inconsistent answers.

Lesson 5: New Techniques

DUST

Interstellar dust is pretty but it is a major problem when trying to measure distances in space, because it makes objects behind it appear fainter. Dust consists of tiny grains, typically created in red giant winds. Dust typically absorbs or scatters blue light more strongly than red light, so it can be possible to determine whether dust is present by seeing if your target appears redder than usual.

If all dust were the same, we could use the reddening to estimate how much light has been lost. Unfortunately all dust is not the same, and depending on the exact composition, the same reddening can be associated with the loss of different amounts of light. Also, you can only measure the reddening if you know what colour your target was before the dust absorbed some of its light (which is hard for Cepheids which vary in their true colours).

One way to minimise this problem is to observe at infra-red wavelengths. Unfortunately, infra-red astronomy is hard. Silicon based detectors do not work past 1.1 microns, so you have to use much more expensive exotic semiconductors such as indium antimonide and mercury cadmium telluride. These detectors are typically built for military purposes, and so can fall foul of export regulations. Also, at infra-red wavelengths the telescope and the sky glow brightly, and diffraction blurs your images.

BIAS

Any distance measurement will have an uncertainty associated with it. Some of this is due to noise in the image, while some comes from the intrinsic diversity of the objects you are studying. This uncertainty means that some objects will appear brighter and others fainter than the average.

If you observe enough objects, you would normally expect to average out this uncertainty. The brighter than average objects would at least partially cancel out the fainter than average objects. And indeed this works for the brightest targets.

But as you study fainter (more distant) targets, things become harder. Any survey will have a detection threshold: objects fainter than this are probably not real, but just artefacts of noise in your detector. If you are looking at targets just a bit brighter than the threshold, those that are brighter than average due to noise or intrinsic variation will easily be seen. While those fainter than average will drop below the threshold and be missed. So when you calculate the average brightness, you will be biassed towards larger values.

The only way to avoid this is to use a distance measurement technique with the smallest possible object-to-object scatter, and to use monte-carlo simulations to try and estimate and correct for the effect.

RECENT IMPROVEMENTS: HUBBLE SPACE TELESCOPE

The Hubble Space Telescope (HST) was deliberately designed to observe cepheid variables in much more distant galaxies than is possible from ground-based telescopes. It can use cepheids to measure distances to galaxies out as far as 30 Mpc. This means you can measure the distances to a few galaxies that have hosted Type Ia supernovae, and also that you can better calibrate other distance indicators such as Tully-Fisher.

In addition, the fine guidance sensor on HST turns out to be a really precise way to measure parallaxes, and has measured parallax distances to a handful of cepheid variables in our own galaxy. These distances allow us to bypass the Large Magellanic Cloud (LMC) in building up our distance ladder, which is important because the LMC has a much lower abundance of heavy elements than most big galaxies, which may bias things.

In addition, infra-red cameras on the HST allow measurements of cepheids at infra-red wavelengths, where dust is less important.

The European Space Agency has launched the GAIA satellite which will measure extremely accurate parallaxes for stars out to kilo-parsec distances, which will greatly extend the first step in the distance ladder.

ECLIPSING BINARY DISTANCES

The distance to the LMC remains a big problem, but several new methods are improving our knowledge of this key distance. One is the use of eclipsing binaries. Microlensing survey telescopes find widely separated binaries with edge-on orbits, so that the stars pass in front of each other, causing dips in the brightness. By combining spectral doppler-effect measurements with brightness measurements, it is possible to measure how fast the binary stars are moving, and hence (from how long it takes them to cross) their radii.

Optical interferometers on Earth allows us to measure the surface brightnesses of nearby stars with the same colours as these eclipsing binaries. Surface brightness is brightness per unit observed solid angle, and is independent of distance. If you assume that the stars in the eclipsing binaries have the same surface brightness as those with the same colours in our galaxy, you can thus work out their luminosity and hence distance. This currently gives distances accurate to about 3%

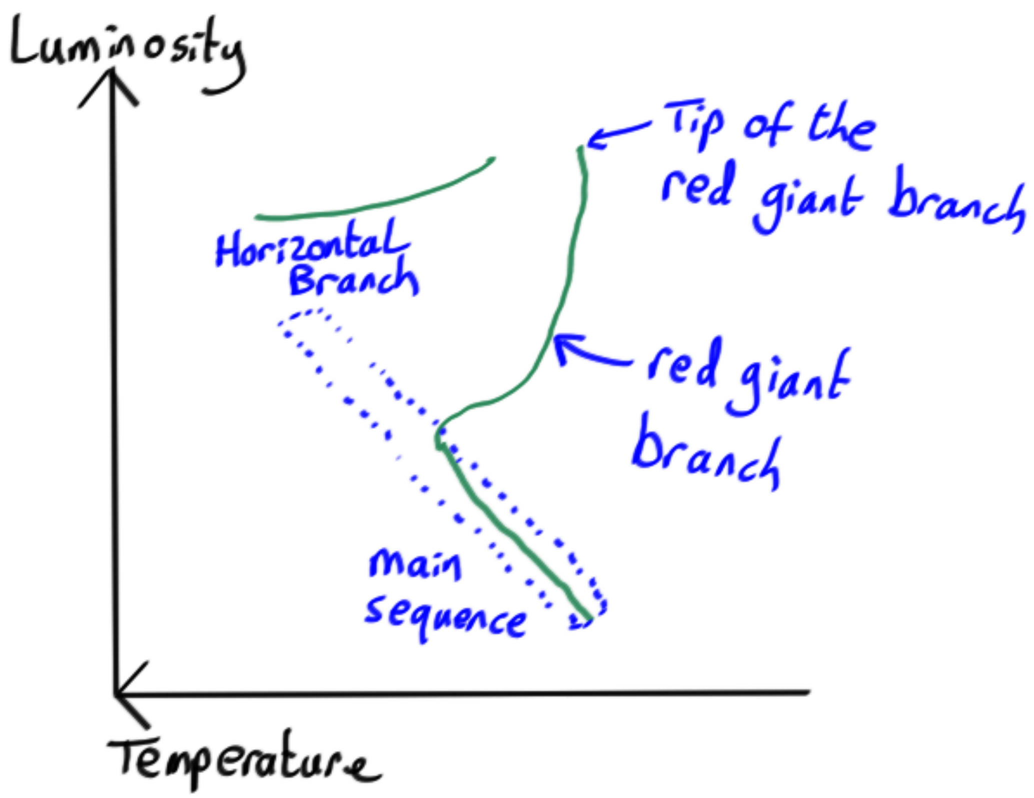
MEGA-MASER DISTANCES

Some active galactic nuclei (giant black holes in the middles of galaxies) are strong sources of maser emission. This maser emission comes from blobs of gas orbiting around the black hole in some sort of disk. Because the maser emission is so strong, it can be observed by radio telescopes using the technique of Very Long Baseline Interferometry (VLBI), in which telescopes all over the world combine their data to get extremely high resolution. This allows us to see the positions, transverse velocities and accelerations of individual blobs. By combining all this information, it is possible to derive the distance to the mega-masers using only geometry and the laws of gravitation.

To do this, we need to observe at least one gas blob off to the side of the disk, and measure how far it appears to be from the black hole (θ), and from the doppler effect, its velocity v. We also need to observe some blobs in front of the black hole, and measure their apparent transverse velocity $\dot{\phi}$ and (from changes in the doppler effect) their acceleration a. Combining these, we get a distance given by

$$D = \sqrt[3]{\frac{\theta v^2 a}{\dot{\phi}^4}}$$

TIP OF THE RED GIANT BRANCH DISTANCES



Stars start on the main sequence, burning hydrogen to form helium. Once they run out of hydrogen in their cores, the core becomes an inert blob of helium surrounded by a hydrogen burning shell, and the stars become red, very large and very bright. As the hydrogen burning continues, the stars become more and more luminous. But all the time, more and more helium is piling up in the inert core, until eventually it becomes so dense that helium abruptly stars to fuse in the core, and the whole star rearranges itself. The star then becomes much bluer and less luminous, moving across to the horizontal branch. This always happens at a very well defined luminosity, giving an abrupt upper limit to the luminosity of red giants. This can be used to measure distances.

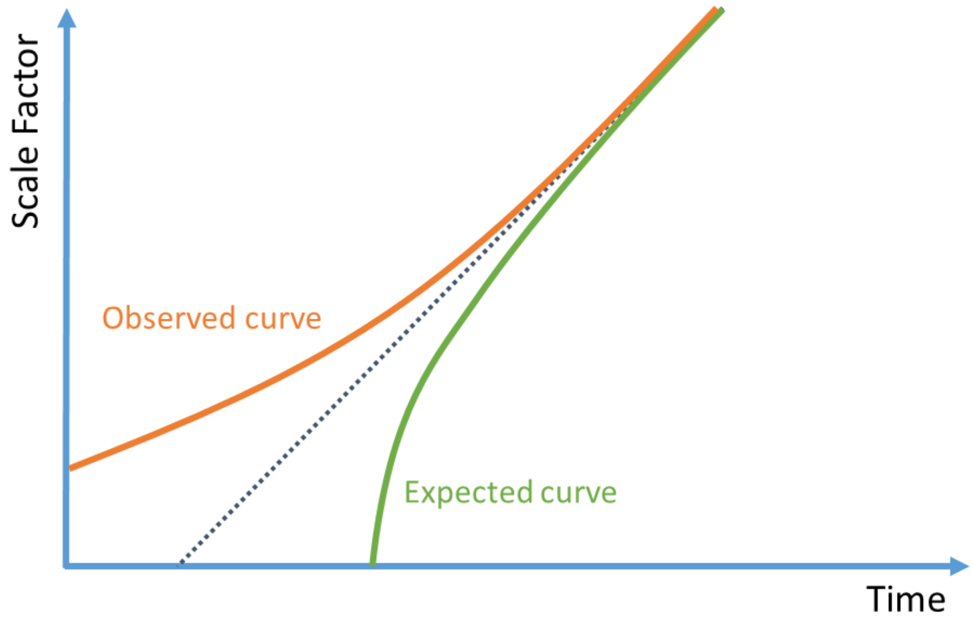
It works at roughly the same distances as cepheids, but can be used in older galaxies, and requires less observing time. It is a useful cross-check of cepheid distances.

Lesson 6: Dark Energy

DARK ENERGY

What was learnt when we applied all these new techniques? It turned out that Hubble's constant was around $70\mathrm{km\ s^{-1}Mpc^{-1}}$. If the universe was matter dominated, that gave an age which was far too young.

Looking further out, supernova measurements allowed the plot of scale factor against time to be explored:



The expansion rate did not slow down with time as expected - instead it is speeding up! None of the cosmological models we've talked about so far can do this.

Back when the Friedmann equation was first derived, Einstein came up with a variant:

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

He added the new term on the end (a constant called Lambda divided by 3) to avoid having an expanding or contracting universe. It was purely a "fudge factor" designed to keep the size of the universe steady. This extra term is sometimes called the "cosmological constant". When Edwin Hubble showed, a few years alter, that space was expanding, Einstein called this his "greatest mistake". But it is just what we need to make the expansion of the universe accelerate. It is equivalent to having a density of mass of

$$\Omega_\Lambda = \frac{\lambda}{3H^2}$$

where H is Hubble's constant.

When this term dominates the modified Friedmann equation, the rate of change of a is proportional to a, which means you get ever accelerating exponential growth (just like for inflation, only not currently as rapid).

What could cause this? It may be another Mexican hat style field. Or it may simply be that the ground state of a vacuum has a small energy, rather than an energy of zero.

LAWRENCE KRAUSS INTERVIEW: COINCIDENCE

It is very strange that the densities of dark matter and dark energy are so similar (within a factor of three or so of each other), and one stays constant as space expands while the other drops rapidly. Simple theoretical models predict dark energy that is 120 orders of magnitude larger than observed, so theorists had always assumed that some symmetry cancelled it out and gave zero. But how come it was mostly but not entirely cancelled out, and happens to have a value within one order of magnitude of Matter? It is a real problem.

One possibility is a theory like "Quintessence" in which the density of dark energy varies with time so as to keep it similar to that of dark matter. Lawrence considers this to be ugly and pooly motivated.

It can't easily be a modification of General Relativity as that would occur at a quite different scale. One possibility is that there are multiple universes with different values of dark energy, but those with high values cannot form galaxies and hence have no astronomers in them wondering about why dark energy is so strong.

GRAVITATIONAL LENSING

Gravitational lensing can be a problem for cosmological measurements, but is also an opportunity.

The problem comes from using distant objects (such as supernovae) as standard candles. If they are far enough away, there is a significant chance that they will either be amplified or de-amplified by gravitational lensing by galaxies and clusters along the line-of-sight. On average the amplification of some objects cancels out the de-amplification of others. But in practice, we see the amplified ones out to larger distances and hence detect more of them, giving us a systematic bias. At present this is not the largest bias, but as surveys get better, it will become more of a challenge.

One opportunity is to use gravitational lensing to get direct distances. If a quasar lies almost directly behind a much nearer foreground galaxy, you will sometimes get multiple images as the light is bent in different ways around the galaxy. The light moving along these different paths takes a slightly different time to reach the Earth. As quasars vary randomly in brightness all the time, it is possible to monitor the different images and work out the delay between them. Comparing this to lensing models gives a distance to the galaxy.

The main uncertainty here is how well the gravity of the foreground galaxy can be modelled. Current models are, however, becoming extraordinarily accurate, and are done using “blind analysis” techniques so the modellers cannot be biassed towards particular answers.

The second opportunity is “Weak Lensing”. Any distribution of mass will modify the apparent shapes of background galaxies by a small amount. Normally this is swamped by the random orientations of galaxies, but if you average over billions of background galaxies you can use this technique to measure the distribution of mass and how it evolves with cosmic time. This is already producing results, but proposed future space missions will turn this into a very powerful technique.

PRIMORDIAL NUCLEOSYNTHESIS

When the universe was between 1 and 3 minutes old, the density and temperature everywhere were suitable for nuclear fusion, and fusion did indeed take place. The main reaction pathway was for neutrons and protons to combine to form deuterium, which then by various pathways ends up as Helium 4. The element ratios produced by this process can be used as a cosmological constraint.

To do this properly requires a full nuclear physics computation, but you can roughly understand one of the key ratios as follows. Deuterium is easily converted into heavier nuclei, so its density tends to settle down at a particular low value: just low enough that deuterium nuclei don't have much chance of meeting any other deuterium nuclei and hence combining to form something else. This means that as the density of baryons increases (a baryon being any particle made up of normal matter such as neutrons and protons), the density of deuterium does not increase.

Thus the deuterium to hydrogen ratio decreases as the density of baryons increases. Measure it, and you would thus know the density of normal (baryonic) matter in the universe.

Unfortunately, deuterium is very easily destroyed in stars. So to measure the ratio of deuterium to hydrogen left over from the big bang, you need to find some gas that has not been processed in a star - perhaps the surface layers of a primordial star (but none have yet been discovered - see course 1) or gas in an interstellar gas cloud, seen in absorption against the light of a background quasar or gamma-ray burst. Even here the measurement is very difficult, as the absorption lines of deuterium lie extremely close to the vastly stronger lines of hydrogen. But as best we can determine, the ratio of deuterium to hydrogen is consistent with 3.5–5% of Ω_c being in the form of baryons.

Lesson 7: Acoustic Peaks

THE COSMIC MICROWAVE BACKGROUND

When the universe was very young (less than around 300,000 years old), it was opaque, because the baryons were ionised, which means they scattered light. Photons were trapped between the atoms, bouncing back and forth. The combination of photons and baryons is called a photon-baryon fluid.

Eventually, as space expanded, the density of the universe dropped low enough that the baryons could become electrically neutral, at which point the universe became transparent and the photons were released. The photons have been flying freely ever since, forming the cosmic microwave background (CMB) which we can detect today. This moment when everything became transparent is called “recombination” (which is curious terminology - electrons and protons did combine at this point, but not “recombine” as they’d never been combined before…).

At first glimpse, the microwave background looks uniform in all directions. More sensitive observations detect the dipole - the motion of the Earth with respect to the early universe, which doppler-shifts . Strong emission is also detected from our own galaxy, from interstellar dust and various other foregrounds. But if all this is subtracted off (using measurements at multiple frequencies to disentangle everything), we see very weak fluctuations in the cosmic microwave background. They are very small - only around one part in 10^5.

It is possible to analyse the typical sizes of these lumps. This is typically done by decomposing the sky into spherical harmonic functions, also known as multipoles. This shows that the dominant lumps have a scale of about one degree, but there are also a number of other scales showing excess lumpiness.

ACOUSTIC WAVES

Where did these characteristic sizes of the lumps come from? We start off with a complicated pattern of density fluctuations, left over from quantum mechanical vacuum fluctuations frozen at the end of inflation. This pattern is complicated, but because the fluctuations are so small, it turns out that we can decompose these fluctations in any way we like, work out how each of the decomposed elements evolves, and then simply add the results together. This is the “linear physics regime” and makes the problem enormously easier.

The simplest way to proceed is to break down the pattern of fluctuations into a set of sine waves of different amplitudes, phases and wavelengths. This decomposition is called the “Fourier Transform” and is very common in many areas of physics (for example, it’s what a spectrum analyser does). We can then ask how each of these different sine-wave patterns evolves.

The evolution of dark matter is easy - it is attracted towards the peaks of the density field, and so falls in, making the peaks get denser and the troughs less dense. Thus basically amplifying the sine-wave. The baryon-photon fluid is more complicated. Initially it too falls in, but as it becomes more and more compressed, the vast numbers of photons act like springs and store up the energy. Once it reaches a peak density, the photons then push the whole fluid back out again - a “bounce”. They bounce out until the fluid is concentrated between the dark matter peaks, and then bounce back in again, and so on. These are the acoustic waves.

Remember that the true matter distribution has been broken up into sine-waves of different wavelengths. Each sine wave is bouncing, but the longer ones will bounce more slowly. The waves move at the speed of sound (57% of the speed of light at this time), so if a wave is larger, it will take longer to bounce in and out. Thus you get a cacophony of different frequency acoustic waves all playing at the same time.

ACOUSTIC PEAKS

How does this produce the characteristic size of the peaks in the microwave background? At recombination, the universe becomes transparent, and all the bouncing stops. The photons fly freely to us, producing the microwave background we see today.

At the longest wavelengths, there isn’t time between the end of inflation and recombination for the baryon-photon fluid to fall in. So on really big scales, the only fluctuations you see are those left over from inflation.

On slightly smaller scales, there has just been enough time for the baryon-photon fluid to fall in and reach peak density, but not to bounce out again. This produces the peak in the fluctuation spectrum - the one-degree scale lumps.

On smaller scales still, the fluid has had time to fall in and bounce half-way out - so it's uniformly spread. There are thus very few fluctuations on this scale - it corresponds to the first trough in the fluctuation spectrum.

Smaller still, and the fluid piles up between the dark matter peaks, producing lots of lumps, and the second peak in the fluctuation spectrum. And so on, down to smaller and smaller scales.

On really small scales, the spectrum is affected by the fact that recombination did not happen instantaneously. There was a period when the fog that filled the universe was beginning to clear, and photons could move further, but the universe was not totally transparent. During this time, photons could move from dense regions to less dense regions and hence partially blur out the patterns. This effect is called "Silk Damping" and suppresses the smallest scale fluctuations (the "damping tail").

MEASURING GEOMETRY

The acoustic peaks allow a very precise measurement of the curvature of spacetime. Because the physics of the peaks is simple and linear (sound waves in an almost perfectly uniform fluid), it is possible to calculate very straightforwardly the physical length of the peaks. It is basically just the horizon length - speed of sound times age of the universe (but factoring in the expansion of space).

So - we know how big these peaks are (physical length L). And we can also measure what angle θ they subtend as viewed from the Earth. If space had a flat geometry (and ignoring the expansion of space), $\theta = L/D$, where D is the distance to the microwave background surface of last scattering. If, however, space has a closed (spherical) geometry, the angle will be larger, while if space has an open (saddle-shaped) geometry, the angle will be smaller. Thus this angle gives us a way to measure the geometry of the universe.

What we find is that space is very close to being geometrically flat.
$\Omega_{\text{matter}} + \Omega_{\Lambda} = 1.0 \pm 0.01.$

This is exactly what inflation predicts.

MEASURING THE MATTER DENSITY

The odd-numbered acoustic peaks (1, 3, 5 …) are due to the baryon-photon fluid falling into dark matter concentrations, while the even-numbered peaks are due to the baryon-photon fluid bouncing back to regions between the peaks. If the universe were mostly made up of photons, these two would have equal strength. If we have more baryons in the fluid, however, their weight will pull them in towards the dark matter and hence make the odd-numbered peaks stronger, relative to the even-numbered ones. This odd-even pattern can be used to place limits on the amount of matter (both baryonic and dark matter) in the universe. From this analysis, we learn that

$$\Omega_{\text{allmatter}} = 0.315 \pm 0.017$$

and

$$\Omega_{\text{Baryons}} = 0.048 \pm 0.0005$$

which is pleasantly consistent with what we deduced from large scale peculiar motions and from primordial nucleosynthesis.

Using simple linear physics, and assuming a flat universe with this matter density, it is possible to predict the whole shape of the fluctuation spectrum with exquisite precision. Any rival cosmological theory would have to do equally well, which is a daunting challenge. The only discrepancy is on very large scales - the measured power in multipoles in the 20s does seem to be a little smaller than predicted. This could just be chance - and there is no way to test it other than find another universe to map.

BARYON ACOUSTIC OSCILLATIONS

At recombination, the photons were set free, and we see their distribution at that time in the microwave background. But with the freeing of the photons, the springs that caused the baryons to bounce were removed, so the baryons were frozen in position. As time continued, the slight density inhomogeneities this produced gravitationally collapsed and became clusters and super-clusters of galaxies. But the imprint of the acoustic peaks should still be visible in the distribution of galaxies today.

Very big galaxy redshift surveys have indeed found that galaxies tend to have excess clustering on scales of around 150 Mpc. This is the first acoustic peak.

This provides us with a new cosmological test. The length at which this excess clustering is found is a “standard ruler” - a fixed length that doesn’t change (apart from due to the overall expansion of space). If you know how long something is and you observe how long it appears to be, that gives you the distance. So this gives us a way to measure distances corresponding to different redshifts, which is quite independent of supernovae and the rest of the distance ladder techniques. Furthermore, because it depends on sizes rather than fluxes, it is unaffected by dust. And you can compare all the lengths you measure directly to the size of the acoustic peak in the microwave background, so you can probe much earlier times than any other method.

This is crucial if we want to see if the density of dark energy is changing with time. Remember that Einstein’s cosmological constant Λ doesn’t vary with time, but if dark energy really were a mexican-hat style field and the universe was rolling down the side of the hat, the density of dark energy could be decreasing with time. Other models are also possible. A possible change in the density of dark energy is parameterised by the equation of state parameter w, where its density ρ varies according to the equation:

$$\rho \propto a^{-3(1+w)}$$

Einstein’s cosmological constant corresponds to $w = -1$. At the moment, observations are consistent with this.

MEASURING COSMOLOGICAL PARAMETERS

Unfortunately, most cosmological observations do not straightfowardly tell you the parameters you are trying to measure (like Hubble's constant, the density of matter or the density of dark energy). Instead, they tell you that certain combinations of these parameters are possible while others are not. For example, microwave background measurements tell you that the geometry of space is flat, but don't tell you whether this is due to lots of dark energy and little matter, or vice versa.

Luckily, different measurements give different constraints, and so by combining (say) microwave background measurements with supernovae and distance ladder Hubble constant measurements, you can pin down the parameters far more accurately than any one method could do by itself.

From this process, we find a set of cosmological parameters that seems to fit all the current observations.

This is a universe which has an age of 13.798 ± 0.037 billion years, made up of 68.3% dark energy (with $w \sim -1$, 26.8% dark matter and 4.9% baryonic matter. We are not sure if the universe is open or closed, because it is right on the boundary line. But regardless, it will expand forever due to dark energy, unless dark energy decays away at some point in the future.

The value of Hubble's constant is a little more uncertain - it is somewhere in the range $H_0 = 67 - 76\mathrm{km\,s^{-1}\,Mpc^{-1}}$. The error bars do overlap on these different measurements, but it is still an uncomfortable situation and may be telling us that there is something we don't understand here.

How secure is this current consensus view? It seems pretty secure, and many different methods calculated by different groups of people seem to give the same answer. More than two separate and independent methods would have to be totally wrong to fool us. And they would have to go wrong in very specific ways. But the herd instinct of astronomers (like all other humans) is always a worry.

Lesson 8: Entropy

THE ARROW OF TIME

All the laws of physics we've discussed so far are time-reversible. They work just as well backwards as forwards. And yet most processes in our universe seem to have a very definite direction in time - they look stupid backwards. This is explained by the Second Law of Thermodynamics: entropy always increases. The simplest case is when you have a hot object in contact with a cold object. Heat always flows from the hot object to the cold one until they reach the same temperature. You never get the cold one cooling down and the hot one warming up, even though that would not violate energy conservation.

But the universe does seem to violate this. We go from the very uniform temperatures of the microwave background to the extreme temperature gradients today. And even on earth, we go from stagnant pools to highly sophisticated life-forms. What is going on here?

QUANTUM OSCILLATORS

Let us look in some detail at the simplest case - two objects with different temperatures. Temperature is a measure of how fast the atoms in an object are jiggling about. We will approximate a solid as a collection of atoms, each of which can move along three axes (x, y and z). In this approximation (due originally to Einstein) we will treat each atom as independent of the others, and the motion along each axis as independent of the other directions. Thanks to quantum mechanics, the motion of the atom in any given direction can only have certain discrete energy levels. Treating the motion of each atom along each axis as a single oscillator, each oscillator can have zero, one, two, three or more quanta of energy. But it cannot have a non-integer number of quanta.

COUNTING STATES

How many ways can you distribute four quanta between the three oscillators (x, y and z) of one atom? You can count them up. You could put all four quanta in x and none anywhere else. Or three in x, one in y. Or three in x, and one in z. Or two in x, two in y and none in z. And so on. In total there are 15 possibilities. We call each possible arrangement a microstate, and the total number of quanta (4) the macrostate. Our basic assumption is going to be that all microstates with a given macrostate are equally likely. This is because as atoms jiggle back and forth, quanta are transferred from one microstate to another, but due to conservation of energy, the macrostate cannot change.

Now, consider these same four quanta of energy, but now they are spread between two connected atoms. One atom could have all four, the second atom could have all four, or they could be distributed more evenly. If atom one has four quanta and atom two has none, then there are 15 microstates for atom one and only one for atom two. The total number of microstates for this configuration is thus 15 times 1 (15). If the quanta are more evenly distributed each atom has less than 15 microstates, but when you multiply the two numbers together you get larger total numbers. When you do the sums, it turns out that there are 126 states, and the largest number of states is for two quanta in each atom. Thus we are beginning to see how counting states and probabilities could lead to objects coming to have the same temperatures.

Counting states one by one rapidly becomes tedious as the number of quanta and atoms grows, so we need an equation. You can work out an equation by looking at the number of permutations and combinations of quanta and boundaries between different oscillators. If you have N oscillators and q quanta, the number of possible microstates Ω is given by:

$$\Omega = \frac{(q + N - 1)!}{q!\,(N - 1)!}$$

Thus is you have 100 quanta distributed between 100 atoms, $N = 300$, $q = 100$ and $\Omega = 1.7 \times 10^{96}$. If all the quanta were in one atom, $N = 3$ and $\Omega = 5.2 \times 10^{3}$. Thus the odds of one atom having all the energy are roughly one in 10^{93} - not zero, but unbelievably small.

If you have two blocks in contact, you can do the same calculation. You will find that the energy tends to even out so that both blocks have the same number of quanta per atom (i.e. the same temperature). It is possible for the temperature to be unbalanced, and if the number of atoms is tiny, small imbalances are in fact quite likely. But for realistic numbers of atoms ($\sim 10^{23}$), imbalances will be tiny and very improbable. It is still possible for all the energy to go into one block, but the odds against it are of the order one in $10^{23}!$ - so while possible, it is very unlikely to ever happen in the observable universe during the whole life of the universe. Note that you can estimate these very small numbers using Stirling's approximation.

ENTROPY

Entropy S is defined as

$$S = k_B \ln(\Omega)$$

where k_B is Boltzmann's constant and has a value of $1.38 \times 10^{-23} \mathrm{J\,K^{-1}}$.

Let us see how this concept can explain an example of an irreversable action. Such as dropping a ball on a table. After a bounce of two, it will come to a rest (having slightly warmed up the table as its kinetic energy was dissipated). But you never see a ball resting on a table spontaneously bounce a couple of times and then fling itself into the air, leaving a marginally cooler table behind. Why not?

When the ball hits the table, the table atoms below it will have a very specific pattern - those directly below the impact point all moving downwards, while those further afield are still moving in random directions. Once things settle down, everything is once again moving randomly. To fling a ball into the air, all the atoms directly below it would have to simultaneously have all their energy in vertical oscillators, and be moving upwards all at once. This is a perfectly possible microstate, but is vastly outnumbered by the number of microstates that do not fling the ball into the air. So while it is possible, the odds are very much against it.

People often invoke entropy as an explanation for chaos in everyday life. For example, does it explain why rooms tend to get messy? Consider all the objects in a room, and all the possible places they could be placed. Any possible arrangement is a microstate. And it is clearly true that microstates that would be considered tiny are an absolutely tiny fraction of the total number of possible microstates - there are far more ways to be messy than to be tidy. So that part of the argument fits. The other assumption we use, however, is that all mcirostates are equally likely, and that the system frequently moves between microstates. Not clear that this applies here.

ENTROPY AND THE UNIVERSE

How then can we explain the universe moving from a state of nearly uniform temperature (the cosmic microwave background) to its very uneven structure today? Nobody is quite sure, but the general consensus is that when gravity is involved, things are different. There have been various attempts to define a “gravitational entropy” (e.g. the Weyl curvature hypothesis). Scientists don’t really know how to do this, but the basic idea is that as the uniform gas gravitationally collapses to form stars and galaxies, the thermal entropy goes down but the gravitational entropy goes up by more.

How can we explain evolution - apparent order (us) coming from apparent chaos (some stagnant pool)? Life forms are capable of maintaining low internal entropy, but they do this by increasing the entropy of the rest of the universe (for example by converting low entropy food into high entropy faeces), so the net effect of life is actually to accelerate the increase in entropy of the universe.

The end of inflation left the universe is a very low entropy state (if you believe the various gravitational entropy conjectures). At time went on stars, planets and life-forms appeared, all taking advantage of this low entropy universe to stay organised, and by so doing, increasing the overall entropy of the universe. Eventually, everything will end up forming black holes, which must therefore be high entropy states. And they in turn will ultimately evaporate into diffuse radiation - the highest entropy state of all.

Lesson 9: The Origin and Fate of the Universe

CAN WE PROBE THE VERY BEGINNING OF OUR UNIVERSE?

Future observations are either going to confirm the model we already have, or break it. The latter would be far more exciting as it would give us a clue to something beyond out current understanding.

There are a few things that don't quite fit our current understanding. One is the weakness of the microwave background fluctuations on certain large scales, but it will probably be impossible ever to confirm this. The other main puzzles at present concern dark matter on small scales. Current theories predict that every big galaxy (like our own) should be surrounded by swarms of small dark matter halos, which are not seen. In addition, models predict that dark matter should probably be more sharply peaked in the centre of galaxies than is observed. It's not clear at present how serious these problems are.

To go back to the very origin of the universe, we must overcome some formidable challenges. The era of inflation reset everything that happened before, so there are no clues left to observe. And to try and probe this era theoretically we'd need a widely accepted theory of quantum gravity, which we do not have.

POSSIBLE WAYS FORWARD

One possible way forward would be if the Large Hadron collider (LHC) identifies supersymmetric particles, and hence confirms supersymmetric theories. An alternative way to confirm these theories would be to find evidence for proton decay - these theories predict that protons should decay on a timescale of 10^{35} years.

If we could detect the imprint of primordial gravity waves on the cosmic microwave background, this would be very direct evidence in favour of inflation. At the time of writing (December 2014) there is a claimed detection (BICEP 2) but a strong suspicion in many quarters that this was an artefact due to dust.

THE VERY BEGINNING

It is entirely possible that the universe has a net energy of zero, and hence could have been produced from a quantum fluctuation. Without a theory of quantum gravity it's hard to know if this is true, but at present it at least appears plausible. We still need there to have been laws of physics existing before, but possibly our concept of "before" is meaningless in this context, so it's not clear that our intuitions about causation are meaningful.

It could be that instead of ever finding a theory of everything, we just continue to peel back layers on an onion - theory after theory each getting closer but never reaching perfection. And it could be that even if we do discover the ultimate physical theory, we'd have no way of knowing this fact.

THE FATE OF OUR UNIVERSE

In two trillion years from now, there will still be main sequence stars in our galaxy (which has long-since merged with Andromeda). Extragalactic astronomy, however, will be at an end as every other galaxy has been redshifted beyond our horizon. But on longer timescales still, eventually everything will collapse to form black holes, which will ultimately merge as gravitational waves radiate their energy away. And these black holes will ultimately evaporate from Hawking radiation, leaving behind ever more redshifted photons. A cold, dark and empty universe. If protons really do decay, residual planets or gas will ultimately turn into photons too.

If $w < -1$ (which is theoretically very unlikely), the universe might end much earlier with a "big rip", with dark energy becoming so strong that it not only pulls other galaxies away but actually rips up our own galaxy and indeed the atoms in our bodies.

If the universe keeps accelerating, then life must inevitably end in our universe. Though it is possible that after life has ended, quantum fluctuations could create the conditions for new life to temporarily form. If eternal inflation models are correct, new universes are always forming, some of them suitable for life, but there is no causal connection between them and us, so it will not save life in this universe from ending.

And if eternal inflation really is true, then anything, no matter how unlikely, must appear somewhere in the infinite series of universes out there. Including planets just like the Earth, with life forms just like us.