5 What We Don't Know

It is often said that each new discovery opens up many more questions than it answers. That's not the case for the Standard Model. The collection of interlinked ideas, bound together in the Standard Model, has brought a synthesis that is unprecedented in science, bringing order to many, seemingly disconnected phenomena and leaving very few threads hanging as a result.

Furthermore, the current experimental situation is one of remarkable harmony. Wielding a broad brush, it is not too inaccurate to say that the Standard Model predicts the correct answer to each and every one of the thousands of particle physics experiments that we've performed.

That's not to say that everything is perfect. There is, as we have recounted in Section 4.4, much still to learn about the neutrino sector. Moreover, if you look in finer detail, then there are a handful of experimental anomalies that seemingly cannot be described by the Standard Model. The most longstanding of these is the magnetic moment of the muon. Recall that the magnetic moment describes how strongly a particle couples to a magnetic field. Our best theoretical result for the muon is

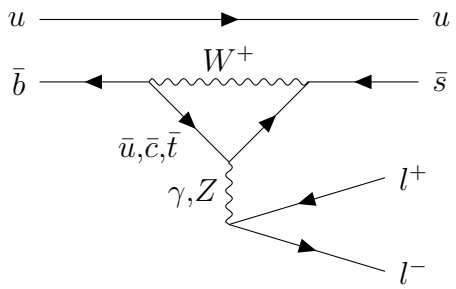
$$g_{\text{theory}} = 2.00233183602$$

while the experimental result is

$$g_{\text{expt}} = 2.00233184122$$

As you can see that, by the time you get to the 9^{th} decimal place, things don't quite match. In any other area of science, you wouldn't care less about a discrepancy in the 9^{th} decimal place. But here it matters. Taken seriously, the deviation between theory and experiment is at the level of 4.2 sigma. Optimistically, this discrepancy might be pointing to extra corrections to g_{expt}, beyond those of the Standard Model. However, there are reasons to be cautious. In particular, the theoretical result involves very difficult numerical simulations to determine contributions from the hadrons and there is some controversy over the accuracy of these results.

There are a number of further niggles too. In nuclear physics, the lifetime of the neutron seems to be slightly different depending on the way in which its measured. In particle physics, a collection of results around B-mesons seem to be slightly discrepant from Standard Model predictions, hinting that electrons, muons and tau leptons may differ in their interactions. The most prominent of these comes from looking at the decays of B-mesons to kaons and a lepton-anti-lepton pair, through a diagram like the following:



These are known as penguin diagrams. (As with the constellations, to see the resemblance you have to squint and reach into the depths of your imagination, before giving up and wondering what these people were smoking. I like to think of the lepton pair as the penguin's beak, but apparently they're the legs.) The quark running in the loop is either (anti) up, charm, or top, while the neutral boson is either a photon or Z. Finally, the end product is, in addition to the kaon, a lepton-anti-lepton pair where either $l = e$ or $l = \mu$.

Since the bottom quark is so heavy, the mass of the leptons is largely irrelevant. This is important because, if we ignore their mass difference, the electron and muon have identical couplings to the weak and electromagnetic forces, a fact that is sometimes called *lepton universality*. This means that the probability to decay to an electron should be the same as the probability to decay to a muon.

The best current measurements suggest that lepton universality is *not* respected in this decay: there is a preference to decay to electrons over muons. Taken at face value, it appears that this because something untoward is going on with muons, rather than electrons.

If these anomalies hold up to further analysis, then they are telling us something extremely important: the Standard Model needs replacing. They may, however, be due to random fluctuations and will disappear as the data improves. The B-meson results are currently 3.1 sigma from Standard Model expectations, somewhat short of the 5 sigma gold standard necessary to claim a discovery. At any given time in history there are always number of such mismatches between theory and experiment. Since the Standard Model was put in place, they have nearly all evaporated upon closer inspection.

These anomalies not withstanding, the current state of affairs is that the Standard Model works extraordinarily well. You have to look very very hard — like the 9^{th} decimal place! — to find clear disagreement between experiment and theory. However, at the same time, it is overwhelmingly clear that the Standard Model is not the last

word in physics and the purpose of this chapter is to describe some of the questions that remain, together with some speculative suggestions for how they may be resolved.

These issues fall into different categories. First, there are a number of unexplained aspects of the Standard Model itself, and these will be described in Section 5.1. Moreover, there is one part of physics that the Standard Model ignores completely: gravity. We will describe how this fits in to the bigger picture in Section 5.2. Finally, if you want some incontrovertible observational evidence that there are things not described by the Standard Model, then we should turn to heavens. In Section 5.3 we describe some of the many puzzles that come from cosmology.

5.1 Beyond the Standard Model

Nearly all the unanswered questions about the Standard Model come from looking more closely at the various constants of Nature and asking why they take the particular values that they do. Here the "constants of Nature" are the parameters that we input into the Standard Model.

For some of these parameters, our understanding is as good as it could possibly be. As explained in Section 4.1.3, the electric charges (or, equivalently, the hypercharges) of the various fermions simply can't be any other way. They are fixed to their values by the stringent requirements of quantum anomaly cancellation. What we would love is to have a similar level of understanding for the other parameters of the Standard Model. Sadly, as we will see, we are a long way from that.

So what are these parameters? Roughly speaking, they fall into three classes (although one could certainly make a more refined classification, especially for the third class). These are:

- Three constants that specify the strengths of the three forces. These are the fine structure constant and its counterpart for the strong and weak force. We'll discuss these in Section 5.1.1.
- Two parameters that specify the Higgs potential. These can be thought of as the mass and expectation value of the Higgs boson. We'll discuss the issues surrounding these in Section 5.1.2.
- Loads of parameters that specify the way the Higgs field interacts with various fermions, usually referred to as the flavour sector. These are the topic of Section 5.1.3.

For quarks, the parameters are six masses (or, equivalently, Yukawa couplings) and a further ten mixing angles that sit in the CKM matrix. (These ten then split further into 9 mixing angles and the phase that gives CP violation.) There is, in addition, one further parameter known as the QCD theta angle that we haven't yet mentioned because, as far as we can tell, it is zero. Nonetheless, zero is a number too and it deserves an explanation.

For leptons the counting is a little more fuzzy. If the neutrinos get a Dirac mass, so that $B-L$ symmetry is preserved, then we again have six masses, or Yukawa, parameters and ten mixing angles in the PMNS matrix. If, however, neutrinos have a component that is a Majorana mass then there are additional parameters (and phases) to specify.

We'll now look at each of these classes of parameters in turn.

5.1.1 Unification

Perhaps the most important fact about all the constants of Nature is that they are very poorly named. They are not constant. Instead, the phenomenon of renormalisation means that the "constants" depend on the energy scale at which you do your experiment. We described this in some detail back in Section 2.3, and again in Section 3.1, when we explained how the fine structure constant, and its counterpart for the strong force, change with energy.

The energy dependence of coupling constants brings important clues when we come to better understand their origin. In particular, we understand well how the coupling constants vary on scales that we've tested – say, up to 10^3 GeV. But we could then extrapolate to further energies. Of course, we don't know what lies ahead at further energies, but to get the ball rolling we could simply assume that there's nothing other than the Standard Model, and then see what we find.

What we find is extremely interesting and shown in Figure 50. First let's explain what we're looking at. The forces of the Standard Model are summarised by $U(1)\times SU(2)\times SU(3)$. Corresponding to each of these is a coupling constant α_i where $i=1,2,3$. Note, in particular, that α_1 is the hypercharge coupling, rather than the fine structure constant of electromagnetism which emerges at low-energies from a combination of hypercharge and the weak force. On the horizontal axis is the energy, here called Q, plotted on a logarithmic scale. The part of the graph that we've measured experimentally is way over to the left, with $\log_{10}(Q/\text{GeV}) \lesssim 3$. Everything else is extrapolation.

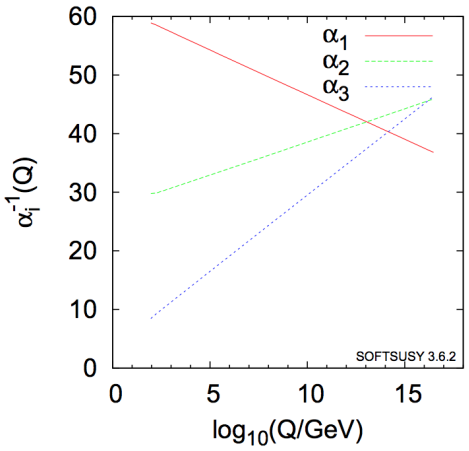


Figure 50. The running of the coupling constants. α_1 is hypercharge; α_2 the weak force and α_3 the strong force. This plot was made by Ben Allanach and taken from the PDG review of GUTs.

On the vertical axis of Figure 50 is the inverse coupling, α_i^{-1}. The red line decreases with energy, while both the blue and green lines increase. This is telling us that α_1 increases with energy, while both α_2 and α_3 decrease, and this is the expected behaviour given our discussion of asymptotic freedom in Section 3.1.

The most striking aspect of the plot is the way the three lines cluster together as we approach higher energies. Obviously, they don't precisely meet, but nonetheless they lie in the same ballpark. This would appear to be hinting that the three forces are not as different as they appear in our world: perhaps, at a much higher energy scale, they are all unified as one. This idea is know as *grand unification*. The weak and strong forces meet at a scale

$$M_{\text{GUT}} \approx 10^{16} \text{ GeV}$$

which is known as the grand unified scale. At this point, the three coupling constants all converge on a value somewhere around

$$\alpha_{\text{GUT}} \approx \frac{1}{40}$$

Even the location of this almost-meeting is important. First, α_{GUT} is nice and small at this point, telling us that the calculation to extrapolate the lines is at least consistent. Second, the scale M_{GUT} lies just below another important scale in nature, namely the

Planck scale

$$M_{\rm pl} \approx 10^{18}\ {\rm GeV}$$

(This is sometimes referred to as the reduced Planck scale to distinguish it from another contender that differs by a factor of $\sqrt{8\pi}$ and so is closer to 10^{19} GeV.) The Planck scale $M_{\rm pl}$ is where the effects of gravity become important in the quantum regime. We'll have more to say about this later. For now we just mention that the fact $M_{\rm GUT} < M_{\rm pl}$ is important. Had it turned out to be the other way round, then there would be no reason to think that $M_{\rm GUT}$ is interesting: it is a scale that was derived by neglecting the effect of gravity, and that's only an acceptable thing to do at energies less than $M_{\rm pl}$.

If we take the existence of three lines not-quite meeting as evidence of unification, what can we do about it? Is it possible to write down a grand unified theory, or GUT, in which the three forces are unified? The answer is yes. In fact, in many ways the Standard Model is just crying out to be packaged into something simpler! Recall that the mathematical way of describing the three forces is as

$$G_{\rm SM} = SU(3) \times SU(2) \times U(1)$$

where each of these can be roughly thought of as electric and magnetic fields whose individual elements are themselves matrices: 3×3 for the strong force, 2×2 for the weak force and just usual numbers for $U(1)$ hypercharge. But all of these can be packaged nicely inside a bigger matrix. (Or, more precisely, a bigger group.) For example, you can put all of them inside a 5×5 matrix with

$$G_{\rm GUT} = SU(5)$$

Alternatively, they can be packaged into a 10×10 matrix, but where the matrix is now based on real numbers rather than complex numbers

$$G_{\rm GUT} = SO(10)$$

Other options are also available.

In all of these cases, there are new gauge bosons that come as part of the unified force. These are usually called *X-bosons*, in analogy with the W and Z bosons of the weak force. There are also new scalar fields that condense, giving rise to a Higgs-like mechanism. Unlike the Higgs mechanism of the weak force, this conjectural GUT-Higgs should get an expectation value at scale of around $M_{\rm GUT}$. Correspondingly, the X-bosons are heavy with a mass also somewhere in the vicinity of $M_{\rm GUT}$. Needless to say, we would not expect to discover such X-bosons in collider experiments any time soon.

It's not just the forces which have to unify. The matter particles and their other interactions must too. Here too things start off looking rosy. Recall that, including the count over colours, there are 16 fermions in one generation of the Standard Model. It must be possible to package these into the groups mentioned above that unify the forces. (Mathematically, we're looking for *representations* of $SU(5)$ or $SO(10)$ and these come only in special numbers, special like the 8 and 10 of the eightfold way are special.) It turns out that the particles of the Standard Model are tailor made to be put together in this way. It seems like everything fits like a glove.

Interestingly, for $SU(5)$ GUTs, the right-handed neutrino remains an outsider, not coupling directly to the forces. Meanwhile, for $SO(10)$ GUTs, the right-handed neutrino also is brought into the fold and, at least at the fundamental level, sits on the same footing as all the other particles.

So far, so good. The last part of grand unified theories is to find a way that the flavour sector drops out nicely, with the Yukawa terms and mixing matrices all falling into place. Here things are less rosy. It's possible, but it's not pretty, typically involving the introduction of yet further fields put together in a fairly baroque way. As with so many other things in the Standard Model, the flavour sector is the one we understand least.

Tweaking the Lines

The coupling constants in Figure 50 get close, but fail to actually meet. However, this plot was made under the assumption that there's nothing new to be found between the energy 10^3 GeV that we've probed experimentally and the GUT scale 10^{16} GeV. And that seems unlikely.

If there are new particles to be found, then they contribute to the renormalisation of coupling constants and so change the way the lines run. One obvious suggestion is that these new particles may correct the lines in such a way that they do, in fact, meet after all.

For this to happen, the new particles shouldn't be too heavy otherwise they come in too late to make a difference. One possibility that has been greatly studied is a theory called *supersymmetry*. We'll say more about this in Section 5.1.2, but for now we'll simply mention that we introduce a bunch of new particles at, say 5×10^3 GeV. This tweaks the running of the couplings, to give the result shown in Figure 51. And ... ta-da. The lines now meet perfectly.

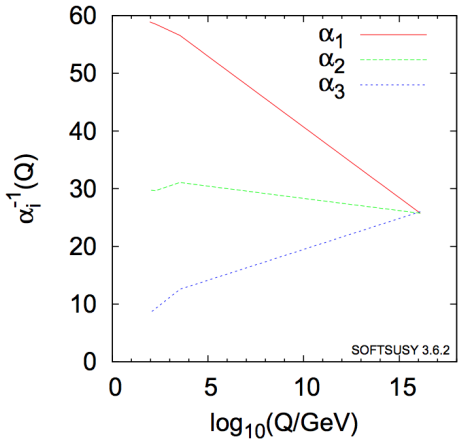


Figure 51. The running of the coupling constants if new supersymmetric particles exist at a low mass. This plot was made by Ben Allanach and taken from the PDG review of GUTs.

However, there's a major problem with this particular supersymmetric scenario. The new particles are so light that some hint of them should have shown up at the LHC by now. They haven't. It's possible to write down theories where supersymmetry still does the job of unification while just evading detection but they look increasingly unlikely.

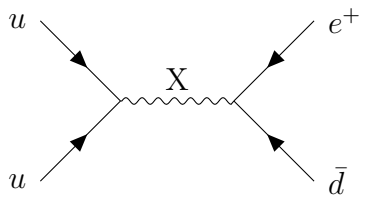
Nonetheless, the idea that there may be further particles out there which point clearly to unification is a tantalising one, and the convergence of the three coupling constants remains one of the major clues we have about physics beyond the Standard Model.

Proton Decay

Although the grand unification scale $M_{\rm GUT}$ is way beyond what we can study experimentally, the idea of grand unification still gives an observable consequence. That is proton decay.

We already discussed proton decay in Section 4.3.5 in the context of the conservation laws. Recall that, within the Standard Model the proton should be absolutely stable. There is a rather exotic process — known as electroweak instantons — that allow three baryons to decay to three leptons but this has never been observed and is unlikely to be any time soon given that calculations give a lifetime of around 10^{173} for a helium nucleus!

However, proton decay is necessarily a consequence of any grand unified theory. This is because the X-bosons gives rise to a Feynman diagram like those of weak decays, but now linking quarks and leptons like so



where, in this example, the X-boson carries electric charge $+4/3$. The $\bar{d}$ quark then combines with the d quark in the proton to form a pion, hence $p \to e^+ + \pi^0$.

Because the mass of the X-boson is so large, the lifetime of the proton is long. But not all that long! Most GUTs predict a lifetime between 10^{31} and 10^{36} years. Current experimental bounds tell us that the proton lifetime is longer than 10^{34} years, already ruling out the simplest GUTs.

Magnetic Monopoles

All magnets are dipoles, with a north pole and a south pole. Cut a magnet in two, and you'll end up with two dipoles. It's impossible to get, say, a single north pole on its own. If one could find such an object – known as a *magnetic monopole* – it would have a distinctive radial magnetic field of the form

$$\mathbf{B} = \frac{g}{4\pi}\frac{\hat{\mathbf{r}}}{r^2} \tag{5.1}$$

where g is the magnetic charge.

At first glance, there seems to be little reason to think that magnetic monopoles exist. Indeed, there is even a law of physics that forbids monopoles! One of the Maxwell equations (2.3), governing the theory of electromagnetism, reads

$$\nabla \cdot \mathbf{B} = 0 \tag{5.2}$$

and its sole purpose is to disallow any solution of the form (5.1).

This makes it somewhat surprising that, as the laws of physics evolve beyond electromagnetism, magnetic monopoles re-emerge as one of the most likely candidates for new particles, finding interesting and creative ways to evade the seemingly insurmountable obstacle (5.2). By the time we get to grand unified theories, monopoles become obligatory, appearing as an entirely novel kind of particle known as a *soliton*. The mass of these magnetic monopoles is

$$M_{\text{monopole}} \approx \frac{M_{\text{GUT}}}{\alpha_{\text{GUT}}}$$

putting them somewhere in the range of 10^{17} GeV, well out of reach of current colliders.

Unlike all heavier particles that we've discussed in these lectures, monopoles would be completely stable. Ironically the same Maxwell equations which once seemed to forbid magnetic monopoles, now forbids them from decaying since they imply the conservation of magnetic charge. Monopoles can only vanish by annihilating with anti-monopoles. This leaves open the possibility that we may, once again, turn to the skies and search for monopoles among cosmic rays. So far, none have been found[18].

It's not just GUTs that give rise to magnetic monopoles. Instead, pretty much any theory that goes beyond the Standard Model will contain magnetic monopoles. They are one of the very few robust predictions for new physics. If you want to learn more about monopoles then you've come to the right place. You can read about their subtle interplay with quantum mechanics in the lectures on Solid State Physics, about their role in quantum field theory in the lectures on Gauge Theory, and about some of their more mathematical aspects in the lectures on Solitons.

5.1.2 The Higgs Potential

Our next pair of parameters are associated to the Higgs potential. Recall from Section 4.2.1 that the Higgs potential $V(\phi)$ determines whether the Higgs boson condenses. In the Standard Model, it takes the very simple form

$$V(\phi) = a|\phi|^2 + b|\phi|^4 \tag{5.3}$$

The two fundamental parameters are a and b.

If $a > 0$ and $b > 0$ then it looks like the graph on the left-hand side of Figure 52. If $a < 0$ and $b > 0$ then it looks like the right-hand side of Figure 52. (If both a and b are less than zero then the potential has no minimum and the Higgs scalar runs away to infinity unless we include further terms of higher powers like $|\phi|^6$.)

As we've seen, the Standard Model has a potential with the shape on the right, meaning that $a < 0$ and $b > 0$. The values of a and b determine both the mass of the Higgs boson m_H and the Higgs expectation value $\langle\phi\rangle$. Roughly speaking, the relationship between these two scales and the parameters in the potential is

$$m_H^2 = |a| \quad \text{and} \quad \langle\phi\rangle^2 = \frac{a}{2b} \tag{5.4}$$

[18] A more correct statement is that exactly one has been found! On Valentine's day, 1982, a single event consistent with a magnetic monopole was observed. Nothing similar has been seen since. Given the importance of replicating scientific results, it's difficult to view this as anything more than a tantalising footnote (literally here) in the story of the monopole

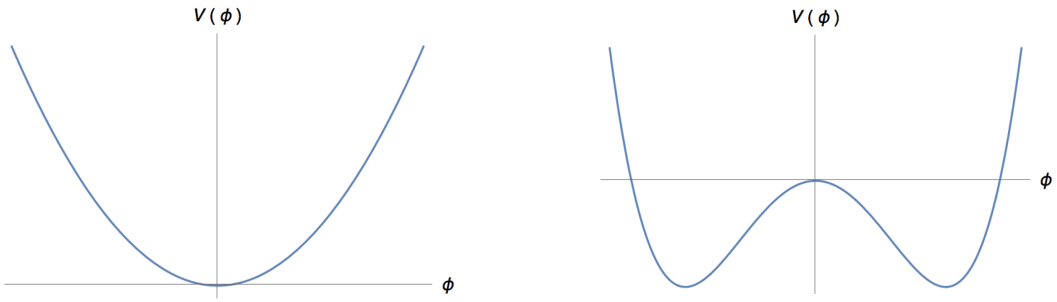


Figure 52. Two possible shapes for the Higgs potential for a scalar field. With $a, b < 0$ we get the shape on the left; with $a < 0$ and $b > 0$ we get the shape on the right.

Much of this section will be devoted to explaining further what the words "roughly speaking" in the previous sentence actually mean. For now, we'll just roll with the equations above and see what they tell us. Experimentally, we know that

$$m_H \approx 125 \text{ GeV} \quad \text{and} \quad \langle \phi \rangle \approx 246 \text{ GeV}$$

This means that the parameters in the potential must take the values

$$a \approx -(125)^2 \text{ GeV}^2 \quad \text{and} \quad b \approx 0.13$$

For now, the important point to note is that a has dimension of energy-squared, while b is dimensionless. Our interest here lies in a. This is the scale of the Higgs sector which, through Yukawa interactions subsequently sets the mass of all the elementary fermions.

As we've stressed earlier in these lectures, most of the mass of the proton and other hadrons doesn't come from the Higgs sector, but instead from $\Lambda_{\rm QCD} \approx 200$ MeV. But this sits on a different footing because $\Lambda_{\rm QCD}$ was, itself, a derived scale: it is the energy at which the dimensionless coupling of the strong force becomes $\alpha_s \approx 1$. This means that $\Lambda_{\rm QCD}$ should be thought of as an emergent energy scale.

In contrast, there is no such story for $\sqrt{|a|} = m_H$. This is an absolute energy scale. In fact, rather remarkably, it is the only fundamental parameter of the Standard Model that is not dimensionless! As we now explain, this means that it comes with certain baggage.

The Higgs Mass

Now we come to the crux of the matter, one that revolves around the "rough speaking" that lead us to the equations (5.4). To understand what the issue is, we need to think a little more deeply about the kind of parameters that make sense in a fundamental theory and how these change at various energy scales due to renormalisation.

To set the scene, I'll introduce some analogies. To this end, here are three objects that should not be viewed as a "fundamental parameters" in a theory.

There sits, in a vault in Paris, a platinum-iridium cylinder that, until 2019, was used as the definition of the kilogram. Clearly it would be ludicrous to take the mass of this object as any kind of input in a fundamental theory of Nature. Even putting aside the facts that the kilogram is very much a human construct, and that the mass of the cylinder is changing over time as it slowly erodes, it is simply silly to think that a huge, complicated object with roughly 10^{24} constituent particles should be important on the tiny distance scales at which our fundamental theory of physics is defined.

This latter criticism can also be levelled at other objects which, at first glance, might appear to be more suitable candidates for fundamental parameters. For example, for many decades the mass of the proton was thought to be a fundamental scale in nature. Of course, we now know that the proton, like the Parisian cylinder, is a horribly complicated object. In particular, the mass of the three valence quarks — two up and a down — contribute a negligible amount to the total mass of the proton. The full mass is often attributed to collection of gluons and sea of quark-anti-quark pairs but, for the purposes of this analogy, we will be better served if we remember what this really means: the mass comes from the wild thrashing of the quantum fields that are excited in complicated ways inside the proton. The mass of the proton is in the ballpark of a few times $\Lambda_{\rm QCD}$ but its exact value is something emergent that depends on lots of messy dynamical processes. For this reason, the mass of the proton is, like the Parisian cylinder, an emergent scale. Neither are good candidates for parameters in a fundamental theory.

And this brings us to our third example of an object which is *not* a good candidate for a fundamental parameter. This is the mass of the Higgs boson!

Quantum fluctuations mean that, even though the Higgs boson is a fundamental particle, its mass depends on all sorts of complicated and messy dynamics and is ultimately determined, like the mass of the proton, by the behaviour of other quantum fields. In particular, if the fundamental theory has a parameter like a in the potential

(5.3), then its relationship to the mass m_H of the Higgs boson is nothing like as simple as the $m_H^2 = |a|$ equation that we used above. Instead the relationship between the two is much more complicated.

To understand what's going on, first recall our discussion of renormalisation in Section 2.3. There we learned that all parameters in a quantum field theory depend on the distance scale or, equivalently, the energy scale at which an experiment takes place. Moreover, we stressed that all quantum field theories should come with a health warning: there is a minimum distance scale, or maximum energy scale, beyond which they shouldn't be used. This energy scale is called the *UV cut-off* and we will denote it as Λ_{UV}.

You should think of the UV cut-off as the energy scale at which a given quantum field theory is defined. When you specify the parameters of the theory, you should specify their value at the scale of the cut-off. This is really just the statement of reductionism in physics: small things determine the behaviour of larger things, and a fundamental theory should be defined on the smallest distance scale at which it applies. As you then look at lower energies – or longer distances – you can use renormalisation to figure out how these parameters change.

Now let's return to the Standard Model. It definitely works up to energy scales of 1 TeV so let's be pessimistic and, with the expectation that the Standard Model will cease to give the right answers very soon, take the cut-off to be $\Lambda_{UV} = 1$ TeV. That means that we take the Standard Model to correctly describe the dynamics of quantum fields down to distance scales as small as 10^{-19} m. The fields may well have fluctuations on scales much smaller than that, but we will just admit ignorance about these and proceed.

What now happens if we put a Higgs boson into the mix. Naively the mass of the Higgs boson is given by the formula that we used previously: $m_H = \sqrt{|a|}$. But the Higgs then gets surrounded by a swarm of quantum fluctuations and these change the mass. The upshot is that if a is the fundamental parameter of the theory, then the mass of the Higgs boson that we measure is actually something like

$$m_H^2 \approx \left|a + \mathcal{O}(\Lambda_{\rm UV}^2)\right| \tag{5.5}$$

where the $\mathcal{O}(\Lambda_{\rm UV}^2)$ means a contribution that is roughly around $\Lambda_{\rm UV}^2$, but where the exact coefficient (including its sign) depends on the nature and dynamics of all the other fields.

Here's a rather nice analogy. Take a ping pong ball and submerse it in water. What's its mass? According to the International Table Tennis Federation, a ping pong ball should have a mass of exactly 2.7 grams. But the federation rarely promotes matches that take place underwater and there is no suggestion on their website that they understand renormalisation. You can use Newton's equation $F = ma$ to experimentally determine the mass of a ping pong ball in water: just apply a force and measure the acceleration. You'll find that the ping pong ball appears to be roughly 11 times heavier than in air, so closer to 30 g. The reason for this is intuitively simple: when the ping pong ball moves in water, it drags a body of fluid with it. This increases its inertial mass.

This same effect is at play for the Higgs boson. Its mass gets a contribution from the fluctuations of all other quantum fields. This additional contribution is of order $\Lambda_{\rm UV}$ which is much larger than the mass that we actually observe.

What are the consequences of this? The formula (5.5) tells us that the fundamental parameter a in the Higgs potential is not equal to the mass of the Higgs boson. Instead, it too must be something on the scale of the UV cut-off $\Lambda_{\rm UV}$. We call the parameter a the *bare mass* (squared), while m_H is physical mass of the Higgs boson that we observe. To get the observed physical mass $m_H \approx 125$ GeV we must add two contributions, a and $\Lambda_{\rm UV}$, both of the same order of magnitude, which then cancel out to leave behind the physical mass.

So what? Well, if we take the cut-off of the Standard Model to be $\Lambda_{\rm UV} \approx 10^3$ GeV, then this seems eminently reasonable. We're just adding two numbers, both of order 1000, to get something of order 100.

But what if the Standard Model holds to higher energy scales? Suppose that it is trustworthy to energies of order $\Lambda_{UV} \approx 10^4$ GeV. Now we add two numbers of order 10,000 to get something left behind that's order 100. Still not preposterous, but you might start to feel a little uneasy.

And what if we really push things? Suppose that the Standard Model actually holds all the way up to the GUT scale of $\Lambda_{UV} \approx M_{\rm GUT} \approx 10^{16}$ GeV. Now we must take the fundamental parameter in the Higgs potential to be $a \approx (10^{16}\ {\rm GeV})^2$ so that this cancels almost precisely the contribution from the quantum fluctuations, leaving behind a measly Higgs boson mass of 125 GeV. If you were to change a by just one part in a billion, you would end up with a Higgs boson mass that is 5 orders of magnitude higher than we observe! By this stage, the situation appears to be ridiculously untenable

We have a bunch of different names for this state of affairs. We say that the parameter a in the Higgs potential should be *finely tuned*. It's a good name and describes the issue well. Alternatively, we say that the Higgs mass m_H that we observe is *unnatural*; the quantum fluctuations want to push the mass to higher and higher energy scales, but this is cancelled — to what may feel like unreasonable accuracy — by the contribution from the bare mass. This seems to be a less useful name as it's hard to see how something in nature can be unnatural. But it does at least stress how jarring the situation is.

This whole set of ideas also goes under the umbrella of the *hierarchy problem*. Why is there a large hierarchy of scales between the Higgs mass and the UV cut-off, when theory suggests that they should be the same order of magnitude?

Recall that the UV cut-off $\Lambda_{\rm UV}$ is really an expression of our ignorance: we confess that our quantum field theory — in this case the Standard Model — is incomplete and does not include relevant physics at energies beyond $\Lambda_{\rm UV}$. The hierarchy problem is really the statement that the mass of the Higgs boson is pushed by quantum corrections to whatever is the highest mass scale in the game. Yet, somehow, the observed mass remains happily at 125 GeV.

The hierarchy problem in this brutal form is only a problem for the parameter a in the Higgs potential. All other parameters in the Standard Model are dimensionless and do not suffer the same fate. They change only very mildly (logarithmically to be precise) under renormalisation. It is only the Higgs mass that is so very sensitive to physics at higher energy scales.

Solutions to the Hierarchy Problem

The hierarchy problem motivated an enormous amount of research in the 1980's, 90's and early 2000's. The favoured explanation was the obvious one: keep the UV cut-off $\Lambda_{\rm UV}$ low enough that you don't need to invoke a silly level of fine-tuning. But this, in turn, means that there should be some new physics, invalidating the Standard Model, that comes in at some low scale, like a few TeV.

This new physics can't just be anything. Add a few new particles at the TeV scale and you'll see that they just make the problem worse, adding yet more quantum fluctuations that increase the mass of the Higgs boson. Instead, you must find a way to add some new fields to the Standard Model that stabilise the mass of the Higgs somewhere in near 100 GeV, solving the problem once and for all. There are a number of ways to achieve this. Here I describe some, roughly in descending order of popularity.

- **Supersymmetry:** This is a proposed, novel symmetry of Nature in which every bosonic field has a fermionic counterpart, and vice versa, with each boson/fermion pair experiencing the same forces.

  Clearly, we don't see supersymmetry among the particles that we know. The idea is that supersymmetry is, like many symmetries, broken so that the additional fields – so called superpartners – only appear when we reach some new mass scale, say a few TeV.

  There are a number of reasons to be enamoured of supersymmetry. First, it solves the hierarchy problem by dint of the fact that bosons and fermions contribute with opposite signs to the mass of the Higgs, ensuring that all quantum fluctuations cancel above the supersymmetry scale. Second, as we saw previously, the presence of supersymmetry causes the three coupling constants to meet perfectly at the unification scale. Moreover, there are reasons to think that supersymmetry may be an important ingredient in quantum gravity, which would mean that it should certainly be present by the time we get to the Planck scale. All in all, TeV scale supersymmetry leads to a nice, comforting story.

- **Technicolour:** The Standard Model contains just one other mass scale, $\Lambda_{\rm QCD}$. But, as we described above, this is associated to the strong coupling dynamics of QCD and so doesn't suffer from a hierarchy problem. Perhaps the mass of the Higgs emerges in a similar way.

  In such scenarios, the Higgs particle is not fundamental at all, but rather appears as a composite of two, new fermions, bound together into a meson by a new force called, in analogy with strong force, *technicolour*.

- **Something Else:** There are quite alternative proposals. Among them is the idea that Higgs as a (pseudo)-Nambu-Goldstone boson (I won't explain what this means!) or, more creative suggestions, such as extra dimensions of space which manage to dilute quantum corrections at the TeV scale.

Each of these ideas provides a viable solution to the hierarchy problem, but only by introducing some observable deviation from the Standard Model at the TeV scale. But now we have a collider – the LHC – that can reach these scales. And nothing is seen. The Higgs boson appears, as far as we can tell, as a genuinely elementary particle and there is, as yet, no hint of new particles that could stabilise its mass. This makes it increasingly unlikely that any of the "natural" solutions described above are implemented in Nature. Of course, it's certainly possible that, say, supersymmetry is still out there, but just pushed up to a higher scale, with an accompanying need

for some amount of fine tuning to the Higgs mass. But the motivation starts to look increasingly shaky.

So what are we to make of this? Admittedly, the hierarchy problem has a different flavour from other major open problems in physics. Is there really anything wrong with just stating that the parameter a is (to give an extreme example) defined at the GUT scale $\Lambda_{\rm UV} = M_{\rm GUT}$ and fine tuned to 15 significant figures so that it perfectly cancels out the contributions to the Higgs mass from quantum fluctuations of order $\Lambda_{\rm UV}$? It certainly seems odd, but perhaps that's just the way it is.

We can look elsewhere in physics to get some guidance. In particular, quantum field theory isn't just useful for particle physics: it is the right language to describe large swathes of solid state physics (also known as condensed matter physics), which is the study of how various materials behave. This arena provides many hundreds, if not thousands, of examples of quantum field theories (or, relatedly, statistical field theories) where we can test our logic. In that framework, we can ask: do we find quantum materials where we have light scalar excitations, like the Higgs boson? Here "light" means with a mass significantly smaller than the UV cut-off which, in solid state physics, is usually supplied by the underlying lattice of the material.

The answer to this question is a resounding no! Or, stated more accurately, within the realm of solid state physics if there is a light scalar excitation then there is always a good reason behind it. These reasons are often similar to the ones invoked for the hierarchy problem, like the scalar is really made of two underling fermions, or it is a (pseudo)-Nambu-Goldstone boson. (I still won't explain what this last phrase means.) The upshot is that in other realms where quantum field theory is useful, the logic of *naturalness* is a very good guide: if you see a scalar field that is unnaturally light, then you should search for an explanation because it will tell you something important and interesting about the system.

At the risk of belabouring this point with a long detour, there is one particular condensed matter system that is worth looking at more closely. This is superconductivity which, as we already mentioned in Section 4.2.1, has a mathematical description that is almost identical to that of the Higgs boson. Usefully, the hierarchy problem also makes an appearance in superconductors, although it's not in the guise of a light scalar field but, as I now explain, something more subtle. As you cool a metal, it undergoes a phase transition to become a superconductor. This happens at the *critical temperature* T_c which is typically a few degrees Kelvin. The phase transition happens discontinuously, meaning that the metal changes abruptly to a superconductor just like water

changes abruptly to ice. This is sometimes called a *first order phase transition*. But theoretical expectations tell us that the phase transition should be smooth and continuous, what's called a *second order phase transition*. This isn't seen in experiment because it turns out that you have to tune the temperature T to ridiculous accuracy before you notice that the transition is actually continuous, something like

$$\frac{T - T_c}{T_c} \approx 10^{-9}$$

Why would you have to tune the temperature so close to T_c before you can see the real physics of the second order phase transition? It's a very unusual state of affairs and in most other systems you only need $(T - T_c)/T_c \approx 1$ before you can see the true nature of the phase transition. The fine-tuning of temperature needed for a superconductor turns out to be entirely analogous to that of the lightness of the Higgs boson. The 10^{-9} accuracy is a small number and deserves an explanation. And, in this case, there is a very good explanation: the would-be Higgs boson in a superconductor is composed of two electrons bound together at an energy scale which emerges mathematically in a manner that is similar to way $\Lambda_{\rm QCD}$ emerges in the strong force. You can trace the existence of the small number 10^{-9} to this underlying dynamics. It's an explanation that is close in spirit to the technicolour proposal for the Higgs boson. But, as far as we can tell, the same story is not what's going on for the Higgs. (If you want more details about the relationship between phase transitions and renormalisation group, you can read about it in the lectures on Statistical Field Theory.)

All of this is to say that our experience with quantum field theory tells us that we should be nervous about the fine-tuning underlying the Higgs. Still, there is a vast difference between meV scale at which effective quantum field theories in condensed matter are valid, and the TeV scale or higher of particle physics. And Nature seems to be telling us very clearly that there is nothing wrong with fine tuning at these very high energy scales, even though it flies in the face of how we understand quantum field theory. It seems to me, and many others, that this is a problem one should take seriously because it tells us that we're not thinking about quantum field theory in the right way. But I don't know a better one!

There is one last word on this. The word is anthropic. It gives me a slight shudder just thinking about it so I'm going to postpone further discussion to Section 5.3.1.

5.1.3 A Bit of Flavour and Strong CP

In the late 1800's, one of the great unsolved mysteries was the spectrum of hydrogen. Only a few lines were known when, in 1885, Balmer noted that their wavelength λ

could be fitted to the remarkably simple formula

$$\lambda \approx 10^{-7} \left(\frac{1}{4} - \frac{1}{n_2^2} \right)^{-1} \text{ m}$$

with $n_2 = 3, 4, 5, 6$. Three years later, Rydberg generalised this to the formula

$$\lambda \approx 10^{-7} \left(\frac{1}{n_1^2} - \frac{1}{n_2^2} \right)^{-1} \text{ m}$$

with n_1 and n_2 both integers. (The simplest $n_1 = 1$ lines lie in the ultra-violet and took another 20 years to discover.)

The Balmer formula provides one of the rare examples in science where fitting data, with no understanding of the underlying physics, gives a strong hint of something important underneath. Indeed, the presence of integers in Balmer's formula foreshadowed one of the greatest paradigm shifts in science: the need for quantum mechanics. Ironically, Bohr would later claim that he had completely forgotten about the Balmer formula when he constructed his quantum model of the atom, and only later realised that the two matched perfectly!

If you were to wonder what collection of today's unexplained experimental data has the best chance of a Balmer-like breakthrough, the answer is obvious: it is everything to do with flavour. There are a few dozen parameters that describe the masses and mixing angles of three generations of quarks and leptons. There are clear patterns among them but, so far, we have little idea what those patterns are telling us.

It is not for want of trying. There are many theories that have tried to explain what's observed, many of them revolving around some kind of new, approximate symmetry, sometimes referred to as family symmetries. Some of these theories are pretty(ish), some baroque. None of them are overwhelmingly compelling, in large part because there are so many options available. For this reason, I won't describe any of these in detail.

The Strong CP Problem

There is, however, one parameter in the Standard Model that deserves special attention. It is a parameter of the strong force known as the QCD theta angle, or θ_{QCD}.

It is not so straightforward to describe, at the level of these lectures, how θ_{QCD} changes the dynamics of the strong force, not least because it is a quantum parameter, in the sense that it has no classical counterpart at all. If you want all the gory details, you can read about theta angles in the lectures on Gauge Theory.

The main reason that the QCD theta angle is interesting is that it doesn't exist! As far as we can tell, $\theta_{\rm QCD} = 0$. Strictly our best experimental bound is

$$\theta_{\rm QCD} < 10^{-10}$$

There are a few questions to unpack here.

First, why should we care about something that doesn't exist?! The point is that the Standard Model is a remarkably restrictive framework. We can't just write down random interactions willy nilly. There are only very special interactions that are allowed, and each of them comes with a parameter. The game that we play is to write down all possible interactions and then determine their associated parameter by experiment. This is what leads us to the couple of dozen parameters or so listed at the beginning of this section. The theta angle is special because it is the only one of these parameters that appears to vanish. And that is crying out for an explanation.

Next: what would the consequences be if $\theta_{\rm QCD}$ were not to vanish? It turns out that this is the one opportunity for the strong force to violate the symmetries of parity P, charge conjugation C and time reversal T.

Recall from Section 4.3.4 that weak force respects neither parity nor CP but, as far as we can tell, both are respected by the strong force and electromagnetism. In a world where $\theta_{\rm QCD} \neq 0$, the strong force would also break these symmetries. It does so by endowing the neutron with an *electric dipole moment*. This means that although the neutron is neutral, the distribution of the quarks inside would be shifted slightly so that one side of the neutron is slightly positively charged with a compensating negative charge on the other side. This means that the neutron carries a little arrow, pointing in the direction of the charge imbalance. This is in addition to another arrow, pointing in the direction in which the neutron spins and it turns out that these two arrows are either aligned or anti-aligned. When one acts with parity, or charge conjugation, or time reversal, one of the arrows flips, while the other does not. This means that these symmetries are broken since, for example, the neutron appears different when viewed in a mirror.

Although all three symmetries are broken when $\theta_{\rm QCD} \neq 0$, physicists tend to focus on CP. The unexplained fact that $\theta_{\rm QCD} = 0$ is known as the *strong CP problem*.

Before we turn to putative explanations for this problem, there is one final question I'd like to address. What does it have to do with flavour physics? The question of flavour is all about how fermions couple to the Higgs boson, while the strong CP

problem appears to be, as the name suggests, firmly in the camp of QCD. Here's where things get a little complicated because the story that I told above isn't entirely accurate.

The fuller truth is that in the Standard Model the masses of the six quarks are actually complex numbers rather than real numbers. Each has a magnitude and a phase

$$M_q = m_q e^{i\chi_q}$$

where the label q runs over the six different quarks. What we observe is the magnitude m_q. The complex phases χ_q have almost no effect on the physics. They do just one thing: they contribute to the QCD theta angle! In fact, what we measure is not the QCD theta angle directly, but rather

$$\bar{\theta} = \theta_{\rm QCD} + \sum_q \chi_q$$

It is this sum of angles that is observed experimentally to vanish: $\bar{\theta} = 0$. In other words, the strong CP problem involves both QCD and flavour physics, entwined in some interesting fashion.

There are a number of prospective solutions to the strong CP problem. Here I'll just mention one, which goes by the name of the Peccei-Quinn theory. The key idea is both simple and dramatic: one takes the parameter $\theta_{\rm QCD}$ and promotes it to a dynamical field.

This means that we no longer get to chose the value of $\theta_{\rm QCD}$. Instead its value can fluctuate. This is only progress if we can explain why we don't observe the fluctuations and why, moreover, the field prefers to sit at the value such that $\bar{\theta} = 0$.

The new field $\theta_{\rm QCD}$ is, like the Higgs boson, a scalar field and its preferred value will be set by some potential, analogous to the $V(\phi)$ potential that we saw before. Here there is a very nice story. It turns out that the rest of the Standard Model fields have something to say about this potential. In particular, the dynamics of the gluons generates a potential for $\theta_{\rm QCD}$. And this potential has the property that its minimum sits at the place where $\bar{\theta} = \theta_{\rm QCD} + \sum_q \chi_q = 0$. In other words, if $\theta_{\rm QCD}$ was dynamical, it would want to arrange itself so that there is no CP violation in the strong force. This, it would seem, is a very compelling solution to the strong CP problem.

What about the second issue? Now that $\theta_{\rm QCD}$ is dynamical, why do we not see its fluctuations? Here things are less rosy. First, it's not clear that the simple dynamical mechanism described above is really sufficient to set $\bar{\theta} = 0$ to one part in 10^{10} as required by experiment. Second, with a new field comes a new particle which, in the case of $\theta_{\rm QCD}$ is referred to as the *axion*. There have been many decades of searches for the axion, with nothing yet seen. There are many further experiments underway, aiming to improve the bounds the axion mass and interaction strength or, in the dream scenario, actually find the thing!

While the Peccei-Quinn theory, and the accompanying axion, remains the most popular explanation for the strong CP problem, the lack of clear experimental support means that it is not the only game in town. Whatever the ultimate reason, the fact that the strong force prefers to respect the discrete symmetries of our world, even though it has a clear opportunity to violate them, is one of the key clues for physics beyond the Standard Model.

5.2 Gravity

There is one force that is obviously missing from the Standard Model. This is gravity, both the most obvious force at play in the world around us and yet, in many ways, the one we understand least.

There are two good reasons why gravity is not included in the Standard Model. The first is that the force of gravity is entirely inconsequential for anything to do with particle physics. To get some sense for this, we can look at the theory of gravity first written down by Newton. Any two objects, with masses m_1 and m_2, sitting a distance r apart, will feel a force

$$F_{\rm Newton} = \frac{Gm_1m_2}{r^2} \tag{5.6}$$

where G is Newton's gravitational constant

$$G \approx 6.67 \times 10^{-11}\ {\rm m^3\,kg^{-1}\,s^{-2}} \tag{5.7}$$

It is a remarkable fact that this takes exactly the same form as the Coulomb force between any two electrically charged objects that we met earlier in (2.4),

$$F_{\rm Coulomb} = \frac{Q_1Q_2}{4\pi\epsilon_0 r^2}$$

This gives us an opportunity to compare the strengths of gravity and electromagnetism on the subatomic scale. For example, a hydrogen atom consists of a single electron

orbiting a proton, held in place by the Coulomb force. The size of the hydrogen atom is about 5×10^{-11} m, known as the *Bohr radius*. But there should also be atoms held together by the gravitational force. We could consider an electron orbiting a single neutron, held in place by Newton's force (5.6). The question is: how big would this gravitationally bound atom be?

The answer is pretty stunning: a gravitationally bound atom would have a size that is substantially larger than our observable universe. Gravity isn't just a little weaker than the other forces. It's much much weaker than the other forces. Another way of saying this is to take two electrons and compare their gravitational attraction to their Coulomb repulsion. The answer is

$$\frac{F_{\rm Newton}}{F_{\rm Coulomb}} \approx 10^{-43}$$

No one cares about the gravitational force acting on a single elementary particle.

However, gravity has a trick up its sleeve. All the other forces of nature have both positive and negative charges which cause particles to attract into effectively neutral objects on small distance scales. This happens most dramatically for the strong force, where quarks are confined into protons and neutrons. It then happens again for electromagnetism where electrons and protons are bound into neutral atoms. This means that by the time you get to the macroscopic world in which we live, these forces have done their job and their effects are no longer manifest. In contrast, there is no negative mass and so nothing to shield us from the effect of gravitation. As you pile more and more particles together, the Coulomb force becomes diluted, while the gravitational force only grows. This is why gravity is the force that seems to dominate our lives.

We do have a good understanding of why gravity is special in this way. The three forces in the Standard Model are associated to spin 1 fields. (The Higgs force, of course, is associated to a spin 0 field.) Gravity is unique because it is associated to a spin 2 field. And in contrast to all other forces, spin 2 fields can only have "positive charge", where the charge is what we call mass.

Spin 2 fields are, it turns out, special in many ways. While we could conceive of theories with many different spin 1 forces, that's not possible for spin 2 fields. There is just one way to introduce a fundamental spin 2 field into the laws of physics and it is utterly remarkable. The spin 2 field, it turns out, must be spacetime itself!

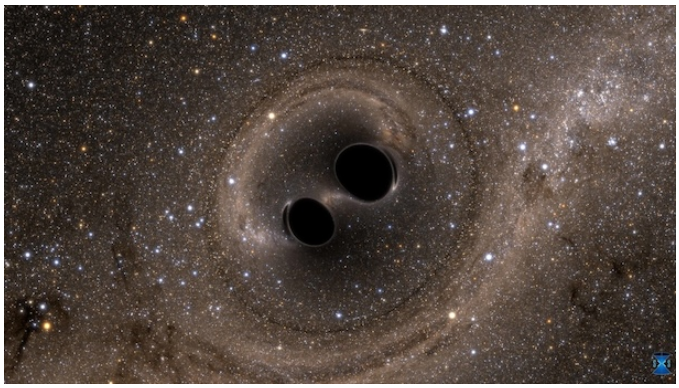
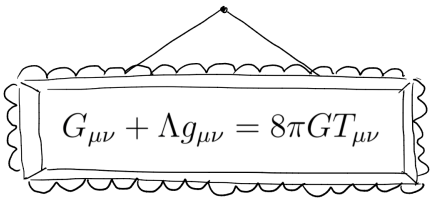


Figure 53. A computer simulation of two black holes, about to collide. The background field of stars is distorted by the curvature of spacetime in the vicinity of the black holes.

The connection between space, time and gravity is Einstein's great insight and the resulting theory goes by the name of General Relativity. It would take us too far from our main narrative to describe general relativity in any great detail, but if any equation deserves being placed in a picture frame, it is Einstein's:



This equation replaces Newton's gravitational force law (5.6). It relates the curvature of spacetime, captured in the $G_{\mu\nu}$ and $g_{\mu\nu}$ on the left-hand side, to the energy distribution of other fields, captured in the $T_{\mu\nu}$ on the right-hand side. There are two gravitational constants in the Einstein equations. The first is Newton's constant G, which retains its place as a fundamental constant of nature in general relativity. The second, Λ, is known as the *cosmological constant*. We will have more to say about this in Section 5.3.1.

The Einstein equations replicate the familiar results of Newton's force law, from apples falling from trees, to the orbits of the planets around the sun. But they do so much more besides. They tell us how light bends as it passes heavy objects, giving rise to the distortion of the background field of stars, how black holes form as the density of matter becomes too great, and how collisions of these black holes can cause detectable ripples of the spacetime continuum known as gravitational waves. Furthermore, the

Einstein equations provide, for the first time, a framework in which we can think about the dynamics of the entire universe, giving rise to the field of Cosmology. For any physics on the very largest scales, the Einstein equations hold sway.

5.2.1 Quantum Gravity

Above I said that there are two reasons that gravity is not included in the Standard Model. The first is that gravity is inconsequential for particle physics. The second is that we don't really know how to do it.

General relativity is a classical theory. It does not incorporate the effects of quantum mechanics. In contrast, everything else that we've described in these lectures is all about quantum mechanics. The question is: can we incorporate general relativity into the quantum world? This is the problem of *quantum gravity*.

It's usually stated that naive attempts to combine quantum mechanics and general relativity fail miserably. That's not quite true. Naive attempts to combine the two theories actually work very well. Up to a point. For example, any simple minded approach to quantum gravity clearly shows that, on some level, spacetime acts just like any other quantum field. Small fluctuations of spacetime are tied up into little knots by the framework of quantum mechanics, resulting in a new kind of massless particle known as a *graviton*. Just as light waves are made of many underlying photons, so too are gravitational waves made of gravitons.

We've never detected an individual graviton. Indeed, because gravity is so very weak we only succeeded in detecting gravitational waves in 2015. Building an experiment to see individual gravitons is not technologically feasible in the forseeable future. To put this in perspective, we first detected light when eyes evolved. Fossil records clearly show the existence of eyes in the Ediacaran period and from then it took something like 550 million years before we were able to detect the underlying photons. I'm not promising that it will take us a similar timescale to detect gravitons, but it's as good a guess as any.

Although it's something of an academic exercise, we can also compute Feynman diagrams for gravitons to see they interact with each other under the rules of quantum mechanics. Again, all is good, but now there's a catch. The calculations make sense provided that the energies of the gravitons are not too large.

We can actually anticipate the problem from the few facts that we know about gravity. The strength of the gravitational force is governed by Newton's constant (5.7). This is what plays the role of the fine structure constant in electromagnetism or

analogous quantities for the strong and weak force. Except there's a crucial difference: the strengths of all the other forces are governed by dimensionless numbers. In contrast, there's no way to get a dimensionless number out of Newton's constant. Instead, if we include some factors of $\hbar$ and c, we get a scale:

$$G = \frac{\hbar c}{8\pi M_{\rm pl}^2}$$

where $M_{\rm pl}$ is the *Planck mass*,

$$M_{\rm pl} \approx 10^{18}\ {\rm GeV}$$

This is telling us that the strength of the gravitational force depends strongly on the energy of the process. If two particles scatter with energy E, then gravitational effects will scale as

$$\text{Strength of Gravity} \sim \left(\frac{E}{M_{\rm pl}}\right)^2 \sim GE^2$$

In some sense, you knew this already. Newton's equation (5.6) tells us that the gravitational force between two objects scales as Gm_1m_2 which is just the formula above with the masses in place of the energy. This formula also gives another perspective on why gravitational forces are so weak in the world of particle physics, where our most powerful accelerators can reach energies of $E \sim 10^3$ GeV, many orders of magnitude below the Planck scale.

This argument means that any naive theory of quantum gravity will give sensible answers providing that we look at energies $E \ll M_{\rm pl}$, where we can use Feynman diagrams and simple perturbation theory. But, as the energy increases to somewhere near the Planck scale, the gravitational interaction becomes strong and our Feynman diagram expansion ceases to work. In this simple minded approach, quantum gravity only becomes challenging when we reach energies close to the Planck scale. In technical, and slightly old-fashioned, terms general relativity is said to be *non-renormalisable*.

Before we proceed, it's worth thinking a little more about the size of the Planck scale. Back at the beginning of these lectures, I commented that the SI unit for energy, the Joule, is not particularly useful in the subatomic realm. Instead we define the much smaller unit of an electronvolt. But, by the time we get to the Planck scale, even the Joule is too small! We have

$$M_{\rm pl} \approx 10^9\ {\rm J}$$

That's a seriously large amount of energy. It is, for example, greater than the kinetic energy of all the cars in a formula one race! In particle physics, when we talk about reaching a certain energy scale we really mean energy density. Physics at the Planck scale involves energies of $M_{\rm pl}$ squeezed into a region of size $L_{\rm pl}^3$, where $L_{\rm pl} \approx 10^{-34}$ m. These are silly numbers. There is, as we will see below, good reason to believe this is the highest energy density allowed by the laws of physics.

An Analogy: Fermi's Theory of the Weak Force

We've encountered a situation very similar to gravity elsewhere in these lectures. Recall from Section 4.2.3 that Fermi's original theory of the weak force included a dimensionful coupling known as the Fermi constant (4.4), whose value is roughly

$$G_F \approx \frac{\alpha_W}{M_W^2} \tag{5.8}$$

where $\alpha_W \approx 1/30$ while $M_W \approx 80$ GeV is the W-boson mass. If you don't know about W- and Z-bosons, and work purely within the Fermi theory then everything is fine provided that you only look at low energy processes. But, from this perspective, the strength of the weak force depends on energy as

$$\text{Strength of Weak Force} \sim \alpha_W \left(\frac{E}{M_W}\right)^2 \sim G_F E^2$$

As you approach energies closer to M_W then Fermi's theory stops working and must be replaced by something else. That something else is, of course, the W- and Z-bosons.

If we take this lesson seriously, it suggests a similar fate for gravity, with general relativity the low-energy approximation to something more fundamental. If we were to push the analogy with (5.8) yet further, we might expect that Newton's constant arises from a combination of a dimensionless coupling $\alpha_{\rm grav}$ and a fundamental scale $M_{\rm grav}$, so that

$$G_N \approx \frac{\alpha_{\rm grav}}{M_{\rm grav}^2}$$

with general relativity breaking down as we approach energies $M_{\rm grav}$. In such a scenario, we would see some new physics emerging at the energy scale $M_{\rm grav}$, possibly manifesting itself as new particles. Note that, if this is the way things pan out, then we can't be sure about the scale $M_{\rm grav}$ because it's related to $M_{\rm pl}$ by a dimensionless coupling $\alpha_{\rm grav}$. Assuming that $\alpha_{\rm grav}$ isn't ridiculously small, it would appear natural that $M_{\rm grav}$ is somewhere in the ballpark of the Planck scale $M_{\rm pl}$.

The discussion above is, appropriately for these lectures, very much a particle physicist's perspective on quantum gravity, focussing on scattering of gravitons and what happens at high energies. But Einstein's theory is, at heart, a theory of geometry and this suggests different approaches to quantum gravity where, for example, spacetime sits in a superposition of all possible geometries, much like a particle in a double slit experiment follows all possible trajectories. This brings new conceptual issues to the table, but the problem described above remains, now in the guise of difficulties when spacetime fluctuates on very small scales,

$$L_{\rm pl} = \frac{1}{M_{\rm pl}} \approx 10^{-34}\ {\rm m}$$

It may well be that the very notion of space and time need replacing when we get to such small distance scales.

The Unreasonable Effectiveness of Classical Gravity

The analogy with Fermi's theory of the weak force gives us a useful way of thinking about what a dimensionful coupling, like G_N, means in quantum theories. But general relativity is a considerably deeper and more subtle theory than Fermi's and it has a number of tricks up its sleeve.

In particular, Fermi's theory of the weak force only works for energies $E \ll M_W$. For any energies higher than that, you need to use the gauge theory of W- and Z-bosons. But that's not the way things work for gravity. General relativity will give you the right answer to any quantum question at energies $E \ll M_{\rm pl}$. But if you throw together two particles at energies $E \gg M_{\rm pl}$, then general relativity will also give you the right answer. That's because, if you throw particles together at very high energies, then you simply form a black hole!

Black holes are nature's way of putting a limit on the amount of matter than can be squeezed into some small amount of space. They have two key features: at the centre of the black hole is the *singularity*. According to general relativity, this is a region where the curvature of spacetime becomes infinitely large. What that really means is that we shouldn't trust general relativity near the singularity and it should be replaced by a theory of quantum gravity. But, happily, we don't need a theory of quantum gravity to understand many features of black holes because the singularity is shielded from view by the second feature of a black hole: the *event horizon*. This is a surface that can be thought of as the edge of the black hole. If you venture through the event horizon of a black hole then you're in trouble: you will never escape and will be dragged inexorably towards the singularity. In this scenario, you will need to very

quickly develop a fully fledged theory of quantum gravity if you want to know what fate awaits you. If, however, you are less foolhardy and remain outside the event horizon then you can get by quite happily with our naive theory of quantum gravity.

That's not to say that there aren't interesting quantum gravity effects associated to the event horizon. If you study quantum field theory in the presence of a black hole, you find *Hawking radiation*, the process in which the black hole emits (mostly) photons and gravitons and slowly evaporates. It is a striking effect, and not without its own puzzles, but these too can seemingly be understood largely within a naive approach to quantum gravity, without the need for the full theory at the Planck scale.

The size of the event horizon is called the *Schwarzschild radius*. A black hole of mass M has Schwarzschild radius

$$R_s = 2GM \sim \frac{M}{M_{\rm pl}^2}$$

This simple equation turns one of the key ideas of particle physics on its head. Throughout these lectures, going to higher and higher energies has been tantamount to looking on smaller and smaller distance scales, with energy and distance related, in natural units, by $\lambda \sim 1/E$. But, when gravitational effects become important, we see that going to higher energies gives rise to bigger black holes. If we scatter two particles at energies $E \gg M_{\rm pl}$, then we make a black hole of size $R \sim E/M_{\rm pl}^2 \gg L_{\rm pl}$. This is how nature evades energy densities smaller than the Planck scale.

This means that, provided we don't do anything stupid, like jump into a black hole, we understand perfectly well what happens in very high energy scattering. You form a big black hole which slowly evaporates over gazillions of years. We never need any knowledge of the fundamental theory of quantum gravity to figure out the physics.

All of this is to say that the applications of a full theory of quantum gravity to scattering appear strangely limited. It is needed only in a small window of energies $E \sim M_{\rm pl}$. For energies $E \ll M_{\rm pl}$, a naive theory of quantum gravity will do the job, and for energies $E \gg M_{\rm pl}$, classical general relativity will do the job, at least until the resulting black hole has had enough time to evaporate down to size $R \sim L_{\rm pl}$. In final stages of the black hole's life, we will again need to resort to a detailed theory of quantum gravity to understand what happens.

Of course, not all of physics is to do with scattering. As we mentioned above, if we really want to understand what happens inside a black hole at the singularity then we surely need quantum gravity. Furthermore, a very similar kind of singularity arises in

general relativity at the Big Bang. If we follow the evolution of the universe back in time, then general theorems of Hawking and Penrose tell us that we will necessarily meet a singularity where general relativity breaks down. This means that if we want to answer the all-important question of how the universe began then we surely need a decent theory of quantum gravity. But it is rather hard to find any lingering effects of the Big Bang in our current universe. This is isn't simply because it happened a long time ago. Rather it is because there was a period of rapid expansion in the early universe known as *inflation* which dilutes any specific effects due to the singularity. It appears, once again, that Nature wishes to thwart us in our attempts to see directly the effects of quantum gravity.

This is not to say that quantum gravity is uninteresting. The questions of how singularities are resolved, both cosmological and those inside black holes, are important ones. Moreover, a closer look at how quantum theory meshes with gravity opens up a number of subtle conceptual issues, including how information escapes from black holes, how we should think about observers in an accelerating universe, and the seeming holographic nature of gravity. Moreover, there is the striking fact that, in string theory, we do have a fully fledged theory of quantum gravity, one that has intricate connections with various quantum field theories. All of these, however, are topics for other lectures.

5.3 Cosmology

> *There are more things in heaven and earth, Horatio,*
> *Than are dreamt of in your Standard Model.*

Hamlet was wrong about the earth. As we've seen, we are painfully short of experiments that cannot be explained by the Standard Model. But he was right about the heavens. When we look to the sky, it becomes clear that there is much we don't understand.

Our shocking lack of understanding becomes apparent when we audit the energy budget of our universe. This reveals that the particles of the Standard Model comprise just 5% of the total energy. All the things that we can see – stars, galaxies, planets, light itself – and all the things that we can detect through more indirect means – inert dust, neutrinos, gravitational waves – make up just a tiny fraction of the energy that is out there.

The vast majority of the energy in the universe is *dark*, which simply means that it doesn't interact (as far as we can tell) with the particles in the Standard Model. This "missing" energy can be further characterised as having two very different properties:

- Roughly 25% of the energy of the universe comprises of some new particles (or, better, new fields) that are not included in the Standard Model. This is called *dark matter*.

- The remaining 70% is a cosmological constant, sometimes called *dark energy*.

For reasons that we won't get into here, there's reason to believe that the short list above accounts for everything. There are very likely surprises in store for us in both dark energy and dark matter, but there is not some extra energy out there that we've missed completely. That's because, in the framework of cosmology, the total energy doesn't actually have to add up to 100%! Any deviation from 100% shows up as some overall curvature to the universe. But the universe is, as far as we can tell, exactly flat and that tallies nicely with our other observations which show that the energy hits the magical 100% threshold.

In this final section, we will elaborate on some of the puzzles of cosmology that are likely to have the biggest impact on particle physics. We will be fairly brief. Many more details can be found in the lectures on Cosmology.

5.3.1 The Cosmological Constant

The vast majority of energy in the universe is rather strange. It is best described as an anti-gravity force field, spread thinly throughout space, causing everything to repel everything else. We refer to this force field as *dark energy*.

We can't detect this dark energy here on earth. Nor can we see its effects within our solar system, nor even within our galaxy. Instead it is only when we look out at vast distance scales when the effects of dark energy become apparent as it manifests itself in the way the universe expands.

We've known that the universe is expanding since the 1920s. This is a straightforward prediction of Einstein's equations of general relativity, albeit one that was not embraced by theorists until the observational data made the case overwhelming. But around the turn of the last century, we learned something striking. The expansion of the universe is *speeding up* over time. We live not just in an expanding universe, but in an accelerating universe. This is despite the fact that the galaxies in the universe are all mutually attracted to each other through gravity which should cause the expansion to slow. But there is something else at play that overwhelms the natural gravitation attraction of galaxies at very large scales. This something else is dark energy.

In some sense, we understand dark energy very well. The Einstein equations of general relativity allow for exactly one additional term, with a single parameter Λ known as the *cosmological constant*. This extra term has exactly the right effect, changing the dynamics of spacetime to give the observed acceleration of the universe. All we have to do is set the cosmological constant to take the value,

$$\Lambda \approx (10^{-29}\text{ eV})^2$$

A better way of thinking about this is in terms of the density of dark energy which, in natural units (where length is equivalent to inverse energy) has dimensions of energy to the power 4. To get this, we should multiply by $M_{\rm pl}$, giving the energy density

$$\rho_\Lambda \approx \Lambda M_{\rm pl}^2 \approx \left(10^{-3}\text{ eV}\right)^4 \tag{5.9}$$

Just plug these values of the cosmological constant into the Einstein equations and you're done: the expansion of the universe then matches perfectly with what we observe.

(Before we go on, I should mention that this last sentence might not be quite true. There is currently a discrepancy between two different methods of measuring the expansion of the universe. The first, and arguably the cleanest, looks at the cosmic microwave background radiation, the afterglow of the Big Bang, and infers how the universe has subsequently evolved. The second looks at distant supernovae which are bright enough (and, apparently, uniform enough) to accurately measure both their distance and their recession velocity. These two results give answers that differ that level of 10% or so and it is difficult to reconcile them in any straightforward way. It may be, of course, that there is some unknown systematic in one of the experiments that gives a skewed result. Or it may be that there is something deep going on that we've missed. This is known as the *Hubble tension*.)

The Energy of the Vacuum

Although we have the equations to describe the accelerated expansion of the universe, there is a level of disquiet about the cosmological constant. This derives from the fact that we actually have a good understanding of the physics underlying the cosmological constant and this seems to be in conflict with the observed value.

The cosmological constant is, it turns out, something very familiar in physics: it is the *vacuum energy*. Recall that we're taught in high school that the overall value of the energy doesn't matter. Instead, only energy differences are important. That's true in all situations except for in cosmology. Gravity responds to all kinds of energy including the energy of empty space and this appears as the cosmological constant in the Einstein equations where it affects the overall expansion of the universe.

In some ways, this is a good thing. The cosmological constant isn't just some random number that we've plucked out of thin air to account for the expansion of the universe. Instead it's something that should arise naturally from our other laws of physics. And this gives us the opportunity to calculate it from what we know about particle physics.

This is where the narrative takes something of a left turn. The vacuum energy in quantum field theory is something interesting. Recall that one of the characteristic features of quantum field theory is that the vacuum isn't a dull place: the quantum fields froth with quantum fluctuations. We showed an example of the quantum field vacuum way back in Figure 4. And all of those fluctuations contribute to the vacuum energy. If the vacuum has fluctuations on some maximum energy scale E, then we typically expect a ground state energy density of order E^4.

That's problematic. We know that the Standard Model of particle physics holds up to the energy scale of a TeV or so. But this strongly suggests that the vacuum energy density should be at least

$$\rho_{\rm SM} = (10^{12}\ {\rm eV})^4$$

This is not particularly close to the observed value. It is 60 orders of magnitude greater than the observed value ρ_Λ. More generally, the contribution of any quantum field to the vacuum energy density is naturally of order $\rho_{\rm QFT} \sim \Lambda_{\rm UV}^4$, where $\Lambda_{\rm UV}$ is the UV cut-off.

In the cosmological context, a vacuum energy density of order a TeV or higher gives ridiculous results. A cosmological constant this large would give a universe that expands so quickly that it is not conducive to forming nuclei or atoms, let alone galaxies and life. The huge discrepancy between the expected value of ρ_{QFT} and the observed value of ρ_Λ is known as the *cosmological constant problem*.

The cosmological constant problem is entirely analogous to the hierarchy problem that surrounds the Higgs mass that we discussed in Section 5.1.2. In both cases, quantum corrections seem to naturally push the value of some quantity to be much higher than we observe. This happens only for the Higgs mass and the cosmological constant because these are the only two dimensionful parameters in the known laws of physics. (Strictly speaking, we also have the Planck mass $M_{\rm pl}$, but this can be thought as setting the scale relative to which all other parameters are measured.)

The cosmological constant problem is sometimes referred to as the worst prediction in the history of physics. The people making this claim clearly haven't studied the

history of physics. (Rayleigh-Jeans law anyone?) Nonetheless, the prediction is clearly nothing to be proud of. The solution to the ultra-violet catastrophe inherent in the Rayleigh-Jeans law ultimately resulted in the greatest paradigm shift in the history of science: the discovery of quantum mechanics. It's not clear if the cosmological constant problem will result in a similar upheaval to our understanding of physics or whether, with some small twist of our head, the question will unravel as we realise that we've been looking at things the wrong way. For now, it's fair to say that we have just one solution to the cosmological constant problem. This is the idea that, in addition to the contribution from quantum field fields, there is another contribution to the cosmological constant, so that the two add up to give the observed value (5.9)

$$\rho_\Lambda = \rho_{\rm SM} + \rho_{\text{something else}}$$

Clearly the additional contribution $\rho_{\text{something else}}$ must cancel the contribution from the Standard Model to 60 significant figures, leaving behind the tiny cosmological constant that we observe. This is another example of fine tuning. It is the same kind of idea that we met in equation (5.5) when discussing the mass of the Higgs boson.

It is quite possible that there is some missing principle that we've failed to grasp that makes fine tuning less silly than it first appears. The task of finding such a mechanism is made considerably harder when we realise that there have been a number of times in the history of the universe when $\rho_{\rm SM}$ abruptly changed while, presumably, $\rho_{\text{something else}}$ did not. This occurs at a *phase transition*. For example, there was a time in the early universe when things were so hot that quarks and gluons were not confined inside baryons and mesons. As the universe cooled, confinement kicked in and, at that time, the vacuum energy of the Standard Model jumped by around $\Delta\rho_{\rm SM} \sim (100\ {\rm MeV})^4$. Still earlier, the electroweak phase transition, where the Higgs boson first condensed, resulted in a change of $\Delta\rho_{QFT} \sim (100\ {\rm GeV})^4$. This means that any putative cancellation mechanism must conspire to give a tiny cosmological constant ρ_Λ at the end of the life of the universe, not at the beginning.

Before we muddy the waters yet further, it's worth mentioning that the observed vacuum energy (5.9) is pretty much in the same ballpark as the neutrino masses. Is this coincidence? Probably. Certainly it's hard to know what to make of it. But, given our evident confusion about the cosmological constant, it's worth bearing in mind.

The A-Word

As we saw above, a naive application of quantum field theory suggests a ludicrous value for the cosmological constant, one that results in an expansion so fast that not even

atoms have a chance to form from their underlying constituents. Given this, we could ask the following question: what is the maximum value of the cosmological constant that still allows complex structures to evolve? For example, what is the maximum allowed value of Λ that allows galaxies to form?

It turns out that the upper bound on Λ depends on the strength of the initial seeds from which the galaxies grew. (We'll mention these briefly later in this section.) However, if we fix this initial condition, then we can ask again: how big can the cosmological constant be?

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

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

The anthropic explanation for the cosmological constant may be correct. But, in the absence of any testable predictions, it is not clear what to make of it and further philosophising tends to be more of a distraction than a help.

A Rebranding: Dark Energy

Given our manifest befuddlement about all things Λ, it is prudent to wonder if perhaps the accelerated expansion of the universe has nothing to do with a cosmological constant at all! It is quite possible that the cosmological constant in the Einstein equations is $\Lambda = 0$. If this is the case, then we need to look for another explanation for the accelerated expansion. It is not difficult to find such explanations, although none of them are particularly compelling. For example, a scalar field with a ridiculously low mass (around 10^{-33} eV or so), rolling down a potential can do the job should we wish.

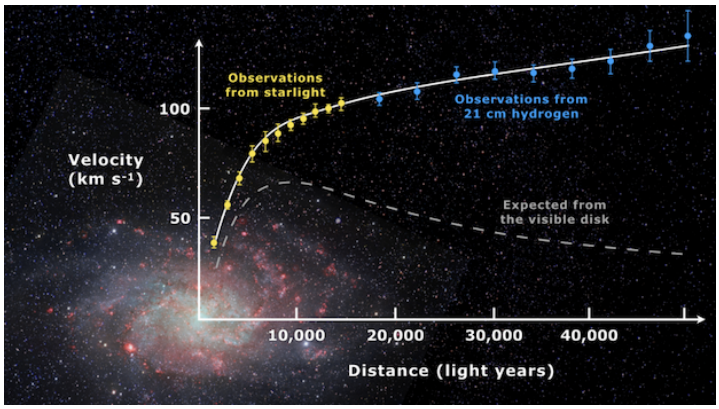


Figure 54. The rotation curve of galaxy M33. Image taken from Wikipedia.

I should stress that these new explanations in no way alleviate the cosmological constant problem. Finding a mechanism that sets the cosmological constant to zero is no easier than finding a mechanism that sets it to the observed value. Furthermore, one typically introduces many more fine tuning issues in whatever new dynamics is then introduced to drive the accelerated expansion.

Nonetheless, the history of particle physics has taught us that we shouldn't be too hasty in following our preconceived ideas about how the universe should be. To avoid committing to the cosmological constant explanation, the mysterious 70% of the energy in the universe is often referred to as *dark energy*.

5.3.2 Dark Matter

Dark matter comprises around 25% of the energy density of the universe. Unlike dark energy, it is conceptually quite straightforward: it is simply some new, heavy particles that are not accounted for in the Standard Model. Or, at a fundamental level, some new quantum fields.

We know very little about the properties of dark matter, beyond the fact that it does not interact with light. We do not know if it is a single species of particle, or many. We do not know if it consists of several decoupled sectors, or just one. Given the wonderful complexity of the Standard Model, it seems reasonable to assume that there is still rather a lot to learn about dark matter.

All we know about dark matter comes from its gravitational interactions. Yet the combined evidence is overwhelming. Here are some highlights:

Figure 55. On the left, the Abell S1063 cluster. The smeared blue lines are background galaxies, distorted by gravitational lensing. On the right, the bullet cluster.

- The beautiful spiral galaxies that we see in the sky seem to be spinning too fast! The attractive gravitational force from all the stars in the galaxy, does not come close to reproducing the necessary centripetal force to stop the galaxy from flying apart. Moreover, if you measure the spectral lines of hydrogen far from the visible edge of the galaxy, you find that it continues to rotate at a roughly constant speed for quite some distance. All of this can be explained by simple Newtonian dynamics, but only if there is much more mass in the galaxy than is visible. To account for the observations, there should be a roughly spherical cloud of dark matter surrounding the galaxy.

  The rotation curve for a nearby galaxy, together with the predicted curve if there is only the visible matter, is shown in Figure 54.

- A *galaxy cluster* is a collection of 100 to 1000 galaxies, bound together by gravity. A clever argument, known as the virial theorem, gives a relationship between the speed of the galaxies and their separation (or, more precisely, their kinetic energy and potential energy). From this, one can extract the mass of the galaxy. The answer is a couple of hundred times greater than the visible mass.

- A classic prediction of general relativity is that light bends as it passes heavy objects. Furthermore, the image gets distorted, a phenomenon known as *gravitational lensing*. Sometimes this happens in a spectacular fashion, as shown in the picture on the left of Figure 55 where the image of a background galaxy is distorted into the blue arcs by the cluster in the foreground. Even small distortions of this kind allow us accurately determine the mass of the cluster in the

foreground. You will not be surprised to hear that the mass greatly exceeds that seen in visible matter.

The bullet cluster, shown in the right of Figure 55, provides a particularly dramatic example of gravitational lensing. This picture shows two sub-clusters of galaxies which are thought to have previously collided. There are three types of matter shown in the picture: stars which you can see, hot gas which is observed in x-rays and is shown in pink, and the distribution of mass detected through gravitational lensing shown in blue. The stars sit cleanly in two distinct sub-clusters because individual galaxies have little chance of collision. In contrast, most of the ordinary matter sits in clouds of hot gas which interact fairly strongly as the clusters collide, slowing the gas and leaving it displaced from the stars as shown in the figure. But most of the matter, as detected through gravitational lensing, is dark and this, like the galaxies, has glided past each other seemingly unaffected by the collision. The interpretation is that dark matter interacts weakly, both with itself and with ordinary matter.

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

  The primary evidence comes from *Big Bang nucleosynthesis*, an impressively accurate theory of how the light elements were forged in the early universe. It turns out that the relative abundance of different elements depends on the total amount of baryon matter. In particular, the amount of deuterium compared to everything else depends in a delicate way on the total amount of ordinary matter. This tells us that the total amount of ordinary matter is just a few percent of the total energy density.

- In Section 5.3.4, we'll describe the cosmic microwave background, a photograph of the fireball that filled the universe when it was very much younger. The flickering of this fireball shows that some spots were hot and others cold. As the universe evolved, these initial fluctuations provided that seeds that later grew into the clusters, galaxies and stars that we see around us.

  It turns out, however, that this cannot be achieved by ordinary matter alone. There simply isn't time, largely because ordinary matter couples to photons and

this causes a pressure which suppresses gravitational collapse. But dark matter doesn't know about the photons, so there is nothing to stop it forming gravitational wells into which visible matter can subsequently fall. The whole story of the formation of galaxies only works because of the existence of dark matter.

Moreover, the existence of dark matter also leaves a distinct imprint in the ripples of the cosmic microwave background (specifically in the relative heights of the first and second peak of the power spectrum).

Taken individually, one might have thought that there could be some alternative explanation for these pieces of evidence. For example, faced just with the galaxy rotation curves, one might try to tinker with Newton's equations of motion to get something that fits. But that will then leave us unable to explain, say, how the light elements were forged in Big Bang nucleosynthesis. And yet all of the problems above are resolved by the simple admission that there are particles (or, strictly speaking, quantum fields) in the universe that are not accounted for in the Standard Model.

Clearly we should try to understand these fields and, ultimately, enlarge the Standard Model to embrace them. This is not so straightforward because all our evidence for dark matter comes from gravitational interactions alone. In an attempt to change this, there are many ongoing experiments designed to detect dark matter here on earth. All of these rely on the hope that there is some non-gravitational interaction between dark matter and Standard Model fields, perhaps through the weak force or perhaps through some new, as yet undiscovered force. These experiments are increasingly impressive at pushing the boundaries and one can only hope that they will one day make a key discovery.

There is one reason for optimism here. This is because the abundance of dark matter and ordinary matter is not wildly different. There is more dark matter by a factor of 5, but not a factor of 500 or 5 billion. The most plausible explanation for this is if dark matter and ordinary matter were in equilibrium together in the early universe, before they subsequently decoupled. But such equilibrium can only be maintained if there are some non-gravitational interactions between them.

In particular, there is one tantalising hint. If one assumes that dark matter has a mass of around the TeV scale associated to the weak force, and moreover interacts with the strength of the weak force, then the relative abundances come up just about right. However, if it were to interact directly through the weak force at this scale then we must ask why it hasn't shown itself at the LHC. Maybe this tantalising hint is merely a red herring.

5.3.3 Baryogenesis

The universe contains lots of matter but very little anti-matter. How did this asymmetry come to be?

One possibility is that it is an initial condition on the universe. Another is that the universe started with equal amounts of matter and anti-matter, but somehow a small dynamical shift took place that preferred one over the other. This latter process is known as *baryogenesis*.

We don't have an established theory of baryogenesis, but there are a set of three criteria that must be obeyed, known as the *Sakharov conditions*. These are:

- The first criterion is the most obvious: particle number cannot be a conserved quantity. Here "particle number" refers to particles minus anti-particles. In a symmetric universe, the total particle number would start off at zero. We want it to end up at something non-zero.

  In the Standard Model, both baryon number and lepton number are almost conserved although, as we saw in Section 4.3.5, in extreme conditions only $B - L$ is strictly conserved. The early universe certainly counts as an extreme condition. The need for baryogenesis suggests that we need interactions that break the symmetry $B - L$. For example, a Majorana mass for neutrinos will do the job.

- The symmetry CP must also be broken. As we've seen in Section 4.3.4, CP relates the behaviour of particles to anti-particles, but for baryogenesis to occur their behaviour must be different.

  We've seen that CP is violated in the quark sector, but this is not enough to give rise to the necessary level of baryogenesis. It remains to be seen whether CP violation in the lepton sector is sufficient to do the job, or whether baryogenesis requires interactions beyond those discussed in these lectures.

- The final criterion is the least obvious: the early universe must deviate from thermal equilibrium.

  A deviation from thermal equilibrium occurs when the universe undergoes a first order phase transition. (You can read more about phase transitions in the lectures on Statistical Physics and Statistical Field Theory.) The Standard Model does not appear to offer the opportunity for such a violent event at the necessary energy scales. This suggests that we should need some new physics to induce baryogenesis.

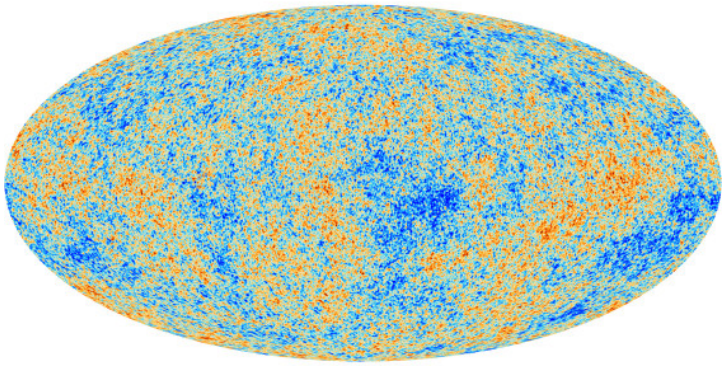


Figure 56. Look and weep, Ansel Adams.

There are many models of baryogenesis on the market, but currently no smoking gun experiment or observation that will determine which, if any, is correct.

5.3.4 Primordial Fluctuations

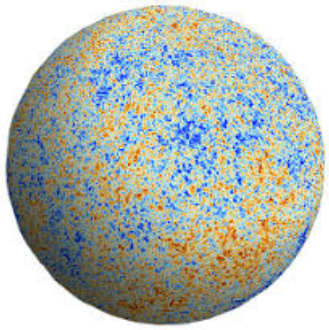
Most reasonable people agree that the greatest movie ever made is Ghostbusters. Sadly the world contains no small number of unreasonable people, those who prefer their movies to have a less intellectual bent, or those who put less stock in powerful acting performances and groundbreaking cinematography. It is difficult to argue that the opinion of these people is any less valid just because it is wrong. Art is not like science. There is no immutable, underlying truth that determines what is the right and wrong.

Until, that is, we come to photography. No one, reasonable or otherwise, can disagree about the greatest photograph ever taken. All other candidates pale into insignificance when faced with the collective endeavours of a bunch of radio horns and a handful of satellites who, between them, achieved the seemingly impossible feat of taking a photograph of the Big Bang.

First, I should tell you what the Big Bang theory entails. It is not a theory that tells us how the universe started. The question "how did the universe start?" has a very straightforward answer which is "we don't know". Instead, the Big Bang theory tells us what the universe was like when it was very much younger. The theory starts with the observation that there was a time — 13.8 billion years ago to be precise – when the universe was so hot that matter, atoms and even nuclei melted and all of space was filled with a fireball. When I say that we've taken a photograph of the Big Bang, I mean that we've taken a photograph of this fireball, capturing the light that

has travelled through the universe uninterrupted for almost 14 billion years. This light is known as the *cosmic microwave background*, or CMB for short.

The clearest photograph that we have was taken by the Planck satellite and is shown in Figure 56. This is a panoramic shot, containing information from each point in the sky which is then depicted in 2d much like a map of the Earth. A better setting for the CMB is shown in figure to the right. We sit in the centre of this sphere. If we look far enough away, roughly 20 billion light years in any direction, then we see the CMB.



In the early universe, the fireball reached extreme temperatures, almost certainly the most extreme temperatures the universe has ever seen. But, as the universe expanded, the fireball cooled and it is now a tepid 2.73 Kelvin. This temperature is almost uniform across the sky, but there are small fluctuations at the level of 1 part in 10^5. These hot and cold spots are depicted in red and blue in the photograph. These fluctuations have been imprinted in the CMB for 14 billion years, and so contain a wealth of information about what the universe was like when it was much younger. This information is usually plotted as a function of angular scale, as shown in Figure 57. From the positions of the peaks and troughs, we can determine much about the age and contents of the universe. For example, the position of the first peak contains information about the age of the universe, while the relative height of the first and second peaks contains information about the amount of dark matter in the universe.

Here, however, our interest is rather different. The question that we wish to ask is: where did those temperature fluctuations come from originally? This is question for which we're fairly confident that we have the answer. And it is nothing short of astonishing.

Inflation

The answer involves a process known as *inflation*, a period of rapid accelerated expansion when the universe was very young. Here "very young" means when the universe was, at most 10^{-11} second old, but most likely it occurred much earlier than this. (In this counting, 10^{-30} seconds counts as "much younger" than 10^{-11} seconds!)

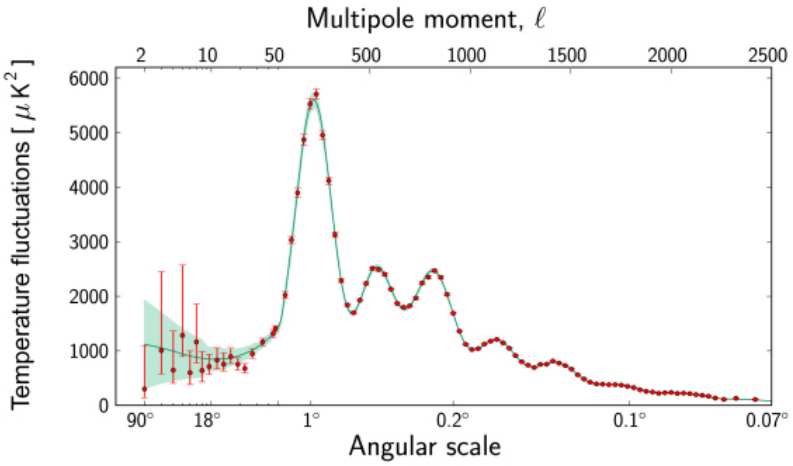


Figure 57. The dependence of the CMB temperature fluctuations with angular scale.

The original reasons for postulating the existence of an inflationary phase had nothing to do with the fluctuations. Instead, such a phase would resolve two unanswered questions about the universe we live in.

The first is: why is the universe so flat? It turns out that a flat universe is dynamically unstable, like a pencil balanced on its tip. Any small overall curvature at the beginning of the universe would have grown over time yet, after 14 billion, the universe appears as flat as a metaphorical and cosmological pancake. Why? Inflation gives an answer to this question. A brief phase of rapid expansion, stretches out any curvature that may have once existed, like pulling a membrane taut. This nicely explains why we find ourselves living in a flat universe.

The second is: why is the CMB so uniform? We can stare out at the sky in one direction and detect photons from the CMB, photons that have travelled roughly 20 billion light years uninterrupted. We can then turn around and see photons that have travelled 20 billion light years from the opposite direction. The properties of these photons are basically the same: in particular, both exhibit identical temperatures of 2.73 Kelvin. The fact that these two far flung reaches of the universe have the same temperature is a puzzle. Usually when two systems sit at the same temperature it's because they've had the opportunity to interact, exchange energy, and settle down to thermal equilibrium. But those two patches of the universe have had no such opportunity simply because they're too far away. Light from one region of space hasn't had

time to reach the other: indeed, we're sitting in the middle and it's only now that we can see both regions. So how could they possibly sit at the same temperature?

There is a more nuanced version of this second issue. This arises, somewhat ironically, when we appreciate that the CMB is not completely uniform after all but contains the tiny fluctuations shown in the photograph. There are fluctuations in both the temperature and the polarisation of light and, importantly, these two different types of fluctuations are correlated. These correlations – which go by the uninspiring name of "TE correlations" – are the kind of the type that would arise through simple and well understood dynamical processes in the early universe, such as photons scattering off electrons. But observations reveal that there are correlations over patches of the sky that were, apparently, never in causal contact with each other. That's a worry. Taken naively, it's telling us that there were dynamical processes in the early universe that occurred faster than the speed of light and that violates one of the key tenets of physics.

The TE correlations are, by far, the strongest argument for inflation. If we want to preserve some of our most cherished notions of physics, like locality and causality, then it tells us very clearly that those far flung patches of the sky *must*, in fact, have been in causal contact back in the day. Inflation is the mechanism that allows this to happen. In the very early universe, two patches of space could be near. But inflation then takes those patches and stretches them to enormous distances, until they sit on opposite sides of the observable universe. Yet they retain, in the CMB, the correlation that gives the game away that they were once playmates.

The arguments given above strongly suggest that inflation happened. The next question is: what caused it? Here we're on less sound footing. The good news is that when general relativity is coupled to quantum fields, one naturally gets the rapid expansion that we need for inflation. Indeed, we've already seen how to do it: the vacuum energy, or cosmological constant, does the job. The accelerated phase of inflation in the early universe is conceptually identical to the accelerated phase that we now find ourselves in, but with two differences. The first difference is quantitative: the effective cosmological constant in the early universe must have been many many orders of magnitude larger than what we see today. The second difference is that the original period of inflation must, ultimately, have come to a halt. The real challenge in constructing a viable model is therefore figuring out how to get inflation to stop!

This, it turns out, is not so difficult. Indeed, if there's bad news in the story of inflation, it is that it's *too* easy to write down models that do the job which means that

every Tong, Dick and Harry has their own theory of inflation, with hundreds now on the market and very little to decide between them. None of these models are particularly exotic and nearly all include the same basic ingredient: a scalar field. The idea is that the inflaton, unlike the Higgs boson, had not yet settled to the minimum of its potential energy in the early universe, so its value changes over time. Correspondingly, so too does the vacuum energy that drives inflation.

None of the models for inflation stand out as being overly compelling. Indeed, in many ways they all look somewhat artificial. One might wonder if perhaps we could identify the inflaton with the Higgs boson itself, and there have been attempts to do so, but it's not a particularly natural fit. This means that while there is good evidence that the process of inflation took place, our knowledge is limited when it comes to the detailed underlying dynamics. Before we beat ourselves up about this too much, it's worth remembering that we're talking about a process that happened something like 10^{-30} seconds after the Big Bang. The fact that we haven't yet got all the details pinned down isn't terribly surprising.

Inflationary Perturbations

All of which brings us to the main topic of this section: where did the ripples in the CMB come from? These ripples contain correlations over large distance scales which means that, if they are to have a local and causal origin, then they must have been laid down during the inflationary period itself. Happily, inflation provides a remarkable origin story for these fluctuations.

The ripples arise simply from the realisation that the inflaton is a quantum field. In fact, we started these lectures by explaining that the vacuum of space is not a dull place since the quantum fields cannot stay still: they froth and bubble with quantum jitters. During inflation, the universe expands so quickly that these quantum fluctuations get caught in the act and are stretched from the microscopic scale, to distances that span the entire visible universe. These are what we see imprinted as hot and cold spots in the CMB: they are nothing less than quantum fluctuations that took place just fractions of a second after the Big Bang and are then frozen in place by the rapid expansion of the universe.

This may be one of the most extraordinary ideas in all of science, connecting our understanding of physics on the very smallest scales with that on the very largest. It passes many checks. A statistical analysis of the CMB fluctuations shows that they agree perfectly with those expected from a weakly interacting quantum field. Moreover, as we get better data on the fluctuations, so we begin to get a handle on

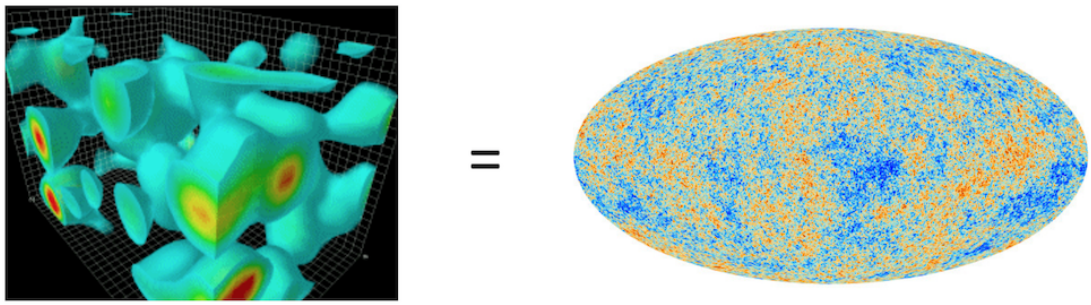


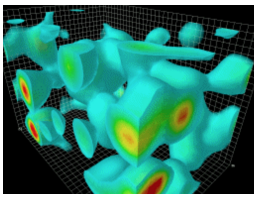
Figure 58. The ripples in the CMB are quantum vacuum fluctuations, laid down in the first few fractions of a second after the Big Bang.

the dynamics of the inflaton field in the very early universe. We're currently at a stage where many putative models of inflation can be ruled out, although there are many that still survive. The hope is that further study of the CMB will yield precious clues about the interactions of these quantum fields in the first few moments after the Big Bang.

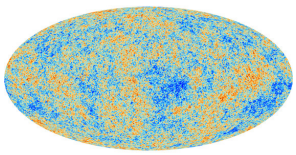
Remarkably, there is one further chapter to this story. Rather than asking where the ripples in the CMB came from, we could ask what subsequently happened to them. Here too we find an astonishing answer. The quantum fluctuations resulted in temperature variations in the CMB, with some places hotter and others colder. As the universe expanded and cooled, these hot and cold spots became the gravitational wells into which matter fell. First protons and electrons, which subsequently bound together into hydrogen dust and a smattering of other light elements. Over time, this dust gathered, and the pressure grew until finally, after 500 million years or so, these balls of dust ignited and became stars.

This means that the quantum fluctuations from when the universe was in its infancy later became the seeds from which galaxies grew. This is backed up by observation: a statistical analysis of the galaxies in our universe matches impressively with an analysis of the CMB fluctuations. For example, the large peak in Figure 57 manifests itself in a particular way in which galaxies cluster in the sky (known, boringly as "baryon acoustic oscillations").

Putting the pieces together, we can draw a direct line from this:



to this:



to this:

It is one of the most remarkable stories in all of science, but there are many details still to be written. Hopefully, in the future we will understand better how the quantum fields involved in this story fit with those of the Standard Model.