The Standard Model

University of Cambridge Part III Mathematical Tripos

David Tong

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

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

Recommended Books and Resources

For a very elementary introduction to the Standard Model, you could take a look at the lectures on Particle Physics that I wrote for the CERN summer school. They cover the subject in a great deal of detail, but without any real mathematical sophistication. If you're completely new to the wonderful world of subatomic particles, this is a good place to get grounded.

Many undergraduate degrees have courses on particle physics that use quantum mechanics and some elementary group theory, without fully embracing quantum field theory. There are a number of good textbooks catering to these courses. Two that I particularly like are:

- Halzen and Martin, "*Quarks and Leptons*"
- David Griffiths, "*Introduction to Elementary Particles*"

More advanced and really excellent books are:

- Cliff Burgess and Guy Moore "*The Standard Model*"
- Mark Thomson, "*Modern Particle Physics*"
- Matt Schwartz, "*Quantum Field Theory and the Standard Model*"

All three have different perspectives. Cliff and Guy's book in particular is closely aligned to the general theme of these lectures. Mark Thomson's book includes many more details about the specifics of particle interactions, while Matt's book is a great all-round QFT book that, as the title suggests, has an increasing focus on the Standard Model as it proceeds.

Finally, if you're serious about particle physics you should acquaint yourself with the all-important Particle Data Group. They have various apps that you can download and, for the more old-fashioned among you, books. Their booklet, available in the download section of the webpage, is particularly useful. They'll even mail you one for free if you ask nicely.

In addition, there are many online lecture notes. You can find links to these on the course webpage.

Contents

Acknowledgements

I'm grateful to Hugh Osborn and Fernando Quevedo who previously lectured a version of this course in Cambridge, and to Wati Taylor for sharing the notes of his MIT course with me. Many thanks to Ben Allanach for explaining various subtle (and less subtle) issues to me. I'd also like to thank Mike Ball and Cory Fletcher for their courageous typo spotting abilities and Elie Hamou for running the examples classes.

This course assumes a familiarity with quantum field theory. You will also need to be comfortable with some group theory.

0 Introduction

The "Standard Model" is the comically inadequate name that physicists give to the greatest scientific theory of all time.

This theory is the poster child for success in reductionist science. It describes the universe on the most fundamental level and correctly predicts the results of every experiment that we have ever done, sometimes with unprecedented levels of accuracy.

There are parts of the theory that are stunningly beautiful, with different facets sliding together like a perfect jigsaw, locked in place with a mathematical rigidity that means large parts of the world we inhabit could not be any other way. But there are other aspects of the theory that appear much less elegant, with a couple of dozen parameters that cannot be predicted from first principles but only by measuring them in experiment. These parameters don't appear to be completely random; there are patterns within them that surely hint at some structure that lies beyond the Standard Model, a structure that we have yet to uncover.

Boiled down to its essence, the Standard Model describes a bunch of particles, interacting with three forces. These forces are the strong nuclear force, the weak nuclear force, and electromagnetism. The force of gravity is not part of the Standard Model but it's straightforward to include it by coupling to a dynamical, curved spacetime. (Claims that the Standard Model is incompatible with general relativity are wildly overblown. The two theories work perfectly well together at all energy scales that we can currently probe by experiment. The difficulties only arise when energies approach the Planck scale.)

Each force in the Standard Model is associated to a Lie group. The upshot is that the Standard Model is built around the group

$$G = U(1) \times SU(2) \times SU(3) \ .$$

Why nature chose the numbers, 1,2, and 3 as the building blocks for her most important theory is not known, but you can't help but smile at the decision. Here $SU(3)$ is associated to the strong force and $SU(2)$ is associated to the weak force and $U(1)$ is not associated to electromagnetism but, instead, to an electromagnetic-like force known as *hypercharge*. It too plays a role in the weak force. The theory of electromagnetism that we know and love can be found hiding within the $SU(2) \times U(1)$ factor.

electron 1	down quark 9	up quark 4	electron neutrino $\sim 10^{-6}$
muon 207	strange quark 186	charm quark 2495	muon neutrino $\sim 10^{-6}$
tau 3483	bottom quark 8180	top quark 340,000	tau neutrino $\sim 10^{-6}$

Table 1. The fermions of the Standard Model

Despite the group theoretic similarities of each force, the resulting physics is wildly different. That's because quantum field theory is cool. It does wonderful and unexpected things. Part of the purpose of this course is to learn about these things and why the dynamics of the strong, weak and electromagnetic forces all play very different roles in our world.

These three forces interact with matter which, in the Standard Model, comes in the form of 15 Weyl fermions which, collectively, go by the name of the electron, the up quark, the down quark, and the neutrino. Why we give just four names to 15 fermions is part of the story that we will unravel, but at heart it is to do with representation theory of the group G.

At this point, one of the deepest facts about nature rears its head. The subtleties of quantum field theory mean that this quartet of particles – the electron, neutrino, and up and down quarks – have to come together as a collective. You don't have a choice. The theory with just, say, an electron and an up quark and no companions makes no sense. On grounds of mathematical consistency alone, we're obliged to have this quartet of particles with their particular properties. This is where some of the most beautiful aspects of the Standard Model can be found.

But then nature has a surprise, one which we've known about for almost a century and yet we are seemingly no closer to understanding. Nature took that collection of four particles and, for mysterious reasons, chose to replicate it twice over. This means that the matter in our world is not made of 15 fermions with four different names, but instead of 45 fermions with twelve different names. The names of these twelve particles are shown in Table 1 together with their masses, relative to the electron mass which is

$$m_e \approx 0.51 \text{ MeV} .$$

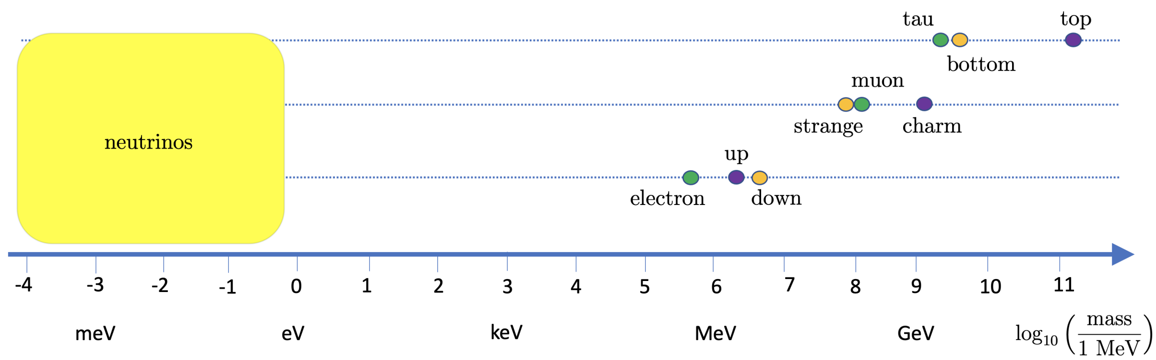


Figure 1. Again, the masses of the fermions of the Standard Model. Note that the ordering of particles in each generation is switched.

Each of the three rows in Table 1 is referred to as a different *generation*. The particles in each generation experience identical forces. So, for example, the electron, muon and tau all have electric charge -1, the down, strange and bottom quarks all have electric charge $-1/3$ and the up, charm and top quarks all have electric charge $+2/3$. All three neutrinos are neutral.

Similarly, the six quarks all experience the strong force in the same way, while the electron, muon, tau and neutrinos (which, collectively are referred to as *leptons*) are all untouched by the strong force.

The masses of the particles are replicated in Figure 1. They span at least 11 orders of magnitude, maybe more. (The masses of the neutrinos are not well constrained, as shown in the figure.) Why these particular masses? Why this ordering of masses? We have no idea. That's one of the outstanding questions that we hope might be answered by a deeper theory.

There is one final piece of the Standard Model that sits, lording over everything. This is the Higgs boson. It is, in many ways, the thing that ties everything together. In particular, all the masses listed above can be traced to the interactions of various fermions with the Higgs field.

The Higgs is simultaneously both the simplest and the most complicated field in the Standard Model. It is the simplest because it is the only fundamental (as far as we can tell!) scalar field that we have so far observed, meaning that it is the only field to carry zero spin. It is the most complicated because, in contrast to fermions and gauge fields, scalar fields don't come with many consistency requirements which means that there are a plethora of interaction terms that we can write down and the only way we have to constrain their values is to go out and measure them. It's here that we find the two dozen or so parameters that we can't yet explain. And it's here that things get messy and interesting.

This, then, is the Standard Model, part beauty, part beast. A glorious and astonishingly successful theoretical edifice that, so far, has stood firm against everything that experimenters have thrown at it. Yet few believe that it can really be the last word in physics. The Standard Model, like the periodic table before it, surely holds clues for what lies beyond. Our duty as physicists is to understand the Standard Model as best we can, to learn its secrets and, if possible, to let it guide us to a still deeper understanding of the world. The purpose of this course is to take you, at least part way, on this journey.

1 Symmetries

A large chunk of the structure of the Standard Model follows from understanding the various symmetries at play. Among these symmetries are

- Poincaré symmetries of spacetime, which restrict us to scalars, fermions, and gauge fields. These are the basic building blocks of the Standard Model.

- Gauge symmetries, better referred to as "gauge redundancies". These dictate the interactions of the spin 1 fields. Indeed, we've already seen that the Standard Model is usually advertised by specifying the gauge group

$$G = U(1) \times SU(2) \times SU(3) \ . \tag{1.1}$$

- Global symmetries. These act on the fermions and include baryon number and lepton number, as well as various approximate flavour symmetries.

- Discrete symmetries. Prominent among these are parity, time-reversal, and charge conjugation. These three symmetries are critically important in the structure of the Standard Model because, we shall see, none of them are actually good symmetries of our universe! But this is one case where not having symmetries puts even stronger constraints on the theory than having symmetries. This is because of something called "anomaly cancellation" that will be described in Section 4.

Of these, the various global symmetries arise because of the specific matter content of the Standard Model and so we will postpone a discussion of them until we have more details in place. (We'll first get there in Section 3 when we describe features of the strong force.) However, the other three symmetries – Poincaré, gauge, and discrete – are ingredients that arise in pretty much all relativistic field theories. For this reason, it makes sense to explore them in some detail in preparation for what's to come.

1.1 Spacetime Symmetries

On the length scales appropriate for particle physics, spacetime is effectively flat. This means that the arena for our story is Minkowski space $\mathbb{R}^{1,3}$, equipped with the Minkowski metric

$$\eta_{\mu\nu} = \text{diag}(+1, -1, -1, -1) \ . \tag{1.2}$$

We label a point in Minkowski space as $x^\mu = (x^0, x^1, x^2, x^3)$. The set of symmetries of Minkowski space include Lorentz transformations of the form $x^\mu \to \Lambda^\mu{}_\nu x^\nu$ where

$$\Lambda^T \eta \Lambda = \eta \ . \tag{1.3}$$

Embedded among these are a couple of discrete transformations: parity with $\Lambda = \text{diag}(1,-1,-1,-1)$ and time reversal with $\Lambda = \text{diag}(-1,1,1,1)$. These are important enough that we will discuss them separately in Section 1.4. The transformations that are continuously connected to the identity have $\det \Lambda = 1$ and $\Lambda^0{}_0 > 0$ and form the *Lorentz group* $SO(1,3)$. (The restriction to $\Lambda^0{}_0 > 0$ is sometimes written as $SO^+(1,3)$.)

Our main goal in this section is to understand some things about the representations of the Lorentz group and its extension to the Poincaré group which also includes spacetime translations. Among these representations, spinors are the most fiddly and subtle and we will describe some of their properties in Section 1.2.

1.1.1 The Lorentz Group

Strictly speaking, the group $SO(1,3)$ doesn't have any spinor representations. However, there is a closely related group called $\text{Spin}(1,3)$ that does admit spinors. This is the double cover, in the sense that

$$SO(1,3) \cong \text{Spin}(1,3)/\mathbb{Z}_2 \tag{1.4}$$

where that $\mathbb{Z}_2$ is related to the famous minus sign that spinors pick up under a 2π rotation, a minus sign that vectors like x^μ are oblivious to. The fact that there are spinors in our world is the statement that the true symmetry group is $\text{Spin}(1,3)$ rather than $SO(1,3)$.

The groups $\text{Spin}(1,3)$ and $SO(1,3)$ share the same Lie algebra $so(1,3)$. A Lorentz transformation acting on a 4-vector can be written as

$$\Lambda = \exp\left(-\frac{i}{2}\omega_{\mu\nu}M^{\mu\nu}\right) \tag{1.5}$$

where $\omega_{\mu\nu}$ are six numbers that specify what Lorentz transformation we're doing, while $M^{\mu\nu} = -M^{\nu\mu}$ are a choice of six 4×4 suitable matrices that generate the different Lorentz transformations. The matrix indices are suppressed in the above expressions; in their full glory we would write $(M^{\mu\nu})^\rho{}_\sigma$. So, for example

$$(M^{01})^\rho{}_\sigma = i\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad (M^{12})^\rho{}_\sigma = i\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}\,. \tag{1.6}$$

(Note that the generators differ by a factor of i from those defined in the Quantum Field Theory lectures. This is compensated by an extra factor of i in the exponent (1.5).) The matrices $M^{\mu\nu}$ generate the algebra $so(1,3)$,

$$[M^{\mu\nu}, M^{\rho\sigma}] = i\left(\eta^{\nu\rho}M^{\mu\sigma} - \eta^{\nu\sigma}M^{\mu\rho} + \eta^{\mu\sigma}M^{\nu\rho} - \eta^{\mu\rho}M^{\nu\sigma}\right)\,. \tag{1.7}$$

The six different Lorentz transformations naturally decompose into three rotations J_i and three boosts K_i, defined by

$$J_i = \frac{1}{2}\epsilon_{ijk}M_{jk} \quad \text{and} \quad K_i = M_{0i} \tag{1.8}$$

where the $j, k = 1, 2, 3$ indices are summed over, and $\epsilon_{123} = +1$. The rotation matrices are Hermitian, with $J_i^\dagger = J_i$ while the boost matrices are anti-Hermitian with $K_i^\dagger = -K_i$. This ensures that the rotations in (1.5) give rise to a compact group while the boosts are non-compact. From the Lorentz algebra, we find that these generators obey

$$[J_i, J_j] = i\epsilon_{ijk}J_k \ , \quad [J_i, K_j] = i\epsilon_{ijk}K_k \ , \quad [K_i, K_j] = -i\epsilon_{ijk}J_k \ . \tag{1.9}$$

The rotations form an $su(2)$ sub-algebra. That, of course, is to be expected and is related to the fact that $SO(3) \cong SU(2)/\mathbb{Z}_2$.

We can, however, find two mutually commuting $su(2)$ algebras sitting inside $so(1, 3)$. For this we take the linear combinations

$$A_i = \frac{1}{2}(J_i + iK_i) \quad \text{and} \quad B_i = \frac{1}{2}(J_i - iK_i) \ . \tag{1.10}$$

Both of these are Hermitian, with $A_i^\dagger = A_i$ and $B_i^\dagger = B_i$. They obey

$$[A_i, A_j] = i\epsilon_{ijk}A_k \ , \quad [B_i, B_j] = i\epsilon_{ijk}B_k \ , \quad [A_i, B_j] = 0 \ . \tag{1.11}$$

But we know all about representations of $SU(2)$: they are labelled by an integer or half-integer $j \in \frac{1}{2}\mathbb{Z}$ which, in the context of rotations, we call "spin". The dimension of the representation is then $2j + 1$. The fact that we can find two $su(2)$ sub-algebras of the Lorentz algebra tells us that all representations must carry two such labels

$$(j_1, j_2) \quad \text{with} \quad j_1, j_2 \in \frac{1}{2}\mathbb{Z} \ . \tag{1.12}$$

Moreover, we know that this representation must have dimension $(2j_1 + 1)(2j_2 + 1)$. We'll flesh out the meaning of these representations more below. But for now, we can identify the simplest such representations just by counting: we have

$$\begin{aligned}
(0,0) :& \quad \text{scalar} \\
(\tfrac{1}{2},0) :& \quad \text{left-handed Weyl spinor} \\
(0,\tfrac{1}{2}) :& \quad \text{right-handed Weyl spinor} \\
(\tfrac{1}{2},\tfrac{1}{2}) :& \quad \text{vector} \\
(1,0) :& \quad \text{self-dual 2-form} \\
(0,1) :& \quad \text{anti-self-dual 2-form}
\end{aligned} \tag{1.13}$$

What we call the physical spin of a particle is the quantum number under rotations $\vec{J}$: this is $j = j_1 + j_2$. The spin-statistics theorem ensures that particles with $j \in \mathbb{Z}$ are bosons, while those with $j \in \mathbb{Z} + \frac{1}{2}$ are fermions.

There's something a little odd about our discovery of two $su(2)$ sub-algebras. After all, it certainly isn't true that the Lorentz group is isomorphic to two copies of $SU(2)$. This is because $SU(2)$ is a compact group: keep doing a rotation and you will eventually get back to where you started. Indeed, two copies of the group $SU(2)$ give the rotation group of Euclidean space $\mathbb{R}^4$:

$$\text{Spin}(4) \cong SU(2) \times SU(2) \quad \text{with} \quad SO(4) \cong \text{Spin}(4)/Z_2 \ . \tag{1.14}$$

In contrast, the Lorentz group is non-compact: keep boosting and you get further and further from where you started. How does this manifest itself in the two $su(2)$ algebras that we've found in (1.11)?

The answer is a little subtle and is to be found in the reality properties of the generators A_i and B_i. Recall that all integer, $j \in \mathbb{Z}$, representations of $SU(2)$ are real, while all half-integer spin, $j \in \mathbb{Z} + \frac{1}{2}$, are pseudoreal (which means that, while not actually real, the representation is isomorphic to its complex conjugate). However, the A_i and B_i in (1.11) do *not* have these properties. You can see in (1.6) that both J_i and K_i are pure imaginary. This, in turn, means that the generators A_i and B_i are complex conjugates of each other

$$(A_i)^\star = -B_i \ . \tag{1.15}$$

This is where the difference lies that distinguishes $SO(4)$ from $SO(1,3)$. The Lie algebra $so(1,3)$ does not contain two, mutually commuting copies of the real Lie algebra $su(2)$, but only after a suitable complexification. This means that certain complex linear combinations of the Lie algebra $su(2) \times su(2)$ are isomorphic to $so(1,3)$. To highlight this, the relationship between the two is sometimes written as

$$so(1,3) \cong su(2) \times su(2)^\star \ . \tag{1.16}$$

For our purposes, it means that the complex conjugate of a representation (j_1, j_2) exchanges the two quantum numbers

$$(j_1, j_2)^\star = (j_2, j_1) \ . \tag{1.17}$$

Both the scalar representation $(0,0)$ and the vector representation $(\frac{1}{2}, \frac{1}{2})$ are real, while the left- and right-handed Weyl spinors $(\frac{1}{2}, 0)$ and $(0, \frac{1}{2})$ are exchanged under complex conjugation. This last statement, which is important, will be elaborated upon in Sections 1.2 and 1.4. In the context of quantum field theory, if a field appears in a theory then so too does its complex conjugate. This means that if you have a left-handed spinor, you also have a right-handed complex conjugated spinor.

1.1.2 The Poincaré Group and its Representations

The continuous symmetries of Minkowski space comprise of Lorentz transformations together with spacetime translations. Combined, these form the *Poincaré group*. Spacetime translations are generated, as usual, by the momentum 4-vector P^μ. Their commutation relations with themselves and with the Lorentz generators $M^{\mu\nu}$ are given by

$$[P^\mu, P^\nu] = 0 \quad \text{and} \quad [M^{\mu\nu}, P^\sigma] = i\left(P^\mu \eta^{\nu\sigma} - P^\nu \eta^{\mu\sigma}\right) \ . \tag{1.18}$$

The latter of these is equivalent to the statement that P^μ transforms as a 4-vector under Lorentz transformations. These commutation relations should be considered in conjunction with the Lorentz algebra (1.7),

$$[M^{\mu\nu}, M^{\rho\sigma}] = i\left(\eta^{\nu\rho} M^{\mu\sigma} - \eta^{\nu\sigma} M^{\mu\rho} + \eta^{\mu\sigma} M^{\nu\rho} - \eta^{\mu\rho} M^{\nu\sigma}\right) \ . \tag{1.19}$$

Together, (1.18) and (1.19) form the algebra of the Poincaré group.

Given an algebra, our next task is to explore its representations. There are different ways that we could approach this. Ultimately, we will be interested in the way that the Poincaré group acts on fields that make up the Standard Model. But first, to build some intuition, we will understand how the Poincaré group acts on single particle states in the Hilbert space.

To set the scene, let's first recall how we construct irreducible representations of the rotation group. We work with the algebra $so(3) \cong su(2)$ rather than the group. This is, of course, defined by the familiar commutation relations

$$[J_i, J_j] = i\epsilon_{ijk} J_k \ . \tag{1.20}$$

To construct representations, the first thing we do is look to the Casimirs. These are operators that commute with all generators of the group. For $su(2)$, there is just a single Casimir,

$$C = \sum_{i=1}^{3} J_i^2 \ . \tag{1.21}$$

Irreducible representations are labelled by the eigenvalue of the Casimir. For $su(2)$, the eigenvalue of J^2 is $j(j+1)$ with the spin j taking values in $j = 0, \frac{1}{2}, 1, \ldots$. Each representation has dimension $2j+1$, with the states within a multiplet identified by their eigenvalue under, say, J_3 whose eigenvalue lies in the range $|j_3| \leq j$. The result is the familiar one from quantum mechanics: states are labelled by two quantum numbers $|j, j_3\rangle$

Now let's turn to the Poincaré group. The irreducible representations are what we call "particles". Again, they are characterised by the Casimirs. I won't tell you how to construct Casimirs, but will instead just present you with the result. First, we introduce the *Pauli-Lubański vector*,

$$W^\mu = \frac{1}{2}\epsilon^{\mu\nu\rho\sigma} P_\nu M_{\rho\sigma} \ . \tag{1.22}$$

This can be thought of as a relativistic version of angular momentum. You can easily check this commutes with momentum $[W_\mu, P_\nu] = 0$. The remaining non-trivial commutation relations are somewhat more laborious to show:

$$[W_\mu, M_{\nu\rho}] = i(\eta_{\mu\nu} W_\rho - \eta_{\mu\rho} W_\nu) \quad \text{and} \quad [W_\mu, W_\nu] = -i\epsilon_{\mu\nu\rho\sigma} W^\rho P^\sigma \ . \tag{1.23}$$

The last of these commutation relations is quadratic on the right-hand side and so we're not looking at a Lie algebra here, but something more complicated. (This is reminiscent of the Runge-Lenz vector which is a conserved quantity for the Kepler problem; there too, the Poisson bracket structure returns something quadratic on the right-hand side.)

The two Casimirs of the Poincaré group are formed from the momentum P_μ and the Pauli-Lubański vector W_μ,

$$C_1 = P_\mu P^\mu \quad \text{and} \quad C_2 = W_\mu W^\mu \ . \tag{1.24}$$

This is our starting point: representations of the Poincaré group are labelled by the eigenvalues of C_1 and C_2, together with the eigenvalues of any other operators that we can find to make a maximally commuting set, analogous to J_3 for the angular momentum.

The most important of these "other operators" is the momentum P^μ itself. All states will be labelled by the eigenvalue p^μ which is simply the 4-momentum of the particle. The first Casimir is then just the rest mass of the particle, $C_1 = p_\mu p^\mu = m^2$. By acting with rotations and boosts $M_{\mu\nu}$, we can change the momentum to take any value subject to the constraint $p_\mu p^\mu = m^2$. In the rotation analogy, the different values of p^μ are like the different values of j_3 in the multiplet. However, in contrast to rotations, representations of the Poincaré group will necessarily be infinite dimensional, labelled (among other things) by the continuous variable p^μ. This difference can be traced to the fact that the Poincaré group is non-compact while the rotation group is compact.

What happens next depends on whether we're dealing with massive or massless particles. We describe each in turn, followed by a somewhat mysterious massless representation that no one really knows what to make of.

Massive Representations

First, consider the situation when $C_1 = m^2 \neq 0$. It's fruitful to pick a representative value of the momentum p^μ and the simplest choice is to boost to the rest frame of the particle so that $p^\mu = (m, 0, 0, 0)$. In this frame, the Pauli-Lubański vector is

$$W^0 = 0 \quad \text{and} \quad W^i = -mJ^i \ . \tag{1.25}$$

with J^i the generators of rotations. Note that the rotation generators J^i are precisely those elements of the Lorentz algebra that *don't* change the value of our chosen momentum $p^\mu = (m, 0, 0, 0)$. That means that these generators J^i must act on whatever other degrees of freedom are carried by the particles. We want to ask: what are the allowed extra degrees of freedom?

But this is a question that we already answered above because our problem has reduced to finding a representation of the Lie algebra $su(2)$, generated by J^i. The second quadratic Casimir of the Poincaré group is $C_2 = -m^2 J^2$ and so is specified by the eigenvalue of J^2 which, as we reviewed above, is $j(j+1)$ for some $j \in \frac{1}{2}\mathbb{Z}$. The full multiplet is then filled out by the different values of j_3 with $|j_3| \leq j$.

We've seen that, if we fix the momentum to the specific value $p^\mu = (m, 0, 0, 0)$, then we're left with finding representations of the rotation group. But, importantly, it doesn't matter which value of the momentum we started with: had we picked a different p^μ (still with $p_\mu p^\mu = m^2$), then we'd have got the same result. This suggests that we can lift the $SU(2)$ representation that we found for our given p^μ to a representation of the full Poincaré group. And, indeed, this is the case.

There is a theorem underlying this result which we won't prove. Instead, I'll just give you some names of things. Once we fix the momentum p^μ, the elements of the Lorentz group that don't change p^μ form a group known as the *little group*. For massive particles, the little group is $SU(2)$. One can then show that representations of the little group uplift to representations of the full Poincaré group. This is what's known as an *induced representation*.

The upshot is something familiar: massive particles are characterised by their mass m and spin j. Given these Casimirs, states in this representation of the Poincaré group are labelled by $|p_\mu, j_3\rangle$.

Massless Representations

The story is slightly different for massless particles, for which the first Casimir vanishes: $C_1 = m^2 = 0$. We again choose a representative momentum. This time we can't boost to the rest frame, but we can choose the momentum to take the form $p^\mu = (E, 0, 0, E)$ where E is the energy of the particle. A short calculation shows that, in this frame, the Pauli-Lubański now takes the form

$$W_\mu = E \begin{pmatrix} -M_{12} \\ M_{23} - M_{02} \\ M_{31} + M_{01} \\ M_{12} \end{pmatrix} = E \begin{pmatrix} -J_3 \\ J_1 - K_2 \\ J_2 + K_1 \\ J_3 \end{pmatrix} \; . \tag{1.26}$$

Here we've replaced the $M_{\mu\nu}$ with the appropriate rotation generator J_i or boost generator K_i defined in (1.8). Once again, each of the components of W_μ leaves our initial momentum $p^\mu = (E, 0, 0, E)$ unchanged, a fact that you can check by looking at the explicit form of the generators (1.6). In other words, these components of W_μ are once again our little group. (This has happened twice now and it is no coincidence: the structure of the Pauli-Lubański vector was designed so that this holds.)

What group do the components of W^μ actually generate? We can look at their commutation relations which, using (1.9), are

$$[W_1, W_2] = 0 \; , \quad [W_3, W_2] = -iEW_1 \; , \quad [W_3, W_1] = iEW_2 \; . \tag{1.27}$$

This is the Euclidean group in $\mathbb{R}^2$, sometimes written as $ISO(2)$, with W_1 and W_2 the generators of translations and W_3 the generator of rotations. Again, the little group doesn't act on our chosen $p^\mu = (E, 0, 0, E)$, but it may act on any other degrees of freedom that our state carries. Said differently, those other degrees of freedom must fall into a representation of the 2d Euclidean group.

Here a subtlety rears its head. For reasons that we will explain below, things turn out to be simplest if we consider representations of the little group on which the translation generators W_1 and W_2 act trivially. If we ignore these translations, the remaining little group is just the $U(1)$ of rotations generated by J_3. Representations of this $U(1)$ are labelled by a single eigenvalue h such that the states transform as

$$e^{i\theta J_3}|h\rangle = e^{ih\theta}|h\rangle \; . \tag{1.28}$$

The eigenvalue h is called the *helicity* and is the analog of spin for massless particles. At times, we'll be lazy and just refer to both as "spin". For a general null p, the helicity tells us the eigenvalue of the state under a rotation along the direction of motion,

$$e^{i\theta\,\hat{\mathbf{p}}\cdot\mathbf{J}}|p_\mu;h\rangle = e^{ih\theta}|p_\mu;h\rangle\ . \tag{1.29}$$

Because the $U(1)$ generated by J_3 was a subgroup $U(1) \subset SU(2)$, we know that this helicity is quantised to take values

$$h \in \frac{1}{2}\mathbb{Z}\ . \tag{1.30}$$

This is the statement that, under a rotation of $\theta = 2\pi$, the states are either left the same (for $h \in \mathbb{Z}$) or pick up a minus sign (for $h \in \mathbb{Z} + \frac{1}{2}$).

There's something missing in the story above. For massive representations, we've seen that the states are labelled by m and j and fill out a multiplet $|p_\mu, j_3\rangle$ with $|j_3| \leq j$. This multiplet has dimension $2j + 1$. (Ok, the multiplet is really infinite dimensional because of the p_μ, but for a fixed p_μ the multiplet has dimension $2j + 1$.)

However, for massless particles there is just a single state $|p_\mu; h\rangle$. This is because the helicity describes the representation of the Abelian group $U(1)$ generated by J_3 rather than the non-Abelian group $SU(2)$ and irreducible representations of Abelian groups are one-dimensional.

The problem with this is that it doesn't fit with what we know about massless particles. For example, the photon has helicity $h = 1$ and has two polarisation states, as does a graviton with $h = 2$. A massless spinor with $h = \frac{1}{2}$ also has two degrees of freedom. Why aren't we seeing this doubling in our representation theory analysis?

What we're missing is the additional requirement that the spectrum of states is invariant under CPT. These are discrete symmetries that we will look at more closely in Section 1.4. For massive particles, this doesn't buy us anything new: the set of states $|p_\mu, j\rangle$ is already invariant under CPT. However, for massless particles CPT flips $h \mapsto -h$ and tells us that massless states must come in pairs

$$|p_\mu; h\rangle \quad \text{and} \quad |p_\mu; -h\rangle\ . \tag{1.31}$$

This is the origin of the two polarisation states of the photon or graviton, or the two helicities of a massless Weyl spinor. Note that a massless scalar has helicity $h = 0$ and so is CPT self-conjugate. This means that there's no requirement from CPT to add an additional degree of freedom in this case.

Weird Continuous Spin Representations

We brushed over something above. When looking at massless representations, we found that the little group coincides with the 2d Euclidean group (1.27). But then, without justification, we restricted ourselves to representations on which the translation generators W_1 and W_2 act trivially. Here we give the justification.

Let's look at representations of the 2d Euclidean group (1.27) for which translations W_1 and W_2 act non-trivially. Because $[W_1, W_2] = 0$, we can simultaneously diagonalise these generators so that they act on states $|w_1, w_2\rangle$ such that

$$W_i|w_1, w_2\rangle = w_i|w_1, w_2\rangle \quad \text{for} \quad i = 1, 2 \; . \tag{1.32}$$

The second Casimir is then

$$C_2 = W^\mu W_\mu = -(w_1^2 + w_2^2) \; . \tag{1.33}$$

For the massless representations above, we assumed that $w_1 = w_2 = 0$. Now we want to understand what happens when they are non-zero. Since C_2 is fixed, we write $w_1 = \rho \cos\alpha$ and $w_2 = \rho \sin\alpha$ with $C_2 = -\rho^2$ and we should think of the collection of states $|w_1, w_2\rangle$ as parameterised by the angle $\alpha \in [0, 2\pi)$ with the action

$$W_1|\alpha\rangle = \rho\cos\alpha|\alpha\rangle \quad \text{and} \quad W_2|\alpha\rangle = \rho\sin\alpha|\alpha\rangle \; . \tag{1.34}$$

It remains to determine the action of $W_3 = EJ_3$ on these states. This is given by

$$e^{i\theta J_3}|\alpha\rangle = e^{ih\theta}|\alpha + \theta\rangle \quad \Longrightarrow \quad J_3|\alpha\rangle = h|\alpha\rangle - i\frac{d}{d\alpha}|\alpha\rangle \; . \tag{1.35}$$

You can check that the actions (1.35) and (1.34) do indeed furnish a representation of the 2d Euclidean algebra (1.27). But, from the perspective of particle physics, it's a very weird representation. This is because particle states $|p_\mu, \alpha; h\rangle$ are labelled by their momentum p_μ and an additional angle $\alpha \in [0, 2\pi)$. This means that for every choice of momentum p_μ, there's still an infinite dimensional Hilbert space, labelled by the continuous parameter α rather than a discrete, bounded parameter like j_3. Said differently, it's as if we have an uncountably infinite number of species of particle. These are known as *continuous spin representations*.

We've certainly never observed particles corresponding to these states and they would have very strange properties (such as infinite heat capacity). Nonetheless, one can't help but wonder if nature may make use of them somewhere.

1.1.3 The Coleman-Mandula Theorem

It's not unusual for quantum field theories to exhibit further continuous symmetries. Say, a global $U(1)$ symmetry that rotates the phase of a complex field, or perhaps a non-Abelian $SU(N)$ symmetry under which a multiplet of fields transforms. The generators of these symmetries – which we'll denote collectively as T – correspond to some conserved charge and are always Lorentz scalars which means that they necessarily commute with the Poincaré generators,

$$[P^\mu, T] = [M^{\mu\nu}, T] = 0 \ . \tag{1.36}$$

One could ask: is it possible for something less trivial to happen, with the new generators transforming in some fashion under the Poincaré group? For example, this would happen if the additional generators T themselves carried some spacetime index. If this were possilble, the Poincaré group would be subsumed into a larger group. And that sounds interesting.

A theorem due to Coleman and Mandula greatly restricts this possibility. Roughly speaking, the theorem states that, in any spacetime dimension greater than $d = 1 + 1$, the symmetry group of any interacting quantum field theory must factorise as

$$\text{Poincaré} \times \text{Internal} \ . \tag{1.37}$$

We won't prove the Coleman-Mandula theorem here. The gist of the proof is to look at 2-to-2 scattering (meaning two incoming particles scatter into two outgoing particles). Poincaré invariance already greatly restricts what can happen, with only the scattering angle left undetermined. Any internal symmetries that factorise, as in (1.37), put restrictions on the kinds of interactions that are allowed, for example enforcing conservation of electric charge. But if the generators T were to carry a spacetime index then they would put further constraints on the scattering angle itself and that would be overly restrictive, at best allowing scattering to occur only at discrete angles. But if one assumes that the scattering amplitudes are analytic functions of the angle then the amplitude must vanish for all angles and the theory is free.

Like all no-go theorems in physics, the Coleman-Mandula theorem comes with a number of underlying assumptions. Some of these are eminently reasonable, such as locality and causality. But it may be possible to relax other assumptions to find interesting loopholes to the Coleman-Mandula theorem. Two such loopholes have proven to be extremely important.

- **Conformal Invariance:** The Coleman-Mandula theorem assumes that the theory has a mass gap, meaning that all particles are massive. Indeed, the theorem is a statement about symmetries of the S-matrix which is really only well defined for massive particles where we don't have to worry about IR divergences. For theories of massless particles something interesting can, and often does, happen.

The first interesting thing is that interacting massless theories typically exhibit scale invariance. This means that physics is unchanged under the symmetry $x^\mu \to \lambda x^\mu$. The associated symmetry generator is called D for "dilatation". This can only be a symmetry of a theory that has no dimensionful parameters, which is the main reason it can occur only for massless theories.

The second interesting thing is more surprising. For reasons that are not entirely understood, theories that exhibit scale invariance also exhibit a further symmetry known as *special conformal transformations* of the form

$$x^\mu \to \frac{x^\mu - a^\mu x^2}{1 - 2a\cdot x + a^2 x^2} \ . \tag{1.38}$$

This transformation depends on a vector parameter a^μ and the associated generator is a 4-vector K^μ. The resulting conformal algebra extends the Poincaré algebra (1.18) and (1.19) with the non-trivial commutators

$$\begin{aligned} [D, K^\mu] = -iK^\mu \quad &, \quad [D, P^\mu] = iP^\mu \\ [K^\mu, P^\nu] &= 2i(D\eta^{\mu\nu} - M^{\mu\nu}) \\ [M^{\mu\nu}, K^\sigma] &= i\left(K^\nu \eta^{\mu\sigma} - K^\mu \eta^{\nu\sigma}\right) \qquad . \end{aligned} \tag{1.39}$$

Interacting conformal field theories crop up in many places in physics. In their Euclidean incarnation, they describe critical points, or second order phase transitions, that were the focus of our lectures on Statistical Field Theory. In $d = 1+1$ dimensions the conformal group has rather more structure and a detailed introduction can be found in the lectures on String Theory.

- **Supersymmetry:** The second loophole to the Coleman-Mandula theorem is supersymmetry. This is a symmetry that relates bosons to fermions. The generator that enacts this magical transformation is denoted as Q_α and carries a spacetime spinor index $\alpha = 1, 2$. (We will learn more about spinors in Section 1.2.) This is exactly the kind of thing that the Coleman-Mandula theorem is supposed to rule out. However, supersymmetry evades the theorem because the generators Q_α do not form a Lie algebra: instead they form what is known as a *super-Lie algebra*, with the commutation relations of the Poincaré group (1.18) and (1.19) augmented by the anti-commutation relation

$$\{Q_\alpha, \bar{Q}_{\dot\alpha}\} = 2\sigma^\mu_{\alpha\dot\alpha} P_\mu \ . \tag{1.40}$$

Here $\sigma^\mu_{\alpha\dot{\alpha}}$ are a collection of 2×2 matrices defined in (1.44). (We'll see a lot more about what the α and $\dot{\alpha}$ spinor indices mean shortly.) You can learn (a lot!) more about this algebra and its consequences for various field theories in the lectures on Supersymmetry.

Neither conformal symmetry nor supersymmetry play a role in the Standard Model. However, both arise in different ways when it comes to ideas for what lies beyond the Standard Model.

1.2 Spinors

Scalars are basic. They have no internal structure and, as such, come with very little baggage. There's a lot of fun that we can have with them, largely by writing down potentials that do interesting things, and we'll see examples of this when we discuss spontaneous symmetry breaking in Section 2. But there's little that is subtle about scalars: what you see is what you get.

In contrast, any field with higher spin is awash with subtleties. For massless spin 1 particles, like photons, these subtleties are all about gauge invariance and we will discuss them in Section 1.3. Here our interest is in spin $\frac{1}{2}$ particles, known as spinors. These are the fields that describe all matter particles in the Standard Model, meaning the quarks and leptons. They are subtle largely because anything that comes back to itself with a minus sign after a 2π rotation is always going to be a little strange.

1.2.1 Dirac vs Weyl Spinors

We start by reviewing some features of spinors that we met in the lectures on Quantum Field Theory. However, our focus is going to be a little different. In particular, to prepare us for the Standard Model, we will need to look more closely at the properties of Weyl spinors.

In the lectures on Quantum Field Theory, we learned about the 4-component Dirac spinor ψ. This comes hand in hand with a collection of gamma matrices that obey the Clifford algebra

$$\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu} \; . \tag{1.41}$$

The Clifford algebra admits a unique irreducible representation, up to conjugation. But that "up to conjugation" caveat hides all manner of headaches as it provides ample opportunity for physicists to use annoying conventions. Here we use the chiral basis of gamma matrices,

$$\gamma^\mu = \begin{pmatrix} 0 & \sigma^\mu \\ \bar{\sigma}^\mu & 0 \end{pmatrix} \quad \text{and} \quad \gamma^5 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{1.42}$$

where we've introduced two collections of 2×2 matrices,

$$\sigma^\mu = (\mathbb{1}, \sigma^i) \quad \text{and} \quad \bar{\sigma}^\mu = (\mathbb{1}, -\sigma^i) \tag{1.43}$$

where σ^i with $i = 1, 2, 3$ are the familiar Pauli matrices,

$$\sigma^1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \, , \quad \sigma^2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \, , \quad \sigma^3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \, . \tag{1.44}$$

The bar on $\bar{\sigma}^\mu$ in (1.43) doesn't denote complex conjugation: these are simply a different collection of 2×2 matrices from σ^μ.

In the Quantum Field Theory lectures, we showed that the generators of Lorentz transformations for a Dirac spinor are

$$S^{\mu\nu} = \frac{i}{4}[\gamma^\mu, \gamma^\nu] = \begin{pmatrix} \sigma^{\mu\nu} & 0 \\ 0 & \bar{\sigma}^{\mu\nu} \end{pmatrix} \, . \tag{1.45}$$

(As with our earlier definition of $M^{\mu\nu}$, this differs by a factor of i from the conventions in the Quantum Field Theory lectures.) Here we've defined

$$\begin{aligned} \sigma^{\mu\nu} &= \frac{i}{4}\left(\sigma^\mu \bar{\sigma}^\nu - \sigma^\nu \bar{\sigma}^\mu\right) \\ \bar{\sigma}^{\mu\nu} &= \frac{i}{4}\left(\bar{\sigma}^\mu \sigma^\nu - \bar{\sigma}^\nu \sigma^\mu\right) \, . \end{aligned} \tag{1.46}$$

Because both of these expressions are anti-symmetrised in μ and ν, each is a collection of six 2×2 matrices.

The generators $S^{\mu\nu}$ defined in (1.45) are block diagonal. This is telling us that they are *not* an irreducible representation of the Lorentz algebra. Instead, it's formed of two distinct representations, one generated by $\sigma^{\mu\nu}$ and the other generated by $\bar{\sigma}^{\mu\nu}$. Indeed, you can check that each of these obeys the Lorentz algebra (1.5)

$$[\sigma^{\mu\nu}, \sigma^{\rho\sigma}] = i\left(\eta^{\nu\rho}\sigma^{\mu\sigma} - \eta^{\nu\sigma}\sigma^{\mu\rho} + \eta^{\mu\sigma}\sigma^{\nu\rho} - \eta^{\mu\rho}\sigma^{\nu\sigma}\right) \tag{1.47}$$

with a similar expression for $\bar{\sigma}^{\mu\nu}$. Correspondingly, the 4-component Dirac spinor ψ also decomposes into two 2-component spinors

$$\psi = \begin{pmatrix} \psi_L \\ \psi_R \end{pmatrix} . \tag{1.48}$$

These are referred to as left-handed and right-handed spinors respectively. In the language of our earlier table of representations (1.13), ψ_L sits in the $(\frac{1}{2}, 0)$ representation while ψ_R sits in the $(0, \frac{1}{2})$ representation. A Dirac spinor is a combination of both representations $(\frac{1}{2}, 0) \oplus (0, \frac{1}{2})$.

Under a Lorentz transformation, a left-handed Weyl spinor transforms as

$$\psi_L \;\rightarrow\; S\psi_L \quad \text{with} \quad S = \exp\left(-\frac{i}{2}\omega_{\mu\nu}\sigma^{\mu\nu}\right) . \tag{1.49}$$

Here $\omega_{\mu\nu}$ are the same set of six numbers that specify the Lorentz transformation (1.5). There is a similar expression for ψ_R, with $\sigma^{\mu\nu}$ replaced by $\bar{\sigma}^{\mu\nu}$.

You can check that $\operatorname{tr}\sigma^{\mu\nu} = 0$ and so, using $\det(e^A) = e^{\operatorname{tr} A}$, we have $\det S = 1$. In fact, $S \in SL(2, \mathbb{C})$, and what we've done in constructing the Weyl spinor representation of the Lorentz group is highlight the group isomorphism $\text{Spin}(1, 3) \cong SL(2, \mathbb{C})$.

(Left-Handed)* = Right-Handed

The two representations – one for a left-handed Weyl spinor, the other for a right-handed Weyl spinor – are related by complex conjugation.

It's not immediately obvious because, as we've seen, the generators are $\sigma^{\mu\nu}$ and $\bar{\sigma}^{\mu\nu}$ and it's not true that these generators are complex conjugates: $(\sigma^{\mu\nu})^\star \neq \bar{\sigma}^{\mu\nu}$. To see the relation, we need an additional conjugation by the anti-symmetric tensor

$$\epsilon = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} . \tag{1.50}$$

You can then check that

$$\epsilon^T (\sigma^{\mu\nu})^\star \epsilon = \bar{\sigma}^{\mu\nu} . \tag{1.51}$$

Operationally, the complex conjugation flips the sign of $(\sigma^2)^\star = -\sigma^2$ leaving the other Pauli matrices alone: $(\sigma^i)^\star = \sigma^i$ for $i = 1, 3$. But the conjugation by $\epsilon = i\sigma^2$ then flips the sign of σ^i with $i = 1, 3$, leaving σ^2 alone.

This simple algebraic relation has an important physical implication. If you have a left-handed particle described by a Weyl spinor ψ_L, then its anti-particle is described by the conjugate spinor $\psi_L^\dagger$ (which we also write as $\bar{\psi}_L$) and is right-handed.

Building Scalars from Spinors

If we're given two left-handed spinors, ψ_L and χ_L, then we can build a scalar. We'll adorn our spinors with indices, so we have $(\psi_L)_\alpha$ and $(\chi_L)_\alpha$ with $\alpha = 1, 2$. We also add indices to our anti-symmetric matrix

$$\epsilon^{\alpha\beta} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} . \tag{1.52}$$

We then define the scalar quantity

$$\psi_L\,\chi_L := \epsilon^{\alpha\beta}(\psi_L)_\beta(\chi_L)_\alpha = (\psi_L)_2(\chi_L)_1 - (\psi_L)_1(\chi_L)_2 \ . \tag{1.53}$$

To see that this does indeed transform as a scalar, we look at

$$\psi_L\,\chi_L \to \ S_\alpha{}^\gamma S_\beta{}^\delta \epsilon^{\alpha\beta}(\psi_L)_\delta(\chi_L)_\gamma = (\det S)\epsilon^{\gamma\delta}(\psi_L)_\delta(\chi_L)_\gamma = \psi_L\,\chi_L \tag{1.54}$$

where, in the first equality we've used the fact that $S_\alpha{}^\gamma S_\beta{}^\delta \epsilon^{\alpha\beta} = \det S\,\epsilon^{\gamma\delta}$, which you can confirm simply by checking all the cases $\gamma, \delta = 1, 2$. In the second equality we've used the fact that $\det S = 1$.

This is an important lesson: you can form a scalar from two left-handed spinors. In terms of the representation theory of the previous section, what we're seeing here is the tensor product $(\frac{1}{2}, 0) \otimes (\frac{1}{2}, 0) = (0, 0) \oplus (1, 0)$, where the scalar (1.53) picks out the singlet $(0, 0)$.

The anti-symmetric tensor $\epsilon^{\alpha\beta}$ is an invariant tensor for the group $SL(2, \mathbb{C})$. In that sense, it plays a role that is similar to the delta function δ^{ab} for the group $SO(N)$, or the Minkowski metric $\eta^{\mu\nu}$ for the group $SO(1, 3)$. In particular, it allows us to form a scalar product between two spinors as in (1.53). The fact that this product is anti-symmetric, rather than symmetric, fits nicely with the fact that, in quantum field theory, spinors are anti-commuting variables whose components are Grassmann-valued. This means that we have,

$$\psi_L\,\chi_L = (\psi_L)_2(\chi_L)_1 - (\psi_L)_1(\chi_L)_2 = -(\chi_L)_1(\psi_L)_2 + (\chi_L)_2(\psi_L)_1 = \chi_L\,\psi_L \ . \tag{1.55}$$

In particular, this means that we can form a scalar from just a single left-handed Weyl spinor

$$\psi_L\,\psi_L = (\psi_L)_2(\psi_L)_1 - (\psi_L)_1(\psi_L)_2 = 2(\psi_L)_2(\psi_L)_1 \ . \tag{1.56}$$

Again, there are similar expressions for right-handed spinors.

There's quite a bit more to say about the two different representations of the Lorentz algebra and their properties. You can read about this (and the corresponding dotted and undotted indices) in the first section of the lectures on Supersymmetry. But the simple summary above will suffice for our purposes.

1.2.2 Actions for Spinors

Our next goal is to understand how to construct Lagrangians for spinors. Again, our starting point will be the Dirac spinor that we met in Quantum Field Theory. There we saw that the Lorentz invariant action is

$$S_{\text{Dirac}} = -\int d^4x \, \left(i\bar{\psi}\gamma^\mu \partial_\mu \psi - M\bar{\psi}\psi \right) \,. \tag{1.57}$$

For a Dirac spinor, the bar notation means $\bar{\psi} = \psi^\dagger \gamma^0$. Decomposed in terms of Weyl fermions (1.48),

$$S_{\text{Dirac}} = -\int d^4x \, \left(i\bar{\psi}_L \bar{\sigma}^\mu \partial_\mu \psi_L + i\bar{\psi}_R \sigma^\mu \partial_\mu \psi_R - M(\bar{\psi}_R \psi_L + \bar{\psi}_L \psi_R) \right) \,. \tag{1.58}$$

First an important, but trivial, notational point: the bar for a Weyl spinor means something different from a bar for a Dirac spinor. It is simply a more elegant way of writing $\bar{\psi}_L = \psi_L^\dagger$.

Second, note that the mass term couples the left- and right-handed Weyl spinors. Combining our observations above, we know that the complex conjugate $\bar{\psi}_R$ is a left-handed spinor, and so in writing $\bar{\psi}_R \psi_L$ we've combined two left-handed spinors into a scalar. Similarly, $\bar{\psi}_L \psi_R$ combines two right-handed spinors into a scalar.

It's worth pausing to look at the symmetries of the action (1.58). Crucially, these symmetries are different for massless and massive fermions. In the absence of the mass term, so $M = 0$, the action has a $U(1)^2$ symmetry, under which the two fermions rotate separately, $\psi_L \to e^{i\alpha}\psi_L$ and $\psi_R \to e^{i\beta}\psi_R$. When we turn on the mass term, only the diagonal combination, with $\alpha = \beta$ survives. This is a general story, and one that will be particularly important for understanding the Standard Model: massless fermions always have more symmetries than massive fermions.

The mass in (1.58) can take values $M \in \mathbb{R}$. (There's no positivity requirement.) Upon quantisation, with $M \neq 0$, we get a particle of spin $+\frac{1}{2}$ and charge $+1$ under the surviving $U(1)$, together with a distinct anti-particle of spin $+\frac{1}{2}$ and charge -1, both with mass $|M|$.

The mass term in (1.58) which combines two different spinors, ψ_L and ψ_R, is known as a *Dirac mass*. It's not the only thing we can write down. Suppose that we have just a left-handed spinor ψ_L. Then it's perfectly possible to write down an action with a mass term,

$$\mathcal{S}_{\text{Weyl}} = -\int d^4x \,\left(i\bar{\psi}_L\bar{\sigma}^\mu\partial_\mu\psi_L + \frac{m}{2}\psi_L\psi_L + \frac{m^\star}{2}\bar{\psi}_L\bar{\psi}_L\right) \ . \tag{1.59}$$

This is known as a *Majorana mass*. Here we can take $m \in \mathbb{C}$.

Again, the massive theory has less symmetry than the massless theory, with the $U(1)$ that rotates the phase of ψ_L broken when $m \neq 0$. This means that there's no $U(1)$ quantum number to distinguish particles from anti-particles and, upon quantisation, the theory describes a single spin $\frac{1}{2}$ particle with mass $|m|$ that is now its own anti-particle.

Because the Majorana mass term explicitly breaks the $U(1)$ symmetry, it is not allowed if the $U(1)$ is gauged. Relatedly, it's not possible to write down such a term for any fermion ψ_L that transforms in a complex representation of a gauge group. It is, however, possible to write down such terms for fermions in real representations.

1.3 Gauge Invariance

In the Standard Model, forces are associated to massless spin 1 particles, known collectively as *gauge bosons*. As we now explain, much of the dynamics of these forces is fixed by gauge invariance.

1.3.1 Maxwell Theory

The key ideas of gauge invariance are familiar from electromagnetism. There, the fundamental field is the 4-vector $A_\mu(x)$, known as the *gauge potential*. Crucially, not all components of $A_\mu(x)$ are physical: instead, we should identify any two gauge potentials that are related by a gauge transformation of the form

$$A_\mu \to A_\mu + \partial_\mu\alpha \tag{1.60}$$

for any function $\alpha(x)$. The transformation (1.60) is sometimes called a *gauge symmetry*. It's not a good name. A "symmetry" describes a situation in which two physically distinct configurations share the same physics. But that's not what's going on in (1.60). Instead, the two configurations related by a gauge transformation describe the *same* physical configuration. A fairly decent analogy is to think of two gauge potentials that are related by (1.60) in the same way as you would view two different coordinate systems. A much better name would be *gauge redundancy*.

As we proceed, we'll see that a great deal of the structure of the Standard Model is determined by the requirements of gauge invariance. Yet, in many ways, this is a strange idea on which to rest our most important theories of physics. Gauge invariance is, at heart, merely an ambiguity in how we choose to present the laws of physics. Why should it play such an important role?

One reason is that the ambiguity allows us to demonstrate various properties that we care about but which, naively, might appear incompatible. These properties include Lorentz invariance and locality and, in the quantum theory, unitarity. We already got a glimpse of this in the lectures on Quantum Field Theory when we quantised Maxwell theory. One choice of gauge makes unitarity manifest while another makes Lorentz invariance manifest. The gauge ambiguity allows us to flit from one choice to another, allowing us to both have our cake and eat it.

Relatedly, we know that the photon has two polarisation states. But try writing down a field which describes the photon that has only two indices and which transforms nicely under the $SO(1,3)$ Lorentz group; its not possible. So instead we introduce the field A_μ which makes Lorentz invariance manifest and then use the gauge symmetry to kill two of four resulting states.

The physical information in A_μ can be found in the *field strength*

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu \ . \tag{1.61}$$

The field strength is invariant under the gauge transformation (1.60). The field strength houses the electric field $\mathbf{E}$ and the magnetic field $\mathbf{B}$. If we write $A^\mu = (\phi, \mathbf{A})$, then we have

$$\mathbf{E} = -\nabla\phi - \frac{\partial \mathbf{A}}{\partial t} \quad \text{and} \quad \mathbf{B} = \nabla \times \mathbf{A} \ . \tag{1.62}$$

The dynamics of the gauge field is described by the action

$$S_{\text{Maxwell}} = -\frac{1}{4}\int d^4x \ F_{\mu\nu}F^{\mu\nu} \ . \tag{1.63}$$

The resulting equations of motion are

$$\partial_\mu F^{\mu\nu} = 0 \ . \tag{1.64}$$

This coincides with two of the Maxwell equations: Gauss' law $\nabla \cdot \mathbf{E} = 0$ and Ampère's law $\nabla \times \mathbf{B} = \partial \mathbf{E}/\partial t$. The other two follow immediately from constructing $F_{\mu\nu}$ in terms of the gauge potential. To see this, we first introduce the *dual field strength*

$${}^\star F^{\mu\nu} = \frac{1}{2}\epsilon^{\mu\nu\rho\sigma}F_{\rho\sigma} \ . \tag{1.65}$$

This is similar to $F_{\mu\nu}$, but with $\mathbf{E}$ and $\mathbf{B}$ swapped (one of them with a minus sign). Then, by the anti-symmetry of $\epsilon^{\mu\nu\rho\sigma}$, together with the definition (1.61), we have the Bianchi identity

$$\partial_\mu {}^\star F^{\mu\nu} = 0 \ . \tag{1.66}$$

Expanding this out gives the remaining two Maxwell equations: the one that says magnetic monopoles don't exist $\nabla\cdot\mathbf{B} = 0$, and the law of induction $\nabla\times\mathbf{E}+\partial\mathbf{B}/\partial t = 0$.

The necessity to keep gauge invariance means that it's not possible to augment the action (1.63) with a mass term of the form $m^2 A_\mu A^\mu$. This would break gauge invariance and cause trouble down the line. Naively, this would appear to guarantee that the photon must always be massless. In fact, there is a way to give the photon a mass, known as the *Higgs mechanism*. This will be discussed in Section 2.3.

Coupling to Matter

Underlying electromagnetism is a $U(1)$ gauge group. That's not so obvious in the description above, where the "symmetry" (really redundancy) manifests itself only as a shift of the gauge field (1.60) depending on a function $\alpha(x)$. However, the $U(1)$ness of electromagnetism becomes more apparent when we couple to charged fields.

Fields that are charged under electromagnetism are necessarily complex. Consider, for example, a complex scalar field $\phi(x)$ of charge e. When the gauge field transforms as (1.60), the scalar field has a corresponding transformation

$$\phi \to e^{ie\alpha}\phi \ . \tag{1.67}$$

Here we see the group emerging more clearly, with $e^{ie\alpha(x)} \in U(1)$. Because the transformation parameter $\alpha(x)$ is a function, we really have a $U(1)$ symmetry/redundancy for each point x in space. This is what it means to have a $U(1)$ "gauge group": it is a much larger group than the global symmetries that appear elsewhere.

We can construct theories that are invariant under the transformation (1.67) by replacing partial derivatives with the covariant derivative

$$\mathcal{D}_\mu\phi = \partial_\mu\phi - ieA_\mu\phi \ . \tag{1.68}$$

This has the nice property that $\mathcal{D}_\mu\phi$ transforms covariantly under a gauge transformation, a fact that requires a couple of quick lines of calculation:

$$\begin{aligned}\mathcal{D}_\mu\phi \to& \left(\partial_\mu - ieA_\mu - ie\partial_\mu\alpha\right) e^{ie\alpha}\phi \\ =& \ e^{ie\alpha}\left(\partial_\mu - ieA_\mu\right)\phi \\ =& \ e^{ie\alpha}\mathcal{D}_\mu\phi \ . \end{aligned} \tag{1.69}$$

The key to this calculation is that the derivative hitting $\partial_\mu(e^{ie\alpha})$ exactly cancels the shift of the gauge field (1.60). Taking the complex conjugate of (1.68), we have

$$\mathcal{D}_\mu \phi^\dagger = (\partial_\mu + ieA_\mu)\phi^\dagger \ . \tag{1.70}$$

From this, we see that the meaning of the covariant derivative $\mathcal{D}_\mu$ depends on the object it's hitting: it's $-ieA_\mu$ for the scalar in (1.68), but $+ieA_\mu$ for the conjugate scalar in (1.70). You can check that, under a gauge transformation, $\mathcal{D}_\mu\phi^\dagger \to e^{-ie\alpha}\mathcal{D}_\mu\phi^\dagger$. This ensures that we can form a gauge invariant action

$$S_{\rm scalar} = \int d^4x \, \Big(\mathcal{D}_\mu\phi^\dagger \mathcal{D}^\mu\phi - V(|\phi|)\Big) \tag{1.71}$$

where we take the potential to depend only on $|\phi|^2 = \phi^\dagger\phi$. In particular, this means that we disallow terms in the potential of the form $\phi^2 + \phi^{\dagger\,2}$ which are real but are not gauge invariant.

If we have multiple scalar fields, then they can carry different charges. When the gauge group is $U(1)$, these charges should be integer multiples of each other, meaning that each field transforms as

$$\phi \to e^{ieq\alpha}\phi \quad \text{with} \quad q \in \mathbb{Z} \ . \tag{1.72}$$

It is possible to write down theories in which the charges q are not integer valued. (For example, one could imagine one scalar field with $q = 1$ and another with $q = \sqrt{2}$.) Strictly, the gauge group should be viewed as $\mathbb{R}$ in this case, rather than $U(1)$. The differences between a $U(1)$ gauge group and an $\mathbb{R}$ gauge group are rather subtle, and manifest themselves only in the presence of magnetic monopoles, or in spacetimes of non-trivial topology. We won't get into these issues here.

Everything that we've said above for scalars also holds for fermions, both Weyl and Dirac. In either case, we replace the partial derivatives in the relevant action (either (1.59) or (1.58)) with covariant derivatives and off we go.

1.3.2 A Refresher on Lie Algebras

There is an important extension of Maxwell theory in which the gauge group $U(1)$ is replaced by a compact Lie group G. Here we give a lightning review of the relevant aspects of Lie groups and Lie algebras.

A Lie group is a group that is also a differentiable manifold[1]. This means, among other things, that a group element is labelled by some continuous parameters. We've already met examples of Lie groups in both the rotation group and the Poincaré group.

Lie groups have the property that, for elements continuously connected to the identity, we can write each $U \in G$ as

$$U = e^{i\theta^A T^A} \tag{1.73}$$

Here the θ^A are just numbers that tell us which group element we're working with, while the T^A are generators of the group. If you like, the T^a tell us the infinitesimal action of the group, with $U \approx \mathbb{1} + i\theta^A T^A + \mathcal{O}(\theta^2)$ when θ is small. A general group element (1.73) can then be constructed by exponentiating the infinitesimal action.

It turns out that, with the exception of some global information, the structure of the Lie group is captured in the behaviour of those infinitesimal generators T^A. They form the associated Lie algebra $\mathfrak{g}$, given by

$$[T^A, T^B] = if^{ABC} T^C \ . \tag{1.74}$$

Here $A, B, C = 1, \ldots, \dim G$ and f^{ABC} are the fully anti-symmetric structure constants which distill the information about the group G. The factor of i on the right-hand side is taken to ensure that the generators are Hermitian: $(T^A)^\dagger = T^A$.

(Mathematicians usually prefer the convention where there is no i on the right-hand side and the generators are anti-Hermitian, largely because there are examples like $SO(N)$ where everything in the game is real and a factor of i makes things needlessly complex. In contrast, physicists tend to include the factor of i on the right-hand side because they're usually working in the realm of quantum mechanics where things will ultimately become complex anyway.)

The T^A in (1.74) are abstract objects but we will shortly want to identify them with matrices. This means, among other things, that we want the commutator in (1.74) to have the same properties as matrix commutation, among them the Jacobi identity

$$[T^A, [T^B, T^C]] + [T^B, [T^C, T^A]] + [T^C, [T^A, T^B]] = 0 \ . \tag{1.75}$$

This puts constraints on the structure constants f^{abc} which must, in turn, obey

$$f^{ADE} f^{BCD} + f^{BDE} f^{CAD} + f^{CDE} f^{ABD} = 0 \ . \tag{1.76}$$

[1]For many physicists, Lie groups are the only groups they know. A mathematician friend of mine told me that a physicist's definition of a finite group is a Lie group without manifold structure.

G	$SU(N)$	$SO(N)$	$Sp(N)$	E_6	E_7	E_8	F_4	G_2
$\dim G$	N^2-1	$\frac{1}{2}N(N-1)$	$N(2N+1)$	78	133	248	52	14
$\dim F$	N	N	$2N$	27	56	248	6	7

Table 2. The classification of compact, semi-simple Lie algebras G, together with their dimension and the dimension of the fundamental representation F.

We will be interested in *simple*, *compact* Lie groups. Here "simple" means that we don't have any trivial $U(1)$ factors floating around that commute with everything else. We can always include such factors if we wish (and we will wish for the Standard Model) but we'll be best served if we ignore them at this stage. Meanwhile, "compact" means that if you continue to rotate in the group then you ultimately come back to where you started from (or close to where you started from). For example, the group of rotations is compact, while the Lorentz group is non-compact because if you keep boosting in a given direction then you just move faster and faster.

There is a classification of simple compact Lie algebras. The possible options for the group G, together with the dimension of the group, are shown in Table 2[2]. All of these groups are referred to as *non-Abelian* meaning that things don't commute with each other. In contrast, $U(1)$ is an Abelian group.

As we mentioned above, the T^A in (1.74) are initially viewed as just abstract objects. But it's interesting to ask when they can take a more concrete form in the guise of matrices. These are the *representations* of the algebra. For each algebra, there is an infinite list of numbers which are the dimensions of the matrices that can be used to represent it. The smallest such (non-trivial) matrix is called the *fundamental representation* and we will denote it as F. The dimension of F for each Lie group G are also shown in Table 2.

In what follows, we will (with a slight abuse of notation) use T^A to refer to the generators of the fundamental representation. When we have occasion to use other representations R, we will refer to the generators as $T^A(R)$ (In later sections, we'll also refer to these as T^A_R.). In fact, for the Standard Model we will only need two different representations: the fundamental and the adjoint. The adjoint is a representation that

[2]We're using the convention $Sp(1) = SU(2)$. Other authors sometimes write $Sp(2N)$, or even $USp(2N)$ to refer to what we've called $Sp(N)$, preferring the argument to refer to the dimension of the fundamental representation F rather than the rank of the Lie algebra $\mathfrak{g}$.

has dimension $\dim(\text{adj}) = \dim G$ with the generators given by

$$T^A(\text{adj})_{BC} = -if^{ABC} \ . \tag{1.77}$$

Don't be lulled into thinking that you don't need to consider other representations: they will appear in other situations, including when we discuss flavour symmetry in QCD in Section 3.

The Lie algebra comes with what, in fancy language, is called a Killing form. But, by the time we're thinking about matrices, this Killing form is just the trace. The generators of any simple Lie algebra obey $\text{Tr}\, T^A = 0$. (This is what it means for the Lie algebra to be "simple".) We take the generators in the fundamental representation F to satisfy

$$\text{Tr}\, T^A T^B = \frac{1}{2}\delta^{AB} \ . \tag{1.78}$$

This can be viewed as tantamount to fixing the normalisation of the structure constants f^{ABC}. Having fixed the normalisation in the fundamental representation, other representations $T^A(R)$ will have different normalisations.

Before we proceed, an example. The simplest non-Abelian Lie group is $SU(2)$, which has $\dim(SU(2)) = 3$ and structure constants given by $f^{ABC} = \epsilon^{ABC}$. In this case, the fundamental representation is (up to an overall normalisation) the 2×2 Pauli matrices

$$T^A = \frac{1}{2}\sigma^A \ . \tag{1.79}$$

These indeed obey $[T^A, T^B] = i\epsilon^{ABC}T^C$, together with the normalisation condition (1.78).

The group $SU(3)$ also plays a prominent role in the Standard Model. (In fact, as we will see, it plays two prominent roles!) We will describe the structure constants and the generators in Section 3.

1.3.3 Yang-Mills Theory

Now we can turn to some physics. Yang-Mills theory is a generalisation of Maxwell theory in which the group $U(1)$ is replaced by a simple, compact Lie group G. To specify the Yang-Mills theory, we need only specify the choice of G together with a coupling constant $g > 0$ that will dictate the strength of the interactions. (The coupling constant g plays the same role as the charge e in Maxwell theory. As we will later see, the phrase "coupling constant" is not particularly accurate because it will turn out not to be constant!)

For each element of the algebra, we introduce a gauge field A_μ^A with $A = 1, \ldots, \dim G$. These are then packaged into the Lie algebra-valued gauge potential

$$A_\mu = A_\mu^A T^A \tag{1.80}$$

A down-to-earth perspective is to think of the T^A as matrices in the fundamental representation. This means, for example, that for $G = SU(N)$, the gauge potential A_μ is a 4-vector where each component is a traceless $N \times N$ matrix.

The fields A_μ^A are collectively referred to as *gauge bosons*. (They have other, more specific, names in the Standard Model when we apply these ideas to the two nuclear forces.) As in Maxwell theory, not all the information in A_μ is physical and any two field configurations related by a gauge transformation should be viewed as equivalent. This time, however, the gauge transformation is a little more intricate.

The action of the gauge symmetry is associated to a Lie group valued function over spacetime,

$$\Omega(x) \in G\ . \tag{1.81}$$

The set of all such transformations is known as the *gauge group*. As in Maxwell theory, we will sometimes be sloppy and refer to the Lie group G as the gauge group, but strictly speaking it is the much bigger group of maps from spacetime into G. The action on the gauge field is

$$A_\mu \to \Omega A_\mu \Omega^{-1} + \frac{i}{g}\Omega\, \partial_\mu \Omega^{-1}\ . \tag{1.82}$$

The first term is the expected transformation for an adjoint-valued field. The second, inhomogeneous, term is an additional piece that is characteristic of gauge transformations.

To make contact with gauge transformations in electromagnetism, suppose that we have $G = U(1)$ and write $\Omega(x) = e^{ie\alpha(x)}$. Then, using the fact that everything commutes, we have

$$\Omega A_\mu \Omega^{-1} + \frac{i}{e}\Omega\partial_\mu\Omega^{-1} = A_\mu + \partial_\mu \alpha \tag{1.83}$$

and the gauge transformation (1.82) reproduces the familiar gauge transformation of Maxwell theory.

As in Maxwell theory, we can construct a field strength. Here too there is an extra ingredient arising from the fact that A_μ is a matrix and the generalisation of (1.61) is

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu - ig[A_\mu, A_\nu] \; . \tag{1.84}$$

In contrast to Maxwell theory, the field strength includes a non-linear term, proportional to the coupling g. This will prove to be important: it is this non-linear term that makes Yang-Mills theory significantly richer and more interesting than Maxwell theory. Like A_μ, the field strength is a Lie algebra-valued field and we could also expand it as $F_{\mu\nu} = F^A_{\mu\nu} T^A$.

So far, I've not explained *why* (1.84) is the right field strength. The main reason is that it transforms nicely under the gauge transformation (1.82)

$$F_{\mu\nu} \rightarrow \Omega\, F_{\mu\nu}\, \Omega^{-1} \; . \tag{1.85}$$

To see this, you could just plug (1.82) into (1.84) but it's mildly laborious; we will offer a shortcut to this result presently.

The transformation (1.85) means that, in contrast to electromagnetism, the Yang-Mills "electric field" $E_i = F_{0i}$ and "magnetic field" $B_i = -\frac{1}{2}\epsilon_{ijk}F_{jk}$ are *not* gauge invariant. To construct something physical, you can multiply together some number of E_i and B_j and then take the trace, which ensures that the Ω and Ω^{-1} in (1.85) cancel and you get something gauge invariant. (You need something that is at least quadratic in $F_{\mu\nu}$ because, for simple Lie groups, $\text{Tr}\, F_{\mu\nu} = 0$.)

The gauge transformations above involve the Lie group valued object $\Omega(x)$. But one of the key properties of Lie groups is that their structure is largely determined by the elements that are infinitesimally close to the identity. This suggests that it's fruitful to look at gauge transformations that are everywhere close to the identity. These can be written as

$$\Omega(x) \approx 1 + ig\alpha^A(x) T^A + \ldots \tag{1.86}$$

where the α^A are taken to be everywhere small. From (1.82), the infinitesimal transformation of the gauge field is $A_\mu \rightarrow A_\mu + \delta A_\mu$ with

$$\delta A_\mu = \partial_\mu \alpha - ig[A_\mu, \alpha] \tag{1.87}$$

where $\alpha = \alpha^A T^A$ is the Lie algebra-valued infinitesimal transformation. It's convenient to write this as $\delta A_\mu = \mathcal{D}_\mu \alpha$ where the covariant derivative is defined to be

$$\mathcal{D}_\mu \alpha = \partial_\mu \alpha - ig[A_\mu, \alpha] \; . \tag{1.88}$$

This is the covariant derivative acting on the Lie algebra-valued (i.e. adjoint) field α. We'll soon see different covariant derivatives acting on other representations.

Now we can check how infinitesimal gauge transformations act on the field strength (1.84). We have

$$\begin{aligned}\delta F_{\mu\nu} &= \partial_\mu \delta A_\nu - \partial_\nu \delta A_\mu - ig[A_\mu, \delta A_\nu] - ig[\delta A_\mu, A_\nu] \\ &= \mathcal{D}_\mu \delta A_\nu - \mathcal{D}_\nu \delta A_\mu \\ &= [\mathcal{D}_\mu, \mathcal{D}_\nu]\alpha \ . \end{aligned} \tag{1.89}$$

We see that we're left with the task of computing the commutator of two covariant derivatives, acting on the adjoint field α. This is a worthwhile and straightforward calculation. We have

$$[\mathcal{D}_\mu, \mathcal{D}_\nu]\alpha = -ig[F_{\mu\nu}, \alpha] \ . \tag{1.90}$$

This gives $\delta F_{\mu\nu} = ig[\alpha, F_{\mu\nu}]$ which is indeed the expected infinitesimal gauge transformation arising from (1.85).

The Yang-Mills Action

The dynamics of the Yang-Mills field is the obvious generalisation of the Maxwell action,

$$S_{\rm YM} = -\frac{1}{2}\int d^4x \ {\rm Tr}\, F^{\mu\nu}F_{\mu\nu} \ . \tag{1.91}$$

Naively, the only difference lies in that overall trace, which ensures that the action is invariant under gauge transformations (1.85). This also accounts for the overall normalisation of the action, which comes with a factor of 1/2 rather than the 1/4 seen in (1.63) because an additional factor of 1/2 comes from the trace in (1.78). This means that the Yang-Mills and Maxwell action come with the same normalisation.

However, the key difference between the two actions is buried in our notation: while the Maxwell action is quadratic in A_μ, the Yang-Mills action includes terms that are cubic and quartic in A_μ, both coming from the commutator in the definition of the field strength (1.84).

The classical equations of motion are derived by minimizing the action with respect to each gauge field A^a_μ. It is a simple exercise to check that they are given by

$$\mathcal{D}_\mu F^{\mu\nu} = 0 \ . \tag{1.92}$$

Here the covariant derivative is defined as in (1.88): $\mathcal{D}_\mu F^{\mu\nu} = \partial_\mu F^{\mu\nu} - ig[A_\mu, F^{\mu\nu}]$. These are the *Yang-Mills equations*. In contrast to the Maxwell equations, they are non-linear. This means that the Yang-Mills fields interact with themselves.

There is also a Bianchi identity that follows from the definition (1.84) of $F_{\mu\nu}$ in terms of the gauge field. This is best expressed by first introducing the dual field strength

$$ {}^\star F^{\mu\nu} = \frac{1}{2}\epsilon^{\mu\nu\rho\sigma}F_{\rho\sigma} \ . \tag{1.93} $$

and noting that this obeys the identity

$$ \mathcal{D}_\mu {}^\star F^{\mu\nu} = 0 \ . \tag{1.94} $$

Both (1.92) and (1.94) are non-linear equations. However, the non-linearities come in the form of commutators like $[A_\mu, A_\nu]$. This means that if we focus on field configurations that sit purely within a subgroup $U(1) \subset G$, then the commutators vanish and the equations reduce to those of Maxwell theory. So although the general solutions to the Yang-Mills equations are surely complicated, we can always import any solution to Maxwell theory and embed it in some $U(1)$. In particular, Yang-Mills theory admits solutions akin to electromagnetic waves that travel at the speed of light.

Although we can always embed solutions of Maxwell theory in the Yang-Mills field, there's nothing that tells us that these solutions are stable. For that, one has to work harder and look at fluctuations of the other fields that do not live in your favourite $U(1)$. (For what it's worth, a constant electric field is stable in Yang-Mills theory, while a constant magnetic field is *unstable*.) We won't discuss these stability issues further in these lectures, largely because our interest lies in what happens in quantum Yang-Mills rather than in the classical theory.

Just as for Maxwell theory, the need to keep gauge invariance means that we can't add a mass term like $A_\mu A^\mu$ or $\text{Tr}\, A_\mu A^\mu$ to the action (1.91). This strongly suggests that quantum Yang-Mills is, like Maxwell theory, a theory of massless particles. This strong suggestion is, it turns out, completely wrong! When we quantise the Yang-Mills action (1.91), we find a theory of interacting massive particles, rather than massless particles. The reason for this can be traced to the interaction terms in Yang-Mills, but is not fully understood. Indeed, proving it from first principles remains one of the most important open problems in mathematical physics. We will discuss this further in section 3.

Coupling to Matter

As with electromagnetism, we can couple the Yang-Mills field to matter. We do this by requiring that the matter fields live in some representation R of the gauge group. This means that the matter fields come in some vector of dimension $\dim R$.

For each such representation, we have generators $T^A(R)$ which we can think of as square matrices of dimension $\dim R$. Dressed resplendent in all their indices, they take the form

$$T^A(R)^a{}_b \quad \text{with } a,b = 1,\ldots,\dim R \;\; \text{and} \;\; A = 1,\ldots,\dim G \; . \tag{1.95}$$

Consider a scalar field in the representation R. Under a gauge transformation $\Omega(x) = e^{ig\alpha^A(x)T^A}$, the scalar transforms as

$$\phi^a \to (\Omega_R)^a{}_b\phi^b \quad \text{with} \quad (\Omega_R)^a{}_b = \Big(\exp\big(ig\alpha^A T^A(R)\big)\Big)^a{}_b \; . \tag{1.96}$$

Some representations R are real, and some are complex. For example, the fundamental representation of $SU(N)$ is complex, and so ϕ must be a complex N-dimensional vector. Meanwhile, the adjoint representation of any group G is always real and, correspondingly, ϕ can be real.

To write down an action for ϕ that is invariant under the gauge transformation (1.96), we follow our Maxwellian noses and construct the covariant derivative,

$$\mathcal{D}_\mu\phi^a = \partial_\mu\phi^a - igA_\mu^A\, T^A(R)^a{}_b\phi^b \; . \tag{1.97}$$

Under a gauge transformation, this covariant derivative transforms, as the name suggests, covariantly, meaning

$$\mathcal{D}_\mu\phi^a \to (\Omega_R)^a{}_b\mathcal{D}_\mu\phi^b \; . \tag{1.98}$$

We will later see that all matter fields in the Standard Model transform in the fundamental representation. For $SU(N)$, this means that we can think of ϕ^a as an N-component complex vector, with $a = 1,\ldots,N$, and write the covariant derivative in terms of the $N\times N$ matrix-valued gauge field $A_\mu = A_\mu^A T^A$,

$$\mathcal{D}_\mu\phi^a = \partial_\mu\phi^a - ig(A_\mu)^a{}_b\phi^b \; . \tag{1.99}$$

This expression differs from our previous covariant derivative (1.88) because ϕ is in the fundamental representation, while α in (1.88) was in the adjoint. This highlights something we've stressed previously: the meaning of the covariant derivative depends on the representation of the object on which it acts. Once again, covariant derivatives do not commute. This time, for covariant derivatives acting on fundamental fields, we find

$$[\mathcal{D}_\mu, \mathcal{D}_\nu] = -igF_{\mu\nu} \; . \tag{1.100}$$

This should be compared to the analogous result (1.90) for covariant derivatives acting on adjoint-valued fields.

As before, it's useful to check some of the formulae for infinitesimal gauge transformations. We have $\delta A_\mu = \mathcal{D}_\mu \alpha$, as in (1.87) and, from (1.96), $\delta\phi = ig\alpha\phi$. Then, suppressing the $a = 1, \ldots, N$ index, the covariant derivative (1.99) transforms as

$$\begin{aligned}\delta(\mathcal{D}_\mu\phi) &= \partial_\mu\delta\phi - ig\,\delta A_\mu\phi - igA_\mu\,\delta\phi \\ &= ig\partial_\mu(\alpha\phi) - ig(\mathcal{D}_\mu\alpha)\phi + g^2A_\mu\alpha\phi \\ &= ig\alpha\,(\partial_\mu\phi - igA_\mu\phi) \\ &= ig\alpha\mathcal{D}_\mu\phi. \end{aligned} \tag{1.101}$$

This is, indeed, the infinitesimal version of the gauge transformation (1.98).

With covariant derivatives that transform nicely, it's straightforward to write down an action for the matter fields. As in electromagnetism, we just need to replace the partial derivatives in the action with covariant derivatives and we have something gauge invariant. This holds for scalars, Weyl fermions, and Dirac fermions.

A Rescaling

Above we've written the action so that the coupling constant g multiplies the non-linear terms. This means, in particular, that it makes an appearance in the field strength (1.84). It also appears, perhaps rather strangely, as the inverse $1/g$ in the gauge transformation (1.82).

There is a different way to normalise the gauge field that, for many purposes, turns out to be more natural. We define the new gauge field

$$\tilde{A}_\mu = gA_\mu \quad \text{and} \quad \tilde{F}_{\mu\nu} = \partial_\mu\tilde{A}_\nu - \partial_\nu\tilde{A}_\mu - i[\tilde{A}_\mu, \tilde{A}_\nu]\ . \tag{1.102}$$

We also define the rescaled gauge parameter $\tilde{\alpha} = g\alpha$, so that the group element is $\Omega = e^{i\tilde{\alpha}}$. This then eliminates the gauge coupling from all kinematic quantities like the field strength and covariant derivatives. The only place that the coupling shows up is in an overall coefficient multiplying the entire action,

$$S_{\rm YM} = -\frac{1}{2}\int d^4x\ \text{Tr}\,F^{\mu\nu}F_{\mu\nu} = -\frac{1}{2g^2}\int d^4x\ \text{Tr}\,\tilde{F}^{\mu\nu}\tilde{F}_{\mu\nu}\ . \tag{1.103}$$

In the first way of writing things, the coupling constant g sits in front of the non-linear terms, making it clear that it governs the strength of interactions. But it also governs the strength of interactions in the second way of writing things. To see this, note that in the Euclidean path integral, we sum over all field configurations weighted by $e^{-S/\hbar}$. With the rescaling above, g^2 sits in the same place in the action as $\hbar$, which suggests that $g^2 \to 0$ will be a classical limit. Heuristically you should think that, for g^2 small, we pay a large price for field configurations that do not minimize the action; in this way, the path integral is dominated by the classical configurations. In contrast, when $g^2 \to \infty$, the Yang-Mills action disappears completely. This is the strong coupling regime, where all field configurations are unsuppressed and contribute equally to the path integral.

The Analogy with General Relativity

General Relativity is rightly lauded for the way it places geometry into the heart of physics. But the other laws of physics, which combine to form the Standard Model, are no less geometrical. Rather than arising from the geometry of spacetime, they instead arise from a slightly more subtle object known as a *fibre bundle*.

We won't describe the mathematics of fibre bundles in any detail in these lectures, but will instead just point out some analogies between the gauge theories discussed above and the differential geometry that underlies general relativity.

One of the key ideas in general relativity is *diffeomorphism invariance*. This is the statement that physical quantities should not depend on the coordinates that we choose to describe them. Such coordinate transformations are analogous to gauge transformations in Yang-Mills theory.

One of the most important objects in general relativity is the Levi-Civita connection $\Gamma^\mu_{\rho\nu}$. Famously, this is *not* a tensor. Under a coordinate transformation $x \to \tilde{x}$, with

$$\Omega^\mu{}_\nu = \frac{\partial x^\mu}{\partial \tilde{x}^\nu} \,, \tag{1.104}$$

the Levi-Civita connection transforms as

$$\Gamma^\mu_{\rho\nu} \to (\Omega^{-1})^\mu{}_\tau \Omega^\sigma{}_\rho \Omega^\lambda{}_\nu \Gamma^\tau{}_{\sigma\lambda} + (\Omega^{-1})^\mu{}_\tau \Omega^\sigma{}_\rho \partial_\sigma \Omega^\tau{}_\nu \,. \tag{1.105}$$

The first term is how a tensor would transform. The second term is independent of Γ and is the characteristic transformation of a connection. But this looks very similar to the transformation of the gauge field (1.82),

$$A_\mu \to \Omega A_\mu \Omega^{-1} + \frac{i}{g} \Omega \, \partial_\mu \Omega^{-1} \tag{1.106}$$

where, again, there is a transformation that befits a tensor, supplemented with the additional derivative term $\partial\Omega$. Indeed, this analogy can be made more precise, and mathematicians refer to the gauge field A_μ as a *connection*. Both connections find their natural home inside covariant derivatives. In gauge theory, this is the $\mathcal{D}_\mu$ that we've already met, while in general relativity it is the object that acts naturally on vector fields Y, with $(\nabla_\nu Y)^\mu = \partial_\nu Y^\mu + \Gamma^\mu_{\nu\rho} Y^\rho$ and is then extended to act on other tensor fields.

Given a Levi-Civita connection, one can construct the Riemann curvature tensor $R^\sigma_{\rho\mu\nu}$. Rearranging some of the indices this can be written as

$$(R_{\mu\nu})^\sigma{}_\rho = \partial_\mu \Gamma^\sigma_{\nu\rho} - \partial_\nu \Gamma^\sigma_{\mu\rho} + \Gamma^\lambda_{\nu\rho}\Gamma^\sigma_{\mu\lambda} - \Gamma^\lambda_{\mu\rho}\Gamma^\sigma_{\nu\lambda} \ . \tag{1.107}$$

Again, we see an immediate similarity with the construction of the field strength in Yang-Mills (1.84) which, including the $a, b = 1, \ldots, \dim F$ indices, reads

$$(F_{\mu\nu})^a{}_b = \partial_\mu (A_\nu)^a{}_b - \partial_\nu (A_\mu)^a{}_b - ig(A_\mu)^a{}_c (A_\nu)^c{}_b + ig(A_\nu)^a{}_c (A_\mu)^c{}_b \ . \tag{1.108}$$

Mathematicians refer to both the Riemann tensor and the field strength $F_{\mu\nu}$ as the *curvature*.

1.4 C,P, and T

Discrete symmetries play a crucial role in understanding the structure of the Standard Model. There are three that are particularly important: parity, charge conjugation, and time reversal. In this section, we describe each of these in turn. We end by explaining why the combination of all three is necessarily a symmetry of any local, relativistic quantum field theory.

1.4.1 Parity

Parity is an inversion of the spatial coordinates,

$$P : (t, \mathbf{x}) \mapsto (t, -\mathbf{x}) \ . \tag{1.109}$$

This can be viewed as a Lorentz transformation, but not one that is continuously connected to the identity. Roughly speaking, the action of parity mimics what a system looks like reflected in the mirror. More precisely, a reflection is implemented by, say, $R : (x, y, z) \mapsto (x, y, -z)$. The parity transformation (1.109), which is a reflection followed by a rotation by 180°, has the advantage that it treats all spatial coordinates on the same footing.

(As an aside: one disadvantage of the parity transformation $P : \mathbf{x} \mapsto -\mathbf{x}$ is that it only works when the number of spatial dimensions is odd. For example, in $d = 2 + 1$ dimensions, the transformation $(x, y) \mapsto (-x, -y)$ is just a rotation by 180°. For this reason, if you're discussing quantum field theories in different dimensions, it's better to talk about reflections which flip the sign of just one spatial direction, rather than parity which flips all of them. In these lectures, we've got no interest in dimension hopping: our interest is strictly in the Standard Model and so we keep with the conventional definition of parity (1.109).)

We would like to understand the circumstances under which a quantum field theory is invariant under parity, and how the fields transform. When we come to discuss the weak force in Section 5, we will find that the laws of our universe are *not* invariant under parity. This is a shocking statement. It means that given a solution to the equations of motion, the parity reflected evolution is *not* a solution!

First, let's ask how electromagnetic fields transform under parity. For this, we can look at the covariant derivative which, regardless of the object it acts on, takes the schematic form

$$\mathcal{D}_\mu = \partial_\mu - iA_\mu \ . \tag{1.110}$$

This ties the behaviour of the gauge field to that of the derivative. Under a parity transformation ∂_0 is left unaffected, while the spatial derivatives ∂_i change sign. This tells us that parity must act as

$$P : A_0(t, \mathbf{x}) \mapsto +A_0(t, -\mathbf{x}) \quad \text{and} \quad P : A_i(t, \mathbf{x}) \mapsto -A_i(t, -\mathbf{x}) \ . \tag{1.111}$$

Tracing this through to the definitions of the electric field $\mathbf{E} = -\nabla\phi - \partial\mathbf{A}/\partial t$ and magnetic field $\mathbf{B} = \nabla \times \mathbf{A}$, we have

$$P : \mathbf{E}(t, \mathbf{x}) \mapsto -\mathbf{E}(t, -\mathbf{x}) \quad \text{and} \quad P : \mathbf{B}(t, \mathbf{x}) \mapsto +\mathbf{B}(t, -\mathbf{x}) \ . \tag{1.112}$$

Vectors like $\mathbf{E}$, which transform under parity in the same way as $\mathbf{x}$ are deemed worthy to keep the name "vector". Meanwhile, vectors like $\mathbf{B}$ which don't pick up a minus sign under parity are said to be *pseudovectors*. The most familiar examples of pseudovectors are the magnetic field and angular momentum $\mathbf{L} = \mathbf{x} \times \mathbf{p}$. These are also the two kinds of vectors that exhibit the most counterintuitive behaviour when we're undergraduates. This is not a coincidence.

In the quantum theory, the parity transformation is enacted by a unitary operator on the Hilbert space that we also call P. The fields $A_\mu(x)$ are now also operators and the transformation (1.111) becomes

$$PA_0(t,\mathbf{x})P^\dagger = A_0(t,-\mathbf{x}) \quad \text{and} \quad PA_i(t,\mathbf{x})P^\dagger = -A_i(t,-\mathbf{x}) \ . \tag{1.113}$$

In what follows, we will flit between the description of parity and other discrete symmetries as a map, as in (1.111), and as an operator acting on a Hilbert space, as in (1.113).

Next, we turn to spinors. It can be somewhat fiddly to figure out how spinors transform under various discrete symmetries, but it's a topic that will play a crucial role as we proceed. The equation of motion for a left-handed massless Weyl spinor ψ_L is

$$\bar{\sigma}^\mu \partial_\mu \psi_L = 0 \tag{1.114}$$

where $\bar{\sigma} = (\mathbb{1}, -\sigma^i)$. Under a parity transformation, the spatial derivative changes sign and the Weyl equation (1.114) is not invariant. This is important: if we have just a single left-handed Weyl spinor ψ_L then this theory is *not* invariant under parity.

We can rescue the situation if, in addition to our left-handed Weyl spinor ψ_L, we also have a right-handed Weyl spinor ψ_R. This obeys the equation of motion

$$\sigma^\mu \partial_\mu \psi_R = 0 \tag{1.115}$$

where $\sigma^\mu = (\mathbb{1}, \sigma^i)$. The different minus signs in σ^μ and $\bar{\sigma}^\mu$ mean that we can compensate for a parity transformation if we also exchange left- and right-handed spinors, so that

$$P\psi_L(t,\mathbf{x})P^\dagger = \psi_R(t,-\mathbf{x}) \quad \text{and} \quad P\psi_R(t,\mathbf{x})P^\dagger = \psi_L(t,-\mathbf{x}) \ . \tag{1.116}$$

There are also options to put different minus signs (and even phases) on the right-hand side as we describe below.

As we've seen in Section 1.2.1, the two spinors ψ_L and ψ_R naturally sit in a Dirac spinor $\psi = (\psi_L, \psi_R)^T$. The action of parity on Weyl spinors (1.116) translates into the action on the Dirac spinor

$$P\psi(t,\mathbf{x})P^\dagger = \gamma^0 \psi(t,-\mathbf{x}) \quad \text{with} \quad \gamma^0 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \ . \tag{1.117}$$

In the lectures on Quantum Field Theory, we saw that a stationary fermion is associated to a solution to the Dirac equation, where the spinor degrees of freedom take the form $\psi = (\xi, \xi)^T$. Here ξ is some 2-component spinor the tells us the orientation of the spin of the particle. Meanwhile, the solution corresponding to an anti-fermion takes the form $\psi = (\xi, -\xi)^T$. This means that the fermion has intrinsic parity $+1$ while the anti-fermion has intrinsic parity -1.

Terms in the action are always constructed out of an even number of fermions. Given the transformation (1.117), we can look at the fate of various fermion bilinears under parity. You can check, for example, that

$$P : \bar{\psi}\psi \mapsto \bar{\psi}\psi \quad \text{and} \quad P : \bar{\psi}\gamma^5\psi \mapsto -\bar{\psi}\gamma^5\psi \tag{1.118}$$

where we've suppressed the all-important spinor indices. We say that $\bar{\psi}\psi$ transforms as a *scalar* while $\bar{\psi}\gamma^5\psi$ transforms as a *pseudoscalar*. Similarly, you can check that $\bar{\psi}\gamma^\mu\psi$ is a vector while $\bar{\psi}\gamma^5\gamma^\mu\psi$ is a pseudovector.

You shouldn't be too dogmatic about insisting that (1.116) and (1.117) are the definitive action of parity. Suppose that you have a Dirac fermion with action

$$S = \int d^4x \, \left(i\bar{\psi}\gamma^\mu\partial_\mu\psi - M\bar{\psi}\psi \right) . \tag{1.119}$$

Then this is invariant under parity with the transformation (1.117). Suppose, in contrast, that you're given the action

$$S = \int d^4x \, \left(i\bar{\psi}\gamma^\mu\partial_\mu\psi - M\bar{\psi}\gamma^5\psi \right) . \tag{1.120}$$

This is not invariant under (1.117) because the mass term is parity odd. Nonetheless, that doesn't mean that the theory doesn't have parity symmetry. We just need to look more carefully. You can check that the action (1.120) is invariant under the redefined parity transformation

$$P'\psi(t, \mathbf{x})P'^{\dagger} = \gamma^5\gamma^0\psi(t, -\mathbf{x}) . \tag{1.121}$$

In terms of Weyl fermions, this inserts an extra minus sign on the right-hand side of one of the transformations in (1.116). Ultimately, given a theory the aim is to find some parity transformation of the fields that leaves the action, and hence the equation of motion, invariant.

So far, we haven't discussed the action of parity on scalar fields. These are more malleable. Given a scalar field ϕ, the kinetic terms are invariant under either

$$P\phi(t,\mathbf{x})P^\dagger = \pm\phi(t,-\mathbf{x}) \ . \tag{1.122}$$

In other words, the kinetic terms don't distinguish between scalar (the plus sign) or pseudoscalar (the minus sign). Typically, this gets fixed when we look at the interaction of the scalar field with fermions. For example, a Yukawa term of the form $\phi\bar{\psi}\psi$ means that the scalar ϕ is parity even under the transformation (1.117) while a Yukawa term of the form $\phi\bar{\psi}\gamma^5\psi$ means that ϕ is parity odd under (1.117).

There are various pay-offs from understanding the way that parity is implemented in a theory. If a theory is invariant under parity then, as we've seen, we can assign transformation laws to the various fields. But, after quantisation, these fields give rise to particles. That means that different species of particles can be thought of as parity even or parity odd. Moreover, this concept of parity is conserved in all interactions and, like all conservation laws, this puts constraints on the kind of things that can happen.

Perhaps surprisingly, it turns out that things are even more constrained when parity is *not* a symmetry of the theory! This is for a much more subtle reason known as an *anomaly*. We will discuss this in Section 4.

1.4.2 Charge Conjugation

Charge conjugation is an operation that switches particles with their anti-particles. If a theory is invariant under charge conjugation, then the laws of physics that govern particles coincide with those that govern anti-particles.

This time we start with a complex scalar field ϕ, coupled to electromagnetism. It will prove simplest to look at actions, rather than equations of motion. Charge conjugation exchanges particles and anti-particles, so we want it to act as

$$C : \phi \mapsto \pm\phi^\dagger \ . \tag{1.123}$$

The $\pm$ ambiguity is like the ambiguity in the action of parity (1.122) and, as in that case, will typically be fixed by the interactions with other fields. In contrast, there's no ambiguity about the action on the gauge field, which is fixed by looking at the covariant derivatives, $\mathcal{D}_\mu\phi = (\partial_\mu - ieA_\mu)\phi$ and $\mathcal{D}_\mu\phi^\dagger = (\partial_\mu + ieA_\mu)\phi^\dagger$. This means that any transformation (1.123) must be accompanied by

$$C : A_\mu \mapsto -A_\mu \ . \tag{1.124}$$

As for parity, we can also think of charge conjugation as a quantum operator C, in which case (1.123) and (1.124) are replaced by $C\phi C^\dagger = \pm\phi^\dagger$ and $CA_\mu C^\dagger = -A_\mu$ respectively. For non-Abelian gauge fields, charge conjugation acts as $CA_\mu C^\dagger = -A_\mu^\dagger$.

Again, the story for spinors is a little more fiddly. We'll start by looking at a Dirac spinor, rather than a Weyl spinor. The Dirac equation is

$$i\gamma^\mu(\partial_\mu - ieA_\mu)\psi - M\psi = 0 \ . \tag{1.125}$$

We will look for an action of charge conjugation that transforms the spinor to

$$C : \psi \mapsto C\psi^\star \ . \tag{1.126}$$

Here C on the right-hand side is a 4×4 matrix that allows for the possibility that the components of the spinor get mixed up under charge conjugation. Note that we've written the transformed spinor as $\psi^\star$, rather than $\psi^\dagger$, to emphasise that it remains a "column vector" rather than a "row vector". (Of course, it's not really a vector at all. It's a spinor!)

The question is: what choice of C ensures that the transformation (1.126), combined with (1.124), is a symmetry? First, we take the complex conjugate of the equation of motion (1.125):

$$-i(\gamma^\mu)^\star(\partial_\mu + ieA_\mu)\psi^\star - M\psi^\star = 0 \ . \tag{1.127}$$

This is the equation that $\psi^\star$ obeys. Next, we compare this to what we get if we act with charge conjugation on the original equation (1.125):

$$\begin{aligned} & i\gamma^\mu(\partial_\mu + ieA_\mu)C\psi^\star - MC\psi^\star = 0 \\ \Longrightarrow \quad & iC^\dagger\gamma^\mu C(\partial_\mu + ieA_\mu)\psi^\star - M\psi^\star = 0 \ . \end{aligned} \tag{1.128}$$

We see that (1.128) coincides with (1.127) provided that the charge conjugation matrix C obeys

$$C^\dagger\gamma^\mu C = -(\gamma^\mu)^\star \ . \tag{1.129}$$

The charge conjugation matrix depends on your chosen basis of gamma matrices. For the chiral basis of gamma matrices (1.42), all gamma matrices are real except for γ^2 which is pure imaginary. This means that we should take $C = \pm i\gamma^2$, and the action of charge conjugation is

$$C : \psi \mapsto \pm i\gamma^2\psi^\star \quad \text{with} \quad \gamma^2 = \begin{pmatrix} 0 & \sigma^2 \\ -\sigma^2 & 0 \end{pmatrix} \ . \tag{1.130}$$

For theories that are invariant under charge conjugation, we can assign an eigenvalue $C = \pm 1$ to each particle, usually referred to as *C-parity*. As with actual parity, P, this new quantum number restricts the possible interactions. For example, it turns out that the neutral pion π^0 has $C = +1$ while, from (1.124), the photon necessarily has $C = -1$. This means that the decay to two photons, $\pi^0 \longrightarrow \gamma + \gamma$, is allowed (and indeed, happens over 98% of the time). But the decay to three photons, $\pi^0 \longrightarrow \gamma+\gamma+\gamma$ is forbidden on symmetry grounds.

If we decompose the Dirac fermion into its two Weyl components, $\psi = (\psi_L, \psi_R)^T$, then we can read off from (1.130) the action of charge conjugation on Weyl spinors,

$$C : \psi_L \mapsto \pm i\sigma^2 \psi_R^\star \quad \text{and} \quad C : \psi_R \mapsto \mp i\sigma^2 \psi_L^\star \ . \tag{1.131}$$

We see that charge conjugation, like parity, involves an exchange of two Weyl spinors.

A theory with just a single Weyl fermion is invariant under neither parity nor charge conjugation. However, there's still hope if we combine the two symmetries. We can take the combined action from (1.116) and (1.131) to be

$$CP : \psi_L(t, \mathbf{x}) \mapsto \mp i\sigma^2 \psi_L^\star(t, -\mathbf{x}) \quad \text{and} \quad CP : \psi_R(t, \mathbf{x}) \mapsto \pm i\sigma^2 \psi_R^\star \ . \tag{1.132}$$

A Weyl fermion coupled to a gauge field is invariant under CP. However, as we will see later, it's quite possible for this symmetry to be violated by other interaction terms (for example, Yukawa interactions between fermions and scalars).

1.4.3 Time Reversal

Our final discrete symmetry is time reversal, which acts on spacetime coordinates as

$$T : (t, \mathbf{x}) \mapsto (-t, \mathbf{x}) \ . \tag{1.133}$$

There's a subtlety in implementing time reversal symmetry in quantum theories. This manifests itself already in the simplest quantum mechanical systems like, say, a free particle moving in $\mathbb{R}^3$. The Schrödinger equation for the wavefunction Ψ takes the form

$$i\frac{\partial \Psi}{\partial t} = -\nabla^2 \Psi \ . \tag{1.134}$$

Now compare this to the heat equation that describes how conserved quantities, such as temperature T, diffuse in a system

$$\frac{\partial T}{\partial t} = \nabla^2 T \ . \tag{1.135}$$

The heat equation most certainly isn't time reversal invariant since the left-hand side picks up a minus sign, while the right-hand side does not. That's to be expected: after all, diffusion is a process that increases entropy and there's a clear arrow of time as things spread out. In contrast, there's no increase in entropy for a single quantum particle and we do expect the physics to be invariant under time reversal. Yet the Schrödinger equation is almost identical to the heat equation in structure. How can one be time reversal invariant, and the other not?

Almost identical, but not quite. The key is that factor of i in the Schrödinger equation that is not there in the heat equation. Suppose that $\Psi(t)$ is a solution to the Schrödinger equation. Then $\Psi(-t)$ is *not* a solution but the factor of i means that $\Psi^\star(-t)$ is. That's the clue that we need: time reversal in quantum mechanics acts as

$$T : \Psi(t) \mapsto \Psi^\star(-t) \; . \tag{1.136}$$

Viewed as an operator acting on the Hilbert space, this complex conjugation translates into the requirement that T is an *anti-unitary* operator, rather than the more familiar unitary operator. This means that, acting on states, we have

$$T(\alpha|\psi_1\rangle + \beta|\psi_2\rangle) = \alpha^\star T|\psi_1\rangle + \beta^\star T|\psi_2\rangle \; . \tag{1.137}$$

In addition, the operator obeys

$$\langle T\psi_1|T\psi_2\rangle = \langle\psi_1|\psi_2\rangle^\star \; . \tag{1.138}$$

See the lectures on Topics in Quantum Mechanics for more discussion of the action of the time reversal in quantum mechanics.

This anti-linear behaviour changes some of the transformation properties of fields. For example, you might naively think, following (1.111), that A_0 would be odd under time reversal and A_i even. But, in fact, it's the opposite way around because there's an additional factor of i in the covariant derivative $\mathcal{D}_\mu = \partial_\mu - ieA_\mu$ which gets conjugated. It means that the action of time reversal on the gauge field is

$$T : A_0(t, \mathbf{x}) \mapsto +A_0(-t, \mathbf{x}) \quad \text{and} \quad T : A_i(t, \mathbf{x}) \mapsto -A_i(-t, \mathbf{x}) \; . \tag{1.139}$$

Tracing this through to the electric field $\mathbf{E} = -\nabla A_0 - \partial\mathbf{A}/\partial t$ and magnetic field $\mathbf{B} = \nabla \times \mathbf{B}$, we have

$$T : \mathbf{E}(t, \mathbf{x}) \mapsto +\mathbf{E}(-t, \mathbf{x}) \quad \text{and} \quad T : \mathbf{B}(t, \mathbf{x}) \mapsto -\mathbf{B}(-t, \mathbf{x}) \; . \tag{1.140}$$

This makes sense: it's the same transformation that we get from the Lorentz force law $m\ddot{\mathbf{x}} = q(\mathbf{E} + \dot{\mathbf{x}} \times \mathbf{B})$.

What about fermions? Once again, the action of time reversal can mix the different components of a Dirac spinor. As we now show, it turns out that (for our chiral basis of gamma matrices (1.42)) the correct transformation is

$$T: \psi(t,\mathbf{x}) \mapsto \Theta\psi(-t,\mathbf{x}) \quad \text{where} \quad \Theta = \gamma^1\gamma^3. \tag{1.141}$$

As for other transformations, we could also include a minus sign on the right-hand side. To see that (1.141) is indeed a symmetry, consider the action of time reversal on the Dirac equation (1.125). Remembering that time reversal also acts by complex conjugation (so, for example, changes γ^μ to $(\gamma^\mu)^\star$), we have

$$\begin{aligned} & -i\big(-(\gamma^0)^\star\mathcal{D}_0 + (\gamma^i)^\star\mathcal{D}_i\big)\Theta\psi - M\Theta\psi = 0 \\ \Longrightarrow \quad & i\Theta^{-1}\big((\gamma^0)^\star\mathcal{D}_0 - (\gamma^i)^\star\mathcal{D}_i\big)\Theta\psi - M\psi = 0. \end{aligned} \tag{1.142}$$

This gives us back the original Dirac equation if the matrix Θ obeys

$$\Theta^{-1}(\gamma^0)^\star\Theta = \gamma^0 \quad \text{and} \quad \Theta^{-1}(\gamma^i)^\star\Theta = -\gamma^i\ . \tag{1.143}$$

It's simple to check that, for the chiral basis of gamma matrices (1.42), $\Theta = \gamma^1\gamma^3$ does the job. We can also translate this to the action on the component Weyl spinors $\psi = (\psi_L, \psi_R)^T$,

$$T: \psi_L(t,\mathbf{x}) \mapsto i\sigma^2\psi_L(-t,\mathbf{x}) \quad \text{and} \quad T: \psi_R(t,\mathbf{x}) \mapsto i\sigma^2\psi_R(-t,\mathbf{x})\ . \tag{1.144}$$

We see that time reversal, like CP, does not mix the left- and right-handed Weyl spinors.

What would it mean for a quantum field theory to break time-reversal invariance? It sounds rather cool. In practice, however, a breaking of time reversal manifests itself in rather mundane ways. One simple example is the presence of an electric dipole moment for particles. Recall from the lectures on Electromagnetism that an electric dipole moment arises from two, equal and opposite, closely separated charges and gives rise to an electric field that drops off as $1/r^3$.

The dipole moment points in a particular direction. For an elementary particle, this direction must align with the spin otherwise the particle would pick a preferred direction in space and so break Lorentz invariance. But the spin and dipole moment transform differently under both parity and time-reversal. To see this, recall that spin $\mathbf{S}$ is a form of angular momentum $\mathbf{L} = m\mathbf{x}\times\dot{\mathbf{x}}$, which is even under parity and odd under time reversal. Hence, we have

$$\begin{aligned} P: \mathbf{S} \mapsto \mathbf{S} \quad & \text{and} \quad T: \mathbf{S} \mapsto -\mathbf{S} \\ P: \mathbf{E} \mapsto -\mathbf{E} \quad & \text{and} \quad T: \mathbf{E} \mapsto \mathbf{E}\ . \end{aligned} \tag{1.145}$$

This means that discovery of a dipole moment for a fundamental particle would imply that the laws of physics break both parity and time reversal invariance. The search for the electric dipole moment of the neutron remains one of the most direct ways to test for time-reversal breaking in the strong nuclear force. So far, no such breaking has been found. (We discuss this further in Section 3.4.) As we will see later, the weak force does break both parity P and, to a lesser extent, time reversal T. This results in a theoretical prediction for the electric dipole moment of the electron, albeit one that is far below current experimental bounds.

1.4.4 CPT

There are theories that are invariant under our three discrete symmetries, C, P and T, and other theories that break them. As we will see, the Standard Model is in the latter class and all three symmetries are broken.

However, there is a theorem that says that all relativistic quantum field theories must necessarily be invariant under the combined action of CPT. In other words, if you look at anti-particles in the mirror, with their motion reversed, then you will have a symmetry on your hands.

One somewhat workaday proof of the CPT theorem is to simply write down all possible Lorentz invariant terms and check that they are indeed invariant under CPT. As we've seen, the most subtle transformations are those of spinors. For example, combining our previous results (1.117), (1.126) and (1.141), we find that a Dirac spinor is transformed by the anti-unitary operation

$$CPT : \psi(x) \mapsto -\gamma^5 \psi^\star(-x) \quad \text{with} \quad \gamma^5 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \,. \tag{1.146}$$

You can check that all fermion bilinears are invariant under this transformation. For example,

$$\bar{\psi}\psi = \psi^\dagger \gamma^0 \psi \mapsto \psi^T \gamma^5 \gamma^0 \gamma^5 \psi^\star = -\psi^T \gamma^0 \psi^\star = \bar{\psi}\psi \tag{1.147}$$

where, in the final equality, we reordered the fermions and picked up a minus sign for our troubles due to their Grassmann nature. The pseudoscalar $\bar{\psi}\gamma^5\psi$ is also invariant by a similar argument, while both $\bar{\psi}\gamma^\mu\psi$ and $\bar{\psi}\gamma^\mu\gamma^5\psi$ transform as vectors, rather than pseudovectors (meaning that they pick up minus signs) which ensures that any kinetic term we write down is invariant. (For this, you will need to use the fact that $\gamma_1^T = -\gamma_1$ and $\gamma_3^T = -\gamma_3$ while $\gamma_0^T = \gamma_0$ and $\gamma_2^T = \gamma_2$.)

A slightly more elegant, but not entirely convincing, demonstration of CPT follows from Wick rotating to Euclidean space. Here we sketch the basic idea. The full Lorentz group in Minkowski space is really $O(1,3)$ and contains four disconnected components, with the actions of parity and time reversal taking us from one component to the other. In contrast, in Euclidean space the group becomes $O(4)$ and this contains only two disconnected components. If you follow the Lorentzian CPT under a Wick rotation, it becomes simply a rotation in $SO(4)$, i.e. a transformation that is connected to the identity. (The need to include C here is roughly because particles are like anti-particles travelling backwards in time.) This means that if your Euclidean theory is to have $SO(4)$ rotational invariance, then your Lorentzian theory must enjoy CPT.

The statement that CPT is a symmetry of all relativistic quantum field theories is something that we can test. Here's an example from neutrino physics. We will learn later that neutrinos oscillate from one flavour to another as they travel through space. So, for example, a muon neutrino ν^μ will have some probability to convert into an electron neutrino ν^e, a process that we write as

$$\nu^\mu \longrightarrow \nu^e \ . \tag{1.148}$$

We could also consider the CP conjugate process, namely

$$\bar{\nu}^\mu \longrightarrow \bar{\nu}^e \ . \tag{1.149}$$

There is no reason for the amplitudes for these two processes to be equal if CP is broken. However, there is also the time reversed process of (1.148)

$$\nu^e \longrightarrow \nu^\mu \ . \tag{1.150}$$

This too may have a different amplitude to (1.148) if time reversal is broken. However, CPT tells us that the amplitude for (1.149) and the amplitude for (1.150) are necessarily equal. Indeed, all experimental tests so far have failed to find any violation of CPT.

2 Broken Symmetries

Global symmetries have two important roles to play in physics. First, they lead to conservation laws through Noether's theorem. Second, if the symmetry is non-Abelian then it leads to a degeneracy in the spectrum, as the states of the theory necessarily furnish a representation of the symmetry. This is familiar from the quantum treatment of the hydrogen atom where states sit in multiplets of the $SO(3)$ rotation group of dimension $2l + 1$ where l is the angular momentum.

But there are other ways in which symmetries can affect the dynamics of a theory. And this happens when symmetries are "broken".

There are actually two different meanings to the phrase "broken symmetry", both of which arise in the context of the Standard Model. The first, sometimes called *explicit breaking*, is when there are terms in the action that are not invariant under the symmetry. Strictly speaking, this is the same as not having a symmetry at all. But the symmetry can still be a useful fiction if the terms that break it are, in some sense, small so that we have an approximate symmetry. In this case, it might be that some quantity is almost conserved, meaning that violations of the conservation law happen rarely. Or it could be that the degenerate multiplets that arose when the symmetry was exact are split by some small amount. This happens, for example, if we place the hydrogen atom in a magnetic field so that the rotation symmetry is broken. Then the $2l+1$ states which were previously all degenerate get slightly split by the Zeeman effect.

In the Standard Model, we will see several examples of approximate symmetries, including isospin and its extension to an $SU(3)$ flavour symmetry known as the *eightfold way*, as well as chiral symmetry. Both of these will be explained in section 3.

The second meaning of the term "broken symmetry" refers to a more subtle and, ultimately, more powerful phenomenon. This arises when the theory is invariant under a symmetry, but the ground state is not. This situation is referred to as *spontaneous symmetry breaking*. The purpose of this section is to explain when this happens and what the consequences are.

Spontaneous symmetry breaking is one of those lovely ideas that crosses into many different areas of physics. It was one of the major themes of the lectures on Statistical Field Theory where it underlies Landau's theory of phase transitions. It also arises in many places in condensed matter physics, from magnets to superconductors. For example, sound waves in a solid can be viewed as the consequence of spontaneous breaking of translation symmetry by the underlying lattice. Spontaneous symmetry breaking also occurs in at least two different contexts in the Standard Model.

2.1 Discrete Symmetries

The idea of spontaneous symmetry breaking is not something new: it appears in some simple classical mechanics systems.

Consider a real, classical degree of freedom $\phi(t)$ with action given by

$$S = \int dt \left(\frac{1}{2}\dot{\phi}^2 - V(\phi) \right) \quad \text{with} \quad V(\phi) = \frac{m^2}{2}\phi^2 + \frac{\lambda}{4}\phi^4 \ . \tag{2.1}$$

In Newtonian mechanics, we would think of $\phi(t)$ as the position of a particle and usually denote it as $x(t)$. We're going to avoid calling the degree of freedom x because we'll soon make the leap to field theory where x becomes an argument of the field, $\phi(\mathbf{x}, t)$. But you should feel free to think of $\phi(t)$ as the position of a particle.

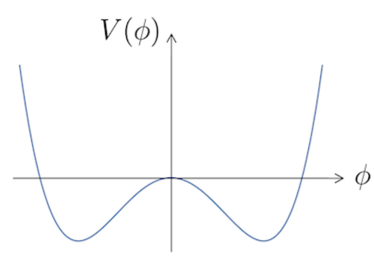
The potential (2.1) enjoys a discrete $\mathbb{Z}_2$ symmetry under which

$$\mathbb{Z}_2 : \phi \mapsto -\phi \ . \tag{2.2}$$

In classical mechanics, where ϕ is the position of the particle, this symmetry is called "parity" but we'll avoid this name because, again, in the context of field theory parity acts differently (as we saw in Section 1.4).

We will assume that $\lambda > 0$. In that case, the issue of spontaneous symmetry breaking is all about the sign of the first term in the potential. When $m^2 > 0$, the potential has a minimum at $\phi = 0$. This is the one point that is invariant under the symmetry $\phi \mapsto -\phi$ and we say that the symmetry is unbroken.

In contrast, if $m^2 < 0$ then the ϕ^2 term in (2.1) comes with a negative coefficient and the point $\phi = 0$ is now a local maximum rather than a minimum, as shown in the figure. This is the *double well potential*. The minimum lies at



$$\phi = \pm v \equiv \pm\sqrt{-\frac{m^2}{\lambda}} \ . \tag{2.3}$$

We see that two related things occur. First, there is not a unique ground state: there are two. Second, neither ground state is invariant under the $\mathbb{Z}_2$ symmetry (2.2). Instead, the symmetry exchanges the two ground states. This is our first, admittedly somewhat trivial, example of spontaneous symmetry breaking. But there is an important lesson that will carry over to more complicated situations: if a discrete symmetry is spontaneously broken, then the theory has multiple, ground states with a potential barrier between them. Acting with the symmetry then transforms us among the ground states.

Suppose that you sit in one of the two ground states, and look only at small oscillations about the minimum. What do you see? We write the potential (2.1) as

$$V(\phi) = \frac{\lambda}{4}(\phi^2 - v^2)^2 + \text{constant} \ . \tag{2.4}$$

We take ourselves to sit near the ground state $\phi = +v$ and expand

$$\phi(t) = v + \sigma(t) \ . \tag{2.5}$$

We can then substitute this back into the potential (2.4) to get

$$V(\sigma) = \frac{\lambda}{4}(2v\sigma + \sigma^2)^2 = \lambda\left(v^2\sigma^2 + v\sigma^3 + \frac{\sigma^4}{4}\right) + \text{constant} \ . \tag{2.6}$$

We see that, while the full potential $V(\phi)$ has the $\mathbb{Z}_2$ symmetry, if you're trapped near one of the minima then you know nothing about it. The action for small oscillations includes the σ^3 term and most certainly isn't invariant under $\sigma \mapsto -\sigma$. This is the sense in which the $\mathbb{Z}_2$ symmetry is hidden, or broken, about any given ground state. The consequence of the symmetry, when broken, is only to generate multiple ground states.

2.1.1 Quantum Tunnelling

The discussion above is straightforward enough and holds for classical particle mechanics. But quantum mechanics brings an extra twist. This is because there is no spontaneous symmetry breaking in quantum mechanics! The ground state is always invariant under the $\mathbb{Z}_2$ symmetry. In fact, all energy eigenstates are invariant under the $\mathbb{Z}_2$ symmetry.

You might be tempted to construct a ground state that is localised near one or other of the minima, say a wavefunction of the form

$$\psi_{\text{left}}(\phi) \approx \exp\left(-\frac{\sqrt{\lambda}v}{2}(\phi + v)^2\right) \quad \text{or} \quad \psi_{\text{right}}(\phi) \approx \exp\left(-\frac{\sqrt{\lambda}v}{2}(\phi - v)^2\right) . \tag{2.7}$$

But neither of these are eigenstates of the $\mathbb{Z}_2$ symmetry, and neither are eigenstates of the Hamiltonian. Indeed, if you were to place the system in, say, $\psi_{\text{left}}(\phi)$ then the wavefunction will leak through the barrier in a process known as *quantum tunnelling*.

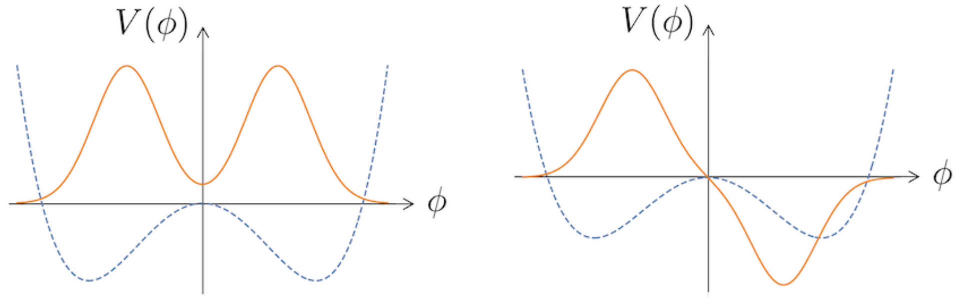


Figure 2. On the left: the ground state of the double well potential. On the right: the first excited state.

Instead, the true ground state wavefunction takes the approximate form

$$\psi_{\text{ground}}(\phi) \approx \psi_{\text{left}}(\phi) + \psi_{\text{right}}(\phi) \ . \tag{2.8}$$

The ground state has no zeros other than at $\phi \to \pm\infty$. Meanwhile, the first excited state is

$$\psi_{\text{excited}}(\phi) \approx \psi_{\text{left}}(\phi) - \psi_{\text{right}}(\phi) \ . \tag{2.9}$$

This has a single node, meaning that it crosses the axis once. The n^{th} excited state has n nodes. (See the lectures on Quantum Mechanics for more discussion of these facts.) The ground state and first excited state are shown in Figure 2.

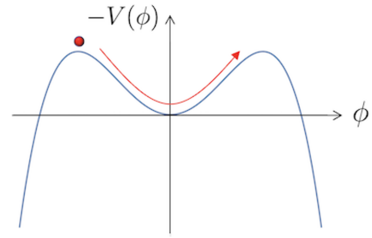
There is another way to see tunnelling that will prove useful when we turn to quantum field theory shortly. We want to compute the amplitude for a particle to start in one minimum, say $\phi = -v$, and end up at the other minimum $\phi = +v$. We can do this using the path integral. After Wick rotating to work with imaginary time $\tau = it$, we have

$$\langle +v|e^{-H\tau}|-v\rangle = \int \mathcal{D}\phi \ e^{-S_E[\phi]} \ . \tag{2.10}$$

Here $S_E[\phi]$ is the "Euclidean action", meaning that is differs from (2.1) by a minus sign.

$$S_E[\phi] = \int d\tau \ \left(\frac{1}{2}\dot{\phi}^2 + V(\phi)\right) \ . \tag{2.11}$$

To compute the amplitude (2.10), we should evaluate the path integral on paths that start in the left-hand vacuum and end up at the right-hand vacuum. We can get some intuition for this by noting that the Euclidean action (2.11) simply flips the sign of the potential term, so if we wished to view it as a classical mechanics system then it describes a particle rolling in the inverted potential $-V(\phi)$. We're then looking for paths that start perched on the left-hand peak, roll down to the minimum, and then rise again to end on the right-hand peak, as shown in the figure.

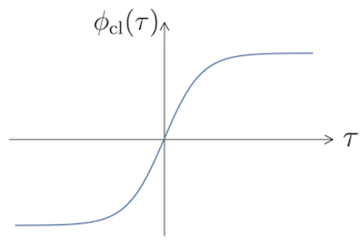


The path integral instructs us to integrate over all such paths. But, in the saddle point approximation, we expect the dominant contribution to come from paths that obey the classical equation of motion,

$$\ddot{\phi} = \lambda\phi(\phi^2 - v^2) \ . \tag{2.12}$$

This equation has a rather nice analytic solution that does what we want, namely

$$\phi_{\rm cl}(\tau) = v \tanh\left(\sqrt{\frac{\lambda v^2}{2}}\tau\right) \ . \tag{2.13}$$



The profile is shown in the figure to the right. It interpolates from $\phi = -v$ to $\phi = +v$, with the interesting stuff happening over a time period $\Delta\tau \sim 1/\sqrt{\lambda v^2} \sim 1/|m|$. We can evaluate the Euclidean action (2.11) on this solution to get

$$\begin{aligned} S_{\rm cl} &= \int_{-\infty}^{+\infty} d\tau \ \left(\frac{1}{2}\dot{\phi}_{\rm cl}^2 + \lambda(\phi_{\rm cl}^2 - v^2)^2\right) \\ &= \frac{\lambda v^4}{2}\int_{-\infty}^{+\infty} d\tau \ \frac{1}{\cosh^4(\sqrt{\lambda v^2/2}\tau)} \\ &= \frac{2}{3}\sqrt{2\lambda}v^3 \ . \end{aligned} \tag{2.14}$$

This can be viewed as a measure of how difficult it is to tunnel under the barrier. As the barrier gets bigger (so λ increases) or the minima get further apart (so v^2 increases), the classical action $S_{\rm cl}$ also increases. This then gives our first guess at the amplitude to tunnel from one minimum to the other,

$$\lim_{\tau\to\infty} \langle +v|e^{-H\tau}| -v\rangle = Ke^{-S_{\rm cl}}. \tag{2.15}$$

Here K is some overall constant that masks all manner of sins that we've swept under the rug. In fact, to do this calculation correctly, we should really be summing over trajectories that bounce back and forth many times. One then finds, in the limit of large T, that you have just as much chance of being in the vacuum $\phi = -v$ as you do of being in the vacuum $+v$. This is the statement that there is no spontaneous symmetry breaking in quantum mechanics. Moreover, you find that the energy difference between the ground state and first excited state is given by

$$E_{\text{excited}} - E_{\text{ground}} \approx \sqrt{\lambda v^2} e^{-S_{\text{cl}}} \ . \tag{2.16}$$

The splitting of the two states is exponentially suppressed.

With these ideas in mind, we can now return to what we really care about: quantum field theory.

2.1.2 Discrete Symmetry Breaking in Quantum Field Theory

We now extend our double well discussion to field theory. Now $\phi(x)$ is a function of spacetime. The action (2.1) is replaced by

$$S = \int d^4x \ \left(\frac{1}{2}\partial_\mu \phi \, \partial^\mu \phi - V(\phi)\right) \quad \text{with} \quad V(\phi) = \frac{m^2}{2}\phi^2 + \frac{\lambda}{4}\phi^4 \ . \tag{2.17}$$

Again, we have a $\mathbb{Z}_2$ symmetry $\phi \mapsto -\phi$ and, when $m^2 < 0$, we have a double well potential with two minima at $\phi = \pm v = \pm\sqrt{-m^2/\lambda}$. We want to ask: is this symmetry spontaneously broken or not?

Quantum field theory is an extension of quantum mechanics (the clue is in the name) so we might think that tunnelling would again mean that there is no spontaneous symmetry breaking. But that's not the way things work. This is one situation where field theory differs from quantum mechanics and our classical intuition is better. The quantum field theory really does have two ground states, in which the vacuum expectation value of the field is given by

$$\langle \phi \rangle = \pm v \ . \tag{2.18}$$

To see why quantum field theory is different from common or garden quantum mechanics, we can return to the tunnelling calculation that we saw above. We can again compute the amplitude to go from one putative ground state to another,

$$\langle +v | e^{-H\tau} | -v \rangle = \int \mathcal{D}\phi \ e^{-S_E[\phi]} \ . \tag{2.19}$$

The Euclidean action $S_E[\phi]$ is now

$$S_E[\phi] = \int d\tau d^3x \ \left(\frac{1}{2}\partial_\mu\phi\,\partial^\mu\phi + V(\phi) \right) \ . \tag{2.20}$$

In the saddle point approximation, the amplitude is dominated by the classical solutions which obey

$$\partial^2\phi = \lambda\phi(\phi^2 - v^2) \ . \tag{2.21}$$

This is the same as (2.12), but with the $\ddot{\phi}$ term replaced by the Laplacian on (Euclidean) spacetime, $\partial^2 = \partial_\tau^2 + \nabla^2$. We still have the same solution as before,

$$\phi_{\rm cl}(\tau) = v \tanh\left(\sqrt{\frac{\lambda v^2}{2}}\tau \right) \ . \tag{2.22}$$

The field varies in (Euclidean) time τ but is constant in space. So far, everything runs in parallel to the quantum mechanics argument. But now we compute the classical action of this solution. It is

$$S = \int d\tau d^3x \ \left(\frac{1}{2}\partial_\mu\phi_{\rm cl}\,\partial^\mu\phi_{\rm cl} + V(\phi_{\rm cl}) \right) = \mathcal{V} S_{\rm cl} \ . \tag{2.23}$$

Here $S_{\rm cl}$ is the quantum mechanical action (2.14) while $\mathcal{V}$ is the volume of space. But, if we're working in uncompactified Minkowski space then $\mathcal{V} = \infty$. This means that both the tunnelling amplitude (2.15) and the energy splitting of the ground states (2.16) are proportional to

$$e^{-\mathcal{V}S_{\rm cl}} \to 0 \quad \text{as} \quad \mathcal{V} \to \infty \ . \tag{2.24}$$

It's obvious what's going on here. In quantum field theory, the ground state of the field in one minimum is, say, $\phi(\mathbf{x}) = +v$ for all $\mathbf{x}$. If you want to tunnel to the other minimum, $\phi(\mathbf{x}) = -v$, then you have to shift the value of the field at every point in space. But that takes effort and quantum tunnelling is not up to the task. It costs an infinite amount of action and so does not occur.

This means that while discrete symmetries cannot be spontaneously broken in quantum mechanics, they can be broken in quantum field theory. The suppression is by the volume factor, so if we're working with quantum field theory on some compact space, rather than infinite volume Minkowski space, then tunnelling reappears. However, if the space is macroscopically large then the suppression factor $e^{-\mathcal{V}S_{\rm cl}}$ may be so tiny that, for all intents and purposes, we can think of the symmetry as broken.

The upshot of this argument is that the quantum field theory (2.17) in $d = 3 + 1$ dimensions (and, indeed, in any dimension greater than $d = 0 + 1$) has two ground states, $|+v\rangle$ and $|-v\rangle$, distinguished by the expectation value of $\phi(x)$ which acts as an *order parameter* to tell us which vacuum we live in,

$$\langle \pm v|\phi(x)| \pm v\rangle = \pm v \quad \text{and} \quad \langle \pm v|\phi(y)| \mp v\rangle = 0. \tag{2.25}$$

This is a story that generalises to other discrete symmetries. For example, if you find yourself with a quantum field theory with $\mathbb{Z}_N$ symmetry which is spontaneously broken, then you will have N ground states that will be permuted into each other by the action of the symmetry.

The Meaning of a Tachyon

Tachyons are mythological beasts in physics. When we first learn special relativity, certain unscrupulous teachers may tell you that a tachyon is a particle with $m^2 < 0$ which is forced forever to travel faster than the speed of light. This is, of course, nonsense.

In field theory, a tachyon is nothing mysterious. Our potential above has $m^2 < 0$ but there is certainly nothing flying around faster than light. Instead, it signals that the point $\phi = 0$ is a maximum of the potential, rather than a minimum. This is the true meaning of a tachyon in field theory: it is telling us that the chosen vacuum is unstable. It's our job to find a better, stable vacuum.

That's not hard in the example above. We just need to expand around one of the minima of the potential, rather than the maximum. In fact, we already did this calculation in (2.6). If we write $\phi(x) = v + \sigma(x)$, then we find a potential for σ given by

$$V(\sigma) = \lambda \left(v^2\sigma^2 + v\sigma^3 + \frac{1}{4}\sigma^4 \right) \ . \tag{2.26}$$

We can read off the mass of particles in the theory from the quadratic term. Any physical excitation has mass $M^2 = 2\lambda v^2$. The mass is real and positive and decidedly not exotic in any way.

Domain Walls

The presence of a spontaneously broken symmetry often implies the existence of some novel excitation in the theory. In the present case, this is a *domain wall*, a field configuration that interpolates from one vacuum to the other.

Indeed, we've already met the classical solution that does the job. We just need to repurpose the tunnelling solution (2.22) by replacing the imaginary time τ with one of the spatial coordinates $\mathbf{x} = (x, y, z)$. For example, the classical field configuration

$$\phi(z) = v \tanh\left(\sqrt{\frac{\lambda v^2}{2}}\, z\right) \tag{2.27}$$

solves the equations of motion of the original Lorentzian action (2.17). This solution interpolates from the vacuum $\phi = -v$ at $z \to -\infty$ to the vacuum $\phi = +v$ at $z \to +\infty$. It describes an excitation of the field, localised around $z = 0$, but extended in the x- and y-directions. This is the domain wall.

The domain wall has finite energy density $\mathcal{E}$ which, it is easy to see, coincides with the action $S_{\rm cl}$ of the same configuration in quantum mechanics. We computed this in (2.14) and found

$$\mathcal{E} = \frac{2}{3}\sqrt{2\lambda}\, v^3 \ . \tag{2.28}$$

Although the domain wall has finite energy density, it has infinite energy because it stretches to infinity in the (x, y)-plane. An exception to this statement is if we are considering domain walls in $d = 1 + 1$ dimensions where there is nowhere else for them to stretch. In this case the domain walls have finite energy and should be viewed as a kind of particle in the theory.

Back in $d = 3 + 1$ dimensions, we can straightforwardly consider variations of this classical configuration (2.27) in which the domain wall forms a sphere of radius R, containing one vacuum $\phi = -v$ inside, and the other vacuum $\phi = +v$ outside. This now has finite energy, given by $E = 4\pi R^2 \mathcal{E}$. However, such a static configuration will no longer solve the equation of motion because the domain wall has tension and will want to contract. To find the classical solution, we will have to solve the full time-dependent partial differential equation.

We can also get some sense for what happens to these configurations in the quantum theory. We can build a Fock space of states above either of the two ground states by exciting the field $\phi(x) = \pm v + \sigma(x)$. As we've noted, this creates particles of mass $M = \sqrt{2\lambda v^2}$. The Hilbert space of the theory decomposes as

$$\mathcal{H} = \mathcal{H}_+ \oplus \mathcal{H}_- \ . \tag{2.29}$$

This is *not* a tensor product, which would mean that we have to choose one state from $\mathcal{H}_+$ and another from $\mathcal{H}_-$ to specify the full state. Instead, it's a tensor sum: we must pick *either* a state from $\mathcal{H}_+$ *or* a state from $\mathcal{H}_-$. The states $|\psi\rangle \in \mathcal{H}_+$ obey

$$\langle\psi|\phi(0,\mathbf{x})|\psi\rangle = +v \quad \text{for} \quad |\mathbf{x}| \to \infty \ . \tag{2.30}$$

This is telling us that we necessarily approach the vacuum $|+v\rangle$ when we're far away. However, this doesn't mean that the excitations about one ground state know nothing about the other ground state. By piling many ϕ excitations on top of each other, it's quite possible to carve out a region of one vacuum inside another, and have excited states $|\psi\rangle \in \mathcal{H}_+$ that obey, for example,

$$\langle\psi|\phi(0,\mathbf{x})|\psi\rangle = \begin{cases} -v & \text{for } |\mathbf{x}| < R \\ +v & \text{as } |\mathbf{x}| \to \infty \end{cases} \ . \tag{2.31}$$

These kind of states are what become of our classical, spherical domain wall.

Cluster Decomposition

We know that the field theory has two ground states $|\pm v\rangle$, but you might wonder why we're necessarily forced to work with these states. What's stopping us taking the linear combinations

$$|0_\pm\rangle = \frac{1}{\sqrt{2}}\Big(|+v\rangle \pm |-v\rangle\Big) \tag{2.32}$$

as our ground states? This is a superposition of a state in $\mathcal{H}_+$ and a state in $\mathcal{H}_-$.

In fact, $|0_\pm\rangle$ are not the right states to work with. There are two arguments for this. The first is a little handwavey. Suppose that we perturb our original Lagrangian by some term $\Delta\mathcal{L}$ that breaks the $\mathbb{Z}_2$ symmetry. This will mean that one of the states $|\pm v\rangle$ has lower energy and is the true ground state. In the limit that we send the coefficient of $\Delta\mathcal{L}$ to zero, we will remain in the ground state, either $|+v\rangle$ or $|-v\rangle$.

This argument seems more compelling for condensed matter systems, where you can well imagine that there are many different perturbations (say, background magnetic fields) that would break the $\mathbb{Z}_2$ symmetry. The argument is less convincing in the context of particle physics where it's not at all clear what these additional terms might be. (Some balm comes from a conjecture that, once we take gravity into account, there are no exact global symmetries so there must, in fact, be some irrelevant symmetry breaking term lurking in the wings.)

There is a second, more important argument for why the states $|0_\pm\rangle$ defined in (2.32) are not the right ground states. This is a property known as *cluster decomposition* which is a way of capturing the locality of field theory. If you sit in some vacuum state $|\text{vac}\rangle$ and compute the two-point function of two operators, $A(x)$ and $B(y)$ then, when x and y are spacelike separated, the expectation value should decompose into

$$\langle\text{vac}|A(x)B(y)|\text{vac}\rangle \rightarrow \langle\text{vac}|A(x)|\text{vac}\rangle\,\langle\text{vac}|B(y)|\text{vac}\rangle \quad \text{as} \quad |x-y|\rightarrow\infty. \tag{2.33}$$

Now, on general grounds you can argue that, when x and y are far separated, we must have

$$\langle\text{vac}|A(x)B(y)|\text{vac}\rangle \rightarrow \sum_n \langle\text{vac}|A(x)|n\rangle\,\langle n|B(y)|\text{vac}\rangle \tag{2.34}$$

where $|n\rangle$ run over all possible vacuum states. But for cluster decomposition to hold, we want this to project onto the specific vacuum state $|n\rangle = |\text{vac}\rangle$ that we started in.

We can check this criterion for our theory with spontaneous symmetry breaking and the choice $A = B = \phi$. If we pick the state $|+v\rangle$ then, using the fact that $\langle +v|\phi(x)|-v\rangle = 0$, we have

$$\langle +v|\phi(x)\phi(y)|+v\rangle \rightarrow \langle +v|\phi(x)|+v\rangle\,\langle +v|\phi(y)|+v\rangle = v^2\ . \tag{2.35}$$

So this indeed obeys cluster decomposition. In contrast, if we work in the state $|0_+\rangle$ defined in (2.32) then you can check that

$$\langle 0_+|\phi(x)|0_+\rangle = \langle 0_-|\phi(x)|0_-\rangle = 0 \quad \text{and} \quad \langle 0_+|\phi(x)|0_-\rangle = v\ . \tag{2.36}$$

We then have

$$\langle 0_+|\phi(x)\phi(y)|0_+\rangle \rightarrow \langle 0_+|\phi(x)|0_-\rangle\,\langle 0_-|\phi(y)|0_+\rangle = v^2\ . \tag{2.37}$$

This does not obey cluster decomposition because the vacuum $|0_-\rangle$ that we need to insert in the middle differs from the vacuum $|0_+\rangle$ that we started with.

2.2 Continuous Symmetries

The story of symmetry breaking is rather different, and more powerful, when the symmetry in question is a continuous symmetry. Here we start by giving a couple of examples before we describe the general result known as Goldstone's theorem.

We'll work in quantum field theory. As in the previous section, there is some tension between spontaneous symmetry breaking in quantum field theory and what we know about the behaviour of wavefunctions in quantum mechanics, but we'll put this on hold for now and return to it in Section 2.2.4.

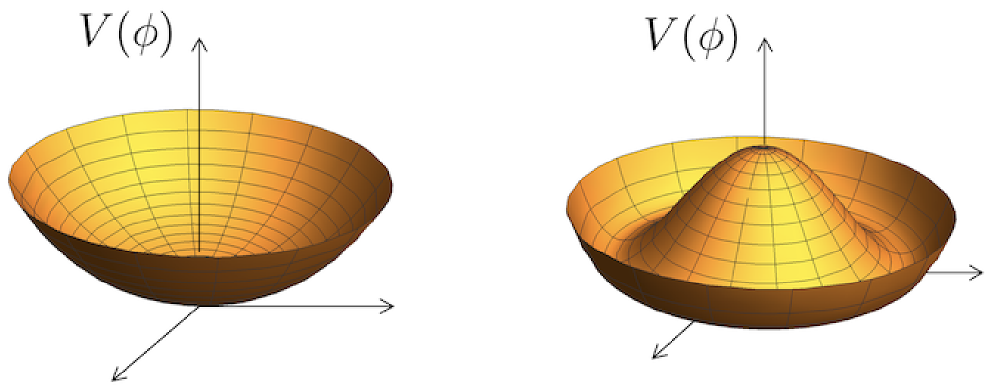


Figure 3. On the left: the potential with $m^2 > 0$. On the right, the Mexican hat potential with $m^2 < 0$.

To start, consider a complex scalar field $\phi(x)$ in $d = 3+1$ dimensions with action

$$S = \int d^4x \left(\partial_\mu \phi^\dagger \partial^\mu \phi - V(\phi, \phi^\dagger)\right) \quad \text{with} \quad V(\phi, \phi^\dagger) = m^2|\phi|^2 + \frac{1}{2}\lambda|\phi|^4 \ . \qquad (2.38)$$

The action is constructed so that it a enjoys $U(1)$ global symmetry which rotates the phase of ϕ,

$$\phi(x) \to e^{i\alpha}\phi(x) \ . \qquad (2.39)$$

Again, the physics depends on the sign of the m^2 term in the potential. The two different cases, with $m^2 > 0$ and $m^2 < 0$ are shown in Figure 3. In the former case, there is little interesting to say: you expand around the vacuum $\phi = 0$ and, after quantisation, find interacting particles of mass m with the $U(1)$ symmetry implying the usual conservation law. Here our interest is in the case $m^2 < 0$.

The potential with $m^2 < 0$ is sometimes called the "Mexican hat potential" because, you know, . It also looks like the bottom of a wine bottle. The defining feature is that there are not isolated minima, but instead an infinite number of ground states, defined by

$$|\phi|^2 = -\frac{m^2}{\lambda} \ . \qquad (2.40)$$

We define the *vacuum manifold* $\mathcal{M}_0$ to be the space of field configurations which have minimum energy. For the double well potential of Section 2.1, the vacuum manifold was just two points. Now, the vacuum manifold is the set of solutions to (2.40) which is a circle,

$$\mathcal{M}_0 = \mathbf{S}^1 \ . \tag{2.41}$$

To see what this buys us, we can write the complex field in polar coordinates, with

$$\phi(x) = r(x)e^{i\theta(x)} \ . \tag{2.42}$$

This is a slightly dangerous thing in quantum field theory, where we usually assume that fields can take any value. In writing (2.42), we need to remember that $r(x) \geq 0$ and $\theta(x) = \theta(x) + 2\pi$. Nonetheless, we can proceed for now and keep this in the back of our minds.

Substituting the polar decomposition into the original action (2.38), and dropping an irrelevant constant that arises when we complete the square, we have

$$S = \int d^4x \ \left(\partial_\mu r \partial^\mu r + r^2 \partial_\mu \theta \partial^\mu \theta - \frac{\lambda}{2}(r^2 - v^2)^2\right) \tag{2.43}$$

where, as in the last section, we've introduced $v^2 = -m^2/\lambda$. Now we can read off the physics. The ground state of the system sits at $r(x) = +v$. If we expand about this vacuum by writing $r(x) = v + \sigma(x)$ then the action becomes

$$S = \int d^4x \ \left(\partial_\mu \sigma \partial^\mu \sigma + (v+\sigma)^2 \partial_\mu \theta \partial^\mu \theta - \frac{\lambda}{2}\sigma^2(\sigma + 2v)^2\right) \ . \tag{2.44}$$

From this, we can read off the physics. In particular, the $\sigma(x)$ excitations have mass $M^2 = 2\lambda v^2$. These are radial oscillations of the field, that go back and forth in the potential.

To pick a vacuum, we also need to specify a value for the angular scalar field $\theta(x)$. But there is no preferred choice here. Once we've set $r(x) = v$, the different constant values of $\theta(x)$ parameterise the vacuum manifold $\mathcal{M}_0 = \mathbf{S}^1$. If this was quantum mechanics, then the wavefunction would simply spread over the $\mathbf{S}^1$. But things are different in quantum field theory, a fact that we will discuss further in Section 2.2.4, and each point on $\mathcal{M}_0$ corresponds to a different ground state of the theory. To specify the ground state, we have to pick one such point. It doesn't matter which point we pick because the physics will be the same in each. But, nonetheless, we have to pick one.

Whatever choice of ground state we make, say $\theta(x) = 0$, will spontaneously break the $U(1)$ symmetry (2.39) which acts as

$$\theta(x) \to \theta(x) + \alpha \ . \tag{2.45}$$

In fact, we see that the symmetry acts by taking us from one point on $\mathcal{M}_0$ to another.

Finally, we can look at the dynamics of the field $\theta(x)$ that parameterises $\mathcal{M}_0$. From the action (2.43), we see that there is no potential term for θ, a fact which simply follows from the $U(1)$ invariance of the potential. If we ignore the coupling to σ, then the θ field is governed by the simple Lagrangian

$$\mathcal{L} = v^2 \partial_\mu \theta \, \partial^\mu \theta \ . \tag{2.46}$$

This is a Lagrangian for a massless scalar field, albeit one that is slightly unusual because θ is a periodic variable. The existence of this massless scalar field is a direct consequence of the spontaneous breaking of the $U(1)$ global symmetry. As we will see, this is a general story: whenever a continuous global symmetry is spontaneously broken, there will be massless scalar fields. These fields are called *Goldstone bosons*.

Goldstone bosons can't have potential terms: only derivative terms. But that's not to say that they're totally boring. There can still be interactions, both among themselves (as we will see in later examples) and with other fields. For example, if we expand out $r(x) = v + \sigma(x)$ in (2.43) then we see that there are interaction terms between the massive scalar σ and the massless Goldstone boson θ that take the form $\sigma(\partial\theta)^2$ and $\sigma^2(\partial\theta)^2$. This means that a σ particle can decay to two Goldstone modes. However, if we look at energies $E \ll \sqrt{\lambda v^2}$, which is the mass of the σ particle, then the only field in town is the massless Goldstone mode, whose dynamics is governed by (2.46).

2.2.1 The $O(N)$ Sigma Model

Here's a generalisation of the ideas above. We take a collection of N real scalar fields $\phi^a(x)$, with $a = 1, \ldots, N$, and consider the following action

$$S = \int d^4x \ \left(\frac{1}{2} \partial_\mu \phi^a \partial^\mu \phi^a - V(\phi) \right) \quad \text{with} \quad V(\phi) = \frac{1}{2} m^2 \phi^a \phi^a + \frac{1}{4} \lambda (\phi^a \phi^a)^2 \ . \tag{2.47}$$

This action is constructed to have an $O(N)$ symmetry, under which the ϕ^a rotate. For $N = 2$, it coincides with the action (2.38) for a complex scalar field whose real and imaginary parts are ϕ^1 and ϕ^2.

Spontaneous symmetry breaking occurs when $m^2 < 0$ and the potential again looks like a Mexican hat but for someone with a higher dimensional head. The minima of the potential obey

$$\phi^a \phi^a = v^2 := -\frac{m^2}{\lambda} \ . \tag{2.48}$$

This is simply the equation for an $(N-1)$-dimensional sphere, and defines the vacuum manifold of the theory

$$\mathcal{M}_0 = \mathbf{S}^{N-1} \ . \tag{2.49}$$

The vacuum of the theory is one point on $\mathcal{M}_0$. It doesn't matter which one. Suppose that we pick the "south pole", so that the vacuum is $\phi^a = (0, 0, \ldots, 0, v)$. Now we can look at fluctuations around this vacuum by writing

$$\phi^a(x) = \big(\pi^1(x), \ldots, \pi^{N-1}(x), v + \sigma(x)\big) \ . \tag{2.50}$$

If we substitute this into the action (2.47), we find

$$S = \int d^4x \ \left(\frac{1}{2}\partial_\mu \pi^a \partial^\mu \pi^a + \frac{1}{2}\partial_\mu \sigma \, \partial^\mu \sigma - V(\pi^a, \sigma)\right) \tag{2.51}$$

with

$$V(\pi^a, \sigma) = \lambda v^2 \sigma^2 + \lambda v \sigma \left(\sigma^2 + \pi^a \pi^a\right) + \frac{1}{4}\lambda(\pi^a \pi^a + \sigma^2)^2 \ . \tag{2.52}$$

We again see that only the σ field has a quadratic term so this gives rise to a massive particle, while quantising the π^a will give $N-1$ massless particles. These are the Goldstone bosons from spontaneous symmetry breaking.

Although the π^a fields are massless, they still appear in the potential (2.52), just in higher order terms. This is in contrast to the case with $U(1)$ symmetry where the potential didn't depend on the Goldstone field $\theta(x)$. There's no mystery here: it's because we've made no attempt to pick our fields to parameterise the vacuum moduli space $\mathcal{M}_0$. Instead, the $\pi^a(x)$ fields are just linear displacements away from the vacuum, and if you move away linearly from a point in $\mathcal{M}_0$, you eventually end up climbing the potential.

To do better, we could write our fields as something akin to the polar ansatz (2.43). Alternatively, if we're at low energies so that we care only about the dynamics of the

Goldstone bosons, and not about their interactions with massive excitations, then we could restrict ourselves to $\mathcal{M}_0$ by insisting that (2.48) is obeyed everywhere, meaning

$$(\pi^a)^2(x) + (\phi^N)^2(x) = v^2 \; . \tag{2.53}$$

We could use this to eliminate $\phi_N(x)$ in our original action (2.47). By construction, the potential term vanishes completely and we're left just with kinetic terms for the Goldstone modes

$$S = \int d^4x \; \frac{1}{2}\left(\partial_\mu \vec{\pi}\cdot\partial^\mu\vec{\pi} + \frac{(\vec{\pi}\cdot\partial_\mu\vec{\pi})(\vec{\pi}\cdot\partial^\mu\vec{\pi})}{v^2 - \vec{\pi}\cdot\vec{\pi}}\right) \; . \tag{2.54}$$

We see that the Goldstone modes now have rather non-trivial interactions between themselves, but these interactions are entirely kinetic. To get a sense for what the action (2.54) is telling us, let's restrict to $N = 3$. In this case, the constraint (2.53) can be solved by the usual polar coordinates on $\mathbb{R}^3$,

$$\pi^1 = v\sin\theta\,\cos\varphi \; , \quad \pi^2 = v\sin\theta\,\sin\varphi \; , \quad \phi^3 = v\cos\theta \; . \tag{2.55}$$

It's important to stress that these are polar coordinates on field space, and both $\theta(x)$ and $\varphi(x)$ are fields that parameterise the vacuum manifold $\mathcal{M}_0 = \mathbf{S}^2$. With this choice of parameterisation, the action (2.54) becomes

$$S = \int d^4x \; \frac{v^2}{2}\Big(\partial_\mu\theta\,\partial^\mu\theta + \sin^2\theta\,\partial_\mu\varphi\,\partial^\mu\varphi\Big) \; . \tag{2.56}$$

We recognise the metric $ds^2 = d\theta^2 + \sin^2\theta d\varphi^2$ on $\mathbf{S}^2$ hiding within this action. More generally, any choice of parameterisation of the constraint (2.53) will give an action for the Goldstone bosons that takes the schematic form

$$S = \int d^4x \; \frac{1}{2} g_{ab}(\pi)\partial_\mu\pi^a\,\partial^\mu\pi^b \tag{2.57}$$

with g_{ab} the round metric on $\mathcal{M}_0$. Actions of this kind, where the fields are themselves coordinates on some manifold $\mathcal{M}$ are known as *non-linear sigma models*. In this context, the manifold $\mathcal{M}$ is sometimes called the *target space*, because the fields $\pi_a(x)$ are maps from spacetime (which is $\mathbb{R}^{1,3}$ for us) to the target manifold $\mathcal{M}$.

Non-linear sigma models like (2.57) are non-renormalisable. That means that they don't make sense up to arbitrarily high energy scales. But that's entirely reasonable! The sigma model (2.57) is constructed so that it describes only the very low energy physics. As we reach energies or order $E \sim \sqrt{\lambda}v$, we will start to be able to climb up the hills of the potential and out of the vacuum manifold $\mathcal{M}_0$. The original theory (2.47) provides a renormalisable, UV completion of the non-linear sigma model.

The origins of the name "sigma model" are somewhat farcical. It comes from the original paper of Gell-Mann and Lévy who did a calculation similar to the one above, eliminating the field $\sigma(x)$ (which, recall, is related to $\phi_N(x) = v + \sigma(x)$) and then naming the resulting Lagrangian after the field they got rid off! We'll see what Gell-Mann and Lévy did, and what the $\sigma(x)$ field describes in our world, when we come to discuss aspects of chiral symmetry breaking in QCD in section 3.

2.2.2 Goldstone's Theorem in Classical Field Theory

With these examples under our belt, we can now look at the general case. We will do this twice: once from the perspective of the classical theory, then again in the quantum theory.

We start classical. Consider a theory with a bunch of scalar fields, which we collectively denote as ϕ, transforming in some representation of a global symmetry group G. We will take G to a be Lie group, so we're dealing with continuous symmetries rather than discrete symmetries.

These fields experience a potential $V(\phi)$ which has some space of minima that define the vacuum manifold of the theory:

$$\mathcal{M}_0 = \left\{ \phi_0 \mid V(\phi_0) = V_{\rm min} \right\} . \tag{2.58}$$

If the ground state is unique – in which case we will assume that it sits at $\phi_0 = 0$ – then $\mathcal{M}_0$ is just a single point and we're back to the usual story in which the symmetry is realised only on excited states.

The more interesting situation is when ϕ_0 is not unique. In this case, acting with some elements of G will typically move us from one point in $\mathcal{M}_0$ to another. Indeed, the generic situation is that all points in $\mathcal{M}_0$ can be reached by a symmetry transformation, meaning that if we take two points $\phi_0, \phi_0' \in \mathcal{M}_0$, then there is a $g \in G$ such that

$$\phi_0' = g\phi_0 \ . \tag{2.59}$$

We can see this, for example, in the $O(N)$ model described above where $\mathcal{M}_0 = \mathbf{S}^{N-1}$ and you can always rotate from one point on the sphere to any other.

While some elements of G will move us around $\mathcal{M}_0$, other elements leave the point ϕ_0 unchanged. It's useful to define the concept of the *stability group* H. If we sit at some point $\phi_0 \in \mathcal{M}_0$, then the group H is defined to be those elements of G which don't change ϕ_0,

$$H = \left\{ h \in G \mid h\phi_0 = \phi_0 \ \right\} . \tag{2.60}$$

The stability group H defined above depends on the choice of $\phi_0 \in \mathcal{M}_0$. Happily, however, if we pick a different point $\phi_0' \in \mathcal{M}_0$ then we will find ourselves with a stability group H' that is isomorphic to H. This is simple to show: if $\phi_0' = g\phi_0$ then then for each $h \in H$ we can construct $h' = ghg^{-1} \in H'$.

Again, we can use the $G = O(N)$ model as an example. For any point in $\mathcal{M}_0 = \mathbf{S}^{N-1}$, the stability group is $H = O(N-1)$. The way in which $O(N-1)$ is embedded in $O(N)$ depends on where we sit in $\mathcal{M}_0$. For example, if we sit in the vacuum $\phi_i = (0, 0, \ldots, v)$ then the surviving $O(N-1)$ resides in the upper-left block of the $N \times N$ matrix, while if we sit in the vacuum $\phi_i = (v, 0, \ldots, 0)$ then $O(N-1)$ resides in the lower-right block. But, wherever we sit, there is always an $O(N-1)$ subgroup that survives.

We say that the group G is *spontaneously broken* to the group H. We usually write this as $G \to H$. The field ϕ is what, in statistical physics, we call an *order parameter* for the symmetry G: its value in the ground state – either zero or non-zero – provides a litmus test for whether the symmetry G is broken. The vacuum manifold $\mathcal{M}_0$ can then be identified as the coset space

$$\mathcal{M}_0 \cong G/H \ . \tag{2.61}$$

Here the coset G/H is defined to be the set of equivalence classes, with $g_1 \sim g_2$ if there exists an $h \in H$ such that $g_1 = hg_2$.

Now we're in a position to state the main result[3]:

Goldstone's Theorem: If a global, continuous symmetry G is spontaneously broken to H then the number of massless Goldstone bosons is given by

$$\dim\,(G/H) = \dim G - \dim H \ . \tag{2.62}$$

In light of the identification (2.61), you can think of these Goldstone bosons as the modes that fluctuate along the vacuum manifold $\mathcal{M}_0$.

Returning, briefly, to our $O(N)$ model, the sphere can be viewed as the coset $\mathbf{S}^{N-1} = O(N)/O(N-1)$. We can do some simple counting. We have $\dim O(N) = \frac{1}{2}N(N-1)$ so $\dim O(N) - \dim O(N-1) = N - 1 = \dim \mathbf{S}^{N-1}$.

[3]Both the classical and quantum versions of Goldstone's theorem were first proved by Goldstone, Salam and Weinberg in a classic 1962 paper entitled "Broken Symmetries". The proof was prompted by specific examples that had been explored by Nambu and by Goldstone.

Proof: The proof of Goldstone's statement is really just a matter of turning our intuition into some equations. Suppose that ϕ sits in a representation R of the symmetry group G. We'll denote the components of ϕ as ϕ^a with $a = 1, \ldots, \dim R$.

Consider how ϕ shifts under an infinitesimal symmetry transformation, $g\phi = \phi + \delta\phi$. If we denote the generators of G in the representation R as $(T^A)^a{}_b$, with $A = 1, \ldots, \dim G$, then we have

$$\delta\phi^a = i\alpha^A (T^A)^a{}_b \phi^b \tag{2.63}$$

with α^A infinitesimal parameters. We know that G is a symmetry of our theory which means, among other things, that the potential must satisfy $V(g\phi) = V(\phi)$. So, for an infinitesimal transformation,

$$V(\phi + \delta\phi) - V(\phi) = i\alpha^A \frac{\partial V}{\partial \phi^a} (T^A)^a{}_b \phi^b = 0 \ . \tag{2.64}$$

We differentiate with respect to ϕ^b to find

$$\left[\frac{\partial V}{\partial \phi^a} (T^A)^a{}_b + \frac{\partial^2 V}{\partial \phi^a \partial \phi^b} (T^A)^a{}_c \phi^c \right] = 0 \tag{2.65}$$

where we've stripped off the α^A on the grounds that they are arbitrary parameters and so this expression must hold for each $A = 1, \ldots, \dim G$. Now we evaluate the result on a ground state ϕ_0. The first term disappears because ϕ_0 is a minimum of the potential and we're left with

$$\left. \frac{\partial^2 V}{\partial \phi^a \partial \phi^b} \right|_{\phi_0} (T^A \phi_0)^a = 0 \quad \text{for} \quad A = 1, \ldots, \dim G \ . \tag{2.66}$$

We recognise the second derivative of the potential as the mass matrix $M^2_{ab} = \partial V / \partial \phi^a \partial \phi^b$; the eigenvalues of this matrix are the physical masses. The result (2.66) is telling us that the mass matrix potentially has a bunch of zero eigenvalues, one for each eigenvector $(T^A \phi_0)^b$.

The "potentially" in the sentence above is there because it may be that the would-be eigenvector $(T^A \phi_0)^b$ actually vanishes. Indeed, this is clearly the case if $\phi_0 = 0$. That's as it should be: if $\phi_0 = 0$ then the symmetry is unbroken and there's no reason to generically expect massless modes. However, even when $\phi_0 \neq 0$, there will be some generators – let us call them $\tilde{T}^A$ – that annihilate the ground state,

$$\tilde{T}^A \phi_0 = 0 \ . \tag{2.67}$$

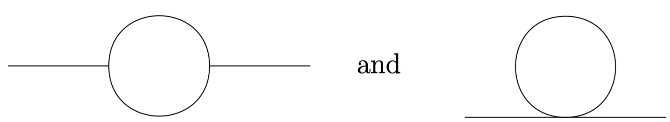
These are precisely the generators of the unbroken stability group H and so there are $\dim H$ of them. We will denote the generators orthogonal to $\tilde{T}^A$ as R^α, with $\alpha = 1, \ldots, \dim(G/H)$. Here, orthogonality means that they obey $\mathrm{Tr}\,(\tilde{T}^A R^\alpha) = 0$. Each of these generators gives a unique eigenstate $(R^\alpha \phi)^b$, and hence a massless mode. We see that there are at least $\dim(G/H)$ massless particles. These are the Goldstone bosons. □

2.2.3 Goldstone's Theorem in Quantum Field Theory

The quantum version of Goldstone's theorem has much more teeth than its classical counterpart. This is not because the theorem itself is very much different – as we'll see, it really involves all the same ingredients that we've seen above, just adapted to life in a Hilbert space. Instead, the importance of the result is due to the environment in which the theorem operates.

In classical field theory, there's no difficulty in writing down a theory for a massless scalar. You literally just need to set $m^2 = 0$ in the potential. So while it's certainly interesting that spontaneous symmetry breaking gives us a mechanism for generating massless scalars, they're not such rare beasts.

But the story is very different for interacting quantum field theories. There, massless scalars (and, indeed scalars that are just "light" in some sense) are very hard to come by. This is because the physical mass is not just the m^2 that you write down in the Lagrangian. Instead, the mass of a scalar picks up extra contributions from the cloud of other fields that accompany the particle. These are captured, at one loop, by Feynman diagrams like this:



Here the external legs are the scalars, while the particle running in the loop is anything that the scalar interacts with, including itself. These diagrams contribute to the mass renormalisation of the scalar and, crucially, are quadratically divergent. Physically, it means that quantum corrections push the mass of a scalar particle up to the UV-cut off of the theory, Λ_{UV}.

The upshot of this is that, if you write down a Lagrangian with $m^2 = 0$, then it won't describe a quantum scalar particle with physical mass zero. Instead, after renormalisation, it will describe a scalar with physical mass $m^2 \sim \Lambda_{UV}^2$. (In some cases,

Λ_{UV} may be some higher energy scale in the theory, rather than the UV-cut off. For example, in QCD we'll see that the masses of scalar mesons typically sit at a scale known as $\Lambda_{\rm QCD}$.) If you want to write down, say a ϕ^4 theory that describes a massless scalar then you will need to tune the mass in the Lagrangian (the so-called "bare mass") to be $m^2 \sim -\Lambda_{UV}^2$, with a coefficient that precisely cancels the contributions from quantum corrections. This is known as *fine tuning* and it is generally agreed to be as tasteless as it sounds. (This same idea also arises in statistical physics, where the mass term is associated to the deviation from a critical temperature. In this case, the fine tuning is physical because you get to turn the temperature up and down at will.)

None of this means that there is some flaw in quantum field theory: instead it's capturing the right physics. Quantum field theories tend not to have massless, or indeed, light, scalar fields. Their mass is typically pushed up to some cut-off scale. This is not true of fermions, which suffer only a logarithmic correction to their mass. This can be traced to the fact that fermions have an extra chiral symmetry when they are massless that protects their mass from being renormalised.

All of this means that things are interesting when you come across a physical system that does have a massless, or inordinately light, scalar field. If you find such a light scalar, then there should be a reason why the preceding arguments fail. In most (but, famously, not all!) cases, that reason is Goldstone's theorem. Spontaneous symmetry breaking provides a robust mechanism to naturally deliver genuinely massless scalars, whose mass is protected against any corrections from renormalisation. And, as we mentioned at the beginning of this section, it is a mechanism that is employed over and over again by nature, from magnets, to phonons to, as we shall see later, pions.

Before we turn to prove Goldstone's theorem in the context of quantum field theory, it's worth commenting on the "famously, not all" remark above. This is a nod to the Higgs boson. It is not particularly light, weighing in at $m_H \approx 126$ GeV. But if we believe that quantum field theory continues to hold at scales significantly higher than m_H, we should ask why the mass of the Higgs boson hasn't been pushed up to higher scales. Or, in other words, why don't the simple arguments that we sketched above apply to the Higgs boson? We don't know the answer to this question. This is known as the *hierarchy problem*.

Broken Symmetries Acting on Hilbert Space

With this preamble in place, we can now see how Goldstone's theorem manifests itself in quantum field theory. We won't work with Lagrangians, or restrict ourselves to perturbation theory. Instead, all the physics can be seen in how symmetries act on the Fock space of particles.

By Noether's theorem, any continuous symmetry G has an associated set of currents J_μ^A, with $A = 1, \ldots, \dim G$. From these we can construct the conserved charges

$$Q^A = \int d^3x \; J_0^A \; . \tag{2.68}$$

One of the lovely features of quantum mechanics (or, indeed, the Hamiltonian version of classical mechanics) is that these charges enact what we might call the "inverse Noether theorem". This means that, given a conserved charge, you can always reconstruct the associated symmetry. This follows from the fact that the charge is the generator of the symmetry, with any operator $\mathcal{O}$ undergoing the infinitesimal transformation

$$\delta_A \mathcal{O} = i[Q^A, \mathcal{O}] \; . \tag{2.69}$$

Comparing to our classical result (2.63), we see that our scalar fields ϕ^a transform as

$$[Q^A, \phi^a] = (T^A)^a{}_b \phi^b \; . \tag{2.70}$$

These are exact operator relations in the quantum theory.

In the classical theory, we saw that ϕ is an order parameter for the symmetry G. The same is true in the quantum theory, although strictly we should talk about the *vacuum expectation value* (or *vev*) of ϕ, as the order parameter,

$$\langle \phi \rangle = \langle \Omega | \phi | \Omega \rangle \tag{2.71}$$

where $|\Omega\rangle$ is the vacuum of the full, interacting theory. If $\langle \phi \rangle \neq 0$ then we say that ϕ *condenses*, a term taken from statistical physics. From (2.70), we have

$$\langle \Omega | [Q^A, \phi^a] | \Omega \rangle = (T^A)^a{}_b \langle \phi^b \rangle \neq 0 \; . \tag{2.72}$$

But this can only be true if

$$Q^A |\Omega\rangle \neq 0 \quad \text{for some } A \; . \tag{2.73}$$

This is what it means for a symmetry to be spontaneously broken in quantum field theory: the symmetry generators do not annihilate the vacuum.

Actually, there's a small caveat that I need to mention here. If we have $Q^A|\Omega\rangle = |\Omega\rangle$ then the commutator does vanish: $\langle\Omega|[Q^A, \phi^a]|\Omega\rangle = 0$. This kind of action on the ground state means that the symmetry is *unbroken* because, when exponentiated, we have $e^{i\alpha Q^A}|\Omega\rangle = e^{i\alpha}|\Omega\rangle$, but just changing the phase of a state in quantum mechanics is the same as leaving the state invariant. So the statement $Q^A|\Omega\rangle \neq 0$ in (2.73) should be better written as $Q^A|\Omega\rangle \neq c|\Omega\rangle$ for some $c \in \mathbb{C}$.

For any symmetry generator, broken or unbroken, we have $[Q^A, H] = 0$ so (2.73) is really telling us that, whenever the symmetry is broken, the vacuum is degenerate. Said slightly differently, in quantum field theory every different choice of $\langle\phi\rangle$ corresponds to a different vacuum of the theory.

Conversely, if $\langle\phi\rangle = 0$ then, from (2.73), we see that the vacuum is annihilated by the symmetry generators: $Q^A|\Omega\rangle = 0$. This is the more familiar case in which the symmetry is unbroken. Excitations above the vacuum then sit in multiplets of G.

When a symmetry is spontaneously broken, the excitations above the vacuum no longer sit in multiplets of the full symmetry group G. To see this, suppose that we have two fields, ϕ^1 and ϕ^2, that are related by a symmetry so there is some conserved charge such that $[Q, \phi^1] = \phi^2$. We can consider excitations of the vacuum by the creation operators associated to ϕ^1, heuristically $|1\rangle = a_1^\dagger|\Omega\rangle$, and similar excitations associated to ϕ^2, $|2\rangle = a_2^\dagger|\Omega\rangle$. We then have

$$|2\rangle = a_2^\dagger|\Omega\rangle = [Q, a_1^\dagger]|\Omega\rangle = Q|1\rangle - a_1^\dagger Q|\Omega\rangle \ . \tag{2.74}$$

We see that the symmetry generator does relate $|1\rangle$ and $|2\rangle$ but only if $Q|\Omega\rangle = 0$. When the symmetry is spontaneously broken, so $Q|\Omega\rangle \neq 0$, the two states $|1\rangle$ and $|2\rangle$ can have different properties. For example, they may have different energies.

So far, we haven't described where the Goldstone bosons come from. Following our classical intuition, we expect them to correspond to fluctuations along the directions of broken symmetry. And that's indeed the case. For each broken symmetry generator, we construct states

$$|\pi^A(\mathbf{p})\rangle \sim \int d^3x \ e^{i\mathbf{p}\cdot\mathbf{x}} J_0^A(x)|\Omega\rangle \ . \tag{2.75}$$

These states carry 3-momentum $\mathbf{p}$. Moreover, in the limit of vanishing momentum, we have

$$\lim_{\mathbf{p}\to 0} |\pi^A(\mathbf{p})\rangle \sim Q^A|\Omega\rangle. \tag{2.76}$$

For those generators that are spontaneously broken, the state $Q^A|\Omega\rangle \neq 0$ has the same energy as the original vacuum $|\Omega\rangle$ because $[Q^A, H] = 0$. This is the statement that the Goldstone boson $|\pi^A(\mathbf{p})\rangle$ has energy $E \to 0$ as $\mathbf{p} \to 0$. In other words, the Goldstone boson is massless.

None of the arguments above rely on perturbation theory: they are all exact statements about the interacting quantum field theory. This means that if we were to write down Lagrangians for these Goldstone bosons then they must remain massless, even after taking into account one-loop effects and so on. In operational terms, this happens because the Goldstone bosons have only derivative couplings.

The argument above is not completely rigorous, not least because $Q|\Omega\rangle$ suffers from divergences and doesn't strictly exist in the Fock space. A better, but more formal, argument uses the Källén-Lehmann spectral decomposition. You can read about this in Volume II of Weinberg's book.

The View From the Effective Potential

There is an alternative proof of Goldstone's theorem in quantum field theory that follows much more closely the classical proof that we saw previously. We first need to review some basic facts about generating functions in quantum field theory. The generating function for connected correlation functions is

$$e^{iW[J]} = \int \mathcal{D}\phi \; e^{i\int d^4x \;\; (\mathcal{L}(\phi)+J\phi)} \; . \tag{2.77}$$

Here $J(x)$ is a source for ϕ and differentiating $W[J]$ successively with respect to $J(x)$ gives the connected correlation functions. In particular, the expectation value of $\phi(x)$ is given by

$$\frac{\delta W[J]}{\delta J(x)} = \langle\Omega|\phi(x)|\Omega\rangle = \phi_{\rm cl}(x) \; . \tag{2.78}$$

In the absence of a source, Lorentz invariance implies that $\phi_{\rm cl}$ is just a number, and coincides with the vev (2.71) that we introduced previously. But, if we turn on a spatially varying source $J(x)$, then the function $\phi_{\rm cl}(x)$ will respond accordingly.

The Legendre transform of $W[J]$ is known as the *one-particle irreducible* (or *1PI* for short) effective action,

$$\Gamma[\phi_{\rm cl}] = W[J] - \int d^4x \; J(x)\phi_{\rm cl}(x) \; . \tag{2.79}$$

As in other examples of Legendre transforms, we should use (2.78) to replace $J(x)$ with $\phi_{\rm cl}(x)$ in the 1PI effective action. We can always return to $W[J]$ (assuming certain convexity properties) using

$$\frac{\delta\Gamma[\phi_{\rm cl}]}{\delta\phi_{\rm cl}(x)} = -J(x) \ . \tag{2.80}$$

The 1PI effective action is not, in general, the same thing as the more physical Wilsonian effective action that we get by integrating out high energy modes to find a description of the low energy physics. Taking derivatives of $\Gamma[\phi_{\rm cl}]$ generates the 1PI Green's functions. In particular, the two derivative term gives the inverse propagator

$$\frac{\delta^2\Gamma}{\delta\phi_{\rm cl}(x)\delta\phi_{\rm cl}(y)} = \Delta^{-1}(x-y) \ . \tag{2.81}$$

In general, $\Gamma[\phi_{\rm cl}]$ can be expressed in terms of a derivative expansion,

$$\Gamma[\phi_{\rm cl}] = \int d^4x \, \Big(-V_{\rm eff}(\phi_{\rm cl}) + \frac{1}{2}Z(\phi_{\rm cl})\partial_\mu\phi_{\rm cl}\partial^\mu\phi_{\rm cl} + \ldots \Big) \tag{2.82}$$

for some functions $V_{\rm eff}(\phi_{\rm cl})$ and $Z(\phi_{\rm cl})$. For our purposes, we're interested only in spatially homogeneous configurations, so we can ignore the derivative terms and the 1PI effective potential becomes

$$\Gamma[\phi_{\rm cl}] = -\mathcal{V}V_{\rm eff}(\phi_{\rm cl}) \tag{2.83}$$

where $\mathcal{V}$ is the (admittedly infinite, but actually irrelevant) volume of spacetime. Restricted to constant configurations, the second derivative of $\Gamma[\phi_{\rm cl}]$ is just the mass matrix, but now for the physical masses as opposed to the classical, bare masses

$$\frac{\partial^2 V_{\rm eff}}{\partial\phi_{\rm cl}\partial\phi_{\rm cl}} = \Delta^{-1}(0) \ . \tag{2.84}$$

Spontaneous symmetry breaking occurs when we have $\phi_{\rm cl} \neq 0$ even when $J = 0$. From (2.80), this translates into the familiar requirement that

$$\phi_{\rm cl} \neq 0 \quad \text{at} \quad \frac{\partial V_{\rm eff}}{\partial\phi_{\rm cl}} = 0 \ . \tag{2.85}$$

Now we may rerun all the arguments of section 2.2.2, but for the effective potential $V_{\rm eff}(\phi)$ rather than the classical potential $V(\phi)$ to again arrive at (2.66),

$$\frac{\partial^2 V_{\rm eff}}{\partial\phi^a_{\rm cl}\,\partial\phi^b_{\rm cl}}(T^A\phi_0)^b = 0 \ . \tag{2.86}$$

As in the classical argument, this is telling us that the mass matrix has a number of zero eigenvalues. (Equivalently, the propagator Δ has poles at $\mathbf{p} \to 0$.) There is one zero eigenvalue for each broken generator.

2.2.4 The Coleman-Mermin-Wagner Theorem

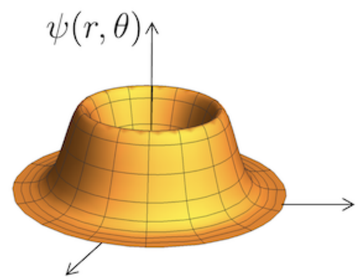
In all our discussions above, we assumed that spontaneous symmetry breaking actually takes place in the quantum theory. For example, we showed that *if* $\langle\phi\rangle \neq 0$ then the ground state must necessarily shift under a symmetry

$$Q|\Omega\rangle \neq 0 \ . \tag{2.87}$$

But how do we know that this actually happens? In particular, there is some tension with what we know from our first courses on quantum mechanics.

Let's return to the simplest example of a Mexican hat potential (2.38), but now think of quantum mechanics, rather than quantum field theory. That means that we have a quantum particle moving in the potential.

It's challenging to write down the exact ground state wavefunction $\psi(r,\theta)$, but it's not difficult to get some idea of what it looks like: it will be peaked in the trough at $r = v$, and be fully delocalised in the angular θ direction. In other words, it will look something like the wavefunction shown in the figure. But, crucially, because the wavefunction spreads around the circle parameterised by θ, there is no spontaneous symmetry breaking.



This begs the question: why is quantum field theory different from quantum mechanics? Why do we expect spontaneous symmetry breaking in the former case, but not in the latter? A similar question arose when we discussed discrete symmetries and there we understood that quantum tunnelling through the barrier was suppressed by the infinite spatial volume. But here there's no barrier to tunnel through. Instead we have a manifold of ground states $\mathcal{M}_0$ and it feels like it should be easier for a wavefunction to spread over $\mathcal{M}_0$ than to tunnel through a barrier. In other words, it should be more difficult to spontaneously break continuous symmetries than to spontaneously break discrete symmetries.

And indeed it is. But in an interesting way. The key physics is captured by the following theorem:

Theorem: A continuous symmetry cannot be broken in quantum theories in $d = 0+1$ (i.e. in quantum mechanics) or $d = 1+1$ dimensions.

This theorem was first proven by Mermin and Wagner for certain spin chains, inspired by previous work by Hohenberg. The proof in the context of quantum field theory is due to Coleman[4]. We see that the story is different for discrete and continuous symmetries. A discrete symmetry can be spontaneously broken in spacetime dimensions $d = 1 + 1$ and higher, but for a continuous symmetry to be spontaneously broken we must be in $d = 2 + 1$ or higher.

Here we offer just a sketch of this theorem. In fact, the basic idea can already be seen in classical field theory. Things are simplest if we work in d-dimensional Euclidean space. Suppose that we have a massless scalar field ϕ with no potential. This means that we have a choice of what we call the vacuum and, for our purposes, we'll decide that $\phi = 0$ is the ground state. Now we excite this scalar field by introducing a delta function source at the origin. That means that we have to solve

$$\nabla^2\phi = \delta(\mathbf{x}) \; . \tag{2.88}$$

This, of course, is the equation for the Green's function of the d-dimensional Laplacian. The solutions take the schematic form (ignoring overall coefficients)

$$\phi(\mathbf{x}) \sim \begin{cases} |x| & \text{for } d = 1 \\ \log|\mathbf{x}| & \text{for } d = 2 \\ 1/|\mathbf{x}|^{d-2} & \text{for } d \geq 3 \end{cases} \tag{2.89}$$

We see that for low dimensions, $d = 1$ and $d = 2$, exciting the scalar field at the origin means that it can no longer take the value $\phi = 0$ asymptotically. Any disturbance at the origin is still felt at $|\mathbf{x}| \to \infty$ where the field continues to grow. In contrast, in $d = 3$ and higher, the field is excited near the origin but then settles back down to $\phi \to 0$ as $|\mathbf{x}| \to \infty$.

The story above is classical. What happens in the quantum theory? We'll stick with the free massless scalar, and continue to work in Euclidean spacetime. Consider the two-point function $\langle\phi(\mathbf{x})\phi(\mathbf{y})\rangle$. We know from the lectures on Quantum Field Theory

[4]The original paper is from 1966, "Absence of Ferromagnetism or Anti-Ferromagnetism in One or Two-Dimensional Heisenberg Models" by Mermin and Wagner and, because of quirk of publication, appeared before the Hohenberg paper which motivated them: "Existence of Long-Range Order in One and Two Dimensions". Sidney Coleman's contribution is from 1973, in the concisely titled "There are no Goldstone Bosons in Two Dimensions".

that this is given by the same Green's function as above, so

$$\langle\phi(\mathbf{x})\phi(\mathbf{y})\rangle \sim \begin{cases} |x-y| & \text{for } d=1 \\ \log|\mathbf{x}-\mathbf{y}| & \text{for } d=2 \\ 1/|\mathbf{x}-\mathbf{y}|^{d-2} & \text{for } d\geq 3 \end{cases} \tag{2.90}$$

Again, we see the infra-red divergence for $d=1$ and $d=2$. Roughly speaking, this is telling us that the wavefunction spreads over all values of ϕ in $d=2$ dimensions, just as it does in $d=1$ quantum mechanics. In both cases, there is no normalisable ground state.

A better way of saying this is that $\phi(\mathbf{x})$ is not a well defined operator in $d=2$ dimensions. In particular, the correlation function $\langle\phi(\mathbf{x})\phi(\mathbf{y})\rangle \sim \log|\mathbf{x}-\mathbf{y}|$ is not positive for all $\mathbf{x}-\mathbf{y}$, which is one of the requirements of a QFT. However, although $\phi(\mathbf{x})$ is not a well-defined operator, its derivatives $\partial_\mu\phi(\mathbf{x})$ are. You can learn more about this 2d theory (which really only makes sense when ϕ is taken to be a periodic variable) in the lectures on String Theory.

No such problems arise for a massless scalar in $d\geq 3$ spacetime dimensions. Here, each value of $\langle\phi\rangle$ specifies a different ground state of the theory. Indeed, for this simple free theory, the massless ϕ field can be viewed as a Goldstone boson for the shift symmetry $\phi\rightarrow\phi+\text{constant}$.

As for the discrete symmetries discussed in Section 2.1, the existence of spontaneous symmetry breaking is due to the infinite volume of space. If we were to take our quantum field theory on a compact spatial manifold, then the long-time behaviour is the same as in quantum mechanics, and the wavefunction will again spread over field space, obviating spontaneous symmetry breaking.

2.3 The Higgs Mechanism

Goldstone's theorem tells us that when a continuous symmetry is spontaneously broken, it results in a massless boson. Here we would like to ask: what happens if that symmetry is gauged?

First, the very concept of a "spontaneously broken gauge symmetry" is a little misleading. As we've stressed, a gauge symmetry is merely a redundancy in the description of a system and there's no way that this redundancy can be "broken" or "lost". This linguistic issue notwithstanding, the physics underlying the spontaneous breaking of gauge symmetries is clear cut. First, there is no massless Goldstone boson. Second, the gauge boson gets a mass. We'll now see, in some detail, how this comes about.

2.3.1 The Abelian Higgs Model

We return to a complex scalar ϕ with the Mexican hat potential of Section 2.2. This time, however, we couple the scalar to a $U(1)$ gauge field. The action is

$$S = \int d^4x \left(-\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \mathcal{D}_\mu\phi^\dagger\mathcal{D}^\mu\phi - \frac{\lambda}{2}(|\phi|^2 - v^2)^2 \right) . \tag{2.91}$$

This is known as the *Abelian Higgs model*. The covariant derivative is $\mathcal{D}_\mu\phi = \partial_\mu\phi - ieA_\mu\phi$. Clearly the ground state sits at

$$|\phi|^2 = v^2 . \tag{2.92}$$

Previously, this meant that we had a vacuum manifold, $\mathcal{M}_0 = \mathbf{S}^1$, parameterised by the phase of ϕ. But now the $U(1)$ that takes us around the $\mathbf{S}^1$ is a gauge symmetry,

$$\phi \to e^{ie\alpha(x)}\phi \quad \text{and} \quad A_\mu \to A_\mu + \partial_\mu\alpha \tag{2.93}$$

and we know that field configurations that are related by gauge symmetries should be considered physically equivalent. This suggests that the gauge theory only has a single ground state, rather than a manifold of ground states. This, it turns out, is the right interpretation.

To see the physics, let's place ourselves in the classical vacuum $\phi = v$ and look at fluctuations that we parameterise as

$$\phi(x) = e^{i\theta(x)}\big(v + \sigma(x)\big) . \tag{2.94}$$

We then have

$$\mathcal{D}_\mu\phi = e^{i\theta}\big(\partial_\mu\sigma + i(v+\sigma)(\partial_\mu\theta - eA_\mu)\big) . \tag{2.95}$$

Substituting this into the action, and expanding out, we have

$$S = \int d^4x \left(-\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \partial_\mu\sigma\partial^\mu\sigma + (v+\sigma)^2(\partial_\mu\theta - eA_\mu)(\partial^\mu\theta - eA^\mu) - V(\sigma) \right)$$

with

$$V(\sigma) = \frac{\lambda}{2}\sigma^2(\sigma + 2v)^2 . \tag{2.96}$$

From this, we can read off the mass spectrum of the theory. First, the scalar σ is reasonably standard: it has a quadratic term that tells us its mass is

$$m_\sigma^2 = 2\lambda v^2 . \tag{2.97}$$

This is the same mass that we calculated for the global symmetry. Later, when we discuss electroweak theory, we will learn that an analogous particle is the Higgs boson.

More interesting is the other scalar field $\theta(x)$. In the absence of the gauge field, this was the Goldstone boson. But now that we've introduced the gauge field, we see something interesting: this field only appears in kinetic terms in the combination

$$\partial_\mu \theta - eA_\mu \ . \tag{2.98}$$

This allows us to eliminate the field $\theta(x)$ completely. We simply define a new gauge field, related to the first by the change of variables

$$A'_\mu = A_\mu - \frac{1}{e}\partial_\mu \theta \ . \tag{2.99}$$

This has the same field strength as A_μ, with $F_{\mu\nu} = \partial_\mu A'_\nu - \partial_\nu A'_\mu$. However, in contrast to A_μ, the new field A'_μ does not change under a gauge transformation since the usual shift $A_\mu \to A_\mu + \partial_\mu \alpha$ is now compensated by $\theta \to \theta + e\alpha$. Said slightly differently, you could also think of the change of variables to A'_μ as analogous to working in $\theta = 0$ gauge, known, in this context, as *unitary gauge*. Either way, the upshot is the same: the field $\theta(x)$ no longer appears in the action

$$S = \int d^4x \ \left(-\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \partial_\mu\sigma\partial^\mu\sigma + e^2(v+\sigma)^2 A'_\mu A'^\mu - V(\sigma) \right) \ . \tag{2.100}$$

We see that we've generated a mass term $e^2v^2A'_\mu A'^\mu$ for the gauge field. This is exactly the kind of term that is usually forbidden by gauge invariance. But such a term arises naturally when we spontaneously break the gauge symmetry and the photon gets a mass

$$m_\gamma^2 = 2e^2v^2 \ . \tag{2.101}$$

This is the *Higgs mechanism*.

There's some interesting interplay of degrees of freedom going on here. Massive spin 1 particles have three degrees of freedom. (This is just the $(2l+1)$-dimensional representation of the little group for $l = 1$.) But massless spin 1 particles have only two degrees of freedom, the two polarisation states. But it's clear where the extra degree of freedom came from because the photon absorbed the would-be Goldstone mode $\theta(x)$. This Goldstone boson breathes life into the longitudinal mode of the photon which is ordinarily killed by the constraints of gauge invariance.

Note that the mass of the Higgs boson (2.97) and the mass of the photon (2.101) have different parameteric dependence on the coupling constants. This means, among other things, that we could always just decouple the Higgs boson by taking $m_\sigma \to \infty$, leaving behind the massive photon at a finite mass m_γ. Given this, you might wonder why we needed all this palava with the Higgs boson. And, in fact, we really don't. We could always just couple the photon directly to the Goldstone mode θ, ignoring the radial mode σ. Said differently, we could just couple the photon to the sigma model with target space $\mathcal{M}_0 = \mathbf{S}^1$ which gives a massive photon and no Higgs boson. However, this option is less viable when we discuss the Higgs mechanism in non-Abelian theories because the corresponding sigma model is non-renormalisable and so should be viewed as an effective low energy theory, breaking down in the UV.

2.3.2 Superconductivity

We will later see that the Higgs mechanism plays a key role in the Standard Model. But there is a glorious unity to physics, and if nature finds a good trick to use in one context, she often recycles it elsewhere. So it is with the Higgs mechanism, which also provides a description of how superconductors work. In that context, it is referred to as the *Anderson-Higgs mechanism*[5].

Superconductivity is a phenomenon exhibited by many metals when they are cooled to a few degrees Kelvin. The metal undergoes a phase transition, and the electrical resistivity promptly plummets. At the same time, any magnetic fields are expelled.

The microscopic explanation for superconductivity is beyond the scope of these lectures. For what it's worth, an attractive coupling mediated by the phonon causes electrons to form an object known as a Cooper pair. For our purposes, all we need to know is that the resulting bound state is described by a complex scalar field ϕ that has charge $-2e$, with the -2 because it's formed of two constituent electrons.

In condensed matter physics, we more commonly work with the free energy, which describes the equilibrium properties of a system at finite temperature, rather than the Lagrangian which describes the zero temperature dynamics. But to avoid taking too much of a detour, here we give a Lagrangian description of superconductivity. This

[5]The history of the Higgs phenomenon is famously murky. Anderson's 1963 paper on superconductivity argues that the would-be Goldstone mode is no longer there and that the photon is gapped. These ideas were extended to the relativistic theory by Brout and Englert and, independently, by Peter Higgs. Only Higgs' paper mentions the existence of an additional massive particle, now called the *Higgs boson*, albeit in what appears to be an afterthought in the final paragraph of the paper. You can decide for yourself whether this was because the existence of the Higgs boson was obvious (as some of the authors later claimed) or because they didn't think to ask the question. Still, the mechanism for giving a photon mass should probably rightly be called the Anderson-Brout-Englert-Higgs mechanism. In line with much of the particle physics community, we choose to unfairly shorten this to simply "Higgs". Meanwhile the term Higgs boson, for the scalar particle, seems more appropriate.

is almost identical to the Abelian Higgs model of the previous section, with just one small difference: the dynamics of the scalar field ϕ is non-relativistic. This means that we should work with the action

$$S = \int dt\, d^3x \, \left(-\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + i\phi^\dagger \mathcal{D}_t\phi - |\mathcal{D}_i\phi|^2 - \frac{\lambda}{2}(|\phi|^2 - v^2)^2\right) \ . \qquad (2.102)$$

In addition, there's an extra factor of -2 buried in the covariant derivatives: $\mathcal{D}_\mu\phi = \partial_\mu\phi + 2ieA_\mu\phi$. (On dimensional grounds, there should be a coefficient with dimension $(\text{mass})^{-1}$ in front of the gradient terms but I've set it to unity to ease comparison with the relativistic Abelian Higgs model (2.91).)

A non-relativistic complex scalar has just a single degree of freedom. (This is true because the kinetic term contains a first order time derivative and so $\phi^\dagger$ is the momentum conjugate to ϕ, rather than a separate degree of freedom.) This means that if we quantise (2.102), we will find a massive photon, but the would-be Higgs boson (what we called σ in the relativistic theory) is missing.

We can read off the charge density and current from the coupling $A_\mu J^\mu$. The charge density is

$$J^0 = -2e|\phi|^2 \ . \qquad (2.103)$$

In the ground state, we have the condensation $|\phi|^2 = v^2$, so the Cooper pairs form a constant background electric charge. (In a real system, this is compensated by the positive electric charge of the underlying lattice of ions.) Meanwhile, assuming that $|\phi|^2 = v^2$, the electric current is

$$\mathbf{J} = 4ev^2\left(\nabla\theta - 2e\mathbf{A}\right) \ . \qquad (2.104)$$

Here, as in the previous section, $\theta(x)$ is the phase of $\phi(x)$. The expression (2.104) is known as the *supercurrent*. It is sometimes denoted as $\mathbf{J}_s$ to distinguish it from the normal current carried by electrons.

Resistance is Futile

The signature of a superconductor is that it conducts electricity without resistance. This follows immediately from the equation of motion for $\phi^\dagger$,

$$i\mathcal{D}_0\phi = -\mathcal{D}^2\phi + \frac{\partial V}{\partial\phi^\dagger} \ . \qquad (2.105)$$

In the lowest energy state, the charge density $|\phi|^2$ is constant. But the phase can vary. Indeed, from (2.104), we see that a spatially varying phase $\nabla\theta \neq 0$ means that an electric current flows.

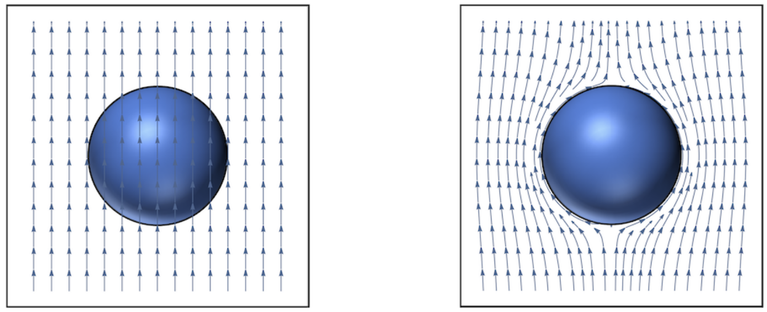


Figure 4. A constant magnetic field can pass through a normal metal, as shown on the left. But when the metal becomes superconducting, as shown on the right, the magnetic field is expelled, a phenomenon known as the Meissner effect.

Suppose that we look at such a configuration with $|\phi|^2 = v^2$. Then the complex equation of motion (2.105) splits into real and imaginary parts, which are

$$\dot{\theta} - 2eA_0 = \frac{1}{(4ev^2)^2}\mathbf{J}^2 \quad \text{and} \quad \nabla \cdot \mathbf{J} = 0\ . \tag{2.106}$$

To see the relevant physics, it's simplest to restrict to the case where $\mathbf{J}$ is constant in space so that $\nabla \mathbf{J}^2 = 0$. Then, taking the time derivative of the (2.104), we have

$$\frac{d\mathbf{J}}{dt} = 4ev^2\left(\nabla\dot{\theta} - 2e\dot{\mathbf{A}}\right) = 2(2ev)^2\left(-\nabla A_0 - \dot{\mathbf{A}}\right) = 2(2ev)^2\mathbf{E}\ . \tag{2.107}$$

This is the *first London equation*. It tells us that an electric field acts to accelerate the current, rather than to maintain the current. But that's not what usually happens in a conductor. Usually, a constant electric field induces a constant current. That's what the famous Ohm's law equation $V = IR$ says. But the resistance R in a normal conductor is due to friction terms, and the London equation (2.107) is telling us that a superconductor has vanishing resistance, $R = 0$.

Meissner Effect

Superconductors don't like magnetic fields very much. If you try to force a magnetic field through a superconductor, then it will resist. This is known as the *Meissner effect*, or sometimes as the *Meissner-Ochsenfeld effect*. A cartoon of this is shown in Figure 4. It has the dramatic consequence that a superconductor, placed above a magnet, is repelled and can levitate in mid-air.

At heart, the Meissner effect arises because the photon gets a mass. The term $\sim v^2\mathbf{A}\cdot\mathbf{A}$ in the action ensures that it is energetically costly to turn on a magnetic field.

We can see this more quantitatively from the form of the supercurrent (2.104). If we take the curl of both sides, we find

$$\nabla \times \mathbf{J} = -2(2ev)^2 \mathbf{B} \ . \tag{2.108}$$

This is the *second London equation*. We can compare it to Ampére's law, $\nabla \times \mathbf{B} = \mu_0 \mathbf{J}$. Taking the curl, and using $\nabla \times \nabla \times \mathbf{B} = -\nabla^2 \mathbf{B}$ (because $\nabla \cdot \mathbf{B} = 0$), we find that the magnetic field inside a superconductor obeys the Helmholtz equation

$$\nabla^2 \mathbf{B} = \frac{1}{\lambda^2} \mathbf{B} \quad \text{with} \quad \lambda^2 = \frac{1}{2(2ev)^2} \ . \tag{2.109}$$

Here λ is the *penetration depth*, a length scale equal to the inverse mass of the photon, $\lambda = 1/m_\gamma$. (The factor of 4 difference with (2.101) can be traced to the fact that, for superconductors, we're dealing with a field with charge $-2e$ rather than e.)

To see why the penetration depth gets it name, we can solve this equation for a constant magnetic field of the form

$$\mathbf{B} = (B(z), 0, 0) \ . \tag{2.110}$$

This configuration automatically obeys $\nabla \cdot \mathbf{B} = 0$. Suppose that the superconductor fills half of space, say the region with $z > 0$. We set up a constant magnetic field $\mathbf{B} = (B_0, 0, 0)$ in the outside region $z < 0$ and ask what becomes of it when it enters the superconductor. There are two solutions to (2.109), but only the decaying one is physical. We find that the magnetic field drops off exponentially quickly inside the superconductor,

$$B(z) = B_0 e^{-z/\lambda} \ . \tag{2.111}$$

This is the Meissner effect: the superconductor does not suffer a magnetic field inside. In most superconductors, $\lambda \approx 10^{-8}$ to 10^{-9} m. This is what allows superconducting materials to levitate above magnets: the magnetic field can't penetrate the superconductor, and has to go around as shown in Figure 4. This squeezes the magnetic field lines which costs energy, making it energetically preferable for the superconductor to remain magically suspended in space, rather than falling like other materials that have more respect for gravity.

Vortices

There's no such thing as an immovable object. If you push hard enough, by cranking up the magnetic field, then the superconductor will eventually relent and let it pass. But the way it does this is interesting.

This follows because of a novel solution to the equations of motion of the action (2.102) known as a *vortex*. (This is also a solution to the relativistic Abelian Higgs model (2.91).) The vortex solution is time-independent, and extends along one spatial direction – say the z-direction – as a string-like object. To this end, we will look for solutions with $\partial_0 = \partial_3 = 0$ as well as $A_0 = A_3 = 0$.

It turns out that no closed form solution to the resulting equations of motion is known (although it is not hard to construct numerically). So rather than try to solve the equations directly, we will instead argue that such a solution must exist. The argument involves a little simple topology.

Consider the (x, y)-plane at $z = 0$. We will work with 2d polar coordinates $x + iy = re^{i\varphi}$. The trick is to look for solutions such that, for any curve C around the origin, we have

$$\oint_C \nabla\theta \cdot d\mathbf{x} \neq 0 \ . \tag{2.112}$$

Our first task is to understand what this means. Usually, the integral of a total derivative is zero, but in the present case there's an opportunity for something more interesting to happen. This is because the field θ started life as a phase of our scalar ϕ and, as such, is periodic, taking values $\theta \in [0, 2\pi)$. For a periodic field θ, the line integral $\oint_C \nabla\theta \cdot d\mathbf{x}$ counts the number of times that θ winds as we traverse the curve C.

For example, if the curve C is parameterised by a coordinate $\varphi \in [0, 2\pi)$ then we could consider field configurations of the form $\theta = k\varphi$. Because θ must be single-valued, this only makes sense for $k \in \mathbb{Z}$ which is acceptable because $\theta = 0$ is equivalent to $\theta = 2\pi$. This, in turn, means that the integral (2.112) is necessarily quantised,

$$\oint_C \nabla\theta \cdot d\mathbf{x} = \int_0^{2\pi} d\varphi \, \frac{d\theta}{d\varphi} = 2\pi k \quad \text{with} \quad k \in \mathbb{Z} \ . \tag{2.113}$$

This quantisation doesn't happen because of anything to do with quantum mechanics. Instead, it's a quantisation imposed upon us by simple topological configurations.

Let's look for configurations in which the phase θ has winding (2.112). If this configuration is to have finite energy (per unit length) then, asymptotically, we must have $\mathcal{D}_i\phi \to 0$. This tells us that

$$\oint_C \nabla\theta \cdot d\mathbf{x} = 2e \oint_C \mathbf{A} \cdot d\mathbf{x} = 2e \int d^2x \ B_3 = 2e\Phi \tag{2.114}$$

with Φ the magnetic flux through the plane. We see that the quantisation of the winding translates into a quantisation of the allowed magnetic flux

$$\Phi = \frac{2\pi}{2e} k \quad \text{with} \quad k \in \mathbb{Z} \ . \tag{2.115}$$

I've not cancelled the factors of 2 here to stress the fact that, by measuring the minimal unit of flux, with $k = \pm 1$, you can determine that the current is carried by particles of charge $\pm 2e$, rather than the electron charge $-e$. (Indeed, this was one of the first experiments to confirm the charge of the condensate in a superconductor.)

The quantisation of winding means that the field configurations in this theory split into distinct topological sectors, labelled by $k \in \mathbb{Z}$. Because this integer is determined by the asymptotic boundary conditions, there's no way that a field configuration in one topological sector can move smoothly into a configuration in another. This means that we can find novel solutions to the equations of motion by minimising the energy (per unit length) in any given sector.

Let's think about how this works for the minimum winding $k = 1$. Because the winding number is quantised, it can't change gradually as we vary the radius of the contour C in (2.113). It must give the same value $k = 1$ for all choices of C. That's all fine until we get to the origin, at which point the phase θ gets something of an identity crisis because it's supposed to point in all directions at once. The only way out is to realise that θ is the phase of the field ϕ, and so there must be a point in the (x, y)-plane where $\phi = 0$ so that the phase is ill-defined. This means that whenever we have winding, there is necessarily a small region of non-superconducting phase, with $\phi = 0$, somewhere inside the contour C. That will be the region where it is energetically preferable for the flux Φ in (2.115) to penetrate.

We can get an estimate for the size of the region over which the condensate varies. For simplicity, we set $A_0 = \mathbf{A} = 0$ and restrict to time-independent configurations $\phi(x, y)$. Then the equation of motion (2.105) reads

$$\nabla^2 \phi = \lambda \phi (|\phi|^2 - v^2) \ . \tag{2.116}$$

This equation contains a natural length scale ξ, given by

$$\xi^2 = \frac{1}{\lambda v^2} \ . \tag{2.117}$$

This is known as the *coherence length*. It is roughly equal to the inverse mass of the scalar (2.97) in the relativistic theory: $\xi = \sqrt{2}/m$. (That factor of $\sqrt{2}$ is just

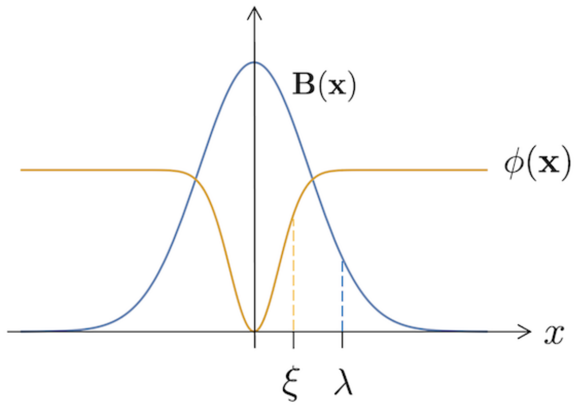


Figure 5. The spatial profile of the magnetic field and condensate for a vortex.

annoying convention.) The coherence length sets the scale over which the condensate ϕ is roughly zero (or, more precisely, exponentially small) in the vortex solution. In most superconductors, the coherence length is within a couple of orders of magnitude of the penetration depth, λ, the analogous quantity for the magnetic field.

We could put more meat on this discussion by explicitly solving the equations of motion for the gauge field and scalar. By making a suitable, rotationally invariant ansatz, you can reduce these equations to two, coupled ordinary non-linear differential equations. There is no solution in closed form, but it is straightforward to solve them numerically. A schematic picture of the resulting condensate and magnetic flux, as a cut-through in the x-direction, is shown in Figure 5 in the case where $\lambda > \xi$, so the magnetic field spills out over the region where $\phi = 0$.

The discussion above took place in the $z = 0$ plane. But we can repeat the story as we move the contour C in the z-direction. The winding can't change, and so the region with $\phi = 0$ and magnetic flux necessarily extends in the z-direction. In other words, we have a magnetic flux tube. This is the *vortex*.

The fact that non-linear equations of motion have novel localised solutions like the vortex is interesting. In particular, the existence of this solution can be traced to the topological nature of the winding. The general name given to solutions of this kind is *soliton*.

For the story above, we restricted attention to the minimal $k = 1$ sector. What happens for higher $k \geq 2$ is also interesting and depends on the ratio of the two length scales ξ/λ. There are three possibilities:

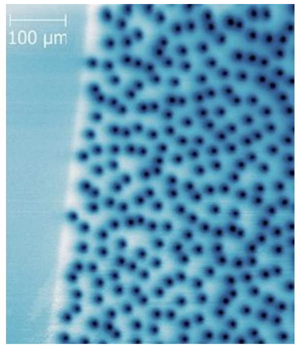


Figure 6. The Abrikosov vortex lattice, observed in the high temperature superconductor YBCO.

- For $\xi > \sqrt{2}\lambda$, the scalar field ϕ spreads out further than the magnetic flux. But there is a general story that magnetic flux repels, while scalar fields attract. (For example, the Yukawa force is always attractive.) This means that two vortices will feel an attractive force, albeit one that is exponentially suppressed on scales $r \gg \xi$. This is what happens in a *Type I superconductor*.

  What actually happens in practice is that, if you apply a magnetic field to a Type I superconductor, then the whole material will transition to the normal, metallic phase at some critical magnetic field B_c. This means that you don't see vortices in this case.

- For $\xi < \sqrt{2}\lambda$, the magnetic field spreads out further than the scalar field, as shown in Figure 5. In this case, two nearby vortices experience a repulsive force. This is known as a *Type II superconductor*.

  If you apply a magnetic field to a Type II superconductor then, initially, the superconductor will resist. But if you crank up the magnetic field suitably high then the superconductor will relent by allowing vortices to penetrate. These vortices repel, and so form a crystal-like structure known as an *Abikosov lattice*.

- The case $\xi = \sqrt{2}\lambda$ is of less relevance physically, because you have to fine tune two length scales, but is the situation with the richest mathematical structure. Now the attractive scalar force and repulsive magnetic force cancel, at least to leading order. Somewhat miraculously, it can be shown that this cancellation persists to all orders and the equations of motion exhibit solutions where k vortices can sit at k arbitrary points on the plane. These are known as *BPS vortices*.

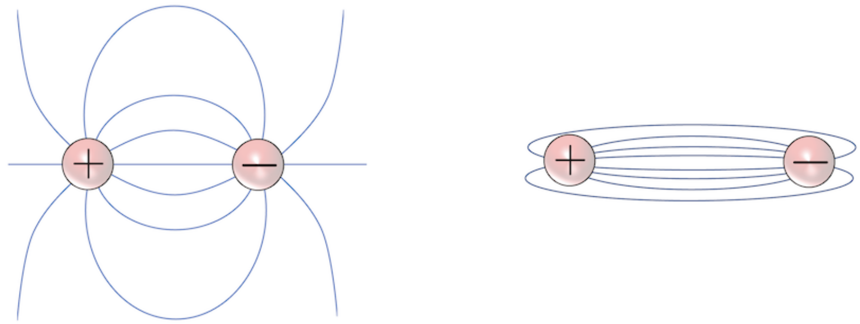


Figure 7. The magnetic field lines between a monopole anti-monopole pair. In a vacuum, the field lines spread out as a dipole configuration as shown on the left. But in a superconductor, the field lines form a flux tube as shown on the right, resulting in the confinement of magnetic monopoles.

Magnetic Monopoles are Confined

There is a lesson to take from the theory of superconductivity that will be important for particle physics. For this, we set up a thought experiment.

Our thought experiment involves a hypothetical object called a *magnetic monopole*, a particle that emits a radial magnetic field

$$\mathbf{B} = \frac{g\hat{\mathbf{r}}}{4\pi r^2} \; . \tag{2.118}$$

Here g is the *magnetic charge*. If you've been told that magnetic monopoles can't exist because the Maxwell equation $\nabla \cdot \mathbf{B} = 0$ is sacrosanct, then you've been lied to. (See, for example, the lectures on Gauge Theory for a discussion of how magnetic monopoles are compatible with everything you know and love.)

Suppose that we have two magnetic monopoles, one with charge $g = 1$ and the other an anti-monopole with charge $g = -1$. If we place these monopoles a distance r apart in the vacuum, then the magnetic field lines will form the kind of dipole configuration that is familiar from our first course on Electromagnetism. This is shown on the left in Figure 7. The potential energy $V(r)$ between two monopoles scales like the Coulomb force,

$$V(r) \sim \frac{g^2}{r} \; . \tag{2.119}$$

Things are more interesting if we put the monopoles inside a superconductor. Now, the Meissner effect means that it's no longer energetically preferable for the magnetic field lines to spread out all over space. Instead, the field lines will clump together to form a magnetic flux tube that, at least far from the monopoles, is described by the vortex solution that we met above. A cartoon of the field lines is shown on the right of Figure 7. Now the potential energy scales linearly with the separation,

$$V(r) \sim \mathcal{E} r \tag{2.120}$$

where $\mathcal{E}$ is the energy per unit length of the vortex. This makes it very difficult to separate the monopole and anti-monopole: the further you want to pull them apart, the more energy it will cost. This is because they are attached by the flux tube which acts a little like an elastic band. (A little like an elastic band, but not a lot. Hooke's law is $V \sim r^2$ while here we have linear potential energy, $V \sim r$, corresponding to a constant force.)

Particles that experience a linear potential, like (2.120), are said to be *confined*. In Section 3, we will see that quarks in QCD exhibit a similar behaviour, albeit for more mysterious reasons.

2.3.3 Non-Abelian Higgs Mechanism

The idea of the Higgs mechanism extends naturally to non-Abelian theories. This is the context in which we will need it when discussing electroweak theory in Section 5.

One novelty is that the gauge group G need not be broken completely, and there could be some surviving massless gauge bosons. We will illustrate this with an example. Consider again the $O(3)$ sigma model that we previously discussed in Section 2.2 in the context of spontaneous symmetry breaking of global symmetries. This time, however, we will promote the $SO(3)$ symmetry to a gauge symmetry.

We have a 3-vector of real scalars, ϕ^a with $a = 1, 2, 3$ and define the covariant derivative

$$\mathcal{D}_\mu \phi^a = \partial_\mu \phi^a + g\epsilon^{abc} A^b_\mu \phi^c \ . \tag{2.121}$$

Here the ϵ symbol appears in its role as the generators for $SO(3)$,

$$T^a_{bc} = -i\epsilon^{abc} \ . \tag{2.122}$$

Alternatively, we could view this as an $SU(2)$ gauge theory with the field ϕ transforming in the adjoint representation. We consider the action

$$S = \int d^4x \ \left(-\frac{1}{4} F^a_{\mu\nu} F^{a\,\mu\nu} + \frac{1}{2} \mathcal{D}_\mu \phi^a \mathcal{D}^\mu \phi^a - \frac{\lambda}{2} (\phi^a \phi^a - v^2)^2 \right) \ . \tag{2.123}$$

Here $F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu + g\epsilon^{abc} A^b_\mu A^c_\nu$. In contrast to our previous Yang-Mills action (1.91), we've written the action in terms of the components of the gauge field, A^a_μ with $a = 1, 2, 3$ rather than packaging them into a 3×3 matrix. (This presentation turns out to be marginally simpler for the case of $SO(3)$.)

In the ground state, we have $\phi \cdot \phi = v^2$. We can make a choice of vacuum, say $\phi = (0, 0, v)$. When we were talking about global symmetries, we saw that this broke $G = SO(3) \to H = U(1)$ (or, equivalently, $SO(2)$), and the same is true now that the symmetries are gauged. This means that we expect a massless photon to remain, corresponding to $H = U(1)$, while the other two gauge bosons should become massive due to the Higgs mechanism. We will now see that this is indeed what happens.

As in the Abelian case, we sit in our chosen vacuum and look at fluctuations. The key is in finding the right parameterisation. We choose

$$\phi^a(x) = e^{i(\xi^1(x)T^1 + \xi^2(x)T^2)} \begin{pmatrix} 0 \\ 0 \\ v + \sigma(x) \end{pmatrix} \tag{2.124}$$

with T^1 and T^2 the appropriate $SO(3)$ generators (2.122). If we were dealing with a global $G = SO(3)$ symmetry, then the fields $\xi^1(x)$ and $\xi^2(x)$ would be the Goldstone bosons. (They are related to the scalars that we called $\theta(x)$ and $\varphi(x)$ in the $O(3)$ sigma-model (2.56).)

Crucially, however, we're now thinking about the situation in which $SO(3)$ is gauged, and the two would-be Goldstones $\xi^1(x)$ and $\xi^2(x)$ can both be removed by an $SO(3)$ gauge transformation which acts on the scalar as $\phi \to e^{i\alpha^a T^a}\phi$ for some choice of $\alpha^i(x)$. In this way, they get eaten by the gauge fields A^1_μ and A^2_μ, just as in the Abelian case. In the resulting unitary gauge, the gauge fields and remaining fluctuating scalar $\sigma(x)$ are then described by the action

$$S = \int d^4x \, \left(-\frac{1}{4} F^a_{\mu\nu} F^{a\mu\nu} + \frac{1}{2}\partial_\mu \sigma \partial^\mu \sigma + \frac{1}{2} g^2 (v + \sigma)^2 (A^1_\mu A^{1\,\mu} + A^2_\mu A^{2\,\mu}) - V(\sigma) \right)$$

with

$$V(\sigma) = \frac{\lambda}{2} \sigma^2 (\sigma + 2v)^2 \ . \tag{2.125}$$

As we anticipated, we have two massive gauge bosons, A^1_μ and A^2_μ, each with mass $m^2_\gamma = g^2 v^2$. But the gauge boson A^3_μ remains massless. This is the photon associated to the unbroken symmetry group $H = U(1)$. There is also the massive Higgs field σ with mass $m^2_\sigma = 4\lambda v^2$.

As we commented previously, the gauge boson and Higgs boson have parametrically different masses, so it naively looks like it's possible to take a limit such that $m_\sigma/m_\gamma \to \infty$ and so we can decouple the Higgs and be left with a theory of only massive interacting gauge bosons. This time, however, the limit turns out to be problematic. This can't be seen in the classical analysis that we're focussing on here, but requires us to look more closely at the quantum amplitudes. Ultimately, it boils down to the fact that the theory of purely Goldstone modes is an interacting sigma-model (2.56) and, as such is non-renormalisable. This contrasts with the Abelian situation where the Goldstone that gets eaten is free before gauging. We will return to this issue in Section 5 when we discuss the Higgs mechanism in the Standard Model.

3 The Strong Force

The full structure of the Standard Model will only become apparent in Section 5, after we understand the implications of parity violation. But, before we get there, there are two self-contained aspects of the theory that we can explore in some detail. These are the electromagnetic and strong forces.

We've already met the former in our first course on Quantum Field Theory. The action is

$$S = \int d^4x \, \left(-\frac{1}{4} F_{\mu\nu}F^{\mu\nu} + i\bar{\psi}\,\not{D}\psi - m\bar{\psi}\psi \right) \; . \tag{3.1}$$

Here $F_{\mu\nu}$ is the field strength of electromagnetism and it's excitations are photons. Meanwhile ψ is a Dirac spinor that describes the electron. We can always add further fields corresponding to any other electrically charged particles, like the muon. Upon quantisation, this theory is known as *quantum electrodynamics*, or *QED* for short.

For QED, what you see is what you get. You can stare at the action and, from your knowledge of perturbative quantum field theory, read off immediately that the theory describes a massless photon, coupled to a charged fermion of mass m. This, it turns out, is the only time we will be able to do this. The rest of the Standard Model is considerably more rich and interesting.

Our goal in this section is to describe the strong force. Remarkably, the action for the strong force is almost identical to that of QED. The only real difference is that the $U(1)$ group of electromagnetism is replaced by the gauge group

$$G = SU(3) \; . \tag{3.2}$$

The theory of the strong force is referred to as *quantum chromodynamics*, or *QCD* for short, and is given by

$$S = \int d^4x \, \left(-\frac{1}{2} \mathrm{Tr}\, G_{\mu\nu}G^{\mu\nu} + i\sum_i \bar{q}_i \,\not{D} q_i - m_i\bar{q}_i q_i \right) \; . \tag{3.3}$$

We'll explain what the various parts of this action mean, before we turn to quantum dynamics.

To avoid confusion with the photon, we denote the gauge field as G_μ. It is, like all Yang-Mills fields, Lie-algebra valued which means that we should think of each G_μ as a 3×3 Hermitian matrix. Replete with its gauge indices, we would write it as $(G_\mu)^a{}_b$ with $a, b = 1, 2, 3$. In the context of QCD, this additional index is referred to as *colour*[6]. The dimension of $SU(N)$ is $\dim SU(N) = N^2 - 1$ so there are 8 gauge bosons contained within the matrix G_μ. These are known, collectively, as *gluons*.

We can decompose G_μ into these gluon fields by writing $G_\mu = G_\mu^A T^A$ where T^A are generators of $SU(3)$ which we take to obey

$$\text{Tr}(T^A T^B) = \frac{1}{2}\delta^{AB} \ . \tag{3.4}$$

A convenient basis is given by

$$T^A = \frac{1}{2}\lambda^A \ . \tag{3.5}$$

Here the λ^A the collection of 3×3 *Gell-Mann matrices*

$$\begin{aligned}
\lambda^1 &= \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \ , \quad \lambda^2 = \begin{pmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \ , \quad \lambda^3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \ , \\
\lambda^4 &= \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \ , \quad \lambda^5 = \begin{pmatrix} 0 & 0 & -i \\ 0 & 0 & 0 \\ i & 0 & 0 \end{pmatrix} \ , \quad \lambda^6 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \ , \\
\lambda^7 &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{pmatrix} \ , \quad \lambda^8 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{pmatrix} \ .
\end{aligned} \tag{3.6}$$

These are to $SU(3)$ what the Pauli matrices are to $SU(2)$. Indeed, you can see the Pauli matrices sitting in the top-left corner of λ^1, λ^2, and λ^3, reflecting the existence of an $SU(2)$ sub-group of $SU(3)$. Because $SU(3)$ has rank 2, there are two diagonal Gell-Mann matrices, λ^3 and λ^8. These span the Cartan sub-algebra.

We define the associated field strength

$$G_{\mu\nu} = \partial_\mu G_\nu - \partial_\nu G_\mu - i g_s [G_\mu, G_\nu] \ . \tag{3.7}$$

[6] Americans prefer to work with the convention $u = 1$.

This too is Lie-algebra valued. Note that the gauge potential and field strength are both called G and are distinguished only by the number of μ, ν spacetime indices that they carry. Buried within the field strength we have the strong coupling constant g_s. This is a dimensionless coupling that characterises the strength of the strong force. We will give its value shortly.

The gluons couple to quarks. These are Dirac spinors that we will call q_α where $\alpha = 1, 2, 3, 4$ is the usual spinor index that adorns a Dirac fermion. The quarks transform in the fundamental 3-dimensional representation of $SU(3)$. In group theoretic language, this is usually denoted as $\mathbf{3}$. This means that, in addition to the spinor index, the quarks also carry a colour index $a = 1, 2, 3$. We should think of this colour degree of freedom as a complex, normalised 3-vector that is rotated by $SU(3)$. To cheer us up, we sometimes refer to these three orthogonal states as red, green and blue. Needless to say, if you prefer to label them by your own favourite choice of colours then the physics remains unchanged.

The covariant derivative for each quark q is given by (now suppressing the spinor index)

$$\mathcal{D}_\mu q^a = \partial_\mu q^a - i g_s (G_\mu)^a{}_b q^b \ . \tag{3.8}$$

Here too we see the strong coupling constant g_s multiplying the interaction term.

Finally, the quarks also come with a *flavour index*, $i = 1, \ldots, N_f$ which simply tells us what kind of quark we're dealing with. The full theory of QCD comes with $N_f = 6$ flavours of quarks which, for reasons that will become clearer only in Section 5, we should think of as three pairs. They are down and up; strange and charm; and bottom and top. These quarks have masses

$$\begin{aligned} m_{\text{down}} &= 5 \text{ MeV} \quad &\text{and}\quad &m_{\text{up}} = 2 \text{ MeV} \\ m_{\text{strange}} &= 93 \text{ MeV} \quad &\text{and}\quad &m_{\text{charm}} = 1.3 \text{ GeV} \\ m_{\text{bottom}} &= 4.2 \text{ GeV} \quad &\text{and}\quad &m_{\text{top}} = 173 \text{ GeV} \ . \end{aligned} \tag{3.9}$$

The most striking aspect of these masses is that they span almost 5 orders of magnitude! In Section 5, we'll get a deeper understanding of how the masses arise from the condensation of the Higgs boson. But we won't get any deeper understanding of the particular values that the masses take: we only know these masses by measuring them experimentally.

The quarks also carry electric charge, and so the theory of QCD (3.3) should be augmented by coupling to electromagnetism. Here we will largely ignore the effects of electromagnetism in the dynamics because, as we will see, it is small compared to the strong force. It will, however, prove useful to just list the electric charges Q of various particles that we come across. For the first generation of quarks they are

$$Q_{\text{down}} = -\frac{1}{3}e \quad \text{and} \quad Q_{\text{up}} = \frac{2}{3}e \ . \tag{3.10}$$

Clearly, these are fractional charges relative to the electron. This pattern then repeats itself: the strange and bottom quark both have $Q = -\frac{1}{3}e$ while the charm and top both have $Q = +\frac{2}{3}e$. Note that, in this regard, the first generation of up and down quarks is the odd one out because the charge $\frac{2}{3}$ quark is lighter than the charge $-\frac{1}{3}$ quark.

This completes our discussion of the various elements in the QCD action (3.3). Now it's time to understand the physics.

3.1 Strong Coupling

If you look naively at the action (3.3), you would think that QCD is a theory of massless gluons interacting with quarks. But that's certainly not what we see in the world around us. Any massless gauge boson would mediate a long range force which drops off, like electromagnetism, as $1/r^2$. Yet we know that the effects of the strong force don't extend beyond the nucleus of the atom, which isn't particularly big. In addition, we don't see quarks wandering around freely. What we see are protons and neutrons. If the weak force didn't exist, these would be joined by light particles called pions. But not quarks.

All of which leads us to ask: why are the particles that we see in the world not directly related to the fields in the fundamental Lagrangian (3.3)?

3.1.1 Asymptotic Freedom

The answer to this question starts with the observation that the coupling constant of the strong force is not at all constant. Like all parameters in quantum field theory, its value depends on the distance scale, or equivalently energy scale, at which you look. This is the essence of renormalisation.

To illustrate the physics, we will briefly step back from QCD and consider the more general theory with $G = SU(N_c)$ gauge group, coupled to N_f *massless* quarks. Hence, N_c is the number of colours, and N_f the number of flavours. The gauge coupling g_s^2

depends on the energy scale μ at which the theory is probed and, at one-loop, is given by

$$\frac{1}{g_s^2(\mu)} = \frac{1}{g_0^2} - \frac{b_0}{(4\pi)^2} \log \frac{\Lambda_{UV}^2}{\mu^2} \ . \tag{3.11}$$

Here g_0^2 is the bare coupling that sits in the Lagrangian. It can be thought of as the coupling evaluated at the cut-off scale Λ_{UV} since $g_s^2(\Lambda_{UV}) = g_0^2$. The coefficient b_0 is given by

$$b_0 = \frac{11}{3} N_c - \frac{2}{3} N_f \tag{3.12}$$

A derivation of this result can be found in the lectures on Gauge Theory.

The running of the coupling constant is often summarised in terms of the one-loop beta function

$$\beta(g) \equiv \mu \frac{dg_s}{d\mu} = -\frac{b_0}{(4\pi)^2} g_s^3 \tag{3.13}$$

whose solution gives the logarithmic behaviour (3.11).

The all-important feature of the beta function is the overall minus sign. The flow of the coupling means that the theory is weakly coupled at high energies, a phenomenon known as *asymptotic freedom*. Conversely, it means that the theory is strongly coupled at low energies. From (3.12), we see that asymptotic freedom persists only if the number of flavours is sufficiently small

$$N_f < \frac{11}{2} N_c \ . \tag{3.14}$$

Clearly this is satisfied by QCD with $N_c = 3$ and $N_f = 6$.

Asymptotic freedom is rare in $d = 3 + 1$ dimensions. In fact, it only happens for non-Abelian gauge theories. Coupling constants in any theory run with scale but all of them – the QED fine structure constant, Yukawa couplings, self-interactions of scalars – get bigger as you go to high energies. It is only non-Abelian gauge theories where the coupling gets bigger as you go to low energies.

The comparison to QED is useful. At distances larger than $r \geq 10^{-12}$ m (which is the Compton wavelength of the lightest charged particle, namely the electron) the fine structure constant stops running and plateaus to the familiar value of $\alpha \approx 1/137$. But as you go to higher energies, or shorter distances, the fine structure constant increases. For example, at $r \approx 10^{-17}$ m, which corresponds to $E \approx 100$ GeV, we have $\alpha(\mu) \approx 1/127$.

Asymptotic freedom means that Yang-Mills theory is simple to understand at high energies, or short distance scales. Here it is a theory of massless, interacting gluon fields whose dynamics are well described by the classical equations of motion, together with quantum corrections which can be computed using perturbation methods. However, it becomes much harder to understand what is going on at large distances where the coupling gets strong. Indeed, the beta function (3.13) itself was computed in perturbation theory and is valid only when $g_s^2(\mu) \ll 1$. This equation therefore predicts its own demise at large distance scales.

We can estimate the distance scale at which we think we will run into trouble. Taking the one-loop beta function at face value, we can ask: at what scale does $g_s^2(\mu)$ diverge? This happens at a finite energy

$$\Lambda_{\rm QCD} = \mu \exp\left(-\frac{8\pi^2}{b_0 g_s^2(\mu)}\right) \ . \tag{3.15}$$

This is known as the *strong coupling scale*, or just the *QCD scale*. It has the property that $d\Lambda/d\mu = 0$. In other words, it is an RG invariant. This is the scale at which the gauge coupling becomes strong.

Viewed naively, there's something very surprising about the emergence of the scale Λ_{QCD}. This is because the classical theory has no dimensionful parameter. Yet the quantum theory has a physical scale, Λ_{QCD}. It seems that the quantum theory has generated a scale out of thin air, a phenomenon which goes by the name of *dimensional transmutation*. In fact, as the definition (3.15) makes clear, there is no mystery about this. Quantum field theories are not defined only by their classical action alone, but also by the cut-off Λ_{UV}. Although we might like to think of this cut-off as merely a crutch, and not something physical, this is misleading. It is not something we can do without. And it is this cut-off which evolves to the physical scale Λ_{QCD}.

$$\Lambda_{\rm QCD} = \Lambda_{UV}\, e^{-8\pi^2/b_0 g_0^2} \ . \tag{3.16}$$

This means that if the bare coupling is small, $g_0 \ll 1$, as it should be then the physical scale $\Lambda_{\rm QCD}$ is exponentially suppressed relative to the UV cut-off: $\Lambda_{\rm QCD} \ll \Lambda_{UV}$. It's a beautiful example of how a low-energy scale can be naturally generated from a high energy scale. (A similar mechanism can be seen in other contexts, including the BCS theory of superconductivity and the Kondo effect.)

The QCD Scale for QCD

So far, our discussion has been for the general theory of $SU(N_c)$ with N_f flavours of massless quarks. What happens for actual QCD?

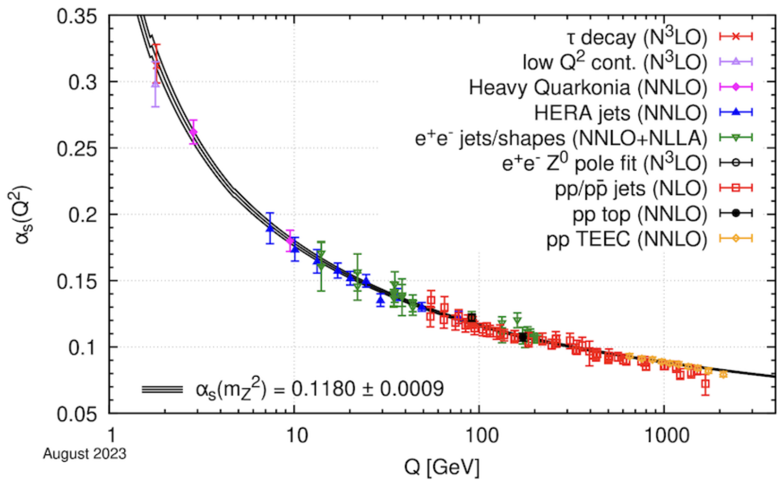


Figure 8. The running of the strong coupling coupling constant $\alpha_s = g_s^2/4\pi$ in terms of energy which is denoted Q in the plot. This is taken from the particle data group's review of QCD.

There is one important modification which is needed because the quarks in QCD are most certainly not massless. This is easy to accommodate. A quark of mass m contributes to the beta function as if it were massless for scales $\mu \gg m$. And it decouples from the physics for scales $\mu \ll m$. For scales $\mu \sim m$ you need to be more careful, but we'll simply duck the issue.

Revisiting the quarks masses in (3.9), we see that the beta function acts as if it has $N_f = 6$ massless quarks for $\mu \gg 173$ GeV. And for 4.2 GeV $\ll \mu \ll 173$ GeV, it acts as if it has $N_f = 5$ massless quarks, and so on. The combined experimental data for the running of $\alpha_s = g_s^2/4\pi$ is shown in Figure 8.

The most important question is: what is the strong coupling scale $\Lambda_{\rm QCD}$? As we will see, this determines the scale at which the interesting physics happens. For the strong force it lies around

$$\Lambda_{\rm QCD} \approx 200 \text{ MeV} . \tag{3.17}$$

This definition isn't precise and you'll also see statements that it is closer to 300 MeV. This could be due to different regularisation schemes, or whether you choose the definition of this scale to be $\alpha_s(\Lambda_{\rm QCD}) = \infty$ or $\alpha_s(\Lambda_{\rm QCD}) = 1$ (which doesn't change things too much). There's no right or wrong answer. As we will see, the point of $\Lambda_{\rm QCD}$ is to give a ballpark energy scale at which much of the physics of QCD takes place.

To give a value for the strength of the coupling g_s itself, we need to specify the energy scale at which we do the measurement. A useful benchmark is the mass of the Z-boson, $M_Z \approx 90$ GeV. Here the strong coupling constant has been measured remarkably accurately

$$\alpha_s(M_Z) = \frac{g_s^2(M_Z)}{4\pi} = 0.1184 \pm 0.0007 \ . \tag{3.18}$$

This is small enough to trust perturbation theory at these scales.

3.1.2 Anti-Screening and Paramagnetism

It's useful to have some intuition for why non-Abelian gauge theories exhibit asymptotic freedom, with a negative beta function, while all other quantum field theories do not. Ultimately, to see this result you just have to roll up your sleeves and do the calculation (and an opportunity will be offered in the sister course on AQFT). Here we give a nice, but slightly handwaving, analogy from condensed matter.

In condensed matter physics, materials are not boring passive objects. They contain mobile electrons, and atoms with a flexible structure, both of which can respond to any external perturbation such as applied electric or magnetic fields. One consequence of this is an effect known as *screening*. In an insulator, screening occurs because an applied electric field will polarise the atoms which, in turn, generate a counteracting electric field. One usually describes this by introducing the electric displacement $\mathbf{D}$, related to the electric field through

$$\mathbf{D} = \epsilon \mathbf{E} \tag{3.19}$$

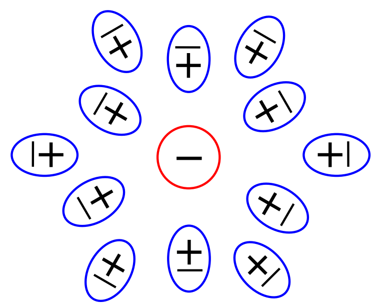
where the permittivity $\epsilon = \epsilon_0(1 + \chi_e)$ with χ_e the electrical susceptibility. For all materials, $\chi_e > 0$. This ensures that the effect of the polarisation is always to reduce the electric field, never to enhance it. You can read more about this in Section 7 of the lecture notes on Electromagnetism.

(As an aside: In a metal, with mobile electrons, there is a much stronger screening effect which turns the Coulomb force into an exponentially suppressed Debye-Hückel, or Yukawa, force. This was described in the final section of the notes on Electromagnetism, but is not the relevant effect here.)

What does this have to do with quantum field theory? In quantum field theory, the vacuum is not a passive boring object. It contains quantum fields which can respond to any external perturbation. In this way, quantum field theories are very much like condensed matter systems. A good example comes from QED. There the one-loop beta function is positive and, at distances smaller than the Compton wavelength of the electron, the gauge coupling runs as

$$\frac{1}{e^2(\mu)} = \frac{1}{e_0^2} + \frac{1}{12\pi^2}\log\left(\frac{\Lambda_{UV}^2}{\mu^2}\right) \ . \tag{3.20}$$

This tells us that the charge of the electron gets effectively smaller as we look at larger distance scales, a phenomenon that is understood in very much the same spirit as condensed matter systems. In the presence of an external charge, electron-positron pairs will polarize the vacuum, as shown in the figure, with the positive charges clustering closer to the external charge. This cloud of electron-positron pairs shields the original charge, so that it appears reduced to someone sitting far away.



The screening story above makes sense for QED. But what about QCD? The negative beta function tells us that the effective charge is now getting larger at long distances, rather than smaller. In other words, the Yang-Mills vacuum does not screen charge: it anti-screens. From a condensed matter perspective, this is weird. As we mentioned above, materials always have $\chi_e > 0$ ensuring that the electric field is screened, rather than anti-screened.

However, there's another way to view the underlying physics. We can instead think about magnetic screening. Recall that in a material, an applied magnetic field induces dipole moments and these, in turn, give rise to a magnetisation. The resulting magnetising field $\mathbf{H}$ is defined in terms of the applied magnetic field as

$$\mathbf{B} = \mu\mathbf{H} \tag{3.21}$$

with the permeability $\mu = \mu_0(1+\chi_m)$. Here χ_m is the magnetic susceptibility and, in contrast to the electric susceptibility, can take either sign. The sign of χ_m determines the magnetisation of the material, which is given by $\mathbf{M} = \chi_m\mathbf{H}$. For $-1 < \chi_m < 0$, the magnetisation points in the opposite direction to the applied magnetic field. Such materials are called *diamagnets*. (A perfect diamagnet has $\chi_m = -1$. This is what happens in a superconductor.) In contrast, when $\chi_m > 1$, the magnetisation points in the same direction as the applied magnetic field. Such materials are called *paramagnets*.

In quantum field theory, polarisation effects can also make the vacuum either diamagnetic or paramagnetic. Except now there is a new ingredient which does not show up in real world materials discussed above: relativity! This means that the product must be

$$\epsilon\mu = 1$$

because "1" is the speed of light. In other words, a relativistic diamagnetic material will have $\mu < 1$ and $\epsilon > 1$ and so exhibit screening. But a relativistic paramagnetic material will have $\mu > 1$ and $\epsilon < 1$ and so exhibit anti-screening. Phrased in this way, the existence of an anti-screening vacuum is much less surprising: it follows simply from paramagnetism combined with relativity.

For free, non-relativistic fermions, we calculated the magnetic susceptibility in the lectures on Statistical Physics when we discussed Fermi surfaces. In that context, we found two distinct contributions to the magnetisation. Landau diamagnetism arose because electrons form Landau levels. Meanwhile, Pauli paramagnetism is due to the spin of the electron. These two effects have the same scaling but different numerical coefficients.

When you dissect the computation of the one-loop beta function in Yang-Mills theory, you can see that the gluons also give two distinct contributions: one diamagnetic, and one paramagnetic. And the paramagnetic contribution wins. Viewed in this light, asymptotic freedom can be traced to the paramagnetic contribution from the gluon spins.

3.1.3 The Mass Gap

When the coupling is small, quantum field theories look similar to their classical counterparts. For example, classical Maxwell theory provides a decent guide to what you might expect from QED. In contrast, when the coupling is large, all bets are off. The quantum theory and classical theory may be completely different. Yang-Mills and QCD provide the archetypal example.

We will start our discussion by ignoring the quarks completely and look just at Yang-Mills theory,

$$S = \int d^4x \ -\frac{1}{2}\mathrm{Tr}\, G_{\mu\nu}G^{\mu\nu} \ . \tag{3.22}$$

For QCD we take gauge group $G = SU(3)$, but everything we're about to say holds for any simple, compact Lie group.

Classically, Yang-Mills describes massless, interacting spin 1 fields. Its solutions include, among other things, waves that propagate at the speed of light. The question that we want to ask is: what is the physics of the quantum theory?

Because the coupling is strong at low energies, we can't answer this question using the traditional perturbative techniques that we learned in our first course on Quantum Field Theory. In fact, if we rely purely on analytic methods we can't answer this question at all! Instead, we rely on numerical simulation and experiment, together with some heuristic ideas and a number of solvable toy models which give us intuition for what quantum field theories can do. But we do have a robust, clear answer:

Quantum Yang-Mills is not a theory of massless particles, Instead, the lightest particle has a mass of $m \sim \Lambda_{\rm QCD}$. This particle is called a *glueball*. We say that the theory is *gapped* which means that there is a gap between the ground state and the first excited state with energy $E = mc^2$. These glueballs also exist in our world, although they mix strongly with various neutral meson states and so don't have a very clean experimental signature.

We don't currently have the ability to prove that Yang-Mills is gapped from first principles. It is generally considered one of the most important and challenging open problems in mathematical physics.

3.1.4 A Short Distance Coulomb Force

The existence of a mass gap goes hand in hand with another phenomenon: this is *confinement*.

To highlight the physics, it's best if we again look at the slightly more general case of $G = SU(N)$ gauge theory. We can ask the kind of questions that we studied in our first course on Electromagnetism. Suppose that you take two test particles, a quark in the fundamental representation $\mathbf{N}$ and an anti-quark in the anti-fundamental $\overline{\mathbf{N}}$. What force do they feel?

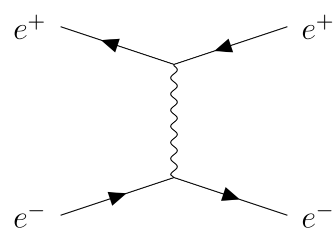
There are two different answers to this question, depending on the separation r between the particles. If they are separated by a short distance $r \ll \Lambda_{\rm QCD}^{-1} \approx 5 \times 10^{-15}$ m, then the coupling g_s^2 is small and we can trust the classical result. However, if the particles are separated by a large distance $r \gg \Lambda_{\rm QCD}^{-1}$, then we're firmly in the regime of strongly coupled physics and we might expect that the classical result is not a good guide.

Here we start by considering the short-distance regime $r \ll \Lambda_{\text{QCD}}^{-1}$. The Compton wavelength of a particle of mass m is $\lambda \sim 1/m$ and it only makes sense to talk about separating two quantum particles a distance r if $r \gg \lambda$. This means that to talk about the short-distance force experienced by two quarks, the quarks must have mass $m \gg \Lambda_{\text{QCD}}$. In the context of QCD, that means that the analysis below is valid only for charm, bottom and top quarks.

Let's remind ourselves of the story in QED. In electromagnetism, two particles of equal and opposite charges $\pm e$, separated by a distance r, experience an attractive Coulomb force, described by the potential energy $V(r)$,

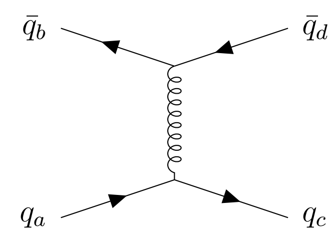
$$V(r) = -\frac{e^2}{4\pi r} \ . \tag{3.23}$$

In the framework of QED, we can reproduce this from the the tree-level exchange of a single photon (where time should be viewed as flowing left-to-right in this diagram)



This computation can be found in the lectures on Quantum Field Theory.

Now we want to do the same calculation in QCD. The diagram is the same, but with a gluon, rather than a photon, as the intermediary. The only difference lies in the fact that quarks carry colour indices, which are the $a, b, c, d = 1, \ldots, N$ indices in the Feynman diagram below



Using the Feynman rules for QCD, the tree level potential between the quarks is given by the same Coulomb force law, dressed with the group theoretic factor

$$V(r) = \frac{g_s^2}{4\pi r} T^A_{ca} T^{\star A}_{db} \ . \tag{3.24}$$

We've still got those colour indices to deal with. At first glance, it looks like there's N^2 different possibilities for the states of the ingoing particles ($a, b = 1, \ldots, N$) and a further N^2 different possibilities for the states of the outgoing particles ($c, d = 1, \ldots, N$). Happily, all of this boils down to some simple group theory. In the present case, we have the tensor product of representations

$$\mathbf{N} \otimes \overline{\mathbf{N}} = \mathbf{1} \oplus \text{adj} \tag{3.25}$$

where the adjoint representation has dimension $N^2 - 1$. The object $T^A T^{\star\, A}$, viewed as a $N^2 \times N^2$ dimensional matrix, will then have two different eigenvalues, one for each of these representations. This will lead to two different coefficients for the forces. Our goal is to determine them. Here we give the general result:

Claim: Suppose that we have two particles in representations $\mathbf{R}_1$ and $\mathbf{R}_2$. For each representation $\mathbf{R} \subset \mathbf{R}_1 \otimes \mathbf{R}_2$, the force experienced by the two particles will be proportional to

$$C(R) - C(R_1) - C(R_2) \tag{3.26}$$

where $C(R)$ is a number that characterises each representation, known as the *quadratic Casimir*, defined as

$$T^A(R) T^A(R) = C(R)\, \mathbb{1}\ . \tag{3.27}$$

Proof: Gluon exchange will result in a Coulomb-like force law (3.24), but with the group theoretic factor $T^A(R_1) T^A(R_2)$. (For $\mathbf{R}_1 = \mathbf{N}$ and $\mathbf{R}_2 = \overline{\mathbf{N}}$, this coincides with the result (3.24).) Consider the operator

$$S^A = T^A(R_1) \otimes \mathbf{1}\ +\ \mathbf{1} \otimes T^A(R_2)\ . \tag{3.28}$$

Squaring and rearranging, we have

$$T^A(R_1) \otimes T^A(R_2) = \frac{1}{2}\left[S^A S^A - T^A(R_1) T^A(R_1) \otimes \mathbf{1} - \mathbf{1} \otimes T^A(R_2) T^A(R_2)\right]\ . \tag{3.29}$$

(This is the same kind of calculation that one does in atomic physics when computing the consequence of the spin orbit coupling $\mathbf{L} \cdot \mathbf{S}$. You can read more about this in the lectures on Topics in Quantum Mechanics.) Each of the final two terms on the right-hand side is a quadratic Casimir (3.27), while the first term decomposes into block diagonal matrices, with components labelled by the irreducible representations $\mathbf{R} \subset \mathbf{R}_1 \otimes \mathbf{R}_2$. We have

$$T^A(R_1) \otimes T^A(R_2)\Big|_R = \frac{1}{2}\left[C(R) - C(R_1) - C(R_2)\right] \tag{3.30}$$

as promised. □

The upshot is that to calculate the force between a quark and anti-quark (or, indeed, between any two representations) we just need to know the quadratic Casimirs. For $G = SU(N)$, the Casimirs for the fundamental, anti-fundamental and adjoint are

$$C(\mathbf{N}) = C(\overline{\mathbf{N}}) = \frac{N^2-1}{2N} \quad \text{and} \quad C(\text{adj}) = N \; . \tag{3.31}$$

We also have $C(\mathbf{1}) = 0$ for the singlet (trivial) representation. This means that a quark-anti-quark pair with their colour degrees of freedom entangled as a singlet experience a force proportional to

$$\frac{1}{2}\left[C(\mathbf{1}) - C(\mathbf{N}) - C(\overline{\mathbf{N}})\right] = -\frac{N^2-1}{2N} \; . \tag{3.32}$$

The minus sign means that this force is attractive. This is what we would have expected from our classical intuition. However, when the quarks sit in the adjoint channel, we have

$$\frac{1}{2}\left[C(\text{adj}) - C(\mathbf{N}) - C(\overline{\mathbf{N}})\right] = \frac{1}{2N} \; . \tag{3.33}$$

Perhaps surprisingly, this is a repulsive force.

We can do the same analysis if we have two quarks, rather than a quark and anti-quark. Now the group theoretic decomposition is

$$\mathbf{N} \otimes \mathbf{N} = \square\square \oplus \begin{smallmatrix}\square\\\square\end{smallmatrix}$$

where $\square\square$ is the Young tableaux representation for the symmetric representation, with $\dim(\square\square) = \frac{1}{2}N(N+1)$ while $\begin{smallmatrix}\square\\\square\end{smallmatrix}$ means the anti-symmetric representation with $\dim(\begin{smallmatrix}\square\\\square\end{smallmatrix}) = \frac{1}{2}N(N-1)$. The relevant Casimirs are

$$C\left(\square\square\right) = \frac{(N-1)(N+2)}{N} \quad \text{and} \quad C\left(\begin{smallmatrix}\square\\\square\end{smallmatrix}\right) = \frac{(N-2)(N+1)}{N}$$

From this we learn that two quarks which sit in the symmetric channel classically repel each other, since

$$\frac{1}{2}\left[C\left(\square\square\right) - C(\mathbf{N}) - C(\mathbf{N})\right] = \frac{N-1}{2N} \; . \tag{3.34}$$

Meanwhile, two quarks that sit in the anti-symmetric channel feel a classical attractive force,

$$\frac{1}{2}\left[C\left(\begin{smallmatrix}\square\\\square\end{smallmatrix}\right) - C(\mathbf{N}) - C(\mathbf{N})\right] = -\frac{N+1}{2N} \; . \tag{3.35}$$

Ultimately, our interest lies in QCD with $G = SU(3)$. Here there's a group theoretic novelty because the anti-symmetric representation is actually the same as the anti-fundamental,

$$\mathbf{3} \otimes \mathbf{3} = \overline{\mathbf{3}} \oplus \mathbf{6} \ . \tag{3.36}$$

This means that two quarks will attract in the anti-symmetric $\bar{\mathbf{3}}$ channel. But we could then add a third quark and, from (3.32), this too will feel an attractive force if all three sit in the singlet. We see that three quarks can feel a mutually attractive force in QCD. Of course, this force is computed classically and it falls off with a $1/r$ potential, just like the Coulomb force of electromagnetism. Nonetheless, this is the first time that we see why it might be energetically preferable for three quarks to form colour singlet bound states.

3.1.5 A Long Distance Confining Force

The analysis above was only for particles separated by very short distances $r \ll \Lambda_{\text{QCD}}^{-1} \approx 5 \times 10^{-15}$ m. But our real interest is in what happens at large distance scales where the Yang-Mills coupling becomes strong.

Previously, we stated (but didn't prove!) that Yang-Mills has a mass gap. This means that, at distances $\gg 1/\Lambda_{QCD}$, the force will be due to the exchange of massive particles rather than massless particles. In many situations, the exchange of massive particles results in an exponentially suppressed Yukawa force, of the form $V(r) \sim e^{-mr}/r$, and you might have reasonably thought this would be the case for Yang-Mills. You would have been wrong.

Let's again consider a quark and an anti-quark, in the $\mathbf{N}$ and $\overline{\mathbf{N}}$ representations respectively. At large distances, the potential energy between the two turns out to grow linearly with distance

$$V(r) = \sigma r \tag{3.37}$$

for some value σ that has dimensions of energy per length. For reasons that we will explain shortly, it is often referred to as the *string tension*. On dimensional grounds, we must have $\sigma \sim \Lambda_{QCD}^2$ since there is no other dimensionful parameter in the game. The force law (3.37) is, to put it mildly, a dramatic departure from what we're used to. The potential energy now *increases* with separation. Indeed, it costs an infinite amount of energy to pull the quark-anti-quark pair to infinity.

For two quarks, both in the fundamental representation, the result is even more dramatic. Now the tensor product of the two representations does not include a singlet (at least this is true for $SU(N)$ with $N \geq 3$). The energy of the two quarks turns out to be infinite. This is a general property of quantum Yang-Mills: the only finite energy states are gauge singlets. The theory is said to be *confining*, meaning that an individual quark cannot survive on its own, but is forced to enjoy the company of friends.

The phenomenon of confinement is, like the mass gap, something that we can't prove from first principles. Once again, however, there is clear numerical evidence together with a plethora of heuristic explanations.

In Section 3.3, we'll look more closely at how quarks and anti-quarks bind together in QCD. Roughly speaking, there are two possibilities. First a quark and anti-quark can bind together to form a colour singlet. The resulting particle is known as a *meson*. But, alternatively, three quarks can bind together to form a colour singlet by dint of the invariant tensor ϵ^{abc} of $SU(3)$. The resulting particle is called a *baryon*, with the proton and neutron being the most obvious examples.

Note that if the strong force was described by $SU(N)$, with $N \neq 3$, then mesons would always be quark-anti-quark pairs and, hence, are always bosons. In contrast, baryons in $SU(N)$ contain N quarks and hence are fermions when N is odd and bosons when N is even.

The QCD Flux Tube

We've already seen an example of a confining potential (3.37) in Section 2.3 when discussing superconductivity. In that context, magnetic monopoles experience a confining force, and the reason was clear: the Meissner effect means that it's energetically preferable for the magnetic field lines to form flux tubes.

No such simple explanation is known for confinement in QCD, but it's clear from numerical simulations that a similar flux tube, or string, does form, now comprised of chromoelectric field lines. Two examples are shown in Figure 9, where we see flux tubes between the quark-anti-quark that form a meson and also between three quarks that form a baryon. In fact, some of the original studies of string theory were motivated by understanding the dynamics of these flux tubes.

However, in contrast to the the Higgs phase of a superconductor, it doesn't make sense to search for a classical solution to the equations of motion that describes the QCD flux tube. Instead the QCD flux tube is very much a quantum effect, arising only after performing the path integral, which involves summing over many different

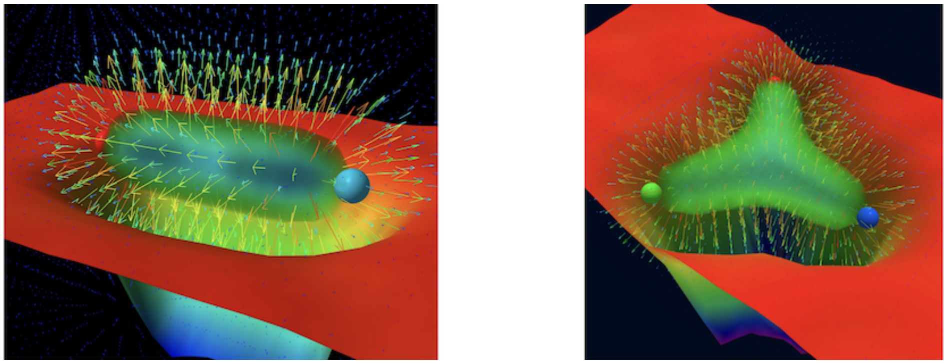


Figure 9. The chromoelectric flux tube between a quark and anti-quark in a meson state, on the left, and between three quarks in a baryon state on the right. From the QCD simulations of Derek Leinweber.

field configurations. To emphasise the physics, it's best to work with the alternative rescaling of the Yang-Mills action (1.103) in which the gauge coupling sits as an overall coefficient, so the path integral over the gauge field takes the schematic form

$$Z = \int \mathcal{D}G_\mu \ \exp\left(-\frac{i}{2g_s^2}\int d^4x \ \mathrm{Tr}\, G_{\mu\nu}G^{\mu\nu}\right) \ . \tag{3.38}$$

At weak coupling, we have $g_s^2 \ll 1$ and we may use saddle-point techniques to show that the path integral is dominated by solutions to the classical equations of motion. But at strong coupling, we have $g_s^2 \to \infty$ which, roughly speaking, is telling us that there's no suppression to the path integral at all. All field configurations, regardless of how wildly they oscillate, contribute equal weight. Among the infinity of different field configurations, those that look like a flux tube seem to dominate. But we don't know why.

Perhaps the best explanation of confinement (although one that falls well short of a proof) comes from an approach that discretises Yang-Mills theory known as *lattice gauge theory*. In that context, you can show that if you naively sum over all field configurations without any weighting, then you do indeed reproduce the confining behaviour. You can find details of this calculation, together with an explanation of why the calculation is not really performed in the physical regime, in the lectures on Gauge Theory.

It's tempting to push the superconductivity analogy further. In a superconductor, electrically charged particles condense (the Cooper pairs) and the result is that magnetic charges confine. Flipping this on its head, if magnetically charged particles were to condense, then electric charges would be confined. This idea goes by the name of the *dual Meissner effect*. It seems right, but it's hard to make it concrete. What are these mysterious chromomagnetic charges that condense in QCD causing quarks to confine? We don't know. However, there are other 4d gauge theories where we can prove confinement analytically and it does happen through the condensation of monopoles. (This is what happens in the famous Seiberg-Witten solution of $\mathcal{N} = 2$ supersymmetric gauge theories.)

The Effect of Light Quarks

As if the problem of confinement wasn't difficult enough, things are actually more complicated than I've sketched above. This is because, in real world QCD, the simple force formula (3.37) that designates a confining theory, simply isn't true!

Here's the deal. Suppose that we have pure Yang-Mills theory. Then, for any choice of non-Abelian gauge group, including $G = SU(3)$, the theory is strongly believed to have a mass gap, determined by its strong coupling scale $\Lambda_{\rm QCD}$, and confine. Here "confinement" means that if you introduce two test particles into the theory – a quark and anti-quark – then the long-distance force law between them will exhibit the linear behaviour (3.37).

Now suppose that you have Yang-Mills theory coupled to a single dynamical quark that has mass $m \gg \Lambda_{\rm QCD}$. For example, you could think of the artificial world in which there is only a charm quark and nothing else. We can again ask what the energy is between two test particles that we take to be a quark-anti-quark pair. At large distances $r \gg \Lambda_{\rm QCD}^{-1}$, we have a confining potential

$$V(r) = \sigma r \ . \tag{3.39}$$

But, this time, it doesn't persist for all r. This is because once we stretch the particles past the point $\sigma r > 2m$, then you can lower the energy of the state by creating a quark-anti-quark pair from the vacuum. The $q\bar{q}$ pair will break the string and you'll be left with two meson-like states, in which your original quark-anti-quark test particles are now bound to the dynamical quarks of the theory.

This means that the regime of the confining force (3.39) is limited. It happens only for long distances, but not too long distances. Using the fact that the string tension scales as $\sigma \sim \Lambda_{\rm QCD}^2$, we see that quarks experience the confining force only in a region

$$\frac{1}{\Lambda_{\rm QCD}} \ll r \ll \frac{m}{\Lambda_{\rm QCD}^2} \ . \tag{3.40}$$

Nonetheless, if we only have dynamical quarks with mass $m \gg \Lambda_{\rm QCD}$, then there's still a window in which we see the confining behaviour.

However, for real world QCD, there is no such window! The lightest quark has mass $m \ll \Lambda_{\rm QCD}$. If you like, the string breaks through the pair creation of up and down $q\bar{q}$ pairs before we even get to the confining regime $r \gg \Lambda_{\rm QCD}^{-1}$. This means that thinking about the confining nature of real world QCD in terms of the linear potential (3.39) is a useful, but not entirely accurate, fiction.

What does survive, however, is the statement that all finite energy states in QCD are necessarily colour singlets. That is the key takeaway that we will need when discussing the observed particle spectrum in Section 3.3.

3.2 Chiral Symmetry Breaking

Here's a general piece of advice. If you want to understand the dynamics of a quantum field theory, first understand the symmetries. They dictate how the dynamics is organised and will often contain clues about the nature of the low-energy physics.

So what are the symmetries of QCD? Well, obviously the theory is based on a $G = SU(3)$ gauge group but, as we've stressed previously, that's really a redundancy rather than a symmetry. Here we are interested in global symmetries.

The actual symmetry group of the QCD action (3.3) is $U(1)^{N_f}$, which rotates the phase of each individual Dirac quark field. That alone doesn't give us much insight. However, there is a much larger *approximate* symmetry of the theory. This emerges if we pretend that the quarks are massless.

First, we should ask: why are we allowed to pretend that quarks are massless? The reason is that QCD comes with its own dynamical scale $\Lambda_{\rm QCD}$. This is the scale at which all the interesting physics happens. This means that if we have any quark with a mass $m \ll \Lambda_{\rm QCD}$, then it's appropriate to first understand the dynamics of the gauge fields in the massless limit, and subsequently figure out how the presence of the mass changes things as corrections of order $m/\Lambda_{\rm QCD}$.

As we've seen, we have $\Lambda_{\rm QCD} \approx 200$ MeV, while the masses of the quarks are

$$\begin{aligned} m_{\rm down} &= 5 \text{ MeV} \quad &&\text{and} \quad m_{\rm up} = 2 \text{ MeV} \\ m_{\rm strange} &= 93 \text{ MeV} \quad &&\text{and} \quad m_{\rm charm} = 1.3 \text{ GeV} \\ m_{\rm bottom} &= 4.2 \text{ GeV} \quad &&\text{and} \quad m_{\rm top} = 173 \text{ GeV} \; . \end{aligned} \tag{3.41}$$

Clearly there's no sense in which the charm, bottom and top quarks are light. In fact, they're so much heavier than the QCD scale that they effectively just decouple from the low-energy dynamics and, for the story that we're about to tell, we can just ignore them. (We'll revisit these heavy quarks in Section 3.3 when we look more closely at the kinds of mesons and baryons that we can form.)

At the other end, no one's going to argue against the statement that $m_{\rm up}, m_{\rm down} \ll \Lambda_{\rm QCD}$ and it's an excellent approximation to treat these as massless and then see how the very small mass changes things. That leaves us with the strange quark. While it's certainly true that $m_{\rm strange} < \Lambda_{\rm QCD}$, you might reasonably complain that it's a bit of stretch to replace $<$ with $\ll$. All of which means that it will certainly be useful to pretend that there are two massless quarks, and it's probably worth seeing what happens if we're more optimistic and pretend that there are three massless quarks.

At this stage, we don't need to commit to the number of massless quarks, and we can work in generality. In fact, we don't even need to commit to the number of colours. Consider $G = SU(N_c)$ Yang-Mills, coupled to N_f flavours of massless fundamental fermions that we will continue to refer to as "quarks".

The additional symmetry comes from the realisation that each 4-component Dirac spinor q decomposes into two 2-component Weyl spinors, as in (1.48),

$$q = \begin{pmatrix} q_L \\ q_R \end{pmatrix} \ . \tag{3.42}$$

Each of the Weyl spinors q_L and q_R carries a colour index that runs over $1, \ldots, N_c$ and a flavour index $i = 1, \ldots, N_f$, as well as it's 2-component spinor index. Written in terms of these Weyl fermions, our generalised but massless, QCD action (3.3) becomes

$$S = \int d^4x \ \left(-\frac{1}{2} {\rm Tr}\, G_{\mu\nu} G^{\mu\nu} + i \sum_{i=1}^{N_f} \bar{q}_{L\,i} \bar{\sigma}^\mu \mathcal{D}_\mu q_{L\,i} + \bar{q}_{R\,i} \sigma^\mu \mathcal{D}_\mu q_{R\,i} \right) \ . \tag{3.43}$$

where we've suppressed both colour and spinor indices in this expression. Written in this way, we see that the classical Lagrangian has a global symmetry

$$G_F = U(N_f)_L \times U(N_f)_R \tag{3.44}$$

which acts on the flavour indices as

$$U(N_f)_L : q_{L\,i} \mapsto L_{ij} q_{L\,j} \quad \text{and} \quad U(N_f)_R : q_{R\,i} \mapsto R_{ij} q_{R\,j} \tag{3.45}$$

where both $L, R \in U(N_f)$. This is known as a *chiral* symmetry because it acts differently on left-handed and right-handed Weyl spinors. This chiral symmetry is a symmetry only of the theory with massless fermions because as soon as we add a mass term like $\bar{q}_L q_R$, it breaks the chiral symmetry to its diagonal subgroup.

As we will see, in the quantum theory different parts of the symmetry group G_F suffer different fates. Perhaps the least interesting is the overall $U(1)_V$, under which both q_L and q_R transform in the same way: $q_{L\,i} \to e^{i\alpha} q_{L\,i}$ and $q_{R\,i} \to e^{i\alpha} q_{R\,i}$. This symmetry survives in the quantum theory and the associated conserved quantity counts the number of quark particles of either handedness. In the context of QCD, this is referred to as *baryon number*, because it counts baryons, but not mesons which have a quark-anti-quark pair.

The other Abelian symmetry is the axial symmetry, $U(1)_A$. Under this, the left-handed and right-handed fermions transform with an opposite phase: $q_{L\,i} \to e^{i\beta} q_{L\,i}$ and $q_{R\,i} \to e^{-i\beta} q_{R\,i}$. This is more subtle. It turns out that although this is a symmetry of the classical Lagrangian, it is *not* a symmetry of the full quantum theory due to a phenomenon known as the *anomaly*. We will explain this in Section 4. For now, you will have to just trust me when I say that $U(1)_A$ is not actually a symmetry and we will not discuss it for the rest of this section.

This means that the global symmetry group of the quantum theory is

$$G_F = U(1)_V \times SU(N_f)_L \times SU(N_f)_R \ . \tag{3.46}$$

The two non-Abelian symmetries act as (3.45), but where L and R are now each elements of $SU(N_f)$ rather than $U(N_f)$. The question that we want to ask is: what becomes of this chiral symmetry?

3.2.1 The Quark Condensate

There are two striking phenomena in QCD-like theories. The first is confinement. The second, which at first glance seems less dramatic, is the formation of a *quark condensate*, also known as a *chiral condensate*.

The quark condensate is a vacuum expectation value of the composite operators $\bar{q}_{L\,i}(x) q_{R\,j}(x)$. (As usual in quantum field theory, one has to regulate coincident operators of this type to remove any UV divergences). It turns out that the strong coupling dynamics of non-Abelian gauge theories gives rise to an expectation value of the form

$$\langle \bar{q}_{L\,i} q_{R\,j} \rangle = -\sigma \delta_{ij} \ . \tag{3.47}$$

Here σ is a constant which has dimension of $[\text{Mass}]^3$ because a free fermion in $d = 3+1$ has dimension $[\psi] = \frac{3}{2}$. (An aside: in Section 3.1 we referred to the string tension as σ; it's not the same object that appears here.) The only dimensionful parameter in our theory is the strong coupling scale Λ_{QCD}, so we expect that parameterically $\sigma \sim \Lambda_{QCD}^3$, although they differ by some order 1 number.

The first question to ask is: why does the condensate (3.47) form? The honest answer is: we don't know. It is, like confinement and many other properties of strongly coupled gauge theories, an open question. It turns out that the formation of the condensate is implied by confinement, a statement that we will prove in Section 4.3. We will also give some very heuristic and hand-waving intuition for the formation of the condensate shortly.

Of more immediate concern are the consequences of the condensate (3.47). This is surprisingly easy to answer because as we now explain, everything is entirely determined by symmetry.

The key point is that, while our theory enjoys the full symmetry group (3.46), the vacuum does not. This is because, under G_F, the condensate (3.47) transforms as

$$\langle \bar{q}_{L\,i} q_{R\,j} \rangle \rightarrow -\sigma (L^\dagger R)_{ij}$$

This means that massless QCD exhibits a dynamical spontaneous symmetry breaking which, in the present context, is known as *chiral symmetry breaking* (sometimes shortened to χSB). We see that the condensate remains untouched only when $L = R$. This tells us that the symmetry breaking pattern is

$$G_F = U(1)_V \times SU(N_f)_L \times SU(N_f)_R \;\rightarrow\; U(1)_V \times SU(N_f)_V \tag{3.48}$$

where $SU(N_f)_V$ is the diagonal subgroup of $SU(N_f)_L \times SU(N_f)_R$.

At this stage, a large part of the physics follows from our general discussion of symmetry breaking in Section 2.2. There will necessarily be a manifold of ground states (2.61), given by the coset

$$\mathcal{M}_0 = [SU(N_f)_L \times SU(N_f)_R] \,/ SU(N_f)_V \ . \tag{3.49}$$

The number of massless Goldstone bosons is given by the dimension

$$\dim \mathcal{M}_0 = N_f^2 - 1 \ . \tag{3.50}$$

This means that, if we pretend that we have $N_f = 2$ massless quarks (up and down), then we should find 3 massless Goldstone bosons in our world. We will soon identify these with light mesons known as *pions*. If we're happy to be bold and think that there are really $N_f = 3$ (up, down, and strange), then we should find 8 massless Goldstone bosons. These additional Goldstone bosons are not-so-light mesons called *kaons* and the *eta*.

In our world the pions are not massless. But this is because the constituent quarks are not exactly massless so the chiral symmetry is not exact. Nonetheless, the chiral symmetry is an approximate symmetry which, in turn, means that the would-be Goldstone bosons are light, but not exactly massless. Indeed, the pions are notably lighter than all other hadrons in QCD. We'll look more closely at the details as this section proceeds.

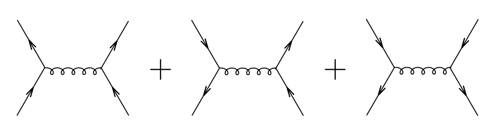
At a more theoretical level, we learn something interesting. Yang-Mills theory has a mass gap. But massless QCD, at least for $N_f \geq 2$ where there is a non-Abelian global symmetry, does not. Even if the theory confines, giving massive baryons and glueballs, chiral symmetry breaking means that there are massless Goldstone bosons.

How to Think About the Quark Condensate

The existence of a quark condensate (3.47) is telling us that the vacuum of space is populated by quark-anti-quark pairs. Again, there is an analogy with superconductivity, albeit with the part of superconductivity that we did not discuss in Section 2.3.2. In a superconductor, the Cooper pairing means that the vacuum is populated by electron pairs. Importantly, these are really electron pairs, rather than electron-hole pairs, which is responsible for the breaking of $U(1)_{\text{em}}$. In contrast, the QCD vacuum contains quark-anti-quark pairs so the overall $U(1)_V$ survives, and it's the chiral symmetry that is broken.

In a superconductor, the instability to formation of an electron condensate is a result of the existence of a Fermi surface, together with a weak attractive force mediated by phonons. In the vacuum of space, however, things are not so easy. The formation of a quark condensate does not occur in weakly coupled theory. Indeed, this follows on dimensional grounds because, as we mentioned above, the only relevant scale in the game is Λ_{QCD}.

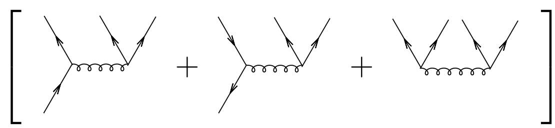
To gain some intuition for why a condensate might form, let's look at what happens at weak coupling $g_s^2 \ll 1$. Here we can work perturbatively and see how the gluons change the quark Hamiltonian. There are two, qualitatively different effects. The first is the kind that we already met in Section 3.1.4; a tree level exchange of gluons gives rise to a force between quarks. This takes the form

$\Delta H_1 = g_s^2 \Big[$  $\Big]$

As we saw in Section 3.1.4, the upshot of these diagrams is to provide a repulsive force between two quarks in the symmetric channel, and an attractive force in the anti-symmetric channel. Similarly, a quark-anti-quark pair attract when they form a colour singlet and repel when they form a colour adjoint.

The second term is more interesting for us. The relevant diagrams take the form

$$\Delta H_2 = g_s^2$$



The novelty of these terms is that they provide matrix elements which mix the empty vacuum with a state containing a quark-anti-quark pair. In doing so, they change the total number of quarks + anti-quarks.

The existence of the quark condensate (3.47) is telling us that, in the strong coupling regime, terms like ΔH_2 dominate. The resulting ground state has an indefinite number of quark-anti-quark pairs. It is perhaps surprising that we can have a vacuum filled with quark-anti-quark pairs while still preserving Lorentz invariance. To do this, the quark pairs must have opposite quantum numbers for both momentum and angular momentum. Furthermore, we expect the condensate to form in the attractive colour singlet channel, rather than the repulsive adjoint.

The handwaving remarks above fall well short of demonstrating the existence of the quark condensate. So how do we know that it actually forms? Historically, it was first realised from experimental considerations since it explains the spectrum of light mesons; we will describe this in some detail in Section 3.3. At the theoretical level, the most compelling argument comes from numerical simulations on the lattice. However, a full analytic calculation of the condensate is not yet possible. (For what it's worth, the situation is somewhat better in certain supersymmetric non-Abelian gauge theories where one has more control over the dynamics and objects like quark condensates can be computed exactly.) Finally, there is a beautiful, but rather indirect, argument which tells us that the condensate (3.47) must form whenever the theory confines. We will give this argument in Section 4.3.

3.2.2 The Chiral Lagrangian

Chiral symmetry breaking implies the existence of Goldstone bosons. Our next task is to construct the theory that describes these massless particles. This too is dictated entirely by the symmetry structure of the theory.

As we've seen, in any theory with a spontaneously broken continuous symmetry, there is a manifold of ground states $\mathcal{M}_0$ which, for us, is given by (3.49). The different points in $\mathcal{M}_0$ are parameterised by the condensate which, in general, takes the form

$$\langle \bar{q}_{L\,i} q_{R\,j} \rangle = -\sigma\, U_{ij}$$

where $U = L^\dagger R \in SU(N_f)$. The Goldstone bosons are long-wavelength ripples of the condensate where its value now varies in space and time: $U = U(x)$. As we've seen, there are $N_f^2 - 1$ such Goldstone bosons, one for each broken generator in (3.48). We parameterise these excitations by writing

$$U(x) = \exp\left(\frac{2i}{f_\pi}\,\pi(x)\right) \quad \text{with} \;\; \pi(x) = \pi^a(x)\, T^a \; . \tag{3.51}$$

Here $\pi(x)$ is valued in the Lie algebra $su(N_f)$. The matrices T^a_{ij} are the generators of the $su(N_f)$. (Note: we've changed notation here: previously we denoted Lie algebra generators as T^A, with a capital A index. But having capital letters as indices is offensive and this particular index will proliferate. Hence the change. To make things worse, in other chapter the index a was used to denote colour. Not so here.)

We will collectively refer to the component fields $\pi^a(x)$, labelled by $a = 1, \ldots, N_f^2 - 1$ as *pions*, although strictly this terminology is only accurate for $N_f = 2$. Indeed, in the case of $N_f = 2$, we can expand the field π in generators of $SU(2)$ and write

$$\pi = \frac{1}{2}\begin{pmatrix} \pi^0 & \sqrt{2}\pi^- \\ \sqrt{2}\pi^+ & -\pi^0 \end{pmatrix} \; . \tag{3.52}$$

We will later identify the field π^0 with the neutral pion, and $\pi^\pm$ with charged pions. (We'll give the extension to $N_f = 3$, for which the Goldstone bosons are pions, kaons, and a meson called the eta, in Section 3.3.)

We have also introduced a constant f_π in the definition (3.51) with mass dimension $[f_\pi] = 1$. For now, this ensures that the pions have canonical dimensions for scalar fields in four dimensions, $[\pi] = 1$. It is called the *pion decay constant*, although this name makes very little sense purely in the context of QCD because the pions are stable excitations and don't decay. We'll see where the name comes from in Section 5 when we look at the weak force. On general grounds, we expect $f_\pi \sim \Lambda_{\text{QCD}}$. In fact, it is measured to be around $f_\pi \approx 130$ MeV.

The Low-Energy Effective Action

We want to construct a theory that governs the Goldstone bosons U. We will require that our theory is invariant under the full global chiral symmetry $G_F = U(1)_V \times SU(N_f)_L \times SU(N_f)_R$, under which

$$U(x) \to L^\dagger U(x) R \; . \tag{3.53}$$

What kind of terms can we add to the action consistent with this symmetry? The obvious term is $\mathrm{tr}\, U^\dagger U$ but this doesn't work because $U \in SU(N_f)$ and so $\mathrm{tr}\, U^\dagger U = 1$. (Here we've denoted the trace over the N_f flavour indices as tr to distinguish from the trace Tr over colour indices that we used in the action (3.43).) Happily, this is consistent with the fact that U is a massless Goldstone field.

Next, we can look at kinetic terms. At first glance, it looks as if there are three different candidates:

$$(\mathrm{tr}\, U^\dagger \partial_\mu U)^2 \ , \quad \mathrm{tr}\, (\partial^\mu U^\dagger \partial_\mu U) \ , \quad \mathrm{tr}\, (U^\dagger \partial_\mu U)^2 \ . \tag{3.54}$$

The first term in (3.54) vanishes because $U^\dagger \partial U$ is an $su(N)$ generator and, hence, traceless. Furthermore, we can use the fact that $U^\dagger \partial U = -(\partial U^\dagger) U$ to write the third term in terms of the second. This means that there is a unique two-derivative Lagrangian that describes the dynamics of pions,

$$\mathcal{L}_{\rm pion} = \frac{f_\pi^2}{4}\, \mathrm{tr}\, (\partial^\mu U^\dagger\, \partial_\mu U) \ . \tag{3.55}$$

This is the *chiral Lagrangian*. Although the Lagrangian is very simple, this is not a free theory because U is valued in $SU(N_f)$. This is a non-linear sigma model of the kind we met in Section 2.2. Indeed, this is really the original non-linear sigma model, first introduced by Gell-Mann and Lévy in 1960.

We've constructed our sigma-model to have both $SU(N_f)_L \times SU(N_f)_R$, given in (3.53), as symmetries. But because U is valued in $SU(N_f)$, we cannot just set $U = 0$. Indeed, our sigma-model describes a degeneracy of ground states, but in each of them $U \neq 0$. This ensures that the chiral Lagrangian spontaneously breaks the $SU(N_f)_L \times SU(N_f)_R$ symmetry, as it must. The field U itself is the Goldstone boson associated to this symmetry breaking.

Pion Scattering

The beauty of the chiral Lagrangian is that it contains an infinite number of interaction terms, packaged in a simple form by the demands of symmetry. To see these interactions more explictly, we rewrite the chiral Lagrangian in terms of the pion fields defined in (3.51). Keeping only terms quadratic and quartic, the chiral Lagrangian $\mathcal{L}_{\rm pion}$ becomes

$$\mathcal{L}_{\rm pion} = \mathrm{tr}\, (\partial_\mu \pi)^2 - \frac{2}{3f_\pi^2} \mathrm{tr}\, \left(\pi^2 (\partial_\mu \pi)^2 - (\pi \partial_\mu \pi)^2\right) + \ldots \tag{3.56}$$

Note that if we use $\mathrm{tr}\, T^a T^b = \frac{1}{2}\delta^{ab}$ for $su(N_f)$ generators, then the kinetic term has the standard normalisation for each pion field: $\mathrm{tr}\, (\partial_\mu \pi)^2 = \frac{1}{2}\, \partial^\mu \pi^a \partial_\mu \pi^a$.

For concreteness, we work with $N_f = 2$ and take the $su(2)$ generators to be proportional to the Pauli matrices: $T^a = \frac{1}{2}\sigma^a$. The quartic interaction terms then read

$$\mathcal{L}_{\rm int} = -\frac{1}{6f_\pi^2}\left(\pi^a\pi^a\partial\pi^b\partial\pi^b - \pi^a\partial\pi^a\pi^b\partial\pi^b\right) \ . \tag{3.57}$$

From this we can read off the tree-level $\pi\pi \to \pi\pi$ scattering amplitude using the techniques that we described in the Quantum Field Theory lectures. We label the two incoming momenta as p_a and p_b and the two outgoing momenta as p_c and p_d. The amplitude is

$$\begin{aligned} i\mathcal{A}^{abcd} = \frac{i}{6f_\pi^2}\Big[\ & \delta^{ab}\delta^{cd}\Big(4(p_a\cdot p_b + p_c\cdot p_d) + 2(p_a\cdot p_c + p_a\cdot p_d + p_b\cdot p_c + p_b\cdot p_d)\Big) \\ & + (b\leftrightarrow c) + (b\leftrightarrow d)\Big] \ . \end{aligned} \tag{3.58}$$

Momentum conservation, $p_a + p_b = p_c + p_d$, ensures that some of these terms cancel. This is perhaps simplest to see using Mandelstam variables which, because all particles are massless, are defined as

$$\begin{aligned} s &= (p_a+p_b)^2 = 2p_a\cdot p_b = 2p_c\cdot p_d \\ t &= (p_a-p_c)^2 = -2p_a\cdot p_c = -2p_b\cdot p_d \\ u &= (p_a-p_d)^2 = -2p_a\cdot p_d = -2p_b\cdot p_c \ . \end{aligned} \tag{3.59}$$

Using the relation $s+t+u=0$, the amplitude takes the particularly simple form,

$$i\mathcal{A}^{abcd} = \frac{i}{f_\pi^2}\Big[\delta^{ab}\delta^{cd}s + \delta^{ac}\delta^{bd}t + \delta^{ad}\delta^{bc}u\Big] \ . \tag{3.60}$$

There are various ways in which we could improve the description of pion scattering. First, we could include higher loop corrections to the amplitude above. The non-linear sigma model is non-renormalisable which means that we need an infinite number of counterterms to regulate divergences. However, this shouldn't be viewed as any kind of obstacle; the theory is designed only to make sense up to a UV cut-off of order f_π. As long as we restrict our attention to low-energies, the theory is fully predictive.

In addition, we could think about adding higher derivative terms to the chiral Lagrangian. These are corrections that are suppressed by E/f_π where E is the energy of the scattering process. At the next order in the derivative expansion, there are three independent terms:

$$\begin{aligned} \mathcal{L}_4 = a_1 & \left({\rm tr}\,\partial^\mu U^\dagger\,\partial_\mu U\right)^2 + a_2\left({\rm tr}\,\partial_\mu U^\dagger\,\partial_\nu U\right)\left({\rm tr}\,\partial^\mu U^\dagger\,\partial^\nu U\right) \\ & + a_3{\rm tr}\left(\partial_\mu U^\dagger\,\partial^\mu U\partial_\nu U^\dagger\,\partial^\nu U\right) \ . \end{aligned} \tag{3.61}$$

Here a_i are dimensionless coupling constants. There is one further, very important term, known as the Wess-Zumino-Witten (WZW) term that appears at the same order, but can't be written in terms of a 4d action. This is the start of a long and gorgeous story that we won't have time to discuss in these lectures. You can read more about it in the lectures on Gauge Theory.

Currents

We started with quarks and gluons in (3.43) and, at low energies, end up with a very different looking theory of pions (3.55). It's interesting to ask how operators get mapped from one theory to the other. This is particularly straightforward when the operators in question are the currents associated to the $SU(N_f)_L \times SU(N_f)_R$ chiral symmetry.

In the microscopic theory, we have flavour currents for $SU(N_f)_L$ and $SU(N_f)_R$, given by

$$J_L^{a\,\mu} = \bar{q}_{L\,i}\bar{\sigma}^\mu T^a_{ij} q_{L\,j} \quad \text{and} \quad J_R^{a\,\mu} = \bar{q}_{R\,i}\sigma^\mu T^a_{ij} q_{R\,j} \tag{3.62}$$

where T^a_{ij} are $su(N_f)$ generators and the colour and spinor indices have been suppressed. If we write these in terms of the vector and axial combinations: $J_V^{a\,\mu} = J_L^{a\,\mu} + J_R^{a\,\mu}$ and $J_A^{a\,\mu} = J_L^{a\,\mu} - J_R^{a\,\mu}$ then we get the familiar expressions

$$J_V^{a\,\mu} = \bar{q}_i T^a_{ij}\gamma^\mu q_j \quad \text{and} \quad J_A^{a\,\mu} = \bar{q}_i T^a_{ij}\gamma^\mu\gamma^5 q_j \ . \tag{3.63}$$

Now we can ask: what are the analogous expressions for $J_L^{a\,\mu}$ and $J_R^{a\,\mu}$ in the chiral Lagrangian?

To answer this, let's start with $SU(N_f)_L$. Consider the infinitesimal transformation

$$L = e^{i\alpha^a T^a} \approx 1 + i\alpha^a T^a$$

Under this $SU(N_f)_L$, we have $U \to L^\dagger U$ so, infinitesimally,

$$\delta_L U = -i\alpha^a T^a U \ . \tag{3.64}$$

We can now compute the current using the standard trick: elevate $\alpha^a \to \alpha^a(x)$. The Lagrangian is no longer invariant but instead transforms as $\delta\mathcal{L} = \partial_\mu \alpha^a J^a_{L\,\mu}$ and the function $J^a_{L\,\mu}$ is the current that we're looking for. Implementing this, we find

$$J^a_{L\,\mu} = \frac{if_\pi^2}{4}\,\mathrm{tr}\Big(U^\dagger T^a \partial_\mu U - (\partial_\mu U^\dagger) T^a U\Big) \ . \tag{3.65}$$

We can also expand this in pion fields (3.51). To leading order we have simply

$$J^a_{L\,\mu} \approx -\frac{f_\pi}{2}\partial_\mu \pi^a \ . \tag{3.66}$$

Similarly, under $SU(N_f)_R$, we have $\delta U = i\alpha^a U T^a$ and

$$J^a_{R\,\mu} = \frac{if_\pi^2}{4}\Big(- T^a U^\dagger \partial_\mu U + (\partial_\mu U^\dagger) U T^a \Big) \approx +\frac{f_\pi}{2}\partial_\mu \pi^a \ . \tag{3.67}$$

Both currents have non-vanishing matrix elements between the vacuum $|0\rangle$ and a one-particle pion state $|\pi^a(p)\rangle$ that carries momentum p. For example

$$\langle 0| J^a_{L\,\mu}(x) |\pi^b(p)\rangle = -\frac{i}{2} f_\pi \delta^{ab}\, p_\mu e^{-ix\cdot p} \ . \tag{3.68}$$

This tallies with our general discussion of symmetry breaking in (2.2) where we saw that the Goldstone bosons are created by acting with the broken symmetry generators on the vacuum (2.75).

Because the Goldstone bosons are associated to the broken symmetry generators for axial current $J^a_{A_\mu}$, which is a pseudovector, the pions must also be pseudoscalars, meaning that they are odd under parity. We'll look more closely at the quark content of the pions in Section 3.3.

Historically, the approach to thinking of chiral symmetry breaking in terms of currents was known as *current algebra*, and predates our understanding of quarks. The equation (3.68) played a starring role in this story. It is telling us that the chiral $SU(N_f)_L \times SU(N_f)_R$ is spontaneously broken, and acting on the vacuum gives rise to the particles that we call pions. In the language of current algebra, we see that the diagonal combination $SU(N_f)_V$ survives since $\langle 0|J^a_{V\,\mu}|\pi^b\rangle = \langle 0|J^a_{L\,\mu} + J^a_{R\,\mu}|\pi^b\rangle = 0$.

Adding Masses

Our discussion so far has been for massless quarks. That's not particularly realistic. Nonetheless, as we stressed in the introduction to this section, there is reason to expect that the massless limit provides a good jumping off point to understand the physics of light quarks. Our next task is to understand how to incorporate masses.

The QCD action is

$$S = \int d^4x \ \left(-\frac{1}{2}\mathrm{Tr}\, G_{\mu\nu}G^{\mu\nu} + \sum_{i=1}^{N_f} \left(i\bar{q}_i \not{D} q_i - m_i \bar{q}_i q_i \right) \right) \ . \tag{3.69}$$

If the masses are large compared to Λ_{QCD}, then the quarks play no role in the low-energy physics. This is the case for the charm, bottom, and top quarks and we continue to ignore them in what follows.

But for the up, down and (optimistically!) strange quarks, we may assume that the quark condensate (3.47)

$$\langle \bar{q}_{L\,i} q_{R\,j} \rangle \approx -\sigma\, U_{ij} \tag{3.70}$$

continues to form at the scale $\sigma \sim \Lambda^3_{\rm QCD}$, with the masses giving small corrections. We can then incorporate the masses in the chiral Lagrangian by introducing the $N_f \times N_f$ mass matrix,

$$M = \text{diag}(m_1, \ldots, m_{N_f}) \ . \tag{3.71}$$

Because we're now dealing with a low-energy effective theory, the masses that appear here should be the renormalised masses, rather than the bare quark masses quoted earlier in (3.41). In the presence of masses, the leading order chiral Lagrangian is then

$$\mathcal{L}_{\rm pion} = \frac{f_\pi^2}{4}\,\text{tr}\,(\partial^\mu U^\dagger\, \partial_\mu U) + \frac{\sigma}{2}\text{tr}\,\left(MU + U^\dagger M^\dagger\right) \ . \tag{3.72}$$

This lifts the vacuum manifold of the theory. It can be thought of as adding a potential to the vacuum moduli space $\mathcal{M}_0$, resulting in a unique ground state. To see the effect in terms of pion fields, we can again expand $U = e^{2i\pi/f_\pi}$, to find

$$\mathcal{L}_2 = \text{tr}\,(\partial\pi)^2 - \frac{\sigma}{f_\pi^2}\,\text{tr}\Big((M + M^\dagger)\pi^2\Big) + \ldots \tag{3.73}$$

and we see that we get a mass term for the pions as expected. These almost-Goldstone bosons are sometimes referred to as *pseudo-Goldstone bosons*.

For example, if we restrict to $N_f = 2$, we have $M = \text{diag}(m_d, m_u)$. Then, expanding the matrix π in terms of the component fields (3.52),

$$\pi = \frac{1}{2}\begin{pmatrix} \pi^0 & \sqrt{2}\pi^- \\ \sqrt{2}\pi^+ & -\pi^0 \end{pmatrix} \ . \tag{3.74}$$

the quadratic terms in (3.73) become

$$\mathcal{L}_2 = \frac{1}{2}\partial_\mu\pi^0\partial^\mu\pi^0 + \partial_\mu\pi^+\partial^\mu\pi^- - \frac{\sigma}{2f_\pi^2}\,(m_d + m_u)\,((\pi^0)^2 + 2\pi^+\pi^-) \ . \tag{3.75}$$

We see that all three pions get an equal mass, given by

$$m_\pi^2 = \frac{\sigma}{f_\pi^2}(m_u + m_d) \ . \tag{3.76}$$

We learn that the square of the pion mass scales linearly with the quark masses. This is known as the *Gell-Mann-Oakes-Renner relation*. The proportionality constant is the (so-far undetermined) ratio σ/f_π^2.

3.2.3 Phases of Massless QCD

Throughout this section, we've couched our discussion in the broader context of a gauge theory with $G = SU(N_c)$ Yang-Mills, coupled to N_f flavours of massless quarks. Obviously, if our interest is in the real world then we can focus on $N_c = 3$ and $N_f = 2$ or 3, depending on taste. But there's a broader theoretical question that we could ask which is: what is the low-energy physics of the theory with general N_c and N_f?

In this section, we take a quick detour to explain what's known. As we will see, there are a number of open questions.

We start with low N_f:

- When $N_f = 0$, we have pure Yang-Mills. The theory sits in the confining phase, with a mass gap.
- When $N_f = 1$, there is no chiral symmetry group (3.46) and so no chiral symmetry breaking. The theory is again thought to have a mass gap, with quarks bound in mesons and baryons.
- When $2 \leq N_f \leq N^\star$ the theory confines and exhibits chiral symmetry breaking. This means that the low energy theory consists of freely interacting Goldstone bosons, parameterising the moduli space (3.49).

  The big question here is: what is the maximum value $N^\star$ for which chiral symmetry breaking occurs? We don't know the answer to this. Various approaches, including numerics, suggest that it is somewhere around

  $$N^\star \approx 4N_c$$

  This means that, for the $N_f = 2$ or 3 of QCD, we are firmly in the chiral symmetry breaking regime. But, in general, our lack of knowledge of this simple question highlights just how poorly we understand strongly interacting field theories.

Now let's jump to high values of N_f and we'll then try to fill in the details in the middle.

- When $N_f \geq \frac{11}{2}N_c$, the beta function is positive. You can see this from the general expression for the beta function (3.12),

  $$b_0 = \frac{11}{3}N_c - \frac{2}{3}N_f \ . \tag{3.77}$$

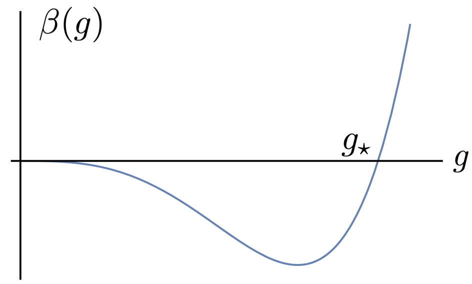


Figure 10. The beta function for N_f slightly below the asymptotic freedom bound has a zero which indicates the existence of an interacting conformal field theory.

This means that the theory is weakly coupled in the infra-red: the low-energy physics consists of massless gluons, weakly interacting with massless quarks. As we go to smaller and smaller energies, the interactions become weaker and weaker. Strictly speaking, in the far IR, the physics is free.

On the flip side, these theories become arbitrarily strongly coupled in the UV, with the gauge coupling diverging at some very high scale. This doesn't mean that we should discard them, but they don't make sense at arbitrarily high energy scales. Said another way, we can't take the UV cut-off Λ_{UV} to infinity while keeping any low-energy interactions. Nonetheless, it's quite possible that these theories may arise as the low-energy limit of some other theory.

That leaves us with the physics in the middle region. We'll keep working down from the asymptotic freedom bound $11N_c/2$.

- When $N^{\star\star} < N_f < \frac{11}{2}N_c$, things are more interesting. To see what happens, we need the two-loop beta function

$$\beta(g) = -\frac{b_0}{(4\pi)^2}g^3 - \frac{b_1}{(4\pi)^4}g^5 + \ldots \tag{3.78}$$

with the one-loop coefficient b_0 given in (3.77) and the two-loop coefficient

$$b_1 = \frac{34N_c^2}{3} - \frac{N_f(N_c^2-1)}{N_c} - \frac{10N_fN_c}{3} \ . \tag{3.79}$$

In the window of interest, $b_0 > 0$ and $b_1 < 0$, so we can play the one-loop contribution against the two-loop contribution to find a zero of the beta function

$$g_\star^2 = -(4\pi)^2\frac{b_0}{b_1} \tag{3.80}$$

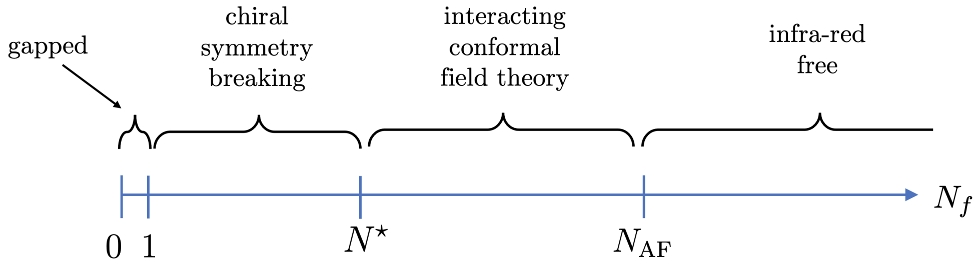


Figure 11. The expected phases of massless QCD. The asymptotic freedom bound is $N_f = \frac{11}{2}N_c$. The lower edge of the conformal window is not known but is expected to be somewhere around $N_f \approx 4N_c$.

with $\beta(g_\star) = 0$. The beta function is shown in Figure 10. The existence of such a fixed point is telling us that we have an interacting conformal field theory: there are massless modes, but they are no longer free in the infra-red. This is known as the *Banks-Zaks fixed point.*

Importantly, when N_f lies just below the asymptotic freedom bound, so $N_f/N_c = 11/2-\epsilon$, this fixed point lies at $g_\star \ll 1$ which means that we can trust the analysis without having to worry about higher order corrections. Moreover, because $g_\star$ is small we can use perturbation theory to calculate anything that we want.

However, as N_f decreases, the value of the fixed point $g_\star$ increases until we can no longer trust the analysis above. The expectation is that we get a conformal field theory only for some range of N_f, lying within $N^{\star\star} < N_f < \frac{11}{2}N_c$. This is known as the *conformal window.* We don't currently know the value of $N^{\star\star}$.

That leaves us with understanding what happens in the middle when $N^\star < N_f \leq N^{\star\star}$. Our best guess is that there is no such regime, and the upper edge of the chiral symmetry breaking phase coincides with the lower edge of the conformal window,

$$N^{\star\star} = N^\star$$

This guess is motivated partly by numerics and partly by a lack of any compelling alternative. For us, the lesson to take away is that strongly interacting quantum field theories are hard and even the most basic questions are beyond our current abilities. A summary of the expected behaviour of massless QCD is shown in Figure 11.

Quark	Charge	Mass (in MeV)
d = down	-1/3	5
u = up	+2/3	2
s = strange	-1/3	93
c = charm	+2/3	1270
b = bottom	-1/3	4200
t = top	+2/3	170,000

Table 3. The quarks

3.3 Hadrons

Confinement means that quarks are bound into colour singlets. There are two group-theoretic possibilities: quark-anti-quark pairs, known as *mesons*, or a collection of three quarks known as *baryons*. Collectively these particles are called *hadrons*[7].

Much of hadron physics is messy and complicated. Some balm comes, once again, from symmetries. Recall that, if we assume that quarks are massless, then the global symmetry exhibits the symmetry breaking pattern

$$U(1)_V \times SU(N_f)_L \times SU(N_f)_R \to U(1)_V \times SU(N_f)_V \ . \tag{3.81}$$

The broken generators give rise to pions and other Goldstone bosons, and we'll see how these arise in terms of quarks shortly. But, for now, our interest lies in the surviving $SU(N_f)_V$ symmetry. This is what we will use to organise the spectrum of hadrons.

We don't need the quarks to be massless to get an $SU(N_f)$ symmetry: we just need their masses to all be equal. Their masses, together with their electric charges, are presented in Table 3.

It seems very reasonable to view $m_{\rm up} \approx m_{\rm down}$, at least to a first approximation. (Remember that we're comparing these values against $\Lambda_{\rm QCD} \approx 200$ MeV.) And, indeed, we will see that there is a clear $SU(2)_V$ symmetry in the hadronic spectrum. This was first identified by Heisenberg, who noted that the proton and neutron have almost identical interactions with the strong force, and is known as *isospin*. (Not a great name as it has nothing to do with "spin".)

[7]I strongly recommend that you take a look, even a brief one, at the booklet published by the Particle Data Group to get a sense for the hadronic world that lies beneath you.

Meanwhile, despite the obvious difference in the strange quark mass, there's also a very visible, albeit approximate, $SU(3)_V$ symmetry in the hadronic spectrum. This was observed, independently, by Gell-Mann and Ne'eman in 1961 and is known as the *eightfold way.* (Because $\dim SU(3) = 8$.) Note that this $SU(3)_V$ has nothing to do with the gauge group $SU(3)$ of QCD. It is an entirely different (and approximate) global $SU(3)_V$ that rotates the different flavours of light quarks.

There are other symmetries of QCD that we can use to assign quantum numbers to particles. These are rotations, corresponding to angular momentum or spin of the particle J, parity, and charge conjugation, both of which are symmetries of QCD, albeit not of the full Standard Model. Particles often come with a label J^{PC}, where $P = \pm$ denotes that the state is even or odd parity and $C = \pm$ denotes even or odd under charge conjugation, which is typically called C-parity in this context.

(As an aside: if you look through the particle data book, you'll sometimes see the additional quantum numbers I^G. Here I is the I_3 eigenvalue of isospin. So for example, particles come in $I = \pm\frac{1}{2}$ pairs if they sit in a double of isospin. Meanwhile G stands for G-parity which is the combination $G = Ce^{i\pi I_2}$ where the isospin rotation is designed to send $I_3 \mapsto -I_3$.)

In the rest of this section, we will describe the hadrons that contain up, down, and strange quarks, and see how they furnish representations of the $SU(3)_V$ flavour symmetry. We then finish by looking at the kinds of particles we can make with heavy charm, bottom, and top quarks.

3.3.1 Mesons

Many hundreds of mesons are observed in nature. A simple model views a meson as a bound state of a quark and an anti-quark, or some linear combination of these states. Each quark is a fermion, so mesons are bosons and, as such, have integer spin. Here we will describe some of the lightest mesons with spin 0 and 1, containing only up, down and strange quarks.

Our three flavours of quarks (d, u, s) transform in the $\mathbf{3}$ of $SU(3)_V$. A little group theory tells us that quark and anti-quark must then transform in

$$\mathbf{3} \otimes \bar{\mathbf{3}} = \mathbf{1} \oplus \mathbf{8} \ . \tag{3.82}$$

So we expect mesons to sit in two representations of $SU(3)_V$: the singlet $\mathbf{1}$ and the adjoint $\mathbf{8}$.

Meson	Quark Content	Mass (in MeV)	Lifetime (in s)
pion π^+	$u\bar{d}$	140	10^{-8}
pion π^0	$\frac{1}{\sqrt{2}}(u\bar{u} - d\bar{d})$	135	10^{-16}
eta η	$\frac{1}{\sqrt{6}}(u\bar{u} + d\bar{d} - 2s\bar{s})$	548	10^{-19}
eta Prime η'	$\frac{1}{\sqrt{3}}(u\bar{u} + d\bar{d} + s\bar{s})$	958	10^{-21}
kaon K^+	$u\bar{s}$	494	10^{-8}
kaon K^0	$d\bar{s}$	498	$10^{-8} - 10^{-11}$

Table 4. The pseudoscalar mesons

Pseudoscalar Mesons

We first look at the lowest mass mesons with spin 0. We get total spin zero if the individual spins of the quarks are anti-aligned, and the particles have zero orbital angular momentum. We saw in Section 1.4 that if a fermion has parity +1 then the anti-fermion has parity −1, which means that the spin 0 meson has odd parity. We write $J^{PC} = 0^{-+}$.

We first give the experimental data for these mesons, and we will then see how they fit into what we know. The names, quark content, masses, and lifetimes of the lightest pseudoscalar mesons are shown in Table 4. The $\pm$ and 0 superscripts tell us the electromagnetic charge of the meson. The charged mesons, π^+ and K^+ both have anti-particles, π^- and K^- respectively. The neutral mesons π^0, η and η' are all their own anti-particles; each is described by a real scalar field. Finally, the neutral K^0 is described by a complex scalar field and its anti-particle is denoted $\bar{K}^0$. This means that there are 9 different meson states in total, in agreement with our simple expectation (3.82).

First, an obvious comment: the masses of the mesons are not equal to the sum of the masses of their constituent quarks! We already anticipated this from our analysis of the chiral Lagrangian and the Gell-Mann-Oakes-Renner relation (3.76). This gets to the heart of what it means to be a strongly coupled quantum field theory. The mesons – and, indeed the baryons – are complicated objects, consisting of a bubbling sea of gluons, quarks and anti-quarks. This is what gives mesons and baryons mass, and also makes these particles hard to understand.

The nine different meson states can be decomposed into the $\mathbf{1} \oplus \mathbf{8}$ multiplets by writing

$$\begin{pmatrix} u \\ d \\ s \end{pmatrix} \otimes (\bar{u}, \bar{d}, \bar{s}) = \begin{pmatrix} u\bar{u} & u\bar{d} & u\bar{s} \\ d\bar{u} & d\bar{d} & d\bar{s} \\ s\bar{u} & s\bar{d} & s\bar{s} \end{pmatrix} = \mu_0 \mathbb{1} + \sum_{a=1}^{8} \mu_a \lambda_a \ . \tag{3.83}$$

Here λ_a are the Gell-Mann matrices (3.6), now in their role as the generators of $SU(3)_V$. We'll ignore the singlet μ_0 for now and focus on the mesons that sit in the $\mathbf{8}$. These are precisely the would-be Goldstone bosons that we met previously. The various fields μ_a naturally rearrange themselves into two real and three complex fields that we call *pions*, *kaons*, and the *eta meson*,

$$\pi^0 = \mu_3 \ , \quad \pi^{\pm} = \frac{1}{\sqrt{2}}(\mu_1 \mp i\mu_2) \tag{3.84}$$

$$K^0 = \frac{1}{\sqrt{2}}(\mu_6 - i\mu_7) \ , \quad K^{\pm} = \frac{1}{\sqrt{2}}(\mu_4 \mp i\mu_5) \ , \quad \eta = \mu_8 \ .$$

The matrix (3.83) is identified with the Goldstone boson matrix that we met in the previous section. We previously wrote this in (3.52) for $N_f = 2$ quarks. The extension to $N_f = 3$ quarks is

$$\pi = \frac{1}{2}\sum_{a=1}^{8} \mu_a \lambda^a = \frac{1}{\sqrt{2}} \begin{pmatrix} \frac{\pi^0}{\sqrt{2}} + \frac{\eta}{\sqrt{6}} & \pi^+ & K^+ \\ \pi^- & -\frac{\pi^0}{\sqrt{2}} + \frac{\eta}{\sqrt{6}} & K^0 \\ K^- & \bar{K}^0 & -\frac{2\eta}{\sqrt{6}} \end{pmatrix} \ . \tag{3.85}$$

You can check that this reproduces the quark content shown in Table 4. If the masses of the three quarks were equal, then these 8 particles would all have the same mass.

The group theoretic underpinnings of these mesons encourages us to draw them on an $SU(3)$ weight diagram, as shown in Figure 12. The charges under the two $U(1)^2 \subset SU(3)_V$ Cartan elements are also shown. These are taken to be isospin $I_3 \subset SU(2)_V \subset SU(3)_V$ and "strangeness" S which effectively counts the number of strange quarks in the meson. A suitable combination, shown on the diagonal, gives the electric charge Q. These are exact quantum numbers in QCD (but not when we include weak interactions) and, historically, it was by observing their conservation in dynamical processes, such as particle decays, that the pattern above was identified.

If we compare pions to kaons, we see from the data that the addition of a strange quark adds about 350 MeV to the mass of a meson. That's significantly more than the

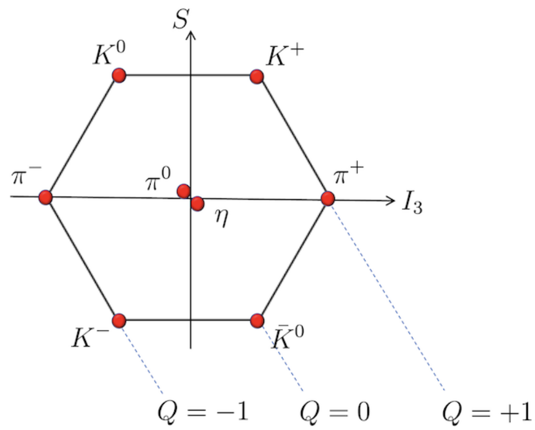


Figure 12. The eightfold way for pseudoscalar (and pseudo-Goldstone) mesons.

bare mass of $\sim$ 100 MeV of a strange quark. Again, this highlights the difficulty of strongly interacting field theories: you don't just read off the physics from the classical Lagrangian.

We can make some progress by looking at the mesons through the lens of the chiral Lagrangian. We return to the massive Lagrangian (3.73), now with the mass matrix $M = \text{diag}(m_u, m_d, m_s)$. Again, I stress that these should be renormalised masses, not bare masses. Expanding out the action using (3.85), we find the masses

$$\mathcal{L}_{\text{mass}} = \frac{-\sigma}{f_\pi^2} \left[\frac{1}{2}(m_u + m_d)\left((\pi^0)^2 + 2\pi^+\pi^-\right) + (m_u + m_s)K^-K^+ \right. \tag{3.86}$$
$$\left. + (m_d + m_s)\bar{K}^0K^0 + \frac{1}{2}\left(\frac{m_u}{3} + \frac{m_d}{3} + \frac{4m_s}{3}\right)\eta^2 + \frac{1}{\sqrt{3}}(m_u - m_d)\pi^0\eta\right] .$$

This generalises our previous result (3.75). Note that there is mixing between π^0 and η, albeit one that disappears when $m_u = m_d$ so that isospin is restored. By taking ratios, we can eliminate the overall scale σ/f_π^2 and relate meson and quark masses directly. For example, we have

$$\frac{m_{K^+}^2 - m_{K^0}^2}{m_\pi^2} = \frac{m_u - m_d}{m_u + m_d} . \tag{3.87}$$

We can also derive expected relationships between the meson masses. For example, we have $3m_\eta^2 + m_\pi^2 = \frac{2\sigma}{f_\pi^2}\left(2(m_u + m_d) + 4m_s\right)$. If we accept that $m_u \approx m_d$, then we get the relation

$$4m_K^2 \approx 3m_\eta^2 + m_\pi^2 . \tag{3.88}$$

This is known as the *Gell-Mann-Okubo relation*. Comparing against the experimentally measured masses, we have $\frac{1}{2}\sqrt{3m_\eta^2 + m_\pi^2} \approx 480$ MeV, which is not far off the measured value of $m_K \approx 495$ MeV.

So far, there is one scalar meson that we've not yet discussed. This is the singlet in the decomposition $\mathbf{3} \otimes \bar{\mathbf{3}} = \mathbf{1} \oplus \mathbf{8}$, associated to the field μ_0 in (3.83). This field corresponds to the meson η', pronounced *eta-prime*,

$$\eta' = \frac{1}{\sqrt{3}}(u\bar{u} + d\bar{d} + s\bar{s}) \ . \tag{3.89}$$

From Table 4, we see that this is by far the heaviest of the scalar mesons. This is because, in contrast to the other mesons, it is not a pseudo-Goldstone boson: if you sent the quark masses to zero, then the pions and kaons and eta all become massless. The eta-prime remains massive.

In fact, there's more to the story of the eta-prime. Recall that back in Section 3.2, we mentioned that the classical Lagrangian of massless QCD also has an axial $U(1)_A$ symmetry. Naively, it appears as if this too is spontaneously broken by the condensate (3.47). If this were true, the eta-prime meson would be the corresponding pseudo-Goldstone boson, in which case we have a puzzle on our hands because it seems too heavy to be Goldstonesque.

The answer to this puzzle will be presented in Section 4 where we'll see that $U(1)_A$, while a symmetry of the classical action, is not a symmetry of the quantum theory because it suffers something called an anomaly. The fact that the eta-prime is inordinately heavy is one consequence of this.

Pseudovector Mesons

This same pattern of $\mathbf{1} \oplus \mathbf{8}$ repeats many more times in excited meson states, in which the spins of the quarks are aligned (rather than anti-aligned) or the quarks have some additional relative orbital angular momentum L. The total parity of these excited meson states is $P = (-1)^{L+1}$.

The first such collection occurs when the spins are aligned, but $L = 0$, giving a collection of 9 pseudovector mesons with $J^{PC} = 1^{--}$, as listed in Table 5. The lightest of these spin 1 mesons are the rhos, $\rho^\pm$ and ρ^0, which can be viewed as excited pions. The heaviest is the phi meson, which is again the singlet $\mathbf{1}$. Note that by the time we get to the excited kaons, some naming exhaustion has set in, and the fact that these are excited states is denoted merely by the addition of a star.

Meson	Quark Content	Mass (in MeV)	Lifetime (in s)
rho ρ^+	$u\bar{d}$	770	10^{-24}
rho ρ^0	$\frac{1}{\sqrt{2}}(u\bar{u} - d\bar{d})$	770	10^{-24}
omega ω	$\frac{1}{\sqrt{2}}(u\bar{u} + d\bar{d})$	780	10^{-22}
phi ϕ	$s\bar{s}$	1020	10^{-22}
kaon $K^{+\star}$	$u\bar{s}$	890	10^{-24}
kaon $K^{0\star}$	$d\bar{s}$	890	10^{-24}

Table 5. The pseudovector mesons

If you look closely at the quark content of the scalar and vector mesons, you'll see that the analogy between them isn't quite perfect. In particular, the excited versions of the η and η' are the ω and ϕ. But the quark content of the pseudoscalar mesons is

$$\eta: \ \frac{1}{\sqrt{6}}(u\bar{u} + d\bar{d} - 2s\bar{s}) \quad \text{and} \quad \eta': \ \frac{1}{\sqrt{3}}(u\bar{u} + d\bar{d} + s\bar{s}) \tag{3.90}$$

while the quark content of the pseudovector mesons is:

$$\omega: \ \frac{1}{\sqrt{2}}(u\bar{u} + d\bar{d}) \quad \text{and} \quad \phi: \ s\bar{s}\ . \tag{3.91}$$

What's going on? Why are these so different?

This is an issue of particle mixing, something that we will see more of when we come to discuss the weak force and neutrinos. First note that the quantum numbers of η and η' are the same (in particular, $I_3 = S = 0$ and hence $Q = 0$ for both). Similarly for the ω and ϕ. In any quantum mechanical system, if you have states with the same quantum numbers then you have to diagonalise the Hamiltonian to find the energy (or in this case, mass) eigenstates. That can lead to linear superpositions of the original states.

That's what's going on here. There are two competing aspects at play. One is the $SU(3)_V$ flavour symmetry that pushes the energy eigenstates to form as $\mathbf{1} \oplus \mathbf{8}$ multiplets, which results in the quark content seen in the pseudoscalars (3.90). The other is the bare mass terms of the quarks, that prefers the energy eigenstates to be the more straightforward $q\bar{q}$. For both pseudoscalar and pseudovector mesons there is some competition between these, meaning that neither (3.90) nor (3.91) is entirely correct. Instead, the honest answer is that the quark content is some linear combination of the two results in both cases, but the group theory dominates for the pseudoscalars, while the mass difference of the strange quark dominates for the pseudovectors.

Of course, this still begs the question of why scalar mesons fall one way, and vectors the other. This is, like many things in QCD, complicated, but it boils down to the fact that the scalar mesons are would-be Goldstone bosons.

Note that masses don't entirely get their own way for the vector mesons. The ρ^0 and ω have constituents $u\bar{u} \pm d\bar{d}$, rather than $u\bar{u}$ and $d\bar{d}$, so the $SU(2)_V$ isospin symmetry is still powerful enough to hold sway over the up/down mass difference.

If you flip through the particle data group booklet, you will find further collections of excitations with $J^{PC} = 0^{++}$ around 1150 MeV. These have orbital angular momentum $L = 1$ and spin $S = 1$ and are given catchy names like a_0, a_1, etc. Then there are states with $J^{PC} = 1^{+-}$ at around 1250 MeV that have $L = 1$ and $S = 0$. These have equally catchy names $b_0, b_1, \ldots$. And so it continues.

3.3.2 Lifetimes

So far we've not said anything about the lifetime of mesons, which we also listed in Tables 4 and 5. This is largely because many of these lifetimes are dictated by the weak force that we haven't yet described. Nonetheless, there are a few straightforward comments that we can make here.

The first is that there is a very wide range of lifetimes exhibited by mesons, from the charged pions and kaons which decay in 10^{-8} seconds to the rho which decays in 10^{-24} seconds. This reflects the different ways in which these particles can decay.

For example, despite their similar masses, the neutral and charged pions have rather different lifetimes. The neutral pion decays through the electromagnetic force to two photons

$$\pi^0 \to \gamma + \gamma \; . \tag{3.92}$$

It has a lifetime of around 10^{-16} seconds. In contrast, the charged pions π^+ and π^- decay only through the weak force. We'll see in Section 5 that they typically decay to a muon and a neutrino

$$\pi^+ \to \mu^+ + \nu_\mu \quad \text{and} \quad \pi^- \to \mu^- + \bar{\nu}_\mu \; . \tag{3.93}$$

They live for 10^{-8} seconds, an eternity in the subatomic world and much longer than any of the other hadrons, except for the proton and neutron.

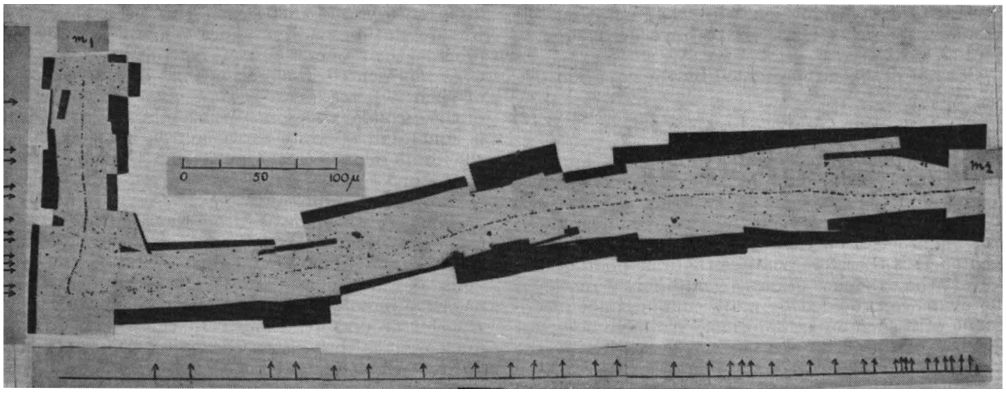


Figure 13. The discovery of the charged pion in 1947. The pion enters in the top left (labelled m_1), slows in the bromide and comes to rest, before decaying into a muon that flies off to the right (labelled m_2) and an anti-neutrino which is invisible in the picture

As a general rule of thumb, each force comes with a characteristic time scale that determines the lifetime of the hadron:

- Strong decay: $\sim 10^{-22}$ to 10^{-24} seconds.
- Electromagnetic decay: $\sim 10^{-16}$ to 10^{-21} seconds.
- Weak decay: $\sim 10^{-7}$ to 10^{-13} seconds.

Where you sit within each range depends on other factors, such as the relative masses of the parent and daughter particles.

In a world with just the strong force, all the pseudoscalar mesons listed in Table 4 would be stable and, despite the fact that some can disappear in 10^{-20} seconds or so, physicists continue to refer to them as stable. In contrast, anything that decays via the strong force is said to be a *resonance*, rather than a particle. All of the vector mesons listed in Table 5 are resonances. For example, the rho decays via the strong force to (predominantly) two pions. If you look through the particle data book, you'll find that resonances are always listed with their mass in brackets. So, for example, you will find $\rho(770)$ in the book but, just above it, η with no brackets.

You'll often find lifetimes quoted in terms of the width, which is an energy scale, rather than a time. The conversion factor is

$$100 \text{ MeV} \approx 10^{-23} \text{ s}^{-1} \ . \tag{3.94}$$

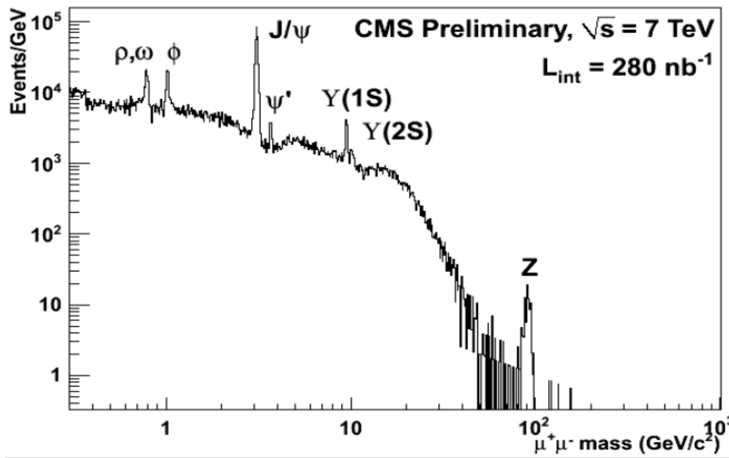


Figure 14. The centre-of-mass energy of $\mu^+\mu^-$ pairs reveals a zoo of mesonic resonances at low energies, with the Z-boson sitting at high energies. This is a plot from 2010 made by the CMS collaboration.

This coincides with what we saw above. The relevant energy scale of the strong force is somewhere around $\Lambda_{\rm QCD} \sim 100$ish MeV and if the strong force does something (like enable a decay), then is typically takes around $T_{\rm QCD} \sim 10^{-23}$ seconds to do it.

Of course, our world has more than the strong force and that means that there's nothing qualitatively different between a particle like the pion and a resonance like the rho. Both will decay in less than the blink of an eye. But it does make a difference for experiments. If something lasts for 10^{-10} seconds then, with good technology, you can take a photograph of the particle's track in a cloud chamber or bubble chamber. For example, the discovery photo of the pion is shown in Figure 13. When a particle leaves such a vivid trace, it's hard to deny its existence. In contrast, we're never going to take a photograph of something that lasts 10^{-20} seconds. But that doesn't mean that it's any less real! It just leaves its signature in more subtle ways, typically as a bump in the cross-section for some process. (See, for example, the chapter on scattering theory in the lectures on Topics in Quantum Mechanics for a discussion of how this comes about.) The glorious plot shown in Figure 14 shows bumps in the number of back-to-back $\mu^+\mu^-$ pairs that were seen in the CMS detector in the early days of the LHC. The resonances start, on the far left, with the ρ, ω and ϕ but then, as the energy increases, there are clear peaks for the J/ψ, which is a charmed meson, the upsilon Υ which is a bottom meson and, far off the right, the Z-boson which is one of the gauge bosons for the weak force.

Finally, hiding within the data are some interesting stories that we will meet again later. For example, the decay of the neutral pion $\pi^0 \to \gamma + \gamma$ is closely tied to the axial anomaly and we will revisit this in Section 4.

The lifetime of the neutral kaons also holds an important lesson. Curiously they appear to have two different lifetimes, either 10^{-7} seconds or 10^{-10} seconds, depending on how you count! That's kind of weird. It turns out to be a manifestation of the fact that the weak force violates time-reversal! We will discuss this in Section 5.

The Elusive Sigma

There is one light scalar meson listed in the particle data book that I have not yet mentioned. It has $J^{PC} = 0^{++}$ and goes by the catchy name of $f_0(500)$ and has a mass which is listed as somewhere between 400 - 550 MeV. The reason that it's so difficult to pin down is that it decays very quickly – via the strong force rather than weak force – to two pions and so has a large width. Moreover, it has vanishing quantum numbers (angular momentum, parity, isospin and strangeness are all zero).

Experimentally, its probably best not to refer to this resonance as a particle at all. However, theoretically it has played a very important role, for this is the "sigma" after which the sigma-model is named. It can be thought of as the excitation that arises from ripples in the value of the quark condensate, $\sigma = \bar{\psi}\psi$, rather than rotations in the quark condensate U.

3.3.3 Baryons

Three quarks can form a gauge singlet by anti-symmetrising over their colour indices $a = 1, 2, 3$ to form a baryon,

$$\mathcal{B} = \epsilon^{abc} q_a q_b q_c \ . \tag{3.95}$$

For baryons constructed of light d, u, and s quarks, these too sit in representations of the $SU(3)_V$ flavour symmetry.

We can again do a little group theory. For two quarks we have

$$\mathbf{3} \otimes \mathbf{3} = \bar{\mathbf{3}} \oplus \mathbf{6} \ . \tag{3.96}$$

Adding the third quark, we have

$$\mathbf{3} \otimes \mathbf{3} \otimes \mathbf{3} = (\bar{\mathbf{3}} \otimes \mathbf{3}) \oplus (\mathbf{6} \otimes \mathbf{3}) = \mathbf{1} \oplus \mathbf{8} \oplus \mathbf{8}' \oplus \mathbf{10} \ . \tag{3.97}$$

Importantly, we want to think of these as representations of the $SU(3)_V$ flavour symmetry rather than the $SU(3)$ gauge symmetry. This tells us that we expect baryons to sit in one of the representations above.

Baryon	Quark Content	Mass (in MeV)	Lifetime (in s)
proton p	uud	938	stable
neutron n	udd	940	10^3
lambda Λ^0	uds	1115	10^{-10}
sigma Σ^+	uus	1189	10^{-10}
sigma Σ^0	uds	1193	10^{-19}
sigma Σ^-	dds	1197	10^{-10}
cascade Ξ^0	uss	1315	10^{-10}
cascade Ξ^-	dss	1321	10^{-10}

Table 6. The octet of spin $\frac{1}{2}$ baryons.

At this point, we have to remember that quarks are fermions and, as such, obey the Pauli exclusion principle. We can look at each of the possibilities above in turn:

- The singlet **1** is fully anti-symmetrised in flavour indices. But any baryon is necessarily fully anti-symmetrised in colour indices, as shown in (3.95), and the Pauli exclusion principle says that the state must be anti-symmetrised overall. We still have the spin degree of freedom to play with, but it's not possible to fully anti-symmetrise in spin so this baryon must have some orbital angular momentum to satisfy Pauli. That makes it heavy and messy. Candidates exist but we won't discuss them.

- At the other end, the decuplet **10** is fully symmetrised in flavour indices and so we can satisfy Pauli by symmetrising over spin degrees of freedom. This means that the decuplet of baryons should have spin $\frac{3}{2}$.

- The **8** and **8′** are a bit more tricky: one is anti-symmetrised only in the first two indices, the other symmetrised in the first two indices, so we have to work a little harder. But it turns out that we can take a suitable linear combination of them that gives a fully anti-symmetrised wavefunction (including colour) when the quarks have total spin $\frac{1}{2}$.

The octet contains the two most famous baryons: protons and neutrons. Collectively, these are called *nucleons*. Others in this multiplet have a mass that differs by about 30% from that of the nucleons. The Σ baryons contain a single strange quark while the Ξ baryons, known either as *xi* or, with a rhetorical flourish, *cascades*, contain two

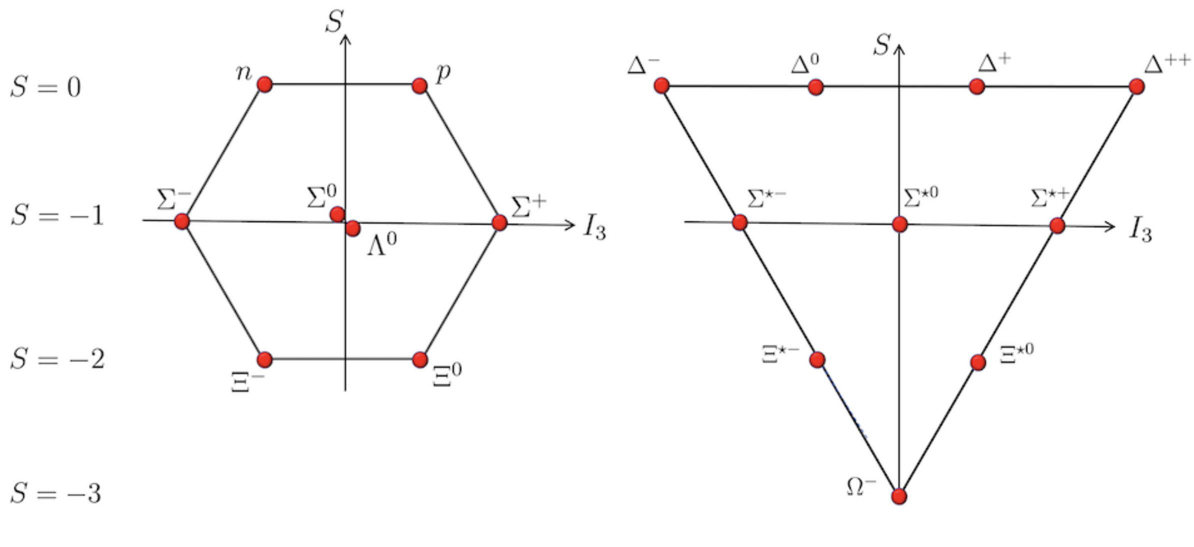


Figure 15. The octet and decuplet of baryons.

strange quarks. The full collection of eight spin $\frac{1}{2}$ baryons are shown in Table 6, and in an $SU(3)$ weight diagram, reflecting their group theoretic origins, in Figure 15.

We saw previously that the octet of pseudoscalar mesons have an interpretation as almost-Goldstone modes. That means, in particular, that if the quarks were massless, then the pions, kaons and eta would all be massless as well. What is the analogous story for the baryons?

Here there is a surprise. If the up and down quark were massless, the mass of the proton and neutron would be more or less unchanged from the values we measure! The mass of the baryons – at least those comprised of light quarks – is not driven by the bare quark mass. Instead, it's driven by the strong coupling scale $\Lambda_{\rm QCD}$. In fact, on general grounds one can argue that the mass of baryons in $SU(N_c)$ QCD scales as $N_c\Lambda_{QCD}$.

That's not to say that the mass of the quarks is entirely unimportant. Crucially, the fact that the down quark is heavier than the up quark is the reason why the neutron is heavier than the proton. If this weren't true, the weak force would allow the proton to decay into the neutron, rather than the other way around, and it's hard to see how atoms and chemistry and physicists could exist.

Similarly, the strange baryons are heavier than the proton and neutron. You can see from the data that each strange quark adds about 140±10 MeV to the baryon mass.

That's smaller than the corresponding amount for mesons, but still bigger than the bare mass $m_s \approx 93$ MeV.

You may have heard it said that the Higgs is responsible for all the mass in the universe. This is a blatant lie. In Section 5, we will see that the Higgs is responsible for the mass of all *elementary particles*, meaning the leptons and quarks. But the overwhelming majority of mass in atoms is contained in the protons and neutrons that make up the nucleus, and this mass has nothing to do with the Higgs boson. It is entirely due to the urgent thrashing of strongly interacting quantum fields.

While we're talking about fairytales that we were subjected to when we were young, here's another one: we are usually told that the strong force is what keeps the nucleus together in the atom. This one is kind of true, but only in an indirect way. The strong force binds quarks together into baryons, which are fermions, and into mesons, which are bosons. But, as described in the lectures on Quantum Field Theory, scalar particles mediate forces. In particular, the pions mediate a force of a Yukawa type, with potential

$$V(r) \sim -\frac{e^{-m_\pi r}}{r} \; . \tag{3.98}$$

This is what binds the protons and neutrons together in the nucleus.

We refer to this force mediated by pions as the strong nuclear force, but it would be better to give it a different name — say "mesonic force", or "Yukawa force" — to highlight the fact that it is really a residual, secondary effect. The upshot is that there are two layers to the strong force: we start with one force and a set of matter particles — gluons interacting with quarks — and end up with a very different force and a new set of matter particles — the mesonic force interacting with protons and neutrons. In this sense, both the particles in the nucleus, and the force that holds them together, are *emergent* phenomena, arising from something more fundamental underneath.

Finally, we briefly look at the spin $\frac{3}{2}$ baryons, that sit in the flavour decuplet. They go by the names Δ (with charges 0, ± 1 and 2), $\Sigma^\star$ (with charges 0 and ± 1), $\Xi^\star$ (with charges -1 and 0) and Ω^- with charge -1. The full list of particles is given in Table 7 and the weight diagram shown in Figure 15.

The real novelties among these baryons are the three outliers, in which all quarks are the same. The Δ^{++} played an important historic role because it was the first particle to be found with charge $+2$ as opposed to 0 or ± 1 and helped enormously in piecing

Baryon	Quark Content	Mass (in MeV)	Lifetime (in s)
Δ^{++}	uuu	1232	10^{-24}
Δ^{+}	uud	1232	10^{-24}
Δ^{0}	udd	1232	10^{-24}
Δ^{-}	ddd	1232	10^{-24}
$\Sigma^{\star -}$	dds	1383	10^{-23}
$\Sigma^{\star 0}$	dus	1384	10^{-23}
$\Sigma^{\star +}$	uus	1387	10^{-23}
$\Xi^{\star -}$	dss	1535	10^{-23}
$\Xi^{\star 0}$	uss	1532	10^{-23}
Ω^{-}	sss	1672	10^{-11}

Table 7. The decuplet of spin $\frac{3}{2}$ baryons.

together the story of the underlying quarks. The Ω^- baryon, meanwhile, holds a special place in the history of science because Gell-Mann used the simple quark model described above to predict its mass and properties before it was discovered experimentally. In that way, he followed Mendeleev and Dirac in predicting the existence of a "fundamental" particle of nature (where, as should by now be clear, the meaning of the word "fundamental" is time-dependent).

One of the lessons to take away from this section is that QCD is complicated. We can make some progress by using symmetries (or approximate symmetries) as organising principles, but that only takes us so far. It is natural to wonder how much of the results above we can calculate from first principles, starting from the Lagrangian of QCD.

If your first principles involve only pen and paper, then the answer is: not much. QCD is hard. But if you extend your first principles to embrace numerical simulations which, in this context, go by the name of lattice QCD, then you can do pretty well. After many decades of work, much of the spectrum described above can be computed to within, say, 5% accuracy. There is now no doubt that the complexity seen in the hadron spectrum can be entirely explained by the dynamics of QCD.

3.3.4 Heavy Quarks

So far, we've only discussed the hadrons constructed from the three lightest quarks. We've still to discuss the heavy ones.

It turns out that there are no hadrons comprised of the top quark. Its extreme high mass means that the top quark decays with a lifetime of around 10^{-25} seconds, which is faster than the characteristic timescale $T_{\rm QCD} \approx 10^{-23}$ seconds of the strong force. This means that such "top hadrons" decay before they even form. Needless to say, none have been observed.

That still leaves us with the charm and bottom. The masses of hadrons containing these quarks are determined more by the bare quark mass than by $\Lambda_{\rm QCD}$. Two sets of these mesons deserve a special mention. The first is *charmonium*, a bound state of charm and anti-charm quark. It also goes by the dual name J-psi (J/ψ),

$$J/\psi\ (\bar{c}c) \quad m \approx 3.1 \text{ GeV}\ . \tag{3.99}$$

Its lifetime is around 10^{-21} seconds. The discovery of this particle in 1974, showing up as a very sharp resonance similar to what is seen in Figure 14, was the first glimpse of the charm quark and played a key role in cementing the Standard Model.

There are a collection of lighter mesons that contain just a single charm quark. These are called (somewhat peculiarly) *D-mesons*. The lightest are:

$$\begin{aligned} D^0\ (c\bar{u}) \qquad & m \approx 1865 \text{ MeV} \\ D^+\ (c\bar{d}) \qquad & m \approx 1869 \text{ MeV}\ . \end{aligned} \tag{3.100}$$

These are remarkably long lived particles, with the D^+ surviving a whopping 10^{-12} seconds, and the D^0 about half this time. The long lifetime is because these particles decay only through a somewhat subtle property of the weak force. We will learn more about this in Section 5.

Similarly, the bottom quark was first discovered in *bottomonium*, also known as the upsilon (Υ)

$$\Upsilon\ (\bar{b}b) \quad m \approx 9.5 \text{ GeV}\ . \tag{3.101}$$

This has a lifetime of 10^{-20} seconds. Once again, it is neither the lightest nor the longest lived meson containing a b-quark. The lightest *B-mesons* are

$$B^+\ (u\bar{b}) \text{ and } B^0\ (d\bar{b}) \quad m \approx 5280 \text{ MeV}\ . \tag{3.102}$$

Despite being significantly heavier, they actually live (very) slightly longer than the D-mesons, with a lifetime of around 1.5×10^{-12} seconds. It's worth stressing how astonishing this is: the ratio of the mass to the width of the B-meson is $m_B/\Gamma_B \sim 10^{13}$ You can compare this to the common or garden mesons, like the ρ, which has $m_\rho/\Gamma_\rho \sim$ 4!. Again, this is down to intricacies of the weak force.

A small comment on terminology. The third generation of quarks was originally termed *beauty* and *truth*. (What can I say? It was the 70s.) Eventually, out of a due sense of embarrassment, these names were phased out in preference for the more boring "bottom" and "top". This has persisted for the top quark, but the term "beauty" lingers. For example, the important experiment LHCb which investigates B-mesons, prefers to be thought of, for obvious reasons, as focussing on the study of beauty, rather than the study of bottoms.

There are also baryons containing charm and bottom quarks. Here the names become increasingly unimaginative, with subscripts c and b denoting the quark content. For example, in addition to the Σ^+, comprised of uus, there is also a Σ_c^+ comprised of uuc and Σ_b^+ comprised of uub, and similar stories for cascades. There are also excited states of all these baryons, in which the quarks orbit each other, not dissimilar to the way in which the electrons orbit the proton in the excited states of the hydrogen atom.

3.4 The Theta Term

For QCD, we've seen that the action is gloriously simple:

$$S = \int d^4x \left(-\frac{1}{2} \mathrm{Tr}\, G_{\mu\nu} G^{\mu\nu} + \sum_{i=1}^{N_f} \left(i\bar{q}_i \not{D} q_i - m_i \bar{q}_i q_i \right) \right) \ . \tag{3.103}$$

The question that we would like to pose is: are there any other interaction terms that we could write down that we've missed?

The answer is that there is one, but that it's rather subtle. This is known as the Yang-Mills *theta term*,

$$S_\theta = \frac{\theta g_s^2}{16\pi^2} \int d^4x \ \mathrm{Tr}\, G_{\mu\nu} {}^\star G^{\mu\nu} \tag{3.104}$$

where ${}^\star G^{\mu\nu} = \frac{1}{2}\epsilon^{\mu\nu\rho\sigma} G_{\rho\sigma}$. Here θ is the eponymous theta angle, and should be viewed as an additional parameter of QCD.

Before we get to the theory underlying the theta term, let me first give some commentary on why we haven't mentioned this term until now. The reason is that, as far as we can tell from experiment, the theta parameter takes the value $\theta = 0$. Said more precisely, we can bound the theta parameter to be

$$\theta < 10^{-10} \ . \tag{3.105}$$

So why should we care about something that doesn't exist? The reason is that zero is a number too! The game that we play in the Standard Model is the same as for all other quantum field theories: after you've figured out what fields you're dealing with, you then write down all possible relevant and marginal interactions that could change the low energy physics. Each of these terms typically comes with a parameter that we have to determine by experiment. These parameters are things like the masses of particles (or, more precisely, Yukawa couplings as we'll see in Section 5.) Out of all these parameters, θ is special because it's the only one that appears to vanish. And that's crying out for an explanation.

What would the consequences be if θ were not to vanish? The answer is pretty dramatic because, in contrast to all other terms in the QCD action (3.103), the theta term violates various discrete symmetries. Written in terms of the chromoelectric and chromomagnetic fields, it takes the form

$$G_{\mu\nu}{}^{\star}G^{\mu\nu} \sim \mathbf{E}\cdot\mathbf{B}\ . \tag{3.106}$$

We've seen in Section 1.4 that, under parity P, charge conjugation C, and time reversal T, the electric and magnetic fields transform as

$$\begin{aligned} P:\mathbf{E}\mapsto -\mathbf{E} \quad &\text{and} \quad P:\mathbf{B}\mapsto +\mathbf{B} \\ C:\mathbf{E}\mapsto -\mathbf{E} \quad &\text{and} \quad C:\mathbf{B}\mapsto -\mathbf{B} \\ T:\mathbf{E}\mapsto +\mathbf{E} \quad &\text{and} \quad T:\mathbf{B}\mapsto -\mathbf{B}\ . \end{aligned} \tag{3.107}$$

This means that the theta term breaks both P and CP or, equivalently, T. As we saw previously, a consequence of CP violation is that particles are endowed with an electric dipole moment. The most precise experimental tests are for the neutron which, experimentally, is found to have an electric dipole moment d_n bounded by

$$d_n < 10^{-26}\ \text{e}\,\text{cm}\ . \tag{3.108}$$

This, ultimately, translates into the bound (3.105). (For what it's worth, the CP violation in the weak sector is predicted to give the neutron a dipole moment around $d_n \approx 10^{-30}$ e cm, somewhat below current experimental bounds.)

So why do we have $\theta = 0$? The answer is: we don't know. One might want to state by fiat that QCD should be invariant under P and CP and that's why the theta term is disallowed. That's a reasonable argument in the context of stand-alone QCD, but not when viewed within the broader framework of the Standard Model which, as we will see, is invariant under neither P nor CP. (Indeed, the fuller story is that the QCD theta term is infected by various other terms in the Standard Model Lagrangian and somehow they collectively conspire to ensure that $\theta = 0$.) The question of why $\theta = 0$ is known as the *strong CP problem*. It is surely one of the most important clues for what lies beyond the Standard Model.

3.4.1 Topological Sectors

The theta term is also special for other reasons. Indeed, of all the terms that we could write down in the Standard Model, it is by far the most subtle. In this sense, it's something of a shame that it vanishes!

We can discuss the physics for a general gauge group G, rather than restricting to QCD and, for that reason, we will revert to the notation of Section 1.3 and refer to the Yang-Mills gauge field as A_μ and the field strength as $F_{\mu\nu}$ (rather than G_μ and $G_{\mu\nu}$ for QCD).

The first important property of the theta term is that it's a total derivative. You can show that

$$S_\theta = \frac{\theta g_s^2}{8\pi^2}\int d^4x\ \partial_\mu K^\mu \quad \text{with} \quad K^\mu = \epsilon^{\mu\nu\rho\sigma}\text{Tr}\left(A_\nu \partial_\rho A_\sigma - \frac{2i}{3}A_\nu A_\rho A_\sigma\right)\ . \tag{3.109}$$

This means that it does not affect the classical equations of motion. Nonetheless, it can affect the quantum dynamics of gauge theories. This arises because the path integral receives contributions from field configurations that have something interesting going on at infinity so that the boundary term S_θ is non-vanishing. This something interesting can be found in the topology of the gauge group.

To explain this, we first Wick rotate so that we work in Euclidean spacetime $\mathbb{R}^4$. Configurations that have a finite action from the Yang-Mills term must asymptote to pure gauge,

$$A_\mu \to \frac{i}{g}\Omega\partial_\mu\Omega^{-1} \quad \text{as } x \to \infty \tag{3.110}$$

with $\Omega \in G$. This means that finite action, Euclidean field configurations involve a map

$$\Omega(x) : \mathbf{S}^3_\infty \mapsto G\ . \tag{3.111}$$

with $\mathbf{S}^3_\infty = \partial\mathbb{R}^4$ the asymptotic boundary of $\mathbb{R}^4$. Maps of this kind fall into disjoint classes. These arise because the gauge transformations can "wind" around the spatial $\mathbf{S}^3$ in such a way that one gauge transformation cannot be continuously transformed into another. Such winding is characterised by *homotopy theory*. In the present case, the maps are labelled by an element of the homotopy group which, for all simple, compact Lie groups G, is given by

$$\Pi_3(G) = \mathbb{Z} . \tag{3.112}$$

This means that the winding of gauge transformations (3.110) at infinity is classified by an integer n.

This statement is most intuitive for $G = SU(2)$ since, viewed as a manifold, $SU(2) \cong \mathbf{S}^3$ and the homotopy group counts the winding from one $\mathbf{S}^3$ to another. For higher dimensional groups, including $G = SU(3)$ relevant for QCD, it turns out that it's sufficient to pick an $SU(2)$ subgroup of G and consider maps which wind within that. You then need to check that these maps cannot be unwound within the larger G.

It can be shown that, in general, the winding $n \in \mathbb{Z}$ is computed by

$$n(\Omega) = \frac{1}{24\pi^2} \int_{\mathbf{S}^3_\infty} d^3S \; \epsilon^{ijk} \mathrm{Tr}\,(\Omega\partial_i\Omega^{-1})(\Omega\partial_j\Omega^{-1})(\Omega\partial_k\Omega^{-1}) \; . \tag{3.113}$$

Evaluated on any configuration that asymptotes to (3.110), the theta term gives

$$S_\theta = \theta n \quad \text{with} \quad n \in \mathbb{Z} \; . \tag{3.114}$$

It is the contribution from configurations with $n \neq 0$ in the path integral that means that observables in quantum gauge theories can depend on θ. In general, all observables are thought to depend on the value of θ. For example, it's expected that the masses of particles in Yang-Mills theory, or indeed, in QCD, depend on θ. (The "expected" in that sentence is because it's very hard to know for sure, largely because it's very difficult to do numerical simulations of these theories when $\theta \neq 0$.)

When exponentiated in the path integral, the theta term contributes to the Euclidean action as $e^{iS_\theta} = e^{i\theta n}$. Importantly, it is a complex phase. The fact that it is complex can be traced to the $\epsilon^{\mu\nu\rho\sigma}$ tensor in S_θ. This means that S_θ contains a single time derivative and so, upon Wick rotation, still sits in the path integral as e^{iS_θ} rather than e^{-S_θ}. The fact that $n \in \mathbb{Z}$ means that θ is a periodic variable, with

$$\theta \in [0, 2\pi) \; . \tag{3.115}$$

For this reason, it's often called the *theta angle*. We see that the role of the theta term is to weight different topological sectors in the path integral with different phases $e^{i\theta n}$.

3.4.2 Instantons

We can say more if we work in a regime in which the theory is weakly coupled. Here the path integral is dominated by the saddle points, which are solutions to the classical equations of motion. This means that any θ dependence should come from field equations that wind at infinity, so $n \neq 0$, and solve the classical equations of motion,

$$\mathcal{D}_\mu F^{\mu\nu} = 0 \ . \tag{3.116}$$

There is a cute way of finding solutions to this equation. The Yang-Mills action is

$$S_{YM} = \frac{1}{2g^2} \int d^4x \ \mathrm{Tr}\, F_{\mu\nu} F^{\mu\nu} \ . \tag{3.117}$$

Note that in Euclidean space, the action comes with a + sign. (This is to be contrasted with the Minkowski space action which comes with a minus sign.) We can write the Euclidean action by completing the square,

$$S_{YM} = \frac{1}{4g^2} \int d^4x \ \mathrm{Tr}\, (F_{\mu\nu} \mp {}^\star F_{\mu\nu})^2 \pm \frac{1}{2g^2} \int d^4x \ \mathrm{Tr}\, F_{\mu\nu} {}^\star F^{\mu\nu} \geq \frac{8\pi^2}{g^2} |n| \ . \tag{3.118}$$

where, in the last inequality, we've used the result (3.114). We learn that in the sector with winding n, the Yang-Mills action is bounded by $8\pi^2 n/g^2$. The action is minimised when the bound is saturated. This occurs when

$$F_{\mu\nu} = \pm^\star F_{\mu\nu} \ . \tag{3.119}$$

These are the (anti)-self-dual Yang-Mills equations. The argument above shows that solutions to these first order (anti)-self-dual equations necessarily minimise the action in a given topological sector and so must solve the equations of motion (3.116). In fact, it's straightforward to see that this is the case since it follows immediately from the Bianchi identity $\mathcal{D}_\mu {}^\star F^{\mu\nu} = 0$.

Solutions to the (anti)-self-dual Yang-Mills equations (3.119) have finite action, which means that any deviation from the vacuum must occur only in localised regions of Euclidean spacetime. In other words, these solutions correspond to point-like objects in Euclidean spacetime $\mathbb{R}^4$. Because they occur for just an "instant of time" they are known as *instantons*. They are very much analogous to the classical tunnelling solutions for the quantum mechanical double well potential that we met in Section 2.1.

There is much to say about instantons. You can read about the role they play in quantum Yang-Mills in the lectures on Gauge Theory and more about the structure of the solutions to (3.119) in the lectures on Solitons. For our purposes, it will suffice to point out that the contributions of instantons to any quantity comes with the characteristic factor

$$e^{-S_{\rm instanton}} = e^{-8\pi^2|n|/g^2} e^{i\theta n} \ . \tag{3.120}$$

Famously, the function $e^{-8\pi^2/g^2}$ has vanishing Taylor expansion about the origin $g^2 = 0$. This is telling us that effects due to instantons are smaller than any perturbative contribution, which takes the form g^{2n}. Nonetheless, that doesn't mean that instantons are useless since they can contribute to quantities that apparently vanish in perturbation theory.

Instantons are usually referred to as *non-perturbative* effects. This is a little bit of a misnomer. The use of instantons requires weak coupling $g^2 \ll 1$, so in this sense they are just as perturbative as usual perturbation theory. The name *non-perturbative* really means "not perturbative around the vacuum". Instead, the perturbation theory occurs around the instanton solution.

An Example: An Instanton in $SU(2)$

It is fairly straightforward to write down the instanton solutions with winding $n = 1$. For $SU(2)$, such a configuration is given by

$$A_\mu = \frac{1}{x^2+\rho^2}\, \eta^a_{\mu\nu} x^\nu \sigma^a \tag{3.121}$$

Here ρ is a parameter whose role we will describe shortly. The $\eta^a_{\mu\nu}$ are usually referred to as 't Hooft matrices. They are three 4×4 matrices which provide an irreducible representation of the $su(2)$ Lie algebra. They are given by

$$\eta^1_{\mu\nu} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{pmatrix} \ , \ \ \eta^2_{\mu\nu} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \ , \ \ \eta^3_{\mu\nu} = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{pmatrix} \ . \tag{3.122}$$

These matrices are self-dual: they obey $\frac{1}{2}\epsilon_{\mu\nu\rho\sigma}\eta^i_{\rho\sigma} = \eta^i_{\mu\nu}$. (Note that we're not being careful about indices up vs down as we are in Euclidean space with no troublesome minus signs.) In the solution (3.121), the 't Hooft matrices intertwine the $su(2)$ group index $a = 1, 2, 3$ with the spacetime index μ and this implements the asymptotic winding of the gauge fields.

The associated field strength is given by

$$F_{\mu\nu} = -\frac{2\rho^2}{(x^2+\rho^2)^2}\eta^a_{\mu\nu}\sigma^a \ . \tag{3.123}$$

This inherits its self-duality from the 't Hooft matrices: $F_{\mu\nu} = {}^\star F_{\mu\nu}$ and therefore solves the Yang-Mills equations of motion, $\mathcal{D}_\mu F_{\mu\nu} = 0$.

We can get some sense of the form of this solution. First, the non-zero field strength is localised around the origin $x = 0$. (By translational invariance, we can shift $x^\mu \to x^\mu - X^\mu$ to construct a solution localised at any other point X^μ.) The solution depends on a parameter ρ which can be thought of as the size of the instanton lump. The fact that the instanton has an arbitrary size follows from the classical conformal invariance of the Yang-Mills action.

4 Anomalies

Our goal in this section is to understand the beautiful and subtle phenomenon known as an *anomaly*[8]. This is one of the deepest ideas in quantum field theory and, as we will see in Section 5, underpins much of the structure of the Standard Model.

Before we jump in, here are two motivating comments.

We already met the theories of QED and QCD in the previous section. Both are described by Lagrangians in which a gauge field is coupled to a bunch of Dirac fermions. But Dirac fermions are not the simplest kind of fermion. Or, said differently, Dirac fermions are not irreducible representations of the Lorentz group. Instead, a Dirac fermion decomposes into two Weyl fermions. So why doesn't nature make use of this more minimal Weyl fermion? And why don't we study the seemingly simpler theory of, say, Yang-Mills coupled to a single Weyl fermion?

The answer, it turns out, is that Yang-Mills coupled to a single Weyl fermion is an inconsistent quantum theory! This is an important and striking statement. There's no problem in writing down a classical Lagrangian, nor indeed a classical Hamiltonian, for this system. But there's no corresponding quantum theory. As we will explain, this is one manifestation of the anomaly.

Here's a second motivation. In the theory of massless QCD, we mentioned that there is a classical $U(1)_A$ axial symmetry which, naively, appears to be spontaneously broken like the non-Abelian chiral symmetry. But there is no associated light meson. The meson that carries the right quantum numbers is the η' and its mass is almost 1 GeV, significantly more than the other pseudo-Goldstone bosons. What's going on?

The answer, it turns out, is that the axial $U(1)_A$ symmetry in massless QED and QCD is a good symmetry of the classical theory, but it is not a symmetry of the quantum theory. This, too, is a manifestation of the anomaly.

Our purpose is to understand these statements and more. There are various ways to understand these features, but the most revealing is through the path integral. As we will see, both of the issues above, and several others, arise from trying to carefully define the path integral for Weyl fermions.

[8]Because these are lectures on the Standard Model, I should mention that there is another, very different meaning to the word "anomaly" in the particle physics community, which is when an experimental result that deviates slightly from the prediction of the Standard Model. Typically, this leads to approximately 10^4 papers being written before the whole things fades away 3 years later. That's not what we're talking about here.

Our First Anomaly

There are a number of different manifestations of anomalies in quantum field theory. Indeed, understanding when such effects arise remains a vibrant research area. Here we will discuss just the simplest kind of anomaly, associated to Weyl fermions.

To set the scene, recall that a Dirac fermion ψ splits into two Weyl fermions

$$\psi = \begin{pmatrix} \psi_L \\ \psi_R \end{pmatrix} \ . \tag{4.1}$$

For our story, we want to take just a single Weyl fermion. We will take a left-handed spinor ψ_L, but everything we're about to say also holds for a single right-handed spinor. The action for a massless Weyl spinor is

$$S = \int d^4x \ i\bar{\psi}_L \bar{\sigma}^\mu \partial_\mu \psi_L \tag{4.2}$$

with $\bar{\sigma}^\mu = (\mathbb{1}, -\sigma^i)$. This action is clearly invariant under the $U(1)$ global symmetry $\psi_L \rightarrow e^{i\alpha}\psi_L$, with the corresponding current $j^\mu = \psi_L^\dagger \bar{\sigma}^\mu \psi_L$. To illustrate the anomaly, we will couple this current to a gauge field A_μ with charge $q \in \mathbb{Z}$. The action is now

$$S = \int d^4x \ i\bar{\psi}_L \bar{\sigma}^\mu \mathcal{D}_\mu \psi_L \tag{4.3}$$

where the covariant derivative contains the coupling to the gauge field $\mathcal{D}_\mu \psi_L = \partial_\mu \psi_L - ieqA_\mu \psi_L$. This action is now invariant under the gauge symmetry

$$\psi \rightarrow e^{ieq\alpha(x)}\psi \quad \text{and} \quad A_\mu \rightarrow A_\mu + \partial_\mu \alpha \ . \tag{4.4}$$

Before we proceed, I should mention that there are two distinct ways to think about the gauge field A_μ and this distinction will be important when we come to look at the various implications of anomalies. They are:

- A_μ could be a *dynamical gauge field*. In the classical theory, this means that we treat it as a dynamical variable, with its own equation of motion, typically after adding a Maxwell term to the action. In the quantum theory, it means that we integrate over A_μ in the path integral.
- A_μ could be a *background gauge field*. This means that it is something fixed, under our control, and should be viewed as a parameter of the theory. Turning it on typically breaks Lorentz symmetry, but could be useful to explore how our system responds to the presence of an electric or magnetic field. In the quantum theory, A_μ appears as a source on which the partition function depends.

We will consider gauge fields of both types in what follows. However, for now, we will consider A_μ to be a background gauge field, whose value is something that we get to decide.

While the classical theory is clearly invariant under the gauge transformation (4.4), the question that we really want to ask is: what happens in the quantum theory? For this, we should turn to the path integral, with the partition function in Euclidean space defined as

$$Z[A] = \int D\psi_L D\bar{\psi}_L \ \exp\left(-\int d^4x \ i\bar{\psi}_L \bar{\sigma}^\mu \mathcal{D}_\mu \psi_L\right) \ . \tag{4.5}$$

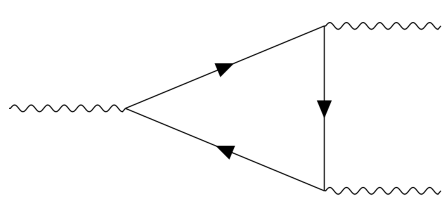
The action in the exponent is designed so that it is invariant under gauge transformations. But now we must also worry about the measure in the path integral and this takes some care to define. The statement of the anomaly is that the measure is *not* invariant under gauge transformations. Instead, it turns out that the measure, and hence the partition function, changes by a phase

$$Z[A] \rightarrow \exp\left(\frac{ie^3q^3}{32\pi^2}\int d^4x \ \alpha F_{\mu\nu}{}^\star F^{\mu\nu}\right) \ Z[A] \tag{4.6}$$

with ${}^\star F^{\mu\nu} = \frac{1}{2}\epsilon^{\mu\nu\rho\sigma}F_{\rho\sigma}$.

This subtlety only happens for fermions. If we have scalar fields charged under a symmetry, then the measure is perfectly invariant. At heart, this is related to the fact that there is no difficulty in giving masses to scalar fields while preserving symmetries, but giving masses for fermions necessarily breaks certain symmetries.

I won't prove the anomaly (4.6) here, but a detailed derivation is given in the lectures on Gauge Theory. In fact, there are two such derivations. The first involves a careful definition of the measure in the path integral to see that it does indeed transform as (4.6). The second derivation works with more conventional perturbation theory. In particular, the anomaly is associated to the following triangle diagram



The external legs are currents associated to the $U(1)$ symmetry, while the fermion runs in the loop. Like most one-loop diagrams, the resulting integral is divergent and has to be regulated. The subtlety arises because of the interplay between regulating the divergence and preserving the $U(1)$ symmetry. It turns out that only diagrams of this kind suffer from this subtlety, and the fact that there are three legs is reflected in the q^3 prefactor of the anomaly in (4.6). Although we won't compute these triangle diagrams here, they will be a useful mnemonic as we describe different kinds of anomalies.

Rather than derive the anomaly, we will instead focus on its implications. Broadly, there are three different implications, depending on whether we think of the gauge field A_μ as background or dynamical. We will address these in turn in Sections 4.1, 4.2, and 4.3.

4.1 Gauge Anomalies

The first implication of the anomaly (4.6) is that it is an obstruction to gauging. Although the action is invariant under the gauge symmetry, the measure is not and neither is the partition function. That means that we cannot promote the gauge field A_μ to a dynamical field, where we integrate over it in the path integral. If we attempted to do this, we would get a sick theory. (Sick as in bad, not sick as in good.)

There are a number of ways to see why the theory is sick but here is a simple one. Recall that when we first attempted to quantise the gauge field A_μ in the lectures on Quantum Field Theory we had some work to do to decouple the negative norm states that arise from quantising A_0. That work ultimately boiled down to using the gauge invariance to remove these states. But in an anomalous theory, we no longer have that gauge invariance at our disposal and the Hilbert space will involve negative norm states. That's bad.

The upshot is that a $U(1)$ gauge theory, coupled to a single Weyl fermion, is a sick theory. If we want to write down a consistent gauge theory, then we must have multiple Weyl fermions so that, combined, the anomaly cancels.

Typically, we think of a given theory in terms of a bunch of left-handed fermions and another bunch of right-handed fermions. But, given a right-handed fermion of charge q, its complex conjugation is a left-handed fermion of charge $-q$. So, we're always at liberty to talk only about left-handed fermions. If we have a bunch of left-handed Weyl fermions $(\psi_L)_i$, each carrying charge q_i under a $U(1)$ gauge field, then the phase in (4.6) is proportional to the sum of q_i^3. The theory is consistent only if

$$\sum_i q_i^3 = 0 \ . \tag{4.7}$$

Alternatively, if we keep the theory written in terms of left-handed and right-handed Weyl fermions, then the anomaly cancellation condition (4.7) becomes

$$\sum_{\text{left}} q_i^3 = \sum_{\text{right}} q_i^3 \ . \tag{4.8}$$

There is a simple way to satisfy (4.7): we just take pairs of Weyl fermions with charges $\pm q$. If we conjugate one of these, then we can equivalently think of one left-handed and one right-handed Weyl fermion, each with charge q. Or, equivalently, we have a single Dirac fermion of charge q. Theories of this kind are called *vector-like*. They enjoy a parity symmetry (at least among the gauge interactions) which, as we saw in Section 1.4, exchanges left- and right-handed fermions. The simplest example is QED.

There are, however, more interesting solutions to (4.7) that do involve $\pm$ pairs. These are known as *chiral gauge theories*. These theories necessarily break parity.

Abelian Chiral Gauge Theories

Can we write down a consistent, Abelian chiral gauge theory? In fact, I'll ask for one more criterion: can we write down a consistent chiral gauge theory with integer charges

$$q_i \in \mathbb{Z} \ . \tag{4.9}$$

I'll say some words below about why we might want to require this.

First, it's clear that for $N = 2$ Weyl fermions, charges obeying (4.7) must come in $\pm$ pairs which is a vector-like theory. What about for $N = 3$ fermions? We must have two positive charges and one negative (or the other way round). Set $q_i = (x, y, -z)$ with x, y, z positive integers. The condition for anomaly cancellation (4.7) then becomes

$$x^3 + y^3 = z^3 \ . \tag{4.10}$$

Rather famously, this equation has no positive integer solutions. (This is the baby version of Fermat's last theorem, proven by Euler.)

What about chiral gauge theories with $N = 4$ Weyl fermions? Now we have two options: we could take three positive charges and one negative and look for positive integers satisfying

$$x^3 + y^3 + z^3 = w^3 \ . \tag{4.11}$$

The simplest integers satisfying this are 3,4,5 and 6. We can also construct chiral gauge theories with $N = 4$ Weyl fermions by having two of positive charge and two of negative charge, so that

$$x^3 + y^3 = z^3 + w^3 \ . \tag{4.12}$$

This equation is closely associated to Ramanujan and the famous story of Hardy's visit to his hospital bed. Struggling for small talk, Hardy commented that the number of his taxicab was particularly uninteresting: 1729. Ramanujan responded that, far from being uninteresting, this corresponds to the simplest four dimensional chiral gauge theory, since it is the first number that can be expressed as the sum of two cubes in two different ways: $1^3 + 12^3 = 9^3 + 10^3$.

There is one further condition that we've not yet met. As we will explain shortly, if you want to be able to couple your theory to gravity (and, let's face it, we do) then the condition (4.7) should be augmented by the requirement

$$\sum_i q_i = 0 \ . \tag{4.13}$$

None of the examples with $N = 4$ Weyl fermions above obey this. The simplest Abelian chiral gauge theory that can be coupled to gravity has $N = 5$ Weyl fermions. For example, the charges $q_i = \{1, 5, -7, -8, 9\}$ do the job.

We see that restricting to integer valued charges $q_i \in \mathbb{Z}$ means that we have to solve Diophantine equations and this breathes a little number theory into the proceedings. But why do we require that $q_i \in \mathbb{Z}$? The answer to this is a little subtle.

Strictly, there are two different Abelian gauge groups. The first is $G = U(1)$ which has only integer charges $q_i \in \mathbb{Z}$. Sometimes, it's useful to rescale the charges (and the Standard Model will be an example) so that you take the charges to be rational, $q_i \in \mathbb{Q}$, but that doesn't change the fact that the charges are quantised. The second is $G = \mathbb{R}$ which have charges that can take any value $q_i \in \mathbb{R}$ so you could have, for example, $q_1 = 1$ and $q_2 = \sqrt{2}$.

The gauge groups $U(1)$ and $\mathbb{R}$ have other differences, beyond the allowed electric charges. In particular, the gauge group $U(1)$ admits magnetic monopoles while the gauge group $\mathbb{R}$ does not (essentially because you can't respect the Dirac quantisation condition with respect to all charges). So one obvious question is: which of these gauge groups describes our world?

Irrep	$\square$	adj	$\square\square$	$\begin{smallmatrix}\square\\\square\end{smallmatrix}$
dim	N	N^2-1	$\frac{1}{2}N(N+1)$	$\frac{1}{2}N(N-1)$
$I(R)$	1	$2N$	$N+2$	$N-2$
$A(R)$	1	0	$N+4$	$N-4$

Table 8. Some group theoretic properties of $SU(N)$ representations. Here $\square\square$ is the symmetric representation and $\begin{smallmatrix}\square\\\square\end{smallmatrix}$ the anti-symmetric. Conjugate representations have $I(\bar{R}) = I(R)$ and $A(\bar{R}) = -A(R)$.

The experimental evidence strongly points to $U(1)$ because all electric charges (and, as we will see in Section 5, all hypercharges) are quantised. Moreover, there are arguments that invoke quantum gravity that we won't describe that are reasonably compelling, but far from rigorous, for why the gauge group in any quantum field theory should be $U(1)$, and not $\mathbb{R}$.

4.1.1 Non-Abelian Gauge Anomalies

So far we've only discussed anomalies for an Abelian gauge field. There is an analogous result for non-Abelian gauge symmetry G. Suppose that we have a single Weyl fermion in the representation R of a group G, with generator T^A_R so that, under a gauge transformation, we have

$$\psi_L \to e^{ig\alpha^A(x)T^A_R}\psi_L \quad \text{and} \quad A_\mu \to \Omega A_\mu \Omega^{-1} + \frac{i}{g}\Omega\partial_\mu\Omega^{-1} \tag{4.14}$$

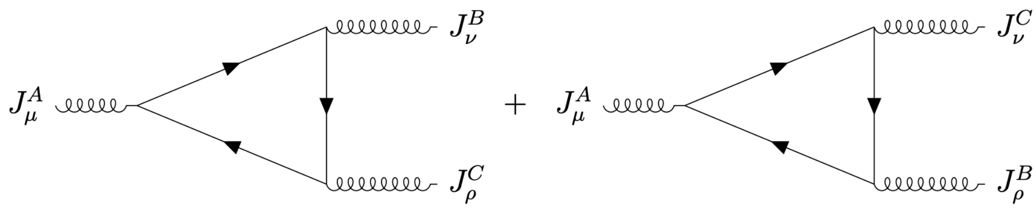
where $\Omega = e^{i\alpha^A T^A}$ with T^A in the fundamental representation. We can define the partition function just as (4.5), but where various fields are now viewed as their non-Abelian avatars. Then, under a gauge transformation, the partition function again changes by a phase

$$Z[A] \to \exp\left(\frac{ig^3 A(R)}{16\pi^2}\int d^4x\ \mathrm{Tr}\left(\alpha F_{\mu\nu}{}^\star F^{\mu\nu}\right)\right)\ Z[A]\ . \tag{4.15}$$

Here $A(R)$ is a group theoretic factor. For the fundamental representation, we have $A(R) = 1$ while, for all other representations, this is defined to be

$$\mathrm{Tr}\, T^A_R\,\{T^B_R, T^C_R\} = A(R)\,\mathrm{Tr}\, T^A\,\{T^B, T^C\}\ . \tag{4.16}$$

The emergence of the anti-commutator can be traced to the requirement to sum over different indices in the triangle diagrams



Some examples of $A(R)$ for $SU(N)$ representations are collected in Table 8. To be consistent, a non-Abelian gauge theory coupled to a bunch of left-handed Weyl fermions must obey

$$\sum_i A(R_i) = 0 \tag{4.17}$$

which is the non-Abelian version of (4.7).

For Abelian anomalies, we could always ensure that things work by taking fermions to come in pairs with charges $\pm q$. A similar result holds for non-Abelian anomalies. This follows from the following result.

Claim: If R is a complex representation, then the conjugate representation $\bar{R}$ has $A(\bar{R}) = -A(R)$.

Proof: If we write a group element as $e^{i\alpha^A T^A_R}$ then, in the conjugate representation, the same group element is given by the complex conjugate $e^{-i\alpha^A T^{A\star}_R}$. This means that the generators for the conjugate representation are $\bar{T}^A_R = -T^{A\star}_R = -(T^A_R)^T$ where the last equality holds because our generators are Hermitian, so $T^A_R = (T^A_R)^\dagger$. Now we have

$$\mathrm{Tr}\,\bar{T}^A_R\,\{\bar{T}^B_R, \bar{T}^C_R\} = -\mathrm{Tr}\,(T^A_R)^T\,\{(T^B_R)^T, (T^C_R)^T\} = -\mathrm{Tr}\,T^A_R\,\{T^B_R, T^C_R\} \quad . \tag{4.18}$$

Here the last equality holds because $\mathrm{Tr}\,A = \mathrm{Tr}\,A^T$. (It's important that we have the anti-commutator inside the trace, because the two terms get exchanged but, happily, they come with a relative plus sign rather than a minus sign.) □

The fact that $A(\bar{R}) = -A(R)$ means that we can always satisfy the anomaly by coupling our gauge field to left-handed fermions that come in R and $\bar{R}$ pairs. Alternatively, instead of working with left-handed fermions in the $\bar{R}$ representation, we could instead view them as right-handed fermions in the R representation. This means that the anomaly cancellation condition (4.17) is satisfied whenever we have a Dirac fermion. That, of course, is what happens for QCD.

One consequence of the relation $A(\bar{R}) = -A(R)$ is that $A(R) = 0$ for any real representation. This means that there is no obstacle to coupling a single Weyl fermion in a real representation to a non-Abelian gauge group. For example, $SU(N)$ coupled to a single adjoint Weyl fermion is a perfectly good field theory. (In fact, it is a very well studied field theory known as *super-Yang-Mills*.) But $SU(N)$ coupled to a single fundamental Weyl fermion does not make sense as a quantum theory.

This highlights a property of anomalies that will become increasingly important as we proceed: only massless fermions contribute to anomalies. Or, said differently, the contribution to the anomaly from any massive fermions will always cancel.

For example, to write down a Dirac mass for a fermion in a complex representation that preserves a symmetry, we need a left-handed ψ_L and a right-handed ψ_R, both transforming in the same representation, so that we can construct the mass term $\bar{\psi}_L\psi_R$. But the contribution to the anomaly from these two Weyl fermions cancels. Meanwhile, if we have a fermion in a real representation, like the adjoint, then we can always write down a Majorana mass $\mathrm{Tr}\,\psi_L\psi_L$ that preserves the symmetry. But now the contribution to the anomaly vanishes. The upshot is that only fermions that cannot get a mass preserving G contribute to the anomaly for G.

The story above also means that the only gauge groups that suffer from potential anomalies are those with complex representations. This already limits the possibilities: we need only worry about gauge anomalies in simply laced groups when

$$G = \begin{cases} SU(N) \text{ with } N \geq 3 \\ SO(4N+2) \\ E_6 \end{cases} . \tag{4.19}$$

We should also add $G = U(1)$ to this list which we discussed previously.

This list is short, but it turns out to be shorter still because all anomaly coefficients $\mathrm{Tr}\,T^A\{T^B, T^C\}$ vanish for E_6 and for $SO(4N+2)$ with $N \geq 2$. (Note that the Lie algebra $so(6) \cong su(4)$ so this one remains.) This means that, when it comes to perturbative anomalies discussed above, we only need to worry when we have gauge groups $G = SU(N)$ with $N \geq 3$.

There is, however, a "non-perturbative anomaly", usually called the *Witten anomaly* that rears its head for $SU(2)$ and, indeed, for all $Sp(N)$. We'll discuss this briefly below.

Non-Abelian Chiral Gauge Theories

We could try to write down chiral non-Abelian gauge theories, in which left-handed and right-handed fermions transform in different representations. This is straightforward to do. For gauge group $G = SU(N)$, from Table 8, the anomaly coefficients for the symmetric $\square\!\square$ and anti-symmetric $\begin{smallmatrix}\square\\\square\end{smallmatrix}$ representations are

$$A(\square\!\square) = N + 4 \quad \text{and} \quad A\left(\begin{smallmatrix}\square\\\square\end{smallmatrix}\right) = N - 4 \ . \tag{4.20}$$

Meanwhile, for the anti-fundamental representation $\bar{\mathbf{N}}$, which we denote as $\overline{\square}$, we have $A(\overline{\square}) = -1$. This means that we can construct a chiral gauge theory by taking, for example $G = SU(N)$ with a $\begin{smallmatrix}\square\\\square\end{smallmatrix}$ and $N - 4$ $\overline{\square}$ left-handed Weyl fermions. The simplest of these theories is $G = SU(5)$ with a $\mathbf{10}$ and a $\bar{\mathbf{5}}$.

Alternatively, we could build a chiral gauge theory by taking either E_6 or $SO(4N + 2)$ with complex representations, for which the anomaly coefficients all vanish. The simplest such example is $SO(10)$ with a single Weyl fermion in the $\mathbf{16}$ representation. This is the spinor representation of $SO(10)$. (Strictly, we should be talking about the double cover $Spin(10)$ as the gauge group, rather than $SO(10)$.) Rather strikingly, both this $SO(10)$ example and the $SU(5)$ example above are prominent candidates for grand unified theories.

One key feature of chiral gauge theories – both non-Abelian and Abelian – is that it's not possible to write down mass terms for fermions. Any such mass term should be of the form $\chi_L \psi_L$ or, equivalently, $\bar{\chi}_R \psi_L$, but these quadratic terms are not gauge invariant.

4.1.2 Mixed Anomalies

Again consider a single Weyl fermion, now coupled to a background non-Abelian gauge field A_μ in some representation R of the global symmetry $G = SU(N)$ and an Abelian gauge field that, for the purposes of this argument, we will call a_μ. The partition function is

$$Z[A; a] = \int D\psi_L D\bar{\psi}_L \ \exp\left(-\int d^4x \ i\bar{\psi}_L \bar{\sigma}^\mu \mathcal{D}_\mu \psi_L\right) \tag{4.21}$$

now with

$$\mathcal{D}_\mu \psi_L = \partial_\mu \psi_L - igA_\mu^A T_R^A \psi_L - ieqa_\mu \psi_L \ . \tag{4.22}$$

Now when we do a $U(1)$ gauge transformation $\psi_L \to e^{ieq\alpha}\psi_L$, the partition function picks up two contributions: one is the phase (4.6) that depends on the $U(1)$ field strength $f_{\mu\nu} = \partial_\mu a_\nu - \partial_\nu a_\mu$, but there is another that depends on the $SU(N)$ field strength,

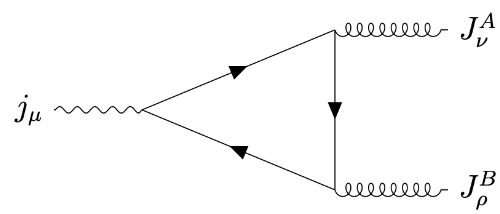
$$Z[A;a] \to \exp\left(\frac{ie^3q^3}{32\pi^2}\int d^4x\ \alpha f_{\mu\nu}{}^\star f^{\mu\nu} + \frac{ieg^2qI(R)}{16\pi^2}\int d^4x\ \alpha \text{Tr}\, F_{\mu\nu}{}^\star F^{\mu\nu}\right) Z[A;a] \quad (4.23)$$

Here $I(R)$ is another group theoretic quantity, known as the *Dynkin index*, defined as

$$\text{Tr}\ T^A_R\, T^B_R = \frac{1}{2} I(R)\, \delta^{AB}\ . \quad (4.24)$$

The Dynkin index is related to the quadratic Casimir $C(R)$, which we previously defined in (3.27) by $T^A_R T^A_R = C(R)\,\mathbb{1}$. You can take the trace of both sides to get $I(R)\dim(G) = 2C(R)\dim(R)$. The fundamental representation has $I(\square) = 1$ and the Dynkin index of the conjugate representation is $I(\bar{R}) = I(R)$. The Dynkin indices for some other common representations of $SU(N)$ are given in Table 8.

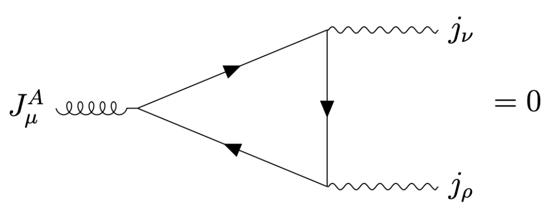
The second term in (4.23) is known as a *mixed anomaly*. It is again cubic in the charges, but this is shared between a single $U(1)$ charge q and two non-Abelian charges. In perturbation theory, it arises from the triangle diagram:



To have a consistent gauge theory, any mixed anomalies must also cancel. For a bunch of left-handed fermions with $U(1)$ charge q_i, sitting in $SU(N)$ representations R_i, the requirement of anomaly cancellation is

$$\sum_i q_i I(R_i) = 0\ . \quad (4.25)$$

You might wonder what happens if we have a single non-Abelian current, and two Abelian currents,



But this vanishes automatically, because it's proportional to the trace of the generator $\text{Tr}\, T^A = 0$.

The Mixed Gauge-Gravitational Anomaly

Something similar plays out if we couple a quantum field theory to gravity. We needn't be bold and talk about quantum gravity here: it's enough just to think about a quantum field theory on a curved spacetime with metric g.

To motivate this, let's first review how to couple spinors to a curved spacetime. The starting point is to decompose the metric in terms of vierbeins,

$$g_{\mu\nu}(x) = e^a_\mu(x)\, e^b_\nu(x) \ . \tag{4.26}$$

There is an arbitrariness in our choice of vierbein, and this arbitrariness introduces an $SO(1,3)$ gauge symmetry into the game. The associated gauge field ω_μ^{ab} is called the *spin connection*. It is determined by the requirement that the vierbeins are covariantly constant

$$\mathcal{D}_\mu e^a_\nu \equiv \partial_\mu e^a_\nu - \Gamma^\rho_{\mu\nu} e^a_\lambda + \omega^a_{\mu\,b} e^b_\nu = 0 \tag{4.27}$$

where $\Gamma^\rho_{\mu\nu}$ are the usual Christoffel symbols. This language makes general relativity look very much like any other gauge theory. In particular, the field strength of the spin connection is

$$(R_{\mu\nu})^a{}_b = \partial_\mu \omega^a_{\nu\,b} - \partial_\nu \omega^a_{\mu\,b} + [\omega_\mu, \omega_\nu]^a{}_b \ . \tag{4.28}$$

This is related to the usual Riemann tensor by $(R_{\mu\nu})^a{}_b = e^a_\rho e^\sigma_b\, R_{\mu\nu}{}^\rho{}_\sigma$.

This machinery is just what we need to couple a Dirac spinor to a background curved spacetime. The appropriate covariant derivative is

$$\mathcal{D}_\mu \psi_\alpha = \partial_\mu \psi_\alpha + \frac{1}{2}\omega_\mu^{ab} (S_{ab})^\beta{}_\alpha \psi_\beta \tag{4.29}$$

where $S_{ab} = \frac{1}{4}[\gamma_a, \gamma_b]$ is the generator of the Lorentz group in the spinor representation. Written in this way, the coupling of spinors to a curved spacetime looks very similar to the coupling to any other gauge field.

This manifests itself in the path integral measure. If we assign the Weyl fermion a charge q and couple it to a $U(1)$ gauge field a transformation, the partition function shifts as

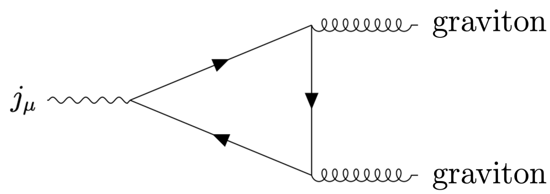
$$Z[a] \rightarrow \exp\left(\frac{eq}{192\pi^2} \int d^4x\ \alpha \epsilon^{\mu\nu\rho\sigma} R_{\mu\nu\lambda\tau} R_{\rho\sigma}{}^{\lambda\tau}\right)\ Z[a] \ . \tag{4.30}$$

with $R_{\mu\nu\lambda\tau}$ the Riemann tensor. This is a mixed $U(1)$-gravitational anomaly. The equivalence principle means that everything couples the same to gravity, so there's no analog of the Dynkin index in (4.25) and the requirement that a $U(1)$ gauge theory is consistent when placed on a curved spacetime becomes

$$\sum_i q_i = 0 \ . \tag{4.31}$$

This is the condition (4.13) that we advertised previously.

Again, this result can also be seen in perturbation theory, this time by a suitable regularisation of the triangle diagram,



This mixed gauge-gravitational anomaly only arises for Abelian gauge groups. There's no corresponding requirement for non-Abelian gauge theories, essentially because $\mathrm{Tr}\, T^A = 0$ for any generator of a simply connected Lie algebra.

It turns out that there is no purely gravitational anomaly, with gravitons on all three legs, in $d = 3 + 1$ dimensions. Such gravitational anomalies do exist in $d = 2 \bmod 8$ dimensions, and there are important implications in $d = 1 + 1$ for condensed matter physics and in $d = 9 + 1$ for string theory.

4.1.3 The Witten Anomaly

Among the $G = SU(N)$ gauge groups, the smallest $G = SU(2)$ stands out as special. This is because all representations of $G = SU(2)$ are either real or pseudoreal. (A *pseudoreal* representation means that, while not actually real, the representation is isomorphic to its complex conjugate.) This means that there are no perturbative gauge anomalies of the kind described above for $G = SU(2)$.

You can check this explicitly for the fundamental representation. This has generators $T^A = \frac{1}{2}\sigma^A$ with σ^A the Pauli matrices. But a little matrix multiplication will convince you that

$$\mathrm{Tr}\, \sigma^A \{\sigma^B, \sigma^C\} = 0 \tag{4.32}$$

for all $A, B, C = 1, 2, 3$. That's the statement that there's no anomaly.

Taken at face value, this suggests that $SU(N)$ coupled to a single fundamental Weyl fermion is inconsistent for all $N \geq 3$ but is fine for $N = 2$. That's a slightly odd state of affairs, not least because the $SU(2)$ theory has a number of strange and hard-to-interpret properties. (The instanton has an odd number of fermion zero modes for example.) However, there's something else at play that we've missed. It turns out that the $SU(2)$ theory suffers from a different kind of anomaly. This is known as the *Witten anomaly*, or sometimes just as the $SU(2)$ anomaly.

The Witten anomaly doesn't show up in perturbation theory. Instead it can be traced to some strange field configurations that we must sum over in the path integral that wind in a non-trivial way around Euclidean spacetime. Mathematically, this follows from the homotopy group

$$\Pi_4(SU(2)) = \mathbb{Z}_2 \ . \tag{4.33}$$

For this anomaly to cancel, an $SU(2)$ gauge theory must have an even number of fundamental Weyl fermions to be consistent. Again, you can find details of this calculation in the lectures on Gauge Theory.

4.2 Chiral (or ABJ) Anomalies

As we stressed at the beginning of this section, the anomaly for a symmetry group G has various avatars depending on whether the symmetry is global or gauged. So far, we've seen one of these avatars: the anomaly provides a collection of consistency conditions on any gauge theory: the charges, or representations, must obey (4.7) and (4.17) and, for mixed anomalies, (4.25) and (4.31).

In this section we discuss the second avatar of anomalies: a perfectly good global symmetry of the classical theory, can fail to be a symmetry of the quantum theory. This was the first place in which anomalies in quantum field theories were discovered. This phenomenon is known as the *ABJ anomaly*, after its discoverer's Adler, Bell and Jackiw, and sometimes as the *chiral anomaly* and sometimes, confusingly, just as the *anomaly*.

The ABJ anomaly can be viewed as a mixed anomaly between a $U(1)$ global symmetry and a gauge symmetry G. As an example, suppose that we have a bunch of left-handed Weyl fermions, transforming in the representation R_i under a $G = SU(N)$ gauge symmetry. Suppose, in addition, that there is a global $U(1)$ symmetry of the classical action, under which the fermions have charges q_i.

The full Euclidean partition function for this theory is, schematically,

$$\mathcal{Z} = \int \mathcal{D}A \ \exp\left(-\frac{1}{2}\int d^4x \ \text{Tr}\, F_{\mu\nu}F^{\mu\nu}\right) Z[A] \tag{4.34}$$

where A is the non-Abelian gauge field and $Z[A]$ is the partition function for the fermions, which are coupled to this gauge field

$$Z[A] = \int D\psi_{L\,i} D\bar{\psi}_{L\,i} \ \exp\left(-\int d^4x \ i\sum_i \bar{\psi}_L\bar{\sigma}^\mu\mathcal{D}_\mu\psi_L\right) \ . \tag{4.35}$$

Note that, in contrast to the previous section, we haven't introduced a background gauge field for the $U(1)$ global symmetry. (This is what we called a_μ in (4.23).)

Now we do a global $U(1)$ transformation

$$\psi_{L\,i} \to e^{i\alpha q_i}\psi_{L\,i} \tag{4.36}$$

for some $\alpha \in \mathbb{R}$. The mixed anomaly (4.23) means that the partition function is not invariant. Instead, the fermionic part of the partition function transforms as

$$Z[A] \to \exp\left(\frac{i\alpha}{16\pi^2}\sum_i q_i I(R_i)\int d^4x \ \text{Tr}\, F_{\mu\nu}{}^\star F^{\mu\nu}\right) Z[A] \ . \tag{4.37}$$

We see that, although the classical action may be invariant under the global $U(1)$ symmetry, for this to persist as a symmetry of the quantum theory we also need the fermionic measure to be invariant. This is true only if

$$\sum_i q_i I(R_i) = 0 \ . \tag{4.38}$$

If this condition does not hold, then the classical symmetry is not a symmetry of the quantum theory. It is said to be *anomalous*.

An Example: The Axial Anomaly in QCD

The most familiar example of this kind of anomaly arises for the (approximate) $U(1)_A$ axial symmetry of QCD. Consider the generalised theory, in which we have a $G = SU(N_c)$, coupled to N_f massless Dirac fermions. The action is

$$S = \int d^4x \ \left(-\frac{1}{2}\text{Tr}\, G_{\mu\nu}G^{\mu\nu} + i\sum_{i=1}^{N_f}\bar{\psi}_i\gamma^\mu\mathcal{D}_\mu\psi_i\right) \ . \tag{4.39}$$

We studied this theory in some detail in Section 3.2 where we learned about the implications of chiral symmetry breaking. Recall that the classical action 4.39 has an $U(N_f)_L \times U(N_f)_R$ global symmetry, with each factor rotating q_L and q_R independently. The $SU(N_f)_L \times SU(N_f)_R$ subgroup is the main character in the story of chiral symmetry breaking. Here we are more interested in the two $U(1)$ subgroups, which we take to act as

$$\begin{aligned} U(1)_V : \ \ \psi_{L\,i} \to e^{i\alpha}\psi_{L\,i} \quad \text{and} \quad \psi_{R\,i} \to e^{i\alpha}\psi_{R\,i} \\ U(1)_A : \ \ \psi_{L\,i} \to e^{i\alpha}\psi_{L\,i} \quad \text{and} \quad \psi_{R\,i} \to e^{-i\alpha}\psi_{R\,i} \ . \end{aligned} \tag{4.40}$$

Here $U(1)_V$ is the "vector-like" symmetry, meaning that it acts the same on left- and right-handed spinors. In the context of the Standard Model, this is also referred to as *baryon number* because it counts the number of baryons in a given state. Meanwhile, the axial symmetry $U(1)_A$ acts differently on the left and right-handed spinors.

The left-handed spinors ψ_L transform in the $\mathbf{N_c}$ of $SU(N_c)$ while the conjugated right-handed spinors $\bar{\psi}_R$ (which, due to the conjugation, are themselves left-handed) transform in the $\overline{\mathbf{N}}_\mathbf{c}$. For both of these, the Dynkin index is $I(\mathbf{N_c}) = I(\overline{\mathbf{N}}_\mathbf{c}) = 1$.

Under $U(1)_V$, the ψ_L have charge $+1$ and the $\bar{\psi}_R$ charge -1, which means that the anomaly (4.38) vanishes. Hence, $U(1)_V$ is a good symmetry of the quantum theory. In contrast, under $U(1)_A$, the ψ_L have charge $+1$ while the $\bar{\psi}_R$ also have charge $+1$. This means that the anomaly (4.38) does not vanish, and $U(1)_A$ is *not* a symmetry of the quantum theory.

We've already seen one consequence of the QCD axial anomaly in Section 3.2: the chiral condensate would naively seem to spontaneously break the $U(1)_A$ axial symmetry, but there's no associated light Goldstone boson in the QCD spectrum. Indeed, the would-be Goldstone boson is the η' which is significantly heavier than the pions. The reason is that $U(1)_A$ was never a symmetry of the quantum theory in the first place and wasn't available to be spontaneously broken.

4.2.1 The Theta Term Revisited

There is another way to think about the chiral anomaly. We see from (4.37), that acting with an anomalous $U(1)$ global symmetry adds a term to the path integral that is proportional to $\text{Tr}\, F_{\mu\nu}{}^\star F^{\mu\nu}$.

But we've met a term like this before. We can always add to the Yang-Mills action (or, indeed, to the Maxwell action) a theta term that takes the form

$$S_\theta = \frac{\theta g^2}{16\pi^2}\int d^4x \ \text{Tr}\, F_{\mu\nu}{}^\star F^{\mu\nu} \ . \tag{4.41}$$

We discussed some properties of this term in Section 3.4. Comparing with the form of the chiral anomaly (4.37), we can interpret the anomaly as saying that the theta parameter is shifted by a $U(1)$ transformation,

$$U(1)_A: \ \theta \to \theta + \alpha \sum_i q_i I(R_i) \ . \tag{4.42}$$

But if a parameter (as opposed to a field) changes under a symmetry, then that means that the symmetry is explicitly broken. This is another way to frame the anomaly.

For example, if we return to our generalised QCD with $G = SU(N_c)$ gauge group and N_f massless Dirac fermions then, under the axial transformation (4.40), the theta angle transforms as

$$U(1)_A : \theta \to \theta + 2N_f\alpha \ . \tag{4.43}$$

Thinking about things in this way makes certain aspects of the physics more transparent. For example, suppose that we have a theory with a single massive Dirac fermion ψ. There are two different Dirac masses that we could write down:

$$\mathcal{L}_{\text{mass}} = m_1\bar{\psi}\psi + im_2\bar{\psi}\gamma^5\psi \ . \tag{4.44}$$

If we decompose the Dirac fermion into Weyl fermions, $\psi = (\psi_L, \psi_R)$, then these masses become

$$\mathcal{L}_{\text{mass}} = m\bar{\psi}_L\psi_R + m^\star\bar{\psi}_R\psi_L \quad \text{with} \quad m = m_1 + im_2 \ . \tag{4.45}$$

Now suppose that we do an axial rotation, $\psi_L \to e^{i\alpha}\psi_L$ and $\psi_R \to e^{-i\alpha}\psi_R$. Then the theory isn't invariant because the mass term shifts by a phase. But, from (4.42), so too does the theta angle. We have

$$U(1)_A: \quad m \to e^{-2i\alpha}m \quad \text{and} \quad \theta \to \theta + 2\alpha \ . \tag{4.46}$$

However, rotating the phase of the fermion can't change the physics of the theory. For example, if we have a free massive fermion (not coupled to a gauge field) then for every value of the mass $m \in \mathbb{C}$ in (4.45), the physical excitation always has mass $|m|$. Now when we couple the fermion to the gauge field, rotating the phase of the fermion changes both the phase of m and the value of θ. This means that the physics depends only on the invariant combination $\theta + \arg(m)$. More generally, with N_f fermions we can have a complex mass matrix M and the quantity $\theta + \arg(\det M)$ remains invariant under chiral rotations.

This, ultimately, is the way in which the strong CP problem in QCD gets its teeth: it's not quite true to say that $\theta = 0$ in QCD. It's more accurate to say that $\theta +$ a bunch of phases of masses $= 0$. And, as we will see in Section 5, those phases of the masses come from rather different physics of the Yukawa couplings.

There is one further observation that follows from the discussion above. Suppose that we have a gauge theory coupled to one, or more, massless fermions. Then rotating the phase of that massless fermion shouldn't affect the physics of the theory, but acts to shift theta as in (4.42). This means that, in a theory with massless fermions, the theta angle isn't physical: it can just be shifted away by an axial rotation. This suggests a rather cute solution to the strong CP problem: perhaps the mass of the up quark is actually zero! In that case, the physics would be independent of the value of θ. Sadly, as numerical simulations have got better, we're now pretty confident that the mass of the up quark is non-zero, and this idea is not a viable solution to the strong CP problem.

4.2.2 Noether's Theorem for Anomalous Symmetries

If a theory has a continuous symmetry, then Noether's theorem tells us that there will be a corresponding conserved current J^μ, obeying the continuity equation

$$\partial_\mu J^\mu = 0 \ . \tag{4.47}$$

What happens if the symmetry is anomalous, so that it's a symmetry of the classical action, but not of the full quantum theory? How does this show up in the conservation of the current?

To answer this, let's first recall how to derive Noether's theorem. To start, we'll work with scalar fields, even though our ultimate interest is in fermions. Consider the transformation of a scalar field ϕ

$$\delta\phi = \alpha X(\phi) \ . \tag{4.48}$$

Here α is a constant, infinitesimally small parameter. This transformation is a *symmetry* if the change in the Lagrangian is

$$\delta\mathcal{L} = 0 \ . \tag{4.49}$$

We can actually be more relaxed than this and allow the Lagrangian to change by a total derivative; this won't change our conclusions below.

The quick way to prove Noether's theorem is to allow the constant α to depend on spacetime: $\alpha = \alpha(x)$. Now the Lagrangian is no longer invariant, but changes as

$$\begin{aligned}\delta\mathcal{L} &= \frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi)}\,\partial_\mu(\delta\phi) + \frac{\partial\mathcal{L}}{\partial\phi}\,\delta\phi \\ &= \frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi)}\,\partial_\mu(\alpha X(\phi)) + \frac{\partial\mathcal{L}}{\partial\phi}\,\alpha X(\phi) \\ &= (\partial_\mu\alpha)\,\frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi)}\,X(\phi) + \left[\frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi)}\partial_\mu X(\phi) + \frac{\partial\mathcal{L}}{\partial\phi}\,X(\phi)\right]\alpha\ . \end{aligned} \tag{4.50}$$

But we know that $\delta\mathcal{L} = 0$ when α is constant, which means that the term in square brackets must vanish. We're left with the expression

$$\delta\mathcal{L} = (\partial_\mu\alpha)J^\mu \quad \text{with} \quad J^\mu = \frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi)}\,X(\phi)\ . \tag{4.51}$$

The action $S = \int d^4x\ \mathcal{L}$ then changes as

$$\delta S = \int d^4x\ \delta\mathcal{L} = \int d^4x\ (\partial_\mu\alpha)J^\mu = -\int d^4x\ \alpha\,\partial_\mu J^\mu \tag{4.52}$$

where we pick $\alpha(x)$ to decay asymptotically so that we can safely discard the surface term.

The expression (4.52) holds for any field configuration ϕ with the specific change $\delta\phi$. However, when ϕ obeys the classical equations of motion then $\delta S = 0$ for *any* $\delta\phi$, including the symmetry transformation (4.48) with $\alpha(x)$ a function of spacetime. This means that *when* the equations of motion are satisfied we have the conservation law

$$\partial_\mu J^\mu = 0\ . \tag{4.53}$$

This is Noether's theorem.

An Example: the Free Fermion

We can apply all of the above ideas to the fermions that we're really interested in. As a warm-up, consider a free, massless Dirac fermion ψ with action

$$S = -\int d^4x\ i\bar{\psi}\gamma^\mu\partial_\mu\psi \tag{4.54}$$

with $\bar{\psi} = \psi^\dagger\gamma^0$. This theory has two symmetries, the vector and axial symmetries of (4.40). Written in terms of the Dirac fermion, the vector symmetry acts as $\psi \to e^{i\alpha}\psi$ and, infinitesimally, this becomes

$$U(1)_V : \delta\psi = i\alpha\psi \quad \text{and} \quad \delta\bar{\psi} = -i\alpha\bar{\psi}\ . \tag{4.55}$$

We can read off the associated current from (4.51): it is

$$J_V^\mu = \bar{\psi}\gamma^\mu\psi \ . \tag{4.56}$$

Meanwhile, the axial symmetry acts as $\psi \to e^{i\alpha\gamma^5}\psi$ and, infinitesimally, this becomes

$$U(1)_A : \delta\psi = i\alpha\gamma^5\psi \quad \text{and} \quad \delta\bar{\psi} = i\alpha\bar{\psi}\gamma^5 \ . \tag{4.57}$$

Here there's an extra minus sign that rears its head in the transformation of $\delta\bar{\psi}$ which arises because the γ^5 has to sneak past the γ^0 that sits in the definition of $\bar{\psi}$. Again, we can read off the associated current from (4.51): this time it is

$$J_A^\mu = \bar{\psi}\gamma^\mu\gamma^5\psi \ . \tag{4.58}$$

As a warm-up to understand the effect of the anomaly, we can see how the currents are affected when we turn on a mass term for the fermion, so

$$S = -\int d^4x \ i\bar{\psi}\gamma^\mu\partial_\mu\psi + m\bar{\psi}\psi \ . \tag{4.59}$$

The action remains invariant under the vector symmetry, and so the current J_V^μ continues to obey $\partial_\mu J_V^\mu = 0$. But the mass term is not invariant under the axial symmetry. Nonetheless, that doesn't mean that we can't say anything. Let's return to our derivation of Noether's theorem and do a transformation with the constant α again promoted to a function of spacetime $\alpha(x)$. We can repeat the steps we did before, except that we need to include an extra term because the action is no longer invariant under the symmetry. Instead, we have

$$\delta S = \int d^4x \ (\partial_\mu\alpha)J_A^\mu + 2im\alpha\bar{\psi}\gamma^5\psi \tag{4.60}$$

with J_A^μ given in (4.58). Now the argument proceeds as before: when the equations of motion are obeyed, we must have $\delta S = 0$ for all transformations, including those with $\alpha(x)$. So whenever the equations of motion are obeyed, the axial current satisfies

$$\partial_\mu J_A^\mu = 2im\bar{\psi}\gamma^5\psi \ . \tag{4.61}$$

This tells us how conservation of axial charge fails when the fermion has a mass.

The Conservation Law for Anomalous Symmetries

Now we can reframe our original question: how is conservation of axial charge affected by the anomaly? We'll consider N_f massless Dirac fermions, coupled to a Yang-Mills theory, with action

$$S_\theta = \int d^4x \ \Big(-\frac{1}{2g^2}\text{Tr}\, F_{\mu\nu}F^{\mu\nu} + \frac{\theta g_s^2}{16\pi^2}\text{Tr}\, F_{\mu\nu}{}^\star F^{\mu\nu} - i\sum_{i=1}^{N_f}\bar{\psi}_i\gamma^\mu\mathcal{D}_\mu\psi_i \Big) \ .$$

We've seen that we can capture the effect of the anomaly by shifting the theta angle, as in (4.43)

$$U(1)_A : \theta \to \theta + 2N_f\alpha \ . \tag{4.62}$$

But now we can think of this as a shift of the classical action, and we're in the same boat as when we looked at massive fermions above. In particular, we find that the axial current obeys

$$\partial_\mu J^\mu_A = \frac{N_f g^2}{8\pi^2}\,\mathrm{Tr}\, F_{\mu\nu}{}^\star F^{\mu\nu} \ . \tag{4.63}$$

This is the effect of the anomaly.

Above, we have derived the anomaly equation (4.63) by thinking about the classical action. But one can also show that this holds as an operator equation in quantum field theory, what's known as a Ward identity. You can read about this in the lectures on Gauge Theory.

The anomaly equation (4.63) tells us that the axial symmetry is not conserved. However, at first glance, it appears that there might be a loophole in this statement. This is because, as we saw in (3.109), the term $\mathrm{Tr}\, F_{\mu\nu}{}^\star F^{\mu\nu}$ is actually a total derivative, with

$$\mathrm{Tr}\, F_{\mu\nu}{}^\star F^{\mu\nu} = 2\partial_\mu K^\mu \quad \text{with} \quad K^\mu = \epsilon^{\mu\nu\rho\sigma}\mathrm{Tr}\left(A_\nu\partial_\rho A_\sigma - \frac{2i}{3}A_\nu A_\rho A_\sigma\right) \ . \tag{4.64}$$

This suggests that we can define a combination of J^μ_A and K^μ to construct a current that is conserved. Indeed that is naively possible, but it's not legal because K^μ is not gauge invariant, even though $\partial_\mu K^\mu$ is.

We can also ask: under what circumstances does the axial charge change? The axial charge is measured by integrating over a spatial slice

$$Q_A = \int d^3x\ J^0_A \ . \tag{4.65}$$

The change in axial charge from time $t \to -\infty$ to time $t \to +\infty$ is (assuming that things drop off suitably fast at spatial infinity)

$$\Delta Q_A = \int dt d^3x\ \frac{\partial J^0_A}{\partial t} = \int d^4x\ \partial_\mu J^\mu_A = \frac{N_f g_s^2}{8\pi^2}\int d^4x\ \mathrm{Tr}\, F_{\mu\nu}{}^\star F^{\mu\nu} \ . \tag{4.66}$$

But we've already seen in section 3.4 that the integral of $\mathrm{Tr}\, F_{\mu\nu}{}^\star F^{\mu\nu}$ is quantised. This means that Q_A can jump by integer amounts. At weak coupling, the violation of axial charge is mediated by instantons.

There is a similar story for the mixed gauge-gravitational anomaly that we discussed previously. For example we saw that a single, free Weyl fermion has a $U(1)$ symmetry that suffers a mixed gravitational anomaly. This shows up because the current for this $U(1)$ is no longer conserved when the theory is placed in a curved background. Instead, it obeys

$$\nabla_\mu j^\mu_A = -\frac{N_f}{384\pi^2}\epsilon^{\mu\nu\rho\sigma}R_{\mu\nu\lambda\tau}R_{\rho\sigma}{}^{\lambda\tau} \tag{4.67}$$

where ∇_μ is the appropriate covariant derivative from differential geometry.

4.2.3 Neutral Pion Decay

The neutral pion, $\pi^0 = \frac{1}{\sqrt{2}}(\bar{u}u - \bar{d}d)$ has a substantially shorter lifespan than its charged cousin. It lasts only around $\sim 10^{-16}$ seconds, decaying primarily to

$$\pi^0 \to \gamma\gamma \ . \tag{4.68}$$

There is an interesting story associated to this. Indeed, it was the effort to understand why this decay occurs at all that first led to the discovery of the anomaly.

To set the scene, first note that, although we've focused on massless QCD above, the axial anomaly also arises in QED coupled to massless fermions. Suppose that we have N_f Dirac fermions ψ_i, each with charge Q_i under a $U(1)$ gauge symmetry. Then the axial symmetry $\psi_i \to e^{i\alpha\gamma^5 q_i}\psi_i$ suffers an ABJ anomaly, and the associated current obeys

$$\partial_\mu J^\mu_A = \left(\sum_i q_i Q_i^2\right)\frac{1}{16\pi^2}F_{\mu\nu}{}^\star F^{\mu\nu} \ . \tag{4.69}$$

Again, this follows from a triangle diagram with one J^μ_A leg, and two photon legs. This is reflected in the charges, which are linear in the axial charge q_i and quadratic in the gauge charge Q_i.

Now let's see the implications of this for QCD. We'll take $N_f = 2$ light quarks, corresponding to the up and down. If we assume that these are massless, we know that the QCD action has a $U(1)_V \times SU(2)_L \times SU(2)_R$ symmetry. Now we introduce the coupling to the photon with charges

$$Q_1 = \frac{2}{3} \quad \text{and} \quad Q_2 = -\frac{1}{3} \ . \tag{4.70}$$

Because the quarks have different electric charges, this breaks the flavour symmetry down to $U(1)_L \times U(1)_R \subset SU(2)_L \times SU(2)_R$. We can combine these into a new vector symmetry $U(1)'_V$ and a new axial symmetry $U(1)'_A$, under which the quarks transform as

$$\begin{aligned} U(1)'_V: &\quad u \to e^{i\alpha} u \quad \text{and} \quad d \to e^{-i\alpha} d \; . \\ U(1)'_A: &\quad u \to e^{i\alpha\gamma^5} u \;\; \text{and} \;\; d \to e^{-i\alpha\gamma^5} d \; . \end{aligned} \tag{4.71}$$

The vector symmetry $U(1)'_V$ is anomaly-free, while the axial symmetry $U(1)'_A$ does not suffer an anomaly due to the QCD gauge field because there is a cancellation between the $q_1 = +1$ charge of the up quark and the $q_2 = -1$ charge of the down quark. However, the axial $U(1)'_A$ *does* suffer an anomaly with the QED gauge field. To compute this, we need to remember that, from the perspective of electromagnetism, each quark comes in $N_c = 3$ different varieties, due to the fact that they also transform under the $SU(3)$ gauge group. This means that the ABJ anomaly (4.69) is

$$\partial_\mu J'^{\mu}_A = N_c \left(\left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 \right) \frac{1}{16\pi^2} F_{\mu\nu}{}^\star F^{\mu\nu} = \frac{N_c}{48\pi^2} F_{\mu\nu}{}^\star F^{\mu\nu} \; . \tag{4.72}$$

where we've left the value of $N_c = 3$ in this formula to highlight that the anomaly coefficient depends on the number of quark colours.

This additional axial current is $J'^{\mu}_A = \bar{u}\gamma^\mu\gamma^5 u - \bar{d}\gamma^\mu\gamma^5 d$ and, from (3.68), is precisely the current that creates the neutral pion π^0,

$$\langle 0 | J'^{\mu}_A(x) | \pi^0(p) \rangle = -i f_\pi \delta^{ab} \, p^\mu e^{-ix\cdot p} \; . \tag{4.73}$$

The anomaly equation then gives an amplitude for $\pi^0 \to \gamma\gamma$. This amplitude is proportional to N_c, the number of colours, and gives an experimental method to determine $N_c = 3$.

There is more to this story which we mention only briefly. This amplitude for $\pi^0 \to \gamma\gamma$ is the same as that which would arise from the coupling in the Lagrangian

$$\mathcal{L} = \frac{N_c e^2}{48\pi^2 f_\pi} \pi^0 \, F_{\mu\nu}{}^\star F^{\mu\nu} \; . \tag{4.74}$$

In other words, the neutral pion field π^0 acts very much like a dynamical theta term! There's something odd in this because π^0 is a Goldstone boson and, as such, should only appear in the action with derivative couplings. But, after an integration by parts, the pion is derivatively coupled in (4.74) if we remember that $F_{\mu\nu}{}^\star F^{\mu\nu} = 2\partial_\mu K^\mu$ as in (4.64). There is a much longer story here, involving the beautiful Wess-Zumino-Witten (WZW) term that you can read about in the lectures on Gauge Theory.

4.2.4 Surviving Discrete Symmetries

Thinking of the anomalous symmetry as shifting the theta angle reveals something novel. That's because the theta angle is, as the name suggests, an angle with $\theta \in [0, 2\pi)$. This means that if we transform by an anomalous $U(1)$ symmetry that maps $\theta \to \theta + 2\pi$, then that hasn't actually changed the value of θ at all. In this way, some discrete subgroup of the $U(1)$ may remain.

We can see this in the case of QCD, although the end result turns out to be a little fiddly and not particularly interesting. From (4.43), we see that a $U(1)_A$ transformation of the form $e^{-i\alpha} = e^{2\pi i/2N_f}$ will send $\theta \to \theta + 2\pi$. By acting with a compensating $U(1)_V$ transformation, there is a surviving $\mathbb{Z}_{N_f}$ subgroup which acts as

$$\mathbb{Z}_{N_f}: \quad \psi_{L\,i} \to e^{2\pi i/N_f}\psi_{L\,i} \quad \text{and} \quad \psi_{R\,i} \to \psi_{R\,i} \ . \tag{4.75}$$

But we recognise this as the centre of the $SU(N_f)_L$ global symmetry. So in this case, the surviving discrete symmetry doesn't tell us anything new.

Here's a different example where things are more interesting. Consider $SU(N)$ Yang-Mills coupled to a single, massless Weyl spinor λ in the adjoint representation. We've already seen that the adjoint representation is real, so this theory doesn't suffer from a gauge anomaly. Indeed, it's a rather famous theory because it secretly has a supersymmetry, exchanging the gauge field and fermion. This theory is known as *super Yang-Mills*. Thankfully, we won't need to know anything about supersymmetry for our discussion. (You can read more in the lectures on Supersymmetry.)

Classically this theory has a global $U(1)$ symmetry which rotates the phase of λ

$$U(1): \quad \lambda \to e^{i\alpha}\lambda \ . \tag{4.76}$$

But quantum mechanically, this theory suffers an anomaly. We need the fact, from Table 8, that $I(\text{adj}) = 2N$ for the adjoint representation. Then, from (4.42), we see that the theta angle shifts under this $U(1)$ symmetry as

$$U(1): \quad \theta \to \theta + 2N\alpha \ . \tag{4.77}$$

This is telling us that the $U(1)$ symmetry is anomalous. But, by the argument above, a discrete $\mathbb{Z}_{2N}$ survives since this shifts $\theta \to \theta + 2\pi$, while the fermion transforms as

$$\mathbb{Z}_{2N}: \quad \lambda \mapsto e^{2\pi i/2N}\lambda \ . \tag{4.78}$$

This discrete symmetry becomes particularly interesting because this theory, like many other non-Abelian gauge theories, flows to strong coupling at some scale $\Lambda_{\rm QCD}$ where it exhibits confinement and the formation of a fermion condensate,

$$\langle \lambda\lambda \rangle \sim \Lambda^3_{\rm QCD} \ . \tag{4.79}$$

In actual QCD, such a condensate breaks the chiral symmetry. And the same is true here, but with the important difference that the chiral symmetry in question is not $U(1)$ but instead just the surviving $\mathbb{Z}_{2N}$. The condensate breaks this to $\mathbb{Z}_{2N} \to \mathbb{Z}_2$, where $\mathbb{Z}_2 : \lambda \mapsto -\lambda$. But we know from our discussion in Section 2.1 that, when a discrete symmetry is spontaneously broken, it means that the theory has multiple, degenerate ground states. Indeed, that's the case here: $SU(N)$ gauge theory, with a single adjoint Weyl fermion, has N degenerate ground states, distinguished by the phase of the fermion condensate $\langle \lambda\lambda \rangle$.

4.3 't Hooft Anomalies

So far we have discussed two manifestations of the anomaly:

- For a gauge symmetry, the anomaly better cancel. Or else.
- A mixed anomaly between a global symmetry and gauge symmetry means that the global symmetry isn't.

But what if we have an anomaly just for a global symmetry? What are the consequences? From what we've discussed above, we know that the symmetry isn't conserved if we couple it to background gauge fields. But nothing compels us to do so. Indeed, if we're in the realm of particle physics then it's a little odd to do so because we're usually interested in relativistic physics in Minkowski space, while turning on a constant background electric or magnetic field breaks Lorentz invariance. So what else can we learn from this?

The answer is both subtle and powerful. The basic idea is that the anomaly provides a way to classify different quantum field theories: two quantum field theories with the same global symmetry group G_F can only be deformed into each other if they share the same anomaly. This is particularly useful when thinking about how theories flow to strong coupling, where we often don't know what happens. The anomalies provide constraints on what the theory can do. Such anomalies in global symmetries are referred to as *'t Hooft anomalies.*

We can flesh out this idea some more. Suppose that we've got some theory with a global symmetry that, for the sake of this argument, I'll call G_F. We can compute the anomaly for this symmetry. This is just a number – say $\sum_i Q_i^3$ if the symmetry is $G_F = U(1)$, or the generalisation if G_F is non-Abelian. As we will now argue, this anomaly is a way to characterise the theory and, provided that the symmetry is not broken, the anomaly remains unchanged under any deformation of the theory. In particular, the anomaly remains unchanged if the theory flows to strong coupling. In fact, this anomaly is one of the few handles that we have on the strong coupling physics of quantum field theories.

We will first explain the basic idea and then give a concrete example. Suppose that we have some quantum field theory – typically a non-Abelian gauge theory – that is weakly coupled in the UV, but flows to strong coupling in the IR. The most important example is, of course, QCD. We will abstractly call the UV theory $\mathcal{T}_{UV}$. We assume that it has some global symmetry G_F. This should be a true symmetry of the quantum theory meaning, in particular, that it has no mixed anomalies with the gauge symmetry.

This UV theory may have a 't Hooft anomaly for G_F. This anomaly is just a number. If G_F is Abelian, this anomaly is simply $\sum Q_i^3$ as in (4.7); if it is non-Abelian the anomaly is $\sum A(R_i)$ as in (4.17). Either way, we will denote this anomaly as $\mathcal{A}_{UV}$ and assume $\mathcal{A}_{UV} \neq 0$.

The theory now flows under RG to a theory $\mathcal{T}_{IR}$ in the IR which will typically be very different. For QCD this is the theory of mesons and baryons. For other quantum field theories, the infra-red physics may be quite mysterious. We have the following result:

Claim: Either the symmetry G_F is spontaneously broken, or the anomalies match meaning

$$\mathcal{A}_{UV} = \mathcal{A}_{IR} \ . \tag{4.80}$$

This is a wonderfully powerful result. If G_F is spontaneously broken then we necessarily have massless Goldstone bosons. But if G_F is unbroken then we must have massless fermions that reproduce the anomaly. This is known as *'t Hooft anomaly matching.*

Proof: The argument for 't Hooft anomaly matching is very slick. Suppose that $\mathcal{A}_{UV} \neq 0$ then we know from the discussion above that we're not allowed to couple G_F to dynamical gauge fields. That would lead to a sick theory.

To proceed, we introduce a bunch of extra massless Weyl fermions transforming under G_F. We call these *spectator fermions*. These won't interact directly with our original fields in $\mathcal{T}_{UV}$, but they are designed so that the total anomaly of the original fields and these new fermions vanishes:

$$\mathcal{A}_{UV} + \mathcal{A}_{\text{spectator}} = 0 \ . \tag{4.81}$$

Now that the anomaly cancels, there's nothing to stop us introducing dynamical gauge fields for G_F. We do so, but with a very (very!) small coupling constant.

Now let's go back to our original theory $\mathcal{T}_{UV}$. It will flow to strong coupling at some scale Λ_{QCD} and we'd like to understand the physics $\mathcal{T}_{IR}$ below this scale. If the gauge coupling for G_F is small enough, then this RG flow takes place entirely unaffected by the presence of the G_F gauge fields. This means that one of two things could have happened. It may be that the strong coupling dynamics of $\mathcal{T}_{UV}$ spontaneously breaks the symmetry G_F. (For example, as we've seen, this is expected to happen if we take G_F to be the chiral symmetry of QCD.) This was the first possibility of our claim. Alternatively, G_F may be unbroken at low-energies. In this case, we're left with $\mathcal{T}_{IR}$, together with the spectator fermions, all coupled to the G_F gauge fields. But this can only be consistent if

$$\mathcal{A}_{IR} + \mathcal{A}_{\text{spectator}} = 0 \ . \tag{4.82}$$

Clearly, this means that we must have $\mathcal{A}_{IR} = \mathcal{A}_{UV}$. □

4.3.1 Confinement Implies Chiral Symmetry Breaking

Anomaly matching has many uses. But the most important is a statement about QCD.

Recall from Section 3 that there are two strong coupling effects that arise in QCD. The first is confinement, the second chiral symmetry breaking. We will now use 't Hooft anomalies to argue that the former implies the latter.

We can work more generally with an $SU(N_c)$ gauge theory, coupled to N_f massless Dirac fermions q_i, each in the fundamental representation. This is a vector-like theory, so doesn't suffer any gauge anomaly. We've already seen that the $U(1)_A$ axial symmetry suffers an ABJ anomaly, so the global symmetry of the theory is

$$G_F = U(1)_V \times SU(N_f)_L \times SU(N_f)_R \ . \tag{4.83}$$

We want to compute the 't Hooft anomalies of this global symmetry group.

This is straightforward if we work in the UV where the theory is weakly coupled. In this case, we can just pretend that the fermions are essentially free and read off the result. There is no 't Hooft anomaly for $U(1)_V^3$, where the subscript 3 means that all three legs in the triangle diagram have $U(1)_V$ currents, because this is a vector-like symmetry. In contrast, there is a 't Hooft anomaly associated to the chiral, $SU(N_f)$ factors. In fact, there are two. The first is the purely non-Abelian anomaly,

$$[SU(N_f)_L]^3 : \quad \mathcal{A} = \sum A(\Box) = N_c \; . \tag{4.84}$$

Here the anomaly arises because each left-handed quark q_L transforms in the fundamental $\Box$ of $SU(N_f)_L$ and $A(\Box) = 1$. But the quarks also come with a colour index which means that there are N_c such fermions. (More generally, you have to sum over any other indices that the fermion carries that aren't themselves involved in the anomaly.) Hence the result $\mathcal{A} = N_c A(\Box) = N_c$. There is a similar anomaly for $SU(N_f)_R$.

In addition, there is a mixed 't Hooft anomaly between $U(1)_V$ and $SU(N_f)$. This is

$$[SU(N_f)_L]^2 \times U(1)_V : \quad \mathcal{A}' = \sum qI(\Box) = N_c \tag{4.85}$$

which again simply counts the number of quark colours.

Now the question is: what happens in the infra-red? For suitably low N_f, we've seen in Section 3 that we expect the chiral symmetry G_F to be broken down to $U(1)_V \times SU(N_f)_{\text{diag}}$, but proving this remains an open problem. Here we will shed some insight. We will assume that the theory confines and, moreover, that in the infra-red, the physics is described by weakly interacting mesons and baryons. (This is in contrast to the conformal field theories that we see at larger N_f.) In such a situation, 't Hooft anomaly matching shows that the chiral symmetry *must* be broken.

Here is the argument. Suppose that G_F is unbroken in the infra-red. Then there must be massless fermions around that can reproduce the anomalies $\mathcal{A}$ and $\mathcal{A}'$. Moreover, by assumption, these massless fermions must be bound states of quarks, either mesons or baryons.

Mesons certainly can't do the job because these are bosons. Baryons, meanwhile, contain N_c quarks so these too are bosons when N_c is even. This is telling us that when N_c is even, a confining theory contains no fermions at low-energies and so certainly can't reproduce the anomalies. We learn that chiral symmetry breaking must occur when N_c is even.

What about N_c odd? Now baryons are fermions. Is it possible that some of these baryons could be massless and reproduce the 't Hooft anomalies? Of course, this doesn't happen in our world: the simplest baryons are the proton and neutron which are certainly not massless. But might it be a theoretical possibility? The answer, it turns out, is no. The basic argument is to figure out what representations of G_F the putative massless baryons must sit in, and then to show that there's no possible combination of baryons that can reproduce the 't Hooft anomalies $\mathcal{A}$ and $\mathcal{A}'$. This means that if QCD confines into weakly interacting colour singlets, then chiral symmetry is necessarily broken. We now present this argument in more detail.

The Representations of Massless Baryons

It turns out that we can make the argument for any number of colours N_c, but it is simplest if we restrict to $N_c = 3$. Which, happily, is the case we care about for QCD.

If the $SU(3)$ gauge group confines, then any massless fermion must be a colour singlet. The only possibility is baryons, comprised of three quarks. Each constituent quark can be either left-handed or right-handed. Under $SU(N_f)_L \times SU(N_f)_R \subset G_F$, the left-handed fermions transform as $(\mathbf{N}_f, \mathbf{1})$, while the right-handed fermions transform as $(\mathbf{1}, \mathbf{N}_f)$. Both of these Weyl fermions have charge $+1$ under $U(1)_V$.

We've already seen in Section 3.3 that baryons in QCD can have either spin $\frac{1}{2}$ or spin $\frac{3}{2}$, depending on how the constituent spins of the quarks are aligned. You might imagine that the same can be true for our putative massless baryons, but there is a theorem by Weinberg and Witten which says that one cannot form massless bound states with helicity $\lambda \geq 1$. So if the massless baryons above do indeed form then they must have helicity $\pm\frac{1}{2}$.

So what representations of $G_F = U(1)_V \times SU(N_f)_L \times SU(N_f)_R$ do the colour singlet baryons sit in? Well, to form a helicity $\frac{1}{2}$ baryon, we should contract the spin indices of two fermions of the same handedness, and then leave the third spinor degree of freedom hanging. There are different ways to do this. For example, we could have three left-handed spinors, so that the indices combine to leave us with a left-handed spinor. In this case, the resulting bound state will transform in one of three possible representations of the $SU(N_f)_L$ symmetry which, in the language of Young diagrams, read

$$\begin{array}{|c|c|c|}\hline L & L & L \\ \hline\end{array} \quad , \quad \begin{array}{|c|}\hline L \\ \hline L \\ \hline L \\ \hline\end{array} \quad , \quad \begin{array}{|c|c|}\hline L & L \\ \hline L & \multicolumn{1}{c}{} \\ \cline{1-1}\end{array} \tag{4.86}$$

The first representation is the totally symmetric, the second the totally anti-symmetric, and the final is some representation whose name I don't know. Some properties of these representations are listed in Table 9. We've labelled the boxes with L to remind us that these are constructed out of three left-handed quarks.

But, alternatively, we could get ourselves a left-handed spinor by combining the indices on two right-handed spinors, and then leaving the final left-handed spinor hanging. These baryons would transform in representations of $SU(N_f)_L \times SU(N_f)_R$ that take the form

$$\begin{array}{|c|}\hline L\\\hline\end{array} \otimes \begin{array}{|c|c|}\hline R & R\\\hline\end{array} \quad , \quad \begin{array}{|c|}\hline L\\\hline\end{array} \otimes \begin{array}{|c|}\hline R\\\hline R\\\hline\end{array} \tag{4.87}$$

Each of these transforms in the fundamental $\square$ of $SU(N_f)_L$, while the first transforms in the symmetric $\square\!\square$ of $SU(N_f)_R$ and the second transforms in the anti-symmetric $\begin{smallmatrix}\square\\\square\end{smallmatrix}$ of $SU(N_f)_R$.

So (4.86) and (4.87) are the possible representations for massless left-handed baryons. But there's also the option for massless right-handed baryons which we get by simply exchanging $L \leftrightarrow R$,

$$\begin{array}{|c|c|c|}\hline R & R & R\\\hline\end{array} \quad , \quad \begin{array}{|c|}\hline R\\\hline R\\\hline R\\\hline\end{array} \quad , \quad \begin{array}{|c|c|}\hline R & R\\\hline R & \multicolumn{1}{|c}{}\\\cline{1-1}\end{array} \quad , \quad \begin{array}{|c|c|}\hline L & L\\\hline\end{array} \otimes \begin{array}{|c|}\hline R\\\hline\end{array} \quad , \quad \begin{array}{|c|}\hline L\\\hline L\\\hline\end{array} \otimes \begin{array}{|c|}\hline R\\\hline\end{array} \tag{4.88}$$

So these are our options for forming massless baryons. Now the question is: which combination of these massless baryons will reproduce the 't Hooft anomalies of the UV theory?

We started with a vector-like theory, in which all fermions came in left/right pairs to make a Dirac fermion. So it seems reasonable to assume that we end up with a vector-like theory. Indeed, a strong constraint comes from the $U(1)_V^3$ anomaly which vanishes. We will assume that we reproduce this by taking left/right pairs, so that if one of the massless baryons in (4.86) or (4.87) arises in the spectrum, then so too does its counterpart from (4.88).

So now we have a well-defined problem on our hands. We take some number $p_\alpha \geq 0$ of each of the $\alpha = 1, 2, 3, 4, 5$ possible baryons above and then see which values of p_α can reproduce the 't Hooft anomalies $\mathcal{A}$ and $\mathcal{A}'$.

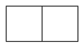
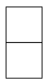
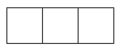
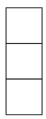
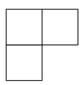
R	$\dim(R)$	$I(R)$	$A(R)$
	N_f	1	1
	$\frac{1}{2}N_f(N_f+1)$	N_f+2	N_f+4
	$\frac{1}{2}N_f(N_f-1)$	N_f-2	N_f-4
	$\frac{1}{6}N_f(N_f+1)(N_f+2)$	$\frac{1}{2}(N_f+2)(N_f+3)$	$\frac{1}{2}(N_f+3)(N_f+6)$
	$\frac{1}{6}N_f(N_f-1)(N_f-2)$	$\frac{1}{2}(N_f-2)(N_f-3)$	$\frac{1}{2}(N_f-3)(N_f-6)$
	$\frac{1}{3}N_f(N_f^2-1)$	N_f^2-3	N_f^2-9

Table 9. Properties of some representations of $SU(N_f)$

Actually, at this point a subtlety raises its head. Above, we confidently asserted that (4.86) and (4.87) were left-handed spinors, while (4.88) were right-handed spinors. That's certainly true if we're dealing with a weakly interacting theory where we can just read off the representations from contracting indices. But things could be more complicated in a strongly interacting theory. In particular, it may be that a massless spin 1 gluon binds with one of the baryons to flip its helicity from $+\frac{1}{2}$ to $-\frac{1}{2}$. So it may be that some of the baryons that we listed in (4.86) and (4.87) are actually right-handed instead of left-handed.

In fact, it's easy to take this subtlety into account. We'll assign an *index*, $p_\alpha \in \mathbf{Z}$, with $\alpha = 1, \ldots, 5$ to each of the five baryons in (4.86) and (4.87). The magnitude $|p_\alpha|$ denotes the number of species of baryon that arise in the massless spectrum. If these baryons are left-handed then we take $p_\alpha > 0$; if they are right-handed then we take $p_\alpha < 0$. Our task is to find which values of p_α will satisfy anomaly matching and reproduce (4.84) and (4.85).

Next, we need a little group theory. For a representation $\mathbf{R}$ of $SU(N_f)$, we will need to know the dimension $\dim(R)$, the anomaly coefficient $A(R)$, as well as the Dynkin index $I(R)$ that we already met in (4.24). The relevant data is shown in Table 9.

We can now compute the infra-red anomalies, assuming that we have p_α massless baryons of each type. For $SU(N_f)^3_L$ with $N_f \geq 3$, the anomaly is

$$\begin{aligned}\mathcal{A} = & \frac{1}{2}(N_f+3)(N_f+6)p_1 + \frac{1}{2}(N_f-3)(N_f-6)p_2 + (N_f^2-9)p_3 \\ & + \left(\frac{1}{2}N_f(N_f+1) - N_f(N_f+4)\right)p_4 + \left(\frac{1}{2}N_f(N_f-1) - N_f(N_f-4)\right)p_5 \ . \end{aligned} \tag{4.89}$$

Note that the baryons with numbers p_4 and p_5 arise from tensor products and have two terms. For example, for p_4 the first term comes from the left-handed baryon $\boxed{L}\otimes\boxed{R}\boxed{R}$, and the second — with the minus sign — from the right-handed baryon $\boxed{R}\otimes\boxed{L}\boxed{L}$.

Meanwhile, for the $SU(N_f)^2 \times U(1)_V$ anomaly, each baryon has charge 3 under the $U(1)_V$. Dividing through by this, we get a contribution proportional to the Dynkin index $I(R)$,

$$\begin{aligned}\frac{\mathcal{A}'}{3} = & \frac{1}{2}(N_f+2)(N_f+3)p_1 + \frac{1}{2}(N_f-2)(N_f-3)p_2 + (N_f^2-3)p_3 \\ & + \left(\frac{1}{2}N_f(N_f+1) - N_f(N_f+2)\right)p_4 + \left(\frac{1}{2}N_f(N_f-1) - N_f(N_f-2)\right)p_5 \ . \end{aligned} \tag{4.90}$$

To match the anomalies, we need to find p_α such that $\mathcal{A} = \mathcal{A}' = 3$.

To start, let's look at $N_f = 3$. Anomaly matching gives

$$\mathcal{A} = 27p_1 - 15p_4 = 3 \quad \text{and} \quad \frac{\mathcal{A}'}{3} = 15p_1 + 6p_3 - 9p_4 = 1 \ . \tag{4.91}$$

We can immediately see that there can be no solutions to the second of these equations since $\mathcal{A}'/3$ in the infra-red theory is necessarily a multiple of 3 and cannot reproduce the ultra-violet anomaly $\mathcal{A}'/3 = 1$. We learn that $G = SU(3)$ gauge theory with $N_f = 3$ massless fermions must spontaneously break the G_F flavour symmetry, as long as the theory confines. You can check that the same argument works whenever N_f is a multiple of 3.

Decoupling Massive Quarks

When N_f is not a multiple of 3, things are not quite so simple. Indeed, we will need one further ingredient to complete the argument. To see this, let's look at the anomaly matching conditions for $G = SU(3)$ gauge theory with $N_f = 4$ flavours. They are:

$$\begin{aligned}\mathcal{A} &= 35p_1 - p_2 + 7p_3 - 22p_4 + 6p_5 = 3 \\ \frac{\mathcal{A}'}{3} &= 21p_1 + p_2 + 13p_3 - 14p_4 - 2p_5 = 1 \ . \end{aligned} \tag{4.92}$$

Now there are solutions. For example $p_2 = 3$ and $p_5 = 1$ with $p_1 = p_3 = p_4 = 0$ does the job. This corresponds to four massless baryons in the representations

$$[3(\bar{\mathbf{4}}, \mathbf{1}) \oplus (\mathbf{4}, \mathbf{6})]_L \oplus [3(\mathbf{1}, \bar{\mathbf{4}}) \oplus (\mathbf{6}, \mathbf{4})]_R \tag{4.93}$$

where the L and R subscripts denote the chirality of these Weyl spinors. Note that the left-handed baryons now transform under both $SU(4)_L$ and $SU(4)_R$ of the chiral flavour symmetry.

Naively, the existence of the solution (4.93) suggests that there is a phase with massless baryons and the chiral symmetry left unbroken. In fact, this cannot happen. The problem comes when we think about giving one of the quarks a mass. We will make the following assumption: when we give a quark a mass, any baryon that contains this quark will also become massive. It is not obvious that this happens, but it turns out to be true, a result known as the Vafa-Witten theorem. (It's one of a number of Vafa-Witten theorems.)

If we give one of the quarks a mass, then the symmetry group is explicitly broken to

$$G_F = U(1)_V \times SU(4)_L \times SU(4)_R \longrightarrow G'_F = U(1)_V \times SU(3)_L \times SU(3)_R \ . \tag{4.94}$$

What happens to our putative massless spectrum (4.93)? A little group decomposition tells us that under G'_F, the left-handed baryons transform as

$$3(\bar{\mathbf{4}}, \mathbf{1}) \rightarrow 3(\bar{\mathbf{3}}, \mathbf{1}) \oplus 3(\mathbf{1}, \mathbf{1}) \quad \text{and} \quad (\mathbf{4}, \mathbf{6}) \rightarrow (\mathbf{3}, \bar{\mathbf{3}}) \oplus (\mathbf{3}, \mathbf{3}) \oplus (\mathbf{1}, \bar{\mathbf{3}}) \oplus (\mathbf{1}, \mathbf{3}) \ . \tag{4.95}$$

The right-handed baryons have their $SU(3)_L \times SU(3)_R$ representations reversed. Of these, the $(\mathbf{1}, \mathbf{1})$ and the $(\mathbf{3}, \bar{\mathbf{3}})$ do not contain the massive fourth quark. By our assumption above, the remainder should become massive.

There is a further constraint however: all of the baryons that contain the fourth quark should become massive while leaving the surviving symmetry G'_F intact. This is because, as the mass becomes large, we should return to the theory with $N_f = 3$ flavours and the symmetry group G'_F. Although we now know that G'_F will ultimately be spontaneously broken by the strong coupling dynamics, this should happen at the scale $\Lambda_{\rm QCD}$ and not at the much higher scale of the fourth quark mass.

So what G'_F-singlet mass terms can we write for the baryons that contain the fourth quark? The left-handed spinors transform as $3(\bar{\mathbf{3}}, \mathbf{1}) \oplus (\mathbf{3}, \mathbf{3}) \oplus (\mathbf{1}, \bar{\mathbf{3}}) \oplus (\mathbf{1}, \mathbf{3})$. Of these, $(\mathbf{3}, \mathbf{3})$ can happily pair up with its right-handed counterpart. Further, one of the $(\bar{\mathbf{3}}, \mathbf{1})$ representations can pair up with the right-handed counterpart of $(\mathbf{1}, \bar{\mathbf{3}})$. But that still leaves us with $2(\bar{\mathbf{3}}, \mathbf{1}) \oplus (\mathbf{1}, \mathbf{3})$ and these have nowhere to go. Any mass term will necessarily break the remaining G'_F chiral symmetry and, as we argued above, this is unacceptable.

The result above should not be surprising. Any baryon that can get a mass without breaking G'_F does not change the 't Hooft anomaly for G'_F. If it were possible for all the baryons containing the massive quark to get a mass without breaking G'_F then the remaining massless baryons should satisfy anomaly matching. Yet we've seen that no such solution is possible for N_f.

The upshot of this argument is that there exists no solution to anomaly matching for $N_f = 4$ which is consistent with the decoupling of massive quarks. It is simple to extend this to all N_f and, indeed, to all N_c. 't Hooft anomaly matching then tells us that the chiral symmetry must be broken for all $N_c \geq 2$ and all $N_f \geq 3$.

Massless Baryons when $N_f = 2$?

There is one situation where it is possible to satisfy the anomaly matching: this is when $N_f = 2$. Since there is no triangle anomaly for $SU(2)$, we need only worry about the mixed $SU(2)_L^2 \times U(1)_V$ 't Hooft anomaly. We can import our results from earlier, although we should be a little bit careful: the anti-symmetric representation $\begin{array}{|c|}\hline R \\ \hline R \\ \hline\end{array}$ is the singlet of $SU(2)$ while the representation $\begin{array}{|c|c|}\hline L & L \\ \hline L & \\ \cline{1-1}\end{array}$ does not exist. The 't Hooft matching condition for gauge group $SU(3)$ now gives

$$\frac{\mathcal{A}'}{3} = 10p_1 - 5p_4 + p_5 = 1 \ . \tag{4.96}$$

This has many solutions. The simplest possibility is $p_1 = p_4 = 0$ and $p_5 = 1$. This means that we can match the anomaly if there are massless baryons which transform under $SU(2)_L \times SU(2)_R \times U(1)_V$ as

$$(\mathbf{2}, \mathbf{1})_3 \oplus (\mathbf{1}, \mathbf{2})_3 \ . \tag{4.97}$$

So for $N_f = 2$ we cannot use 't Hooft anomaly matching to rule out the existence of massless baryons. But it does not mean that they actually arise. To understand what happens, we need to look more carefully at the actual dynamics. The only real tool we have at our disposal is the lattice and this strongly suggests that even for $N_f = 2$ the chiral symmetry is broken and there are no massless baryons.

5 Electroweak Interactions

In this section, we turn to the weak force. But, in contrast to the strong force, if we want to understand the weak force then we really need to take a step back and take in the full structure of the Standard Model. This is because of the single most important feature of the weak force: it breaks parity.

The weak force breaks parity because it is a chiral gauge theory. This means that the gauge bosons interact differently with the left- and right-handed fermions. And, as we saw in Section 4, this forces us to grapple with the issue of gauge anomalies. And this, in turn, means that we must look at all the fermions to check consistency.

5.1 The Structure of the Standard Model

As we advertised in the introduction, the Standard Model is built on the gauge group

$$G = U(1) \times SU(2) \times SU(3)\ . \tag{5.1}$$

Here $U(1)$ is a force known as *hypercharge*. It is not electromagnetism. We will see how electromagnetism emerges from the Standard Model in Section 5.2 when we discuss electroweak symmetry breaking. The group for hypercharge is sometimes denoted as $U(1)_Y$ to distinguish it from electromagnetism. Correspondingly, the charges are usually denoted as Y.

There are a collection of fermions that are charged under this gauge group. The fermions for a single generation are:

	$U(1)$	$SU(2)$	$SU(3)$
Q_L	$\frac{1}{6}$	**2**	**3**
L_L	$-\frac{1}{2}$	**2**	**1**
u_R	$\frac{2}{3}$	**1**	**3**
d_R	$-\frac{1}{3}$	**1**	**3**
e_R	-1	**1**	**1**

(5.2)

What a weird collection of charges and representations! Why these? We'll answer this question below. First some comments.

The hypercharges are taken to be fractional. In some sense, this is merely a convention: we could just have well rescaled the charges so that Q_L has charge $+1$ and e_R charge -6. However, as we will see, the slightly odd fractional scaling above will reproduce our familiar convention for electric charges, in which the electron has charge -1, the up quark charge $\frac{2}{3}$ and the down quark charge $-\frac{1}{3}$.

Each of the fields transforms in either the fundamental representations of $SU(2)$ or $SU(3)$, denoted by $\mathbf{2}$ and $\mathbf{3}$ respectively, or in the singlet representation denoted by $\mathbf{1}$. This means that a bold $\mathbf{1}$ for a non-Abelian group is telling us that a field doesn't experience that force. (In contrast, a charge 1 for the $U(1)$ means that the field very much experiences that force; only charge 0 fields are neutral under $U(1)$.) We will sometimes denote the representations as $(\mathbf{R}_2, \mathbf{R}_3)_Y$, with $\mathbf{R}_2$ and $\mathbf{R}_3$ the representations of $SU(2)$ and $SU(3)$ respective, and Y the hypercharge. So, for example, the field Q_L transforms as $(\mathbf{2}, \mathbf{3})_{1/6}$.

Each of the fields in the table is a Weyl fermion, either left-handed or right-handed as denoted by the L and R subscripts. As we saw in Section 1, the conjugate fermion has the opposite handedness. So, for example, $\bar{Q}_L$ is a right-handed fermion that transforms as $(\mathbf{2}, \bar{\mathbf{3}})_{-1/6}$. (You might have thought that we should have written $\bar{\mathbf{2}}$ but the doublet of $SU(2)$ is pseudoreal, meaning that $\bar{\mathbf{2}} \cong \mathbf{2}$.)

The fermions that transform in the $\mathbf{3}$ of $SU(3)$ are the quarks that we met in Section 3.1. That statement is straightforwardly true for the right-handed quarks, which we've labelled u_R and d_R for the up quark and down quark. But there is just a single left-handed quark Q_L, albeit one that transforms in the $\mathbf{2}$ of $SU(2)$. Indeed, it's only the left-handed fermions that transform in the $\mathbf{2}$ of $SU(2)$. How should we think of the associated $a = 1, 2$ index? In other words, what's the analog of colour for the $SU(2)$ gauge group?

It turns out that the $SU(2)$ index is the names that we give to different particles. We often write the $SU(2)$ gauge structure of the left-handed fermions as

$$Q_L = \begin{pmatrix} u_L \\ d_L \end{pmatrix} \quad \text{and} \quad L_L = \begin{pmatrix} \nu_L \\ e_L \end{pmatrix} . \tag{5.3}$$

For Q_L, we interpret the $SU(2)$ doublet components as the left-handed up quark and left-handed down quark. For L_L, which we refer to as the left-handed lepton, we interpret the $SU(2)$ doublet as the left-handed neutrino ν_L and left-handed electron e_L.

This part of the story is very surprising. For the strong force, the $SU(3)$ gauge symmetry rotates different colours into each other. That's intuitive: we think that the red quark behaves very much like the blue quark. The analogous statement of (5.3) is that the $SU(2)$ gauge symmetry rotates, say, the left-handed neutrino into the left-handed electron. But these particles are nothing like each other, neither in mass nor their interactions! How can they possibly be related by a gauge symmetry? The answer, as we shall see, is that the Higgs field spontaneously breaks the $SU(2)$ gauge symmetry and, when the dust settles, leaves ν_L and e_L with very different properties. Indeed, at this point it's really misleading to write (5.3) because, before we talk about spontaneous symmetry breaking, there's really no sense in which the top component of Q_L is related to the up quark and the second component to the down quark. These properties will only manifest themselves after the Higgs mechanism (and, even then, only when we've made an arbitrary choice of vacuum structure).

Including the gauge degrees of freedom, there are a total of 15 fermions listed above. (The left-handed quark Q_L has $2 \times 3 = 6$. The total number is then 6+2+3+3+1=15.) It is possible that we should augment these 15 fermions with one additional one. This is a right-handed neutrino

	$U(1)$	$SU(2)$	$SU(3)$
ν_R	0	**1**	**1**

(5.4)

Unfortunately, we don't yet know if the right-handed neutrino ν_R exists or not! This is deeply unsatisfactory and the situation will hopefully change in the near future. The main reason for our ignorance is that, as shown above, ν_R doesn't interact with any of the forces. That makes it hard to detect and it is sometimes referred to as a *sterile neutrino*. It's interactions with the other particles are only through the Higgs field and it manifests itself in the way in which neutrinos get masses. We will describe this in Section 7. On aesthetic grounds, things look marginally nicer if ν_R exists, in the sense that each particle has a right-handed fermion and a left-handed counterpart sitting in the doublet of $SU(2)$. But this is not a particularly compelling argument and the situation should ultimately be determined by experiment.

There is one final field in the Standard Model: this is the Higgs boson which we denote as H. It is the only spin 0 particle in the Standard Model and has quantum numbers

	$U(1)$	$SU(2)$	$SU(3)$
H	$\frac{1}{2}$	**2**	**1**

(5.5)

These are the same quantum numbers as $\bar{L}_L$. As we will see, it turns out that there is something magical about this choice which allows the whole jigsaw to fit together.

5.1.1 Anomaly Cancellation

The Standard Model is a chiral gauge theory. The first thing that we have to do is check that it makes sense! As we've seen in Section 4.1, there are a number of stringent consistency checks that any chiral gauge theory must pass. You will probably not be surprised to hear that the Standard Model, and hence our universe, is mathematically consistent. But it should give you a warm fuzzy feeling to check this explicitly.

Only the charged fermions (5.2) contribute to the anomalies. We can go through each anomaly in turn and check that it cancels. Some of these are straightforward. For example, for the $SU(3)^3$ anomaly, we require

$$\sum_{\rm left-handed} A(R) = \sum_{\rm right-handed} A(R)\ . \tag{5.6}$$

All fermions are either singlets with $A(\mathbf{1}) = 0$ or sit in the fundamental representation with $A(\mathbf{3}) = 1$. Clearly there are two right-handed quarks u_R and d_R. There is only the single left-handed quark Q_L but, when computing the anomaly, we should sum over the $SU(2)$ gauge index. (From the perspective of the $SU(3)$ gauge field, the anomaly doesn't know if Q_L is two distinct fields, or a single field transforming as an $SU(2)$ doublet.) The upshot is that $\sum A(R) = 2$ for both left-handed and right-handed quarks.

As we mentioned in Section 4.1, there is no perturbative $SU(2)^3$ anomaly, only the more subtle Witten anomaly which means that we must have an even number of $SU(2)$ doublets. This is achieved because there are three in Q_L (when computing the $SU(2)$ anomaly, we should sum over $SU(3)$ indices) and a single doublet in L_L. Note that the Witten anomaly ties together the quarks and leptons: the theory doesn't make sense with just Q_L alone: we must also have L_L.

The remaining gauge anomalies involve the $U(1)$ factor and are even more intricate. The $U(1)^3$ anomaly requires matching between the sum of the cubes of the charges

$$\sum_{\rm left-handed} Y^3 = \sum_{\rm right-handed} Y^3\ . \tag{5.7}$$

As above, in all of these calculations, we must remember to multiply by the dimension of the representation of the non-Abelian factors. We have

$$\begin{aligned} \sum_{\rm left-handed} Y^3 &= 6 \times \left(\frac{1}{6}\right)^3 + 2 \times \left(-\frac{1}{2}\right)^3 = -\frac{2}{9} \\ \sum_{\rm right-handed} Y^3 &= 3 \times \left(\frac{2}{3}\right)^3 + 3 \times \left(-\frac{1}{3}\right)^3 + (-1)^3 = -\frac{2}{9}\ . \end{aligned} \tag{5.8}$$

So that works.

We also have to check the mixed anomalies between two factors of the gauge group. The $SU(2)^2 \times U(1)$ anomaly requires that

$$\sum_{\text{left-handed}} Y = \sum_{\text{right-handed}} Y \tag{5.9}$$

where the sum is only over those fermions that sit in the **2** of $SU(2)$. This is satisfied by virtue of

$$SU(2)^2 \times U(1): \quad 3 \times \left(\frac{1}{6}\right) + \left(-\frac{1}{2}\right) = 0\ . \tag{5.10}$$

Meanwhile, the $SU(3)^2 \times U(1)$ anomaly requires that (5.9) holds when we sum over the quarks that sit in the **3** of $SU(3)$ which also holds, by virtue of

$$SU(3)^2 \times U(1): \quad 2 \times \left(\frac{1}{6}\right) = \frac{2}{3} - \frac{1}{3}\ . \tag{5.11}$$

Finally, we want to be able to couple our theory consistently to gravity. This requires that (5.9) holds when we sum over all fermions. We have

$$\begin{aligned} \sum_{\text{left-handed}} Y &= 6 \times \frac{1}{6} + 2 \times \left(-\frac{1}{2}\right) = 0 \\ \sum_{\text{right-handed}} Y &= 3 \times \frac{2}{3} + 3 \times \left(-\frac{1}{3}\right) - 1 = 0\ . \end{aligned} \tag{5.12}$$

The sum over left- and right-handed fermions vanish individually, which is stronger than is needed for anomaly cancellation. We see that, happily, our universe makes sense. This is cause for celebration.

This also explains a statement that we made in the introduction to these lectures: there is a remarkable unification in the Standard Model. It is not the usual kind of unification, where seemingly different phenomena are seen to have the same underlying cause. Instead, it is something more subtle: the quarks, electron and neutrino are unified by the need for mathematical consistency. If you remove one of them, then the delicate cancellations that we saw above fail. The whole collection of fermions (5.2) is needed for our theory to hold together.

There are variations on this calculation that we could play. For example, we could keep the matter content of (5.2), but allow the hypercharges Y to be arbitrary. We could then ask: what values of hypercharge are consistent? It turns out that there are two possibilities: one gives a non-chiral theory, the other is (up to rescaling) the world you inhabit. You will be offered the opportunity to do this, and a related calculation, on the examples sheet.

5.1.2 Yukawa Interactions

Because the Standard Model is a chiral gauge theory, it's not possible to write down gauge invariant mass terms for the fermions. That would need left- and right-handed fermions to transform the same way under the gauge symmetry which, as shown in (5.2), they do not. This is striking: it means that all the fermions in the Standard Model are naturally massless! Needless to say, that's not our everyday experience and something must happen along the way to change the situation.

What happens is that all fermions interact with the Higgs boson. We will tell the full story of how they get mass later, but for now we can look at the form of these interactions.

The Higgs field plays no role in the anomaly cancellation story above. But its quantum numbers $(\mathbf{2},\mathbf{1})_{1/2}$ under the gauge group restrict its couplings to the fermions. And, as we now show, the quantum numbers (5.5) are such that it can couple to *all* fermions through Yukawa couplings.

First, consider the quarks. We can form fermion bilinears which are Lorentz scalars and singlets under $SU(3)$ by contracting $\bar{Q}_L$ with either u_R or d_R. From (5.2), we see that $\bar{Q}_L u_R$ has gauge quantum numbers $(\bar{\mathbf{2}},\mathbf{1})_{+1/2}$ and $\bar{Q}_L d_r$ has $(\bar{\mathbf{2}},\mathbf{1})_{-1/2}$. We can then form a gauge invariant Yukawa term by contracting these with either H or $H^\dagger$.

At this point, we need to say a word about how the $SU(2)$ representations combine. Given two $SU(2)$ vectors x^a and z^a, with $a = 1, 2$, each of which transform in the $\mathbf{2}$ of $SU(2)$, there are two ways to form singlets. We can either write $x^\dagger z = \bar{x}_a z^a$ which is what we would call a "meson" in the context of the strong force. Or we can write $xz = \epsilon_{ab} x^a z^b$, making use of the epsilon symbol. This is what we would call a "baryon" for the strong force. The group $SU(2)$ is special because you get to make singlets in two different ways out of just two vectors. More mathematically, this is the statement that the representation $\mathbf{2}$ is pseudoreal because given x^a in the $\mathbf{2}$, we can always form $\epsilon_{ab}x^b$ in the $\bar{\mathbf{2}}$.

For us, $\bar{Q}_L$ naturally sits in the $\bar{\mathbf{2}}$ so we can contract it with H which sits in the $\mathbf{2}$. But we need that epsilon symbol if we are to contract it with $H^\dagger$. To this end, it's common to define

$$\tilde{H}^a = \epsilon^{ab} H_b^\dagger \tag{5.13}$$

with $a, b = 1, 2$ the $SU(2)$ gauge indices. We can then construct gauge invariant Yukawa couplings with the quarks of the form

$$\mathcal{L}_{\text{Yuk}} = -y^d\, \bar{Q}_L H d_R - y^u\, \bar{Q}_L \tilde{H} u_R + \text{h.c.}\ . \tag{5.14}$$

Here y^d and y^u are Yukawa coupling constants. Both of these terms are neutral under hypercharge and, by construction, also singlets under $SU(2) \times SU(3)$.

We can also write down Yukawa interactions with the leptons. This time we have the bilinears $\bar{L}_L e_R$ with quantum numbers $(\bar{\mathbf{2}}, \mathbf{1})_{-1/2}$ and, if the right-handed neutrino exists, $\bar{L}_L \nu_R$ with quantum numbers $(\bar{\mathbf{2}}, \mathbf{1})_{+1/2}$. We can see that both of these also have gauge invariant Yukawa interactions with the Higgs

$$\mathcal{L}_{\text{Yuk}} = -y^e \, \bar{L}_L H e_R - y^\nu \, \bar{L}_L \tilde{H} \nu_R + \text{h.c.} \; . \tag{5.15}$$

Again, y^e and y^ν are Yukawa coupling constants and, as above, the neutrino Yukawa term with $H^\dagger$ should have the $SU(2)$ gauge indices contracted with an ϵ_{ab}.

If we have a right-handed neutrino ν_R, then there is one further term that we can add. This is a Majorana mass of the kind we introduced in (1.59). It's possible only for ν_R because this fermion isn't charged under the gauge group,

$$\mathcal{L}_{\text{Maj}} = M \, \nu_R \nu_R + \text{h.c.} \; . \tag{5.16}$$

We'll discuss this further in Section 7.

5.1.3 Three Generations

For reasons that remain mysterious, the pattern of fermions presented in (5.2) is repeated twice over. Mathematically, it is straightforward to incorporate this: we just add a flavour index $i = 1, 2, 3$ to each of the fermions. We ascribe these additional fields names that we met in the introduction: strange and charm, and bottom and top for the quarks. We write these as

$$\begin{aligned} d^i_R &= \left\{ d_R \; , \; s_R \; , \; b_R \; \right\} \quad &: \quad (\mathbf{1}, \mathbf{3})_{-1/3} \\ u^i_R &= \left\{ u_R \; , \; c_R \; , \; t_R \; \right\} \quad &: \quad (\mathbf{1}, \mathbf{3})_{2/3} \end{aligned} \tag{5.17}$$

and, writing the $SU(2)$ doublets explicitly,

$$Q^i_L = \left\{ \begin{pmatrix} u_L \\ d_L \end{pmatrix} \; , \; \begin{pmatrix} c_L \\ s_L \end{pmatrix} \; , \; \begin{pmatrix} t_L \\ b_L \end{pmatrix} \right\} \quad : \quad (\mathbf{2}, \mathbf{3})_{1/6} \; . \tag{5.18}$$

As before, it's really premature to write this: the labelling only makes sense after we have taken into account the Higgs mechanism.

The names that we give to the leptons are the electron, muon, and tau. We write

$$e^i_R = \left\{ e_R \;,\; \mu_R \;,\; \tau_R \;\right\} \quad : \quad (\mathbf{1},\mathbf{1})_{-1} \tag{5.19}$$

and

$$L^i_L = \left\{ \begin{pmatrix} \nu^e_L \\ e_L \end{pmatrix} \;,\; \begin{pmatrix} \nu^\mu_L \\ \mu_L \end{pmatrix} \;,\; \begin{pmatrix} \nu^\tau_L \\ \tau_L \end{pmatrix} \right\} \quad : \quad (\mathbf{2},\mathbf{1})_{-1/2} \tag{5.20}$$

where, again, the labelling is premature and should be taken with a grain of salt before the Higgs mechanism does its thing.

Meanwhile, the Higgs itself is unaffected by this increase in generations: there is just a unique Higgs.

The fate of the right-handed neutrino ν_R is less certain. It seems tempting to also add an $i = 1, 2, 3$ index to this field too,

$$\nu^i_R = \left\{ \nu^e_R \;,\; \nu^\mu_R \;,\; \nu^\tau_R \;\right\} \quad : \quad (\mathbf{1},\mathbf{1})_0. \tag{5.21}$$

Because each of these is sterile, meaning uncharged under the gauge group, they do not interact directly with any of the forces, nor contribute to anomaly cancellation. It is quite possible there are no right-handed neutrinos or, indeed, any number!

As far as the gauge interactions are concerned, each generation experiences the same forces as the others. In particular, anomaly cancellation happens within each individual generation. There is, as far as we can tell, no necessity to introduce three generations rather than, say, one or seventeen.

The place where the additional generations really add a level of complexity and richness is in the Yukawa couplings. In contrast to the gauge couplings, the Yukawa couplings involve a great deal of inter-generational mixing. The most general Yukawa interactions that we can write down replace each of the coupling constants y^u, y^d, y^e and y^ν with 3×3 matrices,

$$\mathcal{L}_{\text{Yuk}} = -y^d_{ij}\, \bar{Q}^i_L H d^j_R - y^u_{ij}\, \bar{Q}^i_L \tilde{H} u^j_R - y^e_{ij}\, \bar{L}^i_L H e^j_R - y^\nu_{ij}\, \bar{L}^i_L \tilde{H} \nu^j_R + \text{h.c.}\;. \tag{5.22}$$

We will devote Section 6 to understanding the structure of these Yukawa couplings.

5.1.4 The Lagrangian

Usually when introducing a quantum field theory, the first thing that we do is write down an action. But that's not the case here: instead, we've discussed the symmetry structure of the theory. The reason this is sensible is because the symmetries are entirely sufficient to determine the structure of the action.

The game that we play is to write down all possible marginal and relevant terms. These terms must be Lorentz invariant and gauge invariant, but otherwise you write down anything that you want. Despite the plethora of fields, there isn't too much freedom. The full Lagrangian takes the form

$$\mathcal{L}_{\rm SM} = \mathcal{L}_{\rm gauge} + \mathcal{L}_{\rm fermi} + \mathcal{L}_{\rm Higgs} + \mathcal{L}_{\rm Yuk} \ . \tag{5.23}$$

The first two of these are simply kinetic terms for our fields. We will need to give our gauge fields some names. Back in Section 3, we already dubbed the $SU(3)$ gluon field strength $G_{\mu\nu}$. We will call the $SU(2)$ gauge field strength $W_{\mu\nu} = \partial_\mu W_\nu - \partial_\nu W_\mu - ig[W_\mu, W_\nu]$ and the $U(1)$ hypercharge field strength $B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu$. The gauge field kinetic terms are then

$$\mathcal{L}_{\rm gauge} = -\frac{1}{4} B_{\mu\nu}B^{\mu\nu} - \frac{1}{2}{\rm Tr}\, W_{\mu\nu}W^{\mu\nu} - \frac{1}{2}{\rm Tr}\, G_{\mu\nu}G^{\mu\nu} \ . \tag{5.24}$$

The kinetic terms for the fermions are

$$\begin{aligned}\mathcal{L}_{\rm fermi} = -i\sum_{i=1}^{3}\Big(&\bar{Q}^i_L\bar{\sigma}^\mu\mathcal{D}_\mu Q^i_L + \bar{L}^i_L\bar{\sigma}^\mu\mathcal{D}_\mu L^i_L + \bar{u}^i_R\sigma^\mu\mathcal{D}_\mu u^i_R \\ &+ \bar{d}^i_R\sigma^\mu\mathcal{D}_\mu d^i_R + \bar{e}^i_R\sigma^\mu\mathcal{D}_\mu e^i_R + \bar{\nu}^i_R\sigma^\mu\partial_\mu\nu^i_R\Big) \ .\end{aligned} \tag{5.25}$$

The exact form of these kinetic terms depends on the representation of the fermion field. So, for example, Q_L is charged under each of the three gauge fields and has kinetic term

$$\mathcal{D}_\mu Q_L = \partial_\mu Q_L - ig_s G_\mu Q_L - igW_\mu Q_L - \frac{i}{6}g' B_\mu Q_L \ . \tag{5.26}$$

There are similar expressions for all other fields. Buried within these covariant derivatives are the coupling constants: g_s for the $SU(3)$ strong force, g for the $SU(2)$ weak force, and g' for the $U(1)$ hypercharge.

The Lagrangian for the Higgs term includes both its kinetic term and potential

$$\mathcal{L}_{\rm Higgs} = \mathcal{D}_\mu H^\dagger\, \mathcal{D}^\mu H - \lambda\left(H^\dagger H - \frac{v^2}{2}\right)^2 \ . \tag{5.27}$$

The potential is written to emphasise that the minimum will lie away from $H = 0$. We will explore the consequences of this shortly. The Higgs kinetic term also follows from its gauge quantum numbers,

$$\mathcal{D}_\mu H = \partial_\mu H - igW_\mu H - \frac{i}{2}g' B_\mu H \ . \tag{5.28}$$

Finally, the Yukawa terms are given in (5.22).

We can start to count the parameters in the Standard Model. There are three gauge couplings, g_s, g and g', one for each gauge group. And there are two parameters λ and v^2 in the Higgs potential. Then there are the plethora of Yukawa couplings that we will explore further (and count!) in Section 6.

I've omitted two possible terms from the Lagrangian (5.23). One is the theta term for the strong force that we met in Section 3.4. This is omitted on the grounds that, experimentally, $\theta \approx 0$. Still, if we're accounting for parameters of the Standard Model then we should certainly include this one. The second term that I've omitted is the Majorana masses for the right-handed neutrinos, on the slightly weaker grounds that we don't know if they're there or not. We'll discuss this more in Section 7.

There's a lot of repetition in the Standard Model Lagrangian as written. I think that you could be forgiven for advertising it in the more compact form

$$\mathcal{L} = -\frac{1}{4}\sum_a F^a_{\mu\nu}F^{a\,\mu\nu} + i\sum_i \bar{\psi}_i\bar{\sigma}^\mu\mathcal{D}_\mu\psi_i + |\mathcal{D}H|^2 - V(H) - y\psi H\psi + \text{h.c.}\ . \quad (5.29)$$

Admittedly, there's a lot of heavy lifting going on in that $\sum_a$ and $\sum_i$. Still, it's remarkable that everything we know about the universe can be distilled in such a way.

You can sometimes find the Standard Model Lagrangian written out in full component form, in which case it looks something like what's shown in Figure 16. This is usually done by someone trying to convince you that the theory is inelegant (typically because they have their own wares to sell). This always strikes me as being deliberately obtuse, like writing out haiku in binary in an attempt to argue that its beauty is over-rated. The beauty of the Standard Model isn't in the form of the Lagrangian: it's in the consistency conditions inherent in anomaly cancellation that we have taken pains to explain in these lectures.

5.1.5 Global Symmetries

We've built the Standard Model around the gauge group $G = U(1) \times SU(2) \times SU(3)$. But it's natural to ask: what are the global symmetries of the Standard Model?

In the absence of Yukawa terms, this is an easy question to answer: the classical theory has a $U(3)^5$ global symmetry if there are no right-handed neutrinos, and a $U(3)^6$ global symmetry if there are right-handed neutrinos. Here the 3 corresponds to the three generations, and we get a global symmetry group acting on each of Q_L, L_L, u_R, d_R, e_R and (possibly) ν_R.

$$
\begin{gathered}
\mathcal{L}_{SM} = -\tfrac{1}{2}\partial_\nu g^a_\mu \partial_\nu g^a_\mu - g_s f^{abc}\partial_\mu g^a_\nu g^b_\mu g^c_\nu - \tfrac{1}{4}g_s^2 f^{abc} f^{ade} g^b_\mu g^c_\nu g^d_\mu g^e_\nu - \partial_\nu W^+_\mu \partial_\nu W^-_\mu - \\
M^2 W^+_\mu W^-_\mu - \tfrac{1}{2}\partial_\nu Z^0_\mu \partial_\nu Z^0_\mu - \tfrac{1}{2c_w^2} M^2 Z^0_\mu Z^0_\mu - \tfrac{1}{2}\partial_\mu A_\nu \partial_\mu A_\nu - igc_w(\partial_\nu Z^0_\mu (W^+_\mu W^-_\nu - \\
W^+_\nu W^-_\mu) - Z^0_\nu(W^+_\mu \partial_\nu W^-_\mu - W^-_\mu \partial_\nu W^+_\mu) + Z^0_\mu(W^+_\nu \partial_\nu W^-_\mu - W^-_\nu \partial_\nu W^+_\mu)) - \\
igs_w(\partial_\nu A_\mu (W^+_\mu W^-_\nu - W^+_\nu W^-_\mu) - A_\nu(W^+_\mu \partial_\nu W^-_\mu - W^-_\mu \partial_\nu W^+_\mu) + A_\mu(W^+_\nu \partial_\nu W^-_\mu - \\
W^-_\nu \partial_\nu W^+_\mu)) - \tfrac{1}{2}g^2 W^+_\mu W^-_\mu W^+_\nu W^-_\nu + \tfrac{1}{2}g^2 W^+_\mu W^-_\nu W^+_\mu W^-_\nu + g^2 c_w^2 (Z^0_\mu W^+_\mu Z^0_\nu W^-_\nu - \\
Z^0_\mu Z^0_\mu W^+_\nu W^-_\nu) + g^2 s_w^2 (A_\mu W^+_\mu A_\nu W^-_\nu - A_\mu A_\mu W^+_\nu W^-_\nu) + g^2 s_w c_w (A_\mu Z^0_\nu (W^+_\mu W^-_\nu - \\
W^+_\nu W^-_\mu) - 2A_\mu Z^0_\mu W^+_\nu W^-_\nu) - \tfrac{1}{2}\partial_\mu H \partial_\mu H - 2M^2 \alpha_h H^2 - \partial_\mu \phi^+ \partial_\mu \phi^- - \tfrac{1}{2}\partial_\mu \phi^0 \partial_\mu \phi^0 - \\
\beta_h \left(\tfrac{2M^2}{g^2} + \tfrac{2M}{g} H + \tfrac{1}{2}(H^2 + \phi^0\phi^0 + 2\phi^+\phi^-)\right) + \tfrac{2M^4}{g^2}\alpha_h - \\
g\alpha_h M \left(H^3 + H\phi^0\phi^0 + 2H\phi^+\phi^-\right) - \\
\tfrac{1}{8}g^2 \alpha_h \left(H^4 + (\phi^0)^4 + 4(\phi^+\phi^-)^2 + 4(\phi^0)^2\phi^+\phi^- + 4H^2\phi^+\phi^- + 2(\phi^0)^2 H^2\right) - \\
gMW^+_\mu W^-_\mu H - \tfrac{1}{2}g\tfrac{M}{c_w^2} Z^0_\mu Z^0_\mu H - \\
\tfrac{1}{2}ig\left(W^+_\mu(\phi^0\partial_\mu\phi^- - \phi^-\partial_\mu\phi^0) - W^-_\mu(\phi^0\partial_\mu\phi^+ - \phi^+\partial_\mu\phi^0)\right) + \\
\tfrac{1}{2}g\left(W^+_\mu(H\partial_\mu\phi^- - \phi^-\partial_\mu H) + W^-_\mu(H\partial_\mu\phi^+ - \phi^+\partial_\mu H)\right) + \tfrac{1}{2}g\tfrac{1}{c_w}(Z^0_\mu(H\partial_\mu\phi^0 - \phi^0\partial_\mu H) + \\
M\left(\tfrac{1}{c_w}Z^0_\mu \partial_\mu\phi^0 + W^+_\mu\partial_\mu\phi^- + W^-_\mu\partial_\mu\phi^+\right) - ig\tfrac{s_w^2}{c_w}MZ^0_\mu(W^+_\mu\phi^- - W^-_\mu\phi^+) + igs_w MA_\mu(W^+_\mu\phi^- - \\
W^-_\mu\phi^+) - ig\tfrac{1-2c_w^2}{2c_w}Z^0_\mu(\phi^+\partial_\mu\phi^- - \phi^-\partial_\mu\phi^+) + igs_w A_\mu(\phi^+\partial_\mu\phi^- - \phi^-\partial_\mu\phi^+) - \\
\tfrac{1}{4}g^2 W^+_\mu W^-_\mu \left(H^2 + (\phi^0)^2 + 2\phi^+\phi^-\right) - \tfrac{1}{8}g^2\tfrac{1}{c_w^2}Z^0_\mu Z^0_\mu\left(H^2 + (\phi^0)^2 + 2(2s_w^2 - 1)^2\phi^+\phi^-\right) - \\
\tfrac{1}{2}g^2\tfrac{s_w^2}{c_w}Z^0_\mu\phi^0(W^+_\mu\phi^- + W^-_\mu\phi^+) - \tfrac{1}{2}ig^2\tfrac{s_w^2}{c_w}Z^0_\mu H(W^+_\mu\phi^- - W^-_\mu\phi^+) + \tfrac{1}{2}g^2 s_w A_\mu\phi^0(W^+_\mu\phi^- + \\
W^-_\mu\phi^+) + \tfrac{1}{2}ig^2 s_w A_\mu H(W^+_\mu\phi^- - W^-_\mu\phi^+) - g^2\tfrac{s_w}{c_w}(2c_w^2 - 1)Z^0_\mu A_\mu\phi^+\phi^- - \\
g^2 s_w^2 A_\mu A_\mu \phi^+\phi^- + \tfrac{1}{2}ig_s\lambda^a_{ij}(\bar{q}^\sigma_i\gamma^\mu q^\sigma_j)g^a_\mu - \bar{e}^\lambda(\gamma\partial + m^\lambda_e)e^\lambda - \bar{\nu}^\lambda(\gamma\partial + m^\lambda_\nu)\nu^\lambda - \bar{u}^\lambda_j(\gamma\partial + \\
m^\lambda_u)u^\lambda_j - \bar{d}^\lambda_j(\gamma\partial + m^\lambda_d)d^\lambda_j + igs_w A_\mu\left(-(\bar{e}^\lambda\gamma^\mu e^\lambda) + \tfrac{2}{3}(\bar{u}^\lambda_j\gamma^\mu u^\lambda_j) - \tfrac{1}{3}(\bar{d}^\lambda_j\gamma^\mu d^\lambda_j)\right) + \\
\tfrac{ig}{4c_w}Z^0_\mu\{(\bar{\nu}^\lambda\gamma^\mu(1+\gamma^5)\nu^\lambda) + (\bar{e}^\lambda\gamma^\mu(4s_w^2 - 1 - \gamma^5)e^\lambda) + (\bar{d}^\lambda_j\gamma^\mu(\tfrac{4}{3}s_w^2 - 1 - \gamma^5)d^\lambda_j) + \\
(\bar{u}^\lambda_j\gamma^\mu(1 - \tfrac{8}{3}s_w^2 + \gamma^5)u^\lambda_j)\} + \tfrac{ig}{2\sqrt{2}}W^+_\mu\left((\bar{\nu}^\lambda\gamma^\mu(1+\gamma^5)U^{lep}{}_{\lambda\kappa}e^\kappa) + (\bar{u}^\lambda_j\gamma^\mu(1+\gamma^5)C_{\lambda\kappa}d^\kappa_j)\right) + \\
\tfrac{ig}{2\sqrt{2}}W^-_\mu\left((\bar{e}^\kappa U^{lep\dagger}{}_{\kappa\lambda}\gamma^\mu(1+\gamma^5)\nu^\lambda) + (\bar{d}^\kappa_j C^\dagger_{\kappa\lambda}\gamma^\mu(1+\gamma^5)u^\lambda_j)\right) + \\
\tfrac{ig}{2M\sqrt{2}}\phi^+\left(-m^\kappa_e(\bar{\nu}^\lambda U^{lep}{}_{\lambda\kappa}(1-\gamma^5)e^\kappa) + m^\lambda_\nu(\bar{\nu}^\lambda U^{lep}{}_{\lambda\kappa}(1+\gamma^5)e^\kappa\right) + \\
\tfrac{ig}{2M\sqrt{2}}\phi^-\left(m^\lambda_e(\bar{e}^\lambda U^{lep\dagger}{}_{\lambda\kappa}(1+\gamma^5)\nu^\kappa) - m^\kappa_\nu(\bar{e}^\lambda U^{lep\dagger}{}_{\lambda\kappa}(1-\gamma^5)\nu^\kappa\right) - \tfrac{g}{2}\tfrac{m^\lambda_\nu}{M}H(\bar{\nu}^\lambda\nu^\lambda) - \\
\tfrac{g}{2}\tfrac{m^\lambda_e}{M}H(\bar{e}^\lambda e^\lambda) + \tfrac{ig}{2}\tfrac{m^\lambda_\nu}{M}\phi^0(\bar{\nu}^\lambda\gamma^5\nu^\lambda) - \tfrac{ig}{2}\tfrac{m^\lambda_e}{M}\phi^0(\bar{e}^\lambda\gamma^5 e^\lambda) - \tfrac{1}{4}\bar{\nu}_\lambda M^R_{\lambda\kappa}(1-\gamma_5)\hat{\nu}_\kappa - \\
\tfrac{1}{4}\bar{\nu}_\lambda M^R_{\lambda\kappa}(1-\gamma_5)\hat{\nu}_\kappa + \tfrac{ig}{2M\sqrt{2}}\phi^+\left(-m^\kappa_d(\bar{u}^\lambda_j C_{\lambda\kappa}(1-\gamma^5)d^\kappa_j) + m^\lambda_u(\bar{u}^\lambda_j C_{\lambda\kappa}(1+\gamma^5)d^\kappa_j\right) + \\
\tfrac{ig}{2M\sqrt{2}}\phi^-\left(m^\lambda_d(\bar{d}^\lambda_j C^\dagger_{\lambda\kappa}(1+\gamma^5)u^\kappa_j) - m^\kappa_u(\bar{d}^\lambda_j C^\dagger_{\lambda\kappa}(1-\gamma^5)u^\kappa_j\right) - \tfrac{g}{2}\tfrac{m^\lambda_u}{M}H(\bar{u}^\lambda_j u^\lambda_j) - \\
\tfrac{g}{2}\tfrac{m^\lambda_d}{M}H(\bar{d}^\lambda_j d^\lambda_j) + \tfrac{ig}{2}\tfrac{m^\lambda_u}{M}\phi^0(\bar{u}^\lambda_j\gamma^5 u^\lambda_j) - \tfrac{ig}{2}\tfrac{m^\lambda_d}{M}\phi^0(\bar{d}^\lambda_j\gamma^5 d^\lambda_j) + \bar{G}^a\partial^2 G^a + g_s f^{abc}\partial_\mu\bar{G}^a G^b g^c_\mu + \\
\bar{X}^+(\partial^2 - M^2)X^+ + \bar{X}^-(\partial^2 - M^2)X^- + \bar{X}^0(\partial^2 - \tfrac{M^2}{c_w^2})X^0 + \bar{Y}\partial^2 Y + igc_w W^+_\mu(\partial_\mu\bar{X}^0 X^- - \\
\partial_\mu\bar{X}^+ X^0) + igs_w W^+_\mu(\partial_\mu\bar{Y}X^- - \partial_\mu\bar{X}^+ Y) + igc_w W^-_\mu(\partial_\mu\bar{X}^- X^0 - \\
\partial_\mu\bar{X}^0 X^+) + igs_w W^-_\mu(\partial_\mu\bar{X}^- Y - \partial_\mu\bar{Y}X^+) + igc_w Z^0_\mu(\partial_\mu\bar{X}^+ X^+ - \\
\partial_\mu\bar{X}^- X^-) + igs_w A_\mu(\partial_\mu\bar{X}^+ X^+ - \\
\partial_\mu\bar{X}^- X^-) - \tfrac{1}{2}gM\left(\bar{X}^+ X^+ H + \bar{X}^- X^- H + \tfrac{1}{c_w^2}\bar{X}^0 X^0 H\right) + \tfrac{1-2c_w^2}{2c_w}igM\left(\bar{X}^+ X^0\phi^+ - \bar{X}^- X^0\phi^-\right) + \\
\tfrac{1}{2c_w}igM\left(\bar{X}^0 X^-\phi^+ - \bar{X}^0 X^+\phi^-\right) + igMs_w\left(\bar{X}^0 X^-\phi^+ - \bar{X}^0 X^+\phi^-\right) + \\
\tfrac{1}{2}igM\left(\bar{X}^+ X^+\phi^0 - \bar{X}^- X^-\phi^0\right) .
\end{gathered}
$$

Figure 16. If you want to write the Standard Model Lagrangian like this, then you should probably write the Einstein-Hilbert action by expanding out $\mathcal{L} = \sqrt{-g}R$ in terms of the metric $g_{\mu\nu}$.

But the Yukawa terms (5.22) break this symmetry. As we will see later, the values of the Yukawas are different for different generations, ultimately resulting in their different masses. There are some approximate symmetries remaining, like isospin or the eightfold way, but when the dust settles the classical theory has just two exact global symmetries. This is $U(1)_B \times U(1)_L$, corresponding to *baryon number* and *lepton number* respectively. The charges of the various fields under these two $U(1)'s$ are

	Q_L	L_L	u_R	d_R	e_R	ν_R
$U(1)_B$	$\frac{1}{3}$	0	$\frac{1}{3}$	$\frac{1}{3}$	0	0
$U(1)_L$	0	1	0	0	1	1

(5.30)

You can see that $U(1)_B$ acts only on quarks and $U(1)_L$ acts only on leptons. (In fact, $U(1)_B$ is essentially the same as the vector symmetry $U(1)_V$ that we saw when discussing QCD in Section 3.) The normalisation of $\frac{1}{3}$ for the charge of the quarks is just convention: it guarantees that the proton and neutron each have baryon number

+1. These symmetries $U(1)_B$ and $U(1)_L$ act the same on each generation. (The Yukawa interactions include couplings between generations which means that there's no global symmetry which acts on one generation, leaving the others untouched.)

Note that we didn't impose either of these global symmetries $U(1)_L$ and $U(1)_B$ from the outset. Instead, we just wrote down all possible terms consistent with the gauge symmetry and discovered that the end result has $U(1)_L \times U(1)_B$ as a global symmetry. In this sense, we view these symmetries as *accidental*. There is no particular reason to think that they survive to arbitrarily high energies (and, indeed, some reasonably good reasons that we shall explain shortly to think that they do not survive). This means, in particular, that if we were to add irrelevant terms to the Standard Model in an attempt to capture the high energy physics then we should include such terms that break $U(1)_B$ and $U(1)_L$.

ABJ Anomalies Revisited

As we saw in Section 4.2, just because a $U(1)$ symmetry is a good symmetry of the classical theory, doesn't mean that it is necessarily a symmetry of the quantum theory. This is because it may suffer from an ABJ anomaly. And, indeed, both $U(1)_B$ and $U(1)_L$ suffer ABJ anomalies. There is an ABJ anomaly with $SU(2)$ gauge group (because only left-handed fermions carry $SU(2)$ charge), and also with $U(1)$ hypercharge. For the latter, the anomaly for a single generation is given by

$$\sum_{\text{left}} BY^2 - \sum_{\text{right}} BY^2 = \frac{1}{3}\left(6 \times \left(\frac{1}{6}\right)^2 - 3 \times \left(\frac{2}{3}\right)^2 - 3 \times \left(-\frac{1}{3}\right)^2\right) = -\frac{1}{2} \qquad (5.31)$$

and

$$\sum_{\text{left}} LY^2 - \sum_{\text{right}} LY^2 = \left(2 \times \left(-\frac{1}{2}\right)^2 - (-1)^2\right) = -\frac{1}{2}\ . \qquad (5.32)$$

So neither $U(1)_B$ nor $U(1)_L$ are good symmetries of the quantum theory. However, in contrast to the ABJ anomaly of the axial symmetry of the strong force, these ABJ anomalies are associated to the gauge fields of the weak force. And the weak force is, as we shall see, weak. The upshot is that although neither $U(1)_B$ nor $U(1)_L$ are strictly symmetries of the Standard Model, they are both *extremely* good approximate symmetries. Indeed, neither has been observed to be violated!

We can quantify this. If we focus just on the $SU(2)$ anomaly, then the conservation of baryon number picks up a term analogous to (4.63),

$$\partial_\mu J^\mu_B = \frac{12 g^2}{8\pi^2} \mathrm{Tr}\, W^\star_{\mu\nu} W^{\mu\nu}\ . \qquad (5.33)$$

where the factor of 12 arises because there are four $SU(2)$ doublets in each of the three generations, and $3 \times 4 = 12$. There is a similar contribution from $B_{\mu\nu}$.

The kind of process that can violate baryon number is an electroweak instanton. There is a story of fermion zero modes that we will not tell but the end result is that electroweak instantons cannot, for example, allow a proton to decay into a positron: the proton is absolutely stable in the Standard Model. Instead, these instantons can allow a collection of three baryons to decay, where the "three" arises because it's the number of generations. This means, for example, that a ^{3}He nucleus could decay. But the decay is due to instantons and these come with a characteristic suppression factor of $e^{-8\pi^2/g^2}$, as in (3.120). For electroweak instantons, this turns out to give a lifetime of around 10^{173} years! (The age of our universe is roughly 10^{10} years.) That's why baryons seem stable.

All of which means that, for all practical purposes, both baryon number and lepton number are good symmetries. But, if you're a purist (and willing to wait 10^{173} years) then you should accept that neither are good symmetries.

Importantly, however, the ABJ anomalies for both $U(1)_B$ and $U(1)_L$ are the same. This is true both for the mixed anomaly with $U(1)_Y$ shown in (5.31) and (5.32) and also for the mixed anomaly with $SU(2)$. This means that the combination $B - L$ is non-anomalous. This is the one exact global symmetry of the Standard Model.

We still have to check if there is a gravitational contribution to the $B - L$ anomaly. You can check that this vanishes only if there is a right-handed neutrino.

The Weak Theta Term

For the strong force, we can write down a theta term. As we discussed in Section 3.4, this leads to a mystery because, experimentally, $\theta \approx 0$ and we don't know why. This is the strong CP problem.

What about the theta term for the other two gauge groups, $U(1)$ and $SU(2)$?

For Abelian gauge theories, we can write down a theta term but it doesn't affect the local dynamics, such as masses or cross-sections or decay rates. (This is essentially because there are no $U(1)$ instantons.) Instead, the effects are much more subtle. For example, this term would endow magnetic monopoles with electric charge through the Witten effect. We don't have any experimental insight into these features of the theory and so the $U(1)$ theta term remains unknown to us.

That leaves the $SU(2)$ theta term which takes the form

$$S_\theta = \frac{g^2\theta_W}{16\pi^2}\int d^4x\ \mathrm{Tr}\, W_{\mu\nu}{}^\star W^{\mu\nu}\ . \tag{5.34}$$

Is this another term that we should add to the Standard Model action? The answer is no. And the reason is because of the global $U(1)_L$ (or, equivalently $U(1)_B$) ABJ anomaly. As shown in (4.42), if we act with a $U(1)_L$ transformation of $e^{i\alpha L}$, where L is the charge of each fermion, then the anomaly can be re-interpreted as shifting the theta term

$$U(1)_B : \theta_W \to \theta_W + 3\alpha \tag{5.35}$$

where the factor of 3 comes from the existence of three generations. This means that the value of θ_W is unphysical and does not affect the physics. Said differently, we can always use the anomalous $U(1)_L$ symmetry to set $\theta_W = 0$. There is no weak CP problem. In contrast, this mechanism doesn't work for the strong force.

Black Holes

We have seen that the Standard Model has just a single $U(1)$ global symmetry, namely $B-L$. But the standard lore is that there are no global symmetries in the fundamental laws of physics. The main argument for this is black holes.

Black holes aren't black. Hawking taught us long ago that they slowly emit radiation due to quantum effects. While there is much that we don't understand about quantum gravity, the existence of Hawking radiation stands out as one of the few robust and trustworthy calculations that we can do. The prediction of this radiation follows from the known laws of physics and doesn't rely on any speculative ideas about what lies beyond.

If we wait long enough (and we're talking ridiculously long times here), then any black hole will eventually evaporate and disappear. So we can ask: what became of the stuff that we threw in?

First, the black hole can't destroy electric charge. If you throw, say, an electron into a black hole then the black hole itself now carries the electric charge. Moreover, this is visible outside of the event horizon because the black hole emits an electric field and we can detect the electric field by Gauss' law. (This is the Reissner-Nordström solution that we described in the lectures on General Relativity.) That electric field can't just disappear. So, as the black hole evaporates, it must eventually spit out a charged particle – maybe an electron, maybe an anti-proton – which carries the electric charge. The process of black hole evaporation must respect conservation of electric charge.

In contrast, there is nothing to prevent black holes from destroying baryons and leptons. When a black hole forms from the collapse of a star, it will typically contain around 10^{57} protons, and roughly the same number of electrons. But there's no way to detect the baryons from outside the black hole. Furthermore, as the black hole evaporates there's no reason that it should spit back these particles intact. In fact, the vast majority of the mass of a black hole will be emitted in gravitational and electromagnetic radiation rather than baryons or leptons. In this way, we expect black hole evaporation to respect neither baryon number nor lepton number conservation.

This means that, in a full theory of quantum gravity, one doesn't expect any global conservation laws, since one can always construct states in the theory in which the symmetry is violated. What does this mean for our parochial Standard Model? The usual answer is that we shouldn't view $B - L$ as something sacrosanct, but rather just a symmetry that emerges in the infra-red simply because there are no relevant or marginal operators that we can write down that violate it. When we get to high energy scales – and certainly by the time we get to the Planck scale – we expect it to be violated.

5.1.6 What is the Gauge Group of the Standard Model?

The title of this section seems a little daft. After all, we've been running through these lectures safe in the knowledge that the gauge group of the Standard Model is

$$G = U(1) \times SU(2) \times SU(3) \ . \tag{5.36}$$

Or is it?! In fact, there's a subtlety here.

To see this subtlety, consider the action on all fermions by the centre elements $(-1) \in SU(2)$ and $e^{2\pi i/3} \in SU(3)$. A quick check will confirm that

$$Q_L \to \omega^{-1} Q_L \ , \quad L_L \to \omega^3 L_L \ , \quad u_R \to \omega^2 u_R \ , \quad d_R \to \omega^2 d_R \ , \quad e_R \to e_R \tag{5.37}$$

with $\omega = e^{2\pi i/6}$. If we simultaneously act with the $U(1)$ hypercharge transformation $e^{2\pi i Y}$, then the result is that every fermion is either left unchanged, or picks up a minus sign. But a minus sign on a fermion is just part of the Lorentz group. The upshot is that there is a $\mathbb{Z}_6$ subgroup of G that does not act on the fermions (or, indeed, on the Higgs).

This means that it's tempting to say that the gauge group of the Standard Model is

$$G = \frac{U(1) \times SU(2) \times SU(3)}{\Gamma} \tag{5.38}$$

where $\Gamma = \mathbb{Z}_6$. But this too is overly hasty! The honest answer is that we don't know what the gauge group of the Standard Model is. There are four different choices, given by (5.38) where Γ is a subgroup of $\mathbb{Z}_6$, meaning $\Gamma = \mathbb{Z}_6$, $\mathbb{Z}_3$, $\mathbb{Z}_2$ or nothing at all. Strictly, these are all different quantum field theories, although the differences between them are rather subtle and don't show up in correlation functions of local operators. This means, among other things, that the differences between them won't show up in particle colliders like the LHC. Instead, one has to look to more formal aspects of the theories to see the difference, like the spectrum of allowed magnetic monopoles or what happens when the theory is placed on a manifold with non-trivial topology[9] .

5.2 Electroweak Symmetry Breaking

We now have the full Standard Model laid out before us in (5.23). The next question is: how does it give rise to the physics that we know and love? The answer largely lies in the role that the Higgs plays.

The dynamics of the Higgs boson is governed by the action (5.27)

$$\mathcal{L}_{\text{Higgs}} = \mathcal{D}_\mu H^\dagger \, \mathcal{D}^\mu H - \lambda \left(H^\dagger H - \frac{v^2}{2} \right)^2 \; . \tag{5.39}$$

The potential is such that it causes the Higgs to condense. This breaks the $U(1) \times SU(2)$ gauge symmetry under which the Higgs is charged, giving masses to the gauge bosons in the way we saw in Section 2.3. And, through the Yukawa interactions (5.22), it also gives masses to the fermions. In this section, we describe these effects.

Including the Maxwell and Yang-Mills terms for the $U(1) \times SU(2)$ gauge fields, we have the Lagrangian

$$\mathcal{L} = -\frac{1}{4} B_{\mu\nu} B^{\mu\nu} - \frac{1}{2} \text{Tr}\, W_{\mu\nu} W^{\mu\nu} + \mathcal{L}_{\text{Higgs}} \; . \tag{5.40}$$

To understand the physics, we need the Higgs covariant derivative which is given by

$$\mathcal{D}_\mu H = \partial_\mu H - i g W_\mu H - \frac{i}{2} g' B_\mu H \; . \tag{5.41}$$

This reflects the charges (5.5).

[9] For more details on these ideas, see Ofer Aharony, Nati Seiberg, and Yuji Tachikawa's Reading Between the Lines paper. Applications of these ideas to the Standard Model were given in Line Operators in the Standard Model.

In the ground state of the potential (5.27), we have $H^\dagger H = v^2/2$. As usual, we have to pick a direction for the Higgs vacuum expectation value to point in. We choose

$$\langle H \rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix} . \tag{5.42}$$

Then we parameterise the fluctuations of the Higgs as

$$H = e^{i\xi^A(x)T^A} \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + h(x) \end{pmatrix} . \tag{5.43}$$

Here $h(x)$ is a real scalar field, $T^A = \frac{1}{2}\sigma^A$ with $A = 1, 2, 3$ are the generators of $SU(2)$ and $\xi^A(x)$ are the would-be Goldstone bosons. As usual, they are eaten by the gauge bosons as part of the Higgs mechanism. A quick way to say this is to observe that we can just eliminate the factor of $e^{i\xi^A T^A}$ in (5.43) through a gauge transformation. Alternatively, to make contact with what we saw in Section 2.3, we can look at the covariant derivative. If we write $\Omega(x) = e^{i\xi^A(x)T^A}$, then we have

$$\mathcal{D}_\mu H = \frac{1}{\sqrt{2}}\Omega \left(\begin{pmatrix} 0 \\ \partial_\mu h \end{pmatrix} - i \left[g \left(\Omega^{-1} W_\mu \Omega + \frac{i}{g}\Omega^{-1}\partial_\mu \Omega \right) + \frac{g'}{2} B_\mu \right] \begin{pmatrix} 0 \\ v + h \end{pmatrix} \right) \tag{5.44}$$

Here we see that the overall field Ω sits in a way that can be eliminated by a gauge transformation (1.82).

We can always choose to work in unitary gauge in which, through a judicious $SU(2)$ rotation, we simply take $\xi^A(x) = 0$ or, equivalently, $\Omega = \mathbb{1}$. In this case, the Lagrangian (5.40) becomes

$$\begin{aligned} \mathcal{L} = -\frac{1}{4} B_{\mu\nu} B^{\mu\nu} - \frac{1}{2} \text{Tr}\, W_{\mu\nu} W^{\mu\nu} + \frac{1}{2}\partial_\mu h \partial^\mu h - \lambda h^2 \left(v + \frac{h}{2} \right)^2 \\ + \frac{1}{8}(v+h)^2 \left(g^2 (W^1_\mu)^2 + g^2 (W^2_\mu)^2 + (g W^3_\mu - g' B_\mu)^2 \right) . \end{aligned} \tag{5.45}$$

To get the second line, we expanded out $SU(2)$ gauge boson fields W_μ in terms of the generators $T^A = \frac{1}{2}\sigma^A$, and contracted them with the Higgs field. From this we can read off the masses from the quadratic term. There is a $\lambda v^2 h^2$ term that gives a mass for h. This is the particle that, experimentally, we call *the* Higgs boson. It's mass is measured to be

$$M_h = \sqrt{2\lambda} v \approx 125 \text{ GeV} . \tag{5.46}$$

We see that this mass is a combination of the Higgs vev v and the dimensionless coupling λ.

We can also read off the masses of the gauge bosons from the second line in (5.45). Both W^1_μ and W^2_μ have the same mass $M_W = vg/2$. It will prove fruitful to combine them into the complex combination

$$W^\pm_\mu = \frac{1}{\sqrt{2}}(W^1_\mu \mp iW^2_\mu) \ . \tag{5.47}$$

Note the flip of the $\pm$ sign on the right-hand side. We will see shortly that this ensures that $W^\pm$ has electric charge ± 1. The experimentally measured mass of these spin 1 bosons is

$$M_W = \frac{gv}{2} \approx 80 \text{ GeV} \ . \tag{5.48}$$

This mass is set by the Higgs vev v and the $SU(2)$ gauge coupling g.

The final massive gauge boson is slightly more interesting. We see from (5.45) that it is a linear combination of the W^3_μ which is part of $SU(2)$ and B_μ which is associated to the fundamental $U(1)$ hypercharge gauge symmetry. The relevant linear combination is set by the two coupling constants, g and g'. To this end, we define the *Weinberg angle*, also known as the *weak mixing angle*

$$\cos\theta_W = \frac{g}{\sqrt{g^2 + g'^2}} \quad \Longleftrightarrow \quad \sin\theta_W = \frac{g'}{\sqrt{g^2 + g'^2}} \ . \tag{5.49}$$

We then define the two linear combinations of gauge fields

$$\begin{aligned} Z_\mu &= \cos\theta_W \, W^3_\mu - \sin\theta_W \, B_\mu \\ A_\mu &= \sin\theta_W \, W^3_\mu + \cos\theta_W \, B_\mu \ . \end{aligned} \tag{5.50}$$

The first of these has a mass from (5.45) which is experimentally measured to be

$$M_Z = \frac{v}{2}\sqrt{g^2 + g'^2} \approx 91 \text{ GeV} \ . \tag{5.51}$$

We don't have any way to determine any of these masses from first principles. They are combinations of the Higgs vev v, the Higgs coupling λ and the gauge couplings g and g', none of which we know without going out and measuring them. However, the theoretical framework does ensure the mild inequality

$$M_W = M_Z \cos\theta_W < M_Z \tag{5.52}$$

which is indeed observed.

We can do some simple counting here. Our original Higgs boson H was a doublet of $SU(2)$. This means that it has two complex degrees of freedom or, equivalently, four real degrees of freedom. One of these remains as the real scalar h that we call the Higgs boson. The other three got eaten by the three gauge bosons W^1_μ, W^2_μ and Z_μ.

The discovery of the Higgs boson h was announced at CERN in 2013. But in a very real sense, 3/4 of the more fundamental Higgs boson H were discovered when the massive W and Z bosons were first seen in 1983. As we've seen, they get their mass only by eating three of the components of H.

The scales of the masses of the Higgs h and the W and Z bosons are all set by the Higgs expectation value v, multiplied by some dimensionless coupling constant. This is a theme that will continue shortly when we discuss matter particles. These couplings can all be measured directly, through cross-sections or decay rates. We learn that the only dimensionful parameter in the classical Standard Model Lagrangian takes the value

$$v \approx 250 \text{ GeV} . \tag{5.53}$$

We will later see that this is directly related to the *Fermi constant* that governs the strength of weak decays. The dimensionless parameters are

$$\lambda \approx 0.35 \quad \text{and} \quad g \approx 0.64 \quad \text{and} \quad g' \approx 0.34 . \tag{5.54}$$

Each of these runs under RG; the values above are given at the scale $\mu = M_Z$. We also have the Weinberg angle (5.49) which takes the value

$$\cos\theta_W \approx 0.88 \quad \Longrightarrow \quad \theta_W \approx 29^\circ . \tag{5.55}$$

It's common to quote the value $\sin^2\theta_W \approx 0.223$.

5.2.1 Electromagnetism

There is one of the $U(1) \times SU(2)$ gauge bosons that escapes the clutches of the Higgs and remains massless. This is the field A_μ defined in (5.50) and it is the most famous gauge boson of all: the photon.

We can look at this more closely. From a group theoretic perspective, the photon remains massless because the Higgs induces the symmetry breaking

$$U(1)_Y \times SU(2) \to U(1)_{\text{EM}} . \tag{5.56}$$

This is why the $U(1) \times SU(2)$ sector of the Standard Model is referred to as *electroweak theory*.

We can identify this unbroken $U(1)$ symmetry by looking at how the Higgs vev (5.42) transforms under a general $U(1) \times SU(2)$ transformation, with parameters α^A and β,

$$\langle H \rangle = \begin{pmatrix} 0 \\ v \end{pmatrix} \longrightarrow e^{ig\alpha^A T^A} e^{ig'\beta Y} \begin{pmatrix} 0 \\ v \end{pmatrix} . \tag{5.57}$$

The Higgs has hypercharge $Y = \frac{1}{2}$ so, writing the $SU(2)$ generators $T^A = \frac{1}{2}\sigma^A$, we have

$$g\alpha^A T^A + g'\beta Y = \frac{g}{2} \begin{pmatrix} \alpha^3 + g'\beta/g & \alpha^1 - i\alpha^2 \\ \alpha^1 + i\alpha^2 & -\alpha^3 + g'\beta/g \end{pmatrix} . \tag{5.58}$$

We see that the choice of parameters that leaves $\langle H \rangle$ invariant is $\alpha^1 = \alpha^2 = 0$ and $g\alpha^3 = g'\beta$. This means that the unbroken generator is the combination

$$Q = T^3 + Y . \tag{5.59}$$

We identify this with the generator of the unbroken $U(1)_{\rm EM}$ subgroup which, in more everyday terms, means that Q determines the electric charge of the fields. We'll see how this works in practice for all the fermion fields below.

The electroweak theory also sets the electromagnetic coupling constant e. This is simplest to see if we look at the general covariant derivative for a field that transforms in the fundamental of $SU(2)$ and with hypercharge Y,

$$\mathcal{D}_\mu = \partial_\mu - igW_\mu^A T^A - ig'YB_\mu . \tag{5.60}$$

We work with the fields $W_\mu^\pm$ defined in (5.47) and the corresponding generators $T^\pm = (T^1 \pm iT^2)/\sqrt{2}$. We also work with the fields Z_μ and A_μ defined in (5.50) to get

$$\mathcal{D}_\mu = \partial_\mu - ig(W_\mu^+ T^+ + W_\mu^- T^-) - i(g\cos\theta_W T^3 - g'\sin\theta_W Y)Z_\mu - ieQA_\mu . \tag{5.61}$$

For our immediate interests, it's that last term that's important. It involves the charge Q, together with the coupling

$$e = g\sin\theta_W = g'\cos\theta_W . \tag{5.62}$$

The electromagnetic coupling takes value

$$e \approx 0.30 . \tag{5.63}$$

This particular coupling constant is better known in the form $\alpha = e^2/4\pi$ which is called the *fine structure constant* and takes the famous value $\alpha \approx 1/137$.

The bosons of the electroweak sector are the Higgs, and the W and Z bosons. The Higgs h is electrically neutral. This must be the case simply because it's a real scalar field, but we can check explicitly by noting that it sits in the lower component of the doublet (5.43) which has $T^3 = \frac{1}{2}\sigma^3$ eigenvalue $-\frac{1}{2}$. The Higgs also has hypercharge $Y = +\frac{1}{2}$ ensuring that $Q = T^3 + Y = 0$.

The Z boson is similarly neutral. Again, this must be the case because it is a real field. Operationally, this follows because it carries no hypercharge and commutes with the $SU(2)$ generator T^3.

That leaves us with the W bosons. Under an $SU(2)$ transformation with $\alpha^1 = \alpha^2 = 0$ and α^3 constant, we have, from (1.87),

$$\delta W_\mu = -ig[W_\mu, \alpha^3 T^3] = g\alpha^3(-W_\mu^1 T^2 + W_\mu^2 T^1) \ . \tag{5.64}$$

We can write this as $\delta W_\mu^1 = g\alpha^3 W_\mu^2$ and $\delta W_\mu^2 = -g\alpha^3 W_\mu^1$. We think of this $SU(2)$ transformation as part of the $U(1)_{\rm EM}$ transformation, with $g\alpha^3 = e\alpha$. Then, written in terms of our fields $W_\mu^\pm$ defined in (5.47), we have

$$\delta W_\mu^\pm = \pm ie\alpha W_\mu^\pm \ . \tag{5.65}$$

This is telling us that the W boson $W_\mu^\pm$ has electric charge $Q = \pm 1$.

5.2.2 Running of the Weak Coupling

The gauge couplings of the electroweak sector run with energy scale. Because hypercharge is a $U(1)$ gauge theory, the associated coupling g' gets smaller as we flow to the infra-red.

But for the non-Abelian $SU(2)$ gauge symmetry, we have to be more careful. We gave the general formula for $SU(N_c)$ gauge theory coupled to N_f massless Dirac fermions in (3.11) when discussing QCD. Now we need the generalisation to include N_s scalars in the fundamental representation. The result is

$$\frac{1}{g^2(\mu)} = \frac{1}{g_0^2} - \frac{b_0}{(4\pi)^2}\log\frac{\Lambda_{UV}^2}{\mu^2} \tag{5.66}$$

with the coefficient given by

$$b_0 = \frac{11}{3}N_c - \frac{2}{3}N_f - \frac{1}{6}N_s \ . \tag{5.67}$$

Applied to electroweak theory, we clearly have $N_c = 2$ and $N_s = 1$, corresponding to the Higgs doublet. But what about N_f? We saw in (5.2) that each generation of fermions has an $SU(2)$ doublet of quarks Q_L and a doublet of leptons L_L. This is $3+1=4$ Weyl fermions. But the N_f in (5.67) counts *Dirac* fermions, so each generation has $N_f = 2$ Dirac fermions as far as the beta function is concerned. And, of course, we have three generations. So the coefficient of the one-loop beta function for the weak force is $b_0 = b_{\rm weak}$ with

$$b_{\rm weak} = \frac{11}{3}\times 2 - \frac{2}{3}\times 6 - \frac{1}{3} = 3\ . \tag{5.68}$$

With $b_{\rm weak} > 0$, we see that the $SU(2)$ sector of the Standard Model is, like QCD, asymptotically free. It flows to strong coupling in the infra-red.

This begs the question: do we have to worry about strong coupling effects in the weak sector, like we did for QCD? The answer is no. And the reason is that the Higgs mechanism gives masses to the gauge bosons and, in doing so, freezes the running of the coupling g at the scale $\mu \sim M_W$. This is where the quoted value of $g \approx 0.64$ in (5.54) is measured.

It's worth commenting that, although we call the weak nuclear force "weak", the actual value of the coupling is not small. Indeed, $\alpha_W = g^2/4\pi \approx 1/30$, which is almost 5 times bigger than the fine structure constant! The reason that the weak force is actually weak has nothing to do with the strength of the coupling and everything to do with the mass of the W and Z bosons (or, equivalently, the scale of the Higgs vev). As we will see in Section 5.3, particles that decay through the weak force do so by the emission of an intermediate W or Z boson. The large mass of these bosons translates to a small decay rate.

It's also fruitful to compare the couplings for the weak and strong force. Measured at the weak scale M_Z, we have

$$\alpha_s(M_Z) \approx 0.12 \quad \text{and} \quad \alpha_w(M_Z) \approx 0.034\ . \tag{5.69}$$

So the weak force is indeed weaker than the strong force.

Asymptotic freedom ensures that both g_s and g_w get smaller as we look at higher energies. But they do so at different speeds. The running of the strong coupling (assuming three massless generations) is dictated by

$$b_{\rm strong} = \frac{11}{3}\times 3 - \frac{2}{3}\times 6 = 7\ . \tag{5.70}$$

Because we have $b_{\rm strong} > b_{\rm weak}$, the two couplings will converge as we go to higher energies. And it's natural to ask: where does this convergence take place?

You have to be a little bold to do this calculation. We will take $\Lambda_{UV} = M_W$ in (5.66) and then extrapolate the equation to energy scales $\mu \gg M_W$ and, moreover, to energy scales beyond those that we've probed experimentally. There's nothing wrong with this per se, since the equation is invertible: if you know the coupling at one scale, then we can always determine it at any other scale, whether lower or higher. But we are assuming that there's no additional matter to discover which would change the coefficient b_0 as we go to higher energies. That seems like a rather big assumption.

With these health warnings in place, the two couplings meet at a scale μ given by

$$\frac{1}{g_s^2} - \frac{b_{\text{strong}}}{(4\pi)^2} \log \frac{M_W^2}{\mu^2} = \frac{1}{g_w^2} - \frac{b_{\text{weak}}}{(4\pi)^2} \log \frac{M_W^2}{\mu^2} \ . \tag{5.71}$$

Solving, we find

$$\mu = M_W \exp\left(\frac{2\pi}{b_{\text{strong}} - b_{\text{weak}}}\left(\frac{1}{\alpha_w} - \frac{1}{\alpha_s}\right)\right) \approx 2 \times 10^{16}\,\text{GeV} \ . \tag{5.72}$$

So the two couplings do indeed meet, although it takes them a long time because the running is only logarithmic.

Nonetheless, the couplings meet in an intriguing place. The Planck scale sits at about $M_{\text{pl}} \sim 10^{19}$ GeV (or a bit less depending on where you put factors of 8π.) Had the two couplings converged at a scale $\mu \gg M_{\text{pl}}$ then we could have simply discarded this computation. We did it assuming that there was nothing new to find as we went to higher energies but as soon as quantum gravity effects kick in there's certainly no reason to trust the formula (5.66). The fact that the two lines meet at a scale just below M_{pl} is, if nothing else, telling us that we don't have an immediate reason to discard it. It also suggests that perhaps something more interesting is going on.

That something is the idea of *unification*. Is it perhaps possible that the two coupling constants are meeting because the $SU(2)$ and $SU(3)$ forces sit within a larger gauge group? The answer is: we don't know. But it is a compelling idea. Proposals for this larger gauge group include $SU(5)$ and $SO(10)$ (strictly $Spin(10)$).

There is, of course, a third coupling constant in the Standard Model. This is the hypercharge coupling g'. This is the smallest of the three couplings and it too runs, now getting bigger as we go to higher energies. This means that it must also meet the other two. But where? A similar calculation shows that $\alpha_Y = g'^2/4\pi$ meets the strong and weak couplings at

$$\begin{aligned} \alpha_Y = \alpha_s \quad &\text{at} \quad \mu \approx 5 \times 10^{19}\ \text{GeV} \\ \alpha_Y = \alpha_w \quad &\text{at} \quad \mu \approx 10^{21}\ \text{GeV} \ . \end{aligned} \tag{5.73}$$

We see that the three lines don't meet. Things aren't as clean as that. Moreover, the unification of the hypercharge coupling seems to be in the regime where quantum gravity comes into play. Nonetheless, it's still in the same ballpark. So, while not perfect, this also lends credence to the idea of unification. Needless to say, we don't know if unification does indeed take place. But if we're searching the Standard Model for clues for what lies beyond, this is certainly one of the most striking.

5.2.3 A First Look at Fermion Masses

The Higgs gives mass to the W and Z boson. But it also gives masses to all the fundamental fermions in the Standard Model. These arise through the Yukawa interactions.

First, a repeat of a comment that we made previously: it's not possible to write down straightforward mass terms for the fermions in the Standard Model. This is because it is a chiral theory, with left- and right-handed fermions transforming differently under the gauge group. This means that any mass term necessarily violates gauge symmetry. The Yukawa terms are the gauge invariant interaction terms and give a mass only once the Higgs field gets an expectation value.

To kick things off, let's ignore the fact that we have three generations of fermions and focus only on the first. This will allow us to see how the basic structure of particles arises. We will then see the complications that arise from having multiple generations in Section 6.

The Yukawa couplings for a single generation were given in (5.14) and (5.15),

$$\mathcal{L}_{\text{Yuk}} = -y^d\,\bar{Q}_L H d_R - y^u\,\bar{Q}_L \tilde{H} u_R - y^e\,\bar{L}_L H e_R - y^\nu\,\bar{L}_L \tilde{H} \nu_R + \text{h.c.}\ . \tag{5.74}$$

Here H is the Higgs doublet that transforms in the **2** of the $SU(2)$ gauge group, and $\tilde{H}$ is the conjugated Higgs doublet, contracted with an ϵ so that it too transforms in the **2**,

$$\tilde{H}^a = \epsilon^{ab} H_b^\dagger \quad \text{with} \quad a,b = 1,2\ . \tag{5.75}$$

Meanwhile, y^d, y^u, y^e and y^ν are dimensionless Yukawa couplings. We'll give their values in Section 6. (This is one place where we really should include all three generations to appreciate the values.) Recall, also, that we're not sure if there is a right-handed neutrino field ν_R, so we might have to dispense with the final term in (5.74).

Our immediate interest is to understand the implications of the Higgs vev (5.42)

$$\langle H\rangle = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 \\ v \end{pmatrix} \quad \Longrightarrow \quad \langle\tilde{H}\rangle = \frac{1}{\sqrt{2}}\begin{pmatrix} v \\ 0 \end{pmatrix} \ . \tag{5.76}$$

This will distinguish the two components of the $SU(2)$ doublets Q_L and L_L, giving them different masses and, as we will see, different charges under the unbroken symmetry of electromagnetism. For this reason, it's useful to introduce different names for the two components of these doublets. We write

$$Q_L = \begin{pmatrix} u_L \\ d_L \end{pmatrix} \quad \text{and} \quad L_L = \begin{pmatrix} \nu_L \\ e_L \end{pmatrix} \ . \tag{5.77}$$

(We already introduced these names in (5.18) and (5.20) although, as we noted at the time, it was premature before we discussed electroweak symmetry breaking.)

Now we can look at the Yukawa couplings (5.74), focussing only on the role of the vev v and ignoring the interactions with the fluctuations of the Higgs boson h. We have

$$\mathcal{L}_{\text{Yuk}} = -\frac{v}{\sqrt{2}}\Big(y^d\,\bar{d}_L d_R + y^u\,\bar{u}_L u_R + y^e\,\bar{e}_L e_R + y^\nu\,\bar{\nu}_L \nu_R\Big) \ . \tag{5.78}$$

We see that each of the fermions gets a mass, given by

$$m^X = \frac{1}{\sqrt{2}} y^X v \tag{5.79}$$

where $X = d, u, e, \nu$ labels the appropriate Yukawa coupling y^X. The scale of all these masses is, like all particles in the Standard Model, set by Higgs vev. If the Higgs did not condense, all fermions would be massless.

This is the source of the oft-repeated claim that the Higgs boson is responsible for all mass in the Standard Model. It is, as we stressed in Section 3, a lie. It is true that the Higgs vev v is the only dimensionful scale in the Standard Model Lagrangian and that all fundamental particles would be massless if it were to vanish. But there is another, more subtle, scale in the Standard Model itself which is Λ_{QCD}, the scale at which the strong force lives up to its name. And this scale would exist even in the absence of the Higgs vev and would continue to give a mass to the proton and neutron. Of course, that's not to say that the Higgs is unimportant: in this hypothetical world in which $v = 0$, electrons would be massless so physics, atoms, and life would be vastly different.

We can also determine the electric charges of each of the fermions using the formula (5.59)

$$Q = T^3 + Y \ . \tag{5.80}$$

We listed the hypercharges Y of all particles previously. They are

$$\begin{array}{c|ccccc} & Q_L & L_L & u_R & d_R & e_R \\ \hline Y & \frac{1}{6} & -\frac{1}{2} & \frac{2}{3} & -\frac{1}{3} & -1 \end{array} \tag{5.81}$$

Each of the right-handed fermions is uncharged under the $SU(2)$ gauge group and so we have simply $Q = Y$. Indeed, we recognise the hypercharge as the usually advertised electric charge of these particles.

For the $SU(2)$ doublets Q_L and L_L, we have a small calculation to do. The T^3 eigenvalues are $\pm\frac{1}{2}$, with $+$ for the upper component and $-$ for the lower component. This means that the electric charges $Q = T^3 + Y$ are:

$$\begin{aligned} u_L: \quad Q = \frac{1}{2} + \frac{1}{6} = \frac{2}{3} \quad &\text{and} \quad d_L: \quad Q = -\frac{1}{2} + \frac{1}{6} = -\frac{1}{3} \\ \nu_L: \quad Q = \frac{1}{2} - \frac{1}{2} = 0 \quad &\text{and} \quad e_L: \quad Q = -\frac{1}{2} - \frac{1}{2} = -1 \ . \end{aligned} \tag{5.82}$$

We see that the electric charges of the left-handed fermions coincide with those of the right-handed fermions in (5.81), as indeed they must so that the mass terms (5.78) are invariant under the surviving $U(1)_{\rm EM} \subset SU(2) \times U(1)_Y$.

The upshot of symmetry breaking is that we are left with four *Dirac* fermions. These are the up quark u with charge $+2/3$, the down quark d with charge $-1/3$, the electron e with charge -1, and the neutral neutrino ν. If the right-handed neutrino ν_R doesn't exist then the neutrino is a Weyl fermion and cannot get a mass through the simple mechanism described above. We will discuss the issue of neutrino masses further in Section 7.

The collection of electric charges of fermions in the Standard Model looks kind of random. And, viewed as a low-energy vector-like theory, they are! But, as we have seen, there is a deeper reason underlying this choice that only becomes apparent when you realise that the Standard Model is a chiral theory, subject to the stringent constraints of anomaly cancellation.

5.3 Weak Decays

Since the time of Newton, we've tended to think of forces as things that push and pull. That's an intuition that holds well for QED and the Coulomb force, and also for QCD which binds quarks together into hadrons. But it's not the best way to think about the weak force. Instead, the weak force is an instrument of decay.

One of the consequences of the weak force is that it rents asunder what the strong force so carefully put together. We saw in Section 3 that quarks are bound into baryons and mesons. In a world of just QCD, the baryon octet that contains, among other things, the proton and neutron would be stable. So too would the octet of pseudoscalar mesons that includes the pions and kaons. But in our world, only the proton is stable. (Admittedly, we can also have stable nuclei consisting of bound states of protons and neutrons.) Everything else decays through the weak force.

In this section, we will start to understand how these decay processes take place. We will start by better understanding what fermions the W and Z bosons couple to and constructing the relevant Feynman diagrams.

5.3.1 Electroweak Currents

To start, we understand how the various gauge bosons couple to the fermions. For now, we will again stick with just a single generation. (There is an interesting twist to the story when we introduce multiple generations that we describe in Section 6.)

The fermion kinetic terms are

$$\mathcal{L}_{\text{fermi}} = -i\Big(\bar{Q}_L\bar{\sigma}^\mu\mathcal{D}_\mu Q_L + \bar{L}_L\bar{\sigma}^\mu\mathcal{D}_\mu L_L + \bar{u}_R\sigma^\mu\mathcal{D}_\mu u_R + \bar{d}_R\sigma^\mu\mathcal{D}_\mu d_R + \bar{e}_R\sigma^\mu\mathcal{D}_\mu e_R\Big)\ . \tag{5.83}$$

We haven't included the right-handed neutrino ν_R because it is neutral under all gauge symmetries. We'll ignore the gluon fields for now, and just focus on the terms that involve interactions with the electroweak gauge bosons. These are

$$\begin{aligned}\mathcal{L}_{\text{kin}}\Big|_{\text{weak}} = & -\frac{g}{2}W^3_\mu\left(\bar{u}_L\bar{\sigma}^\mu u_L - \bar{d}_L\bar{\sigma}^\mu d_L + \bar{\nu}_L\bar{\sigma}^\mu\nu_L - \bar{e}_L\bar{\sigma}^\mu e_L\right)\\ & - \frac{g}{\sqrt{2}}W^+_\mu(\bar{u}_L\bar{\sigma}^\mu d_L + \bar{\nu}_L\bar{\sigma}^\mu e_L) - \frac{g}{\sqrt{2}}W^-_\mu(\bar{d}_L\bar{\sigma}^\mu u_L + \bar{e}_L\bar{\sigma}^\mu\nu_L)\\ & - g'B_\mu\Big(\frac{1}{6}\bar{u}_L\bar{\sigma}^\mu u_L + \frac{1}{6}\bar{d}_L\bar{\sigma}^\mu d_L - \frac{1}{2}\bar{\nu}_L\bar{\sigma}^\mu\nu_L - \frac{1}{2}\bar{e}_L\bar{\sigma}^\mu e_L\\ & \qquad + \frac{2}{3}\bar{u}_R\sigma^\mu u_R - \frac{1}{3}\bar{d}_R\sigma^\mu d_R - \bar{e}_R\sigma^\mu e_R\Big)\ .\end{aligned} \tag{5.84}$$

If we replace W^3_μ and B_μ with the Z boson and photon fields, as in (5.50), these terms can be written as

$$\mathcal{L}_{\rm kin}\Big|_{\rm weak} = -\frac{e}{\sqrt{2}\sin\theta_W}(W^+_\mu J^\mu_+ + W^-_\mu J^\mu_-) - \frac{e}{\sin\theta_W\,\cos\theta_W}Z^\mu J^Z_\mu - eA^\mu J^{\rm EM}_\mu\ . \quad (5.85)$$

Here we've replaced the two coupling constants g and g' with the Weinberg angle $\tan\theta_W = g'/g$ and the electromagnetic coupling $e = g\sin\theta_W = g'\cos\theta_W$ and we've introduced various currents that interact with the gauge fields. The electromagnetic current that couples to the photon is given by

$$\begin{aligned} J^{\rm EM}_\mu &= \frac{2}{3}(\bar{u}_L\bar{\sigma}_\mu u_L + \bar{u}_R\sigma^\mu u_R) - \frac{1}{3}(\bar{d}_L\bar{\sigma}_\mu d_L + \bar{d}_R\sigma^\mu d_R) - (\bar{e}_L\bar{\sigma}_\mu e_L + \bar{e}_R\sigma^\mu e_R) \\ &= \Big(\frac{2}{3}\bar{u}\gamma_\mu u - \frac{1}{3}\bar{d}\gamma_\mu d - \bar{e}\gamma_\mu e\Big)\ . \end{aligned} \quad (5.86)$$

This takes the expected form, with each fermion multiplied by its electric charge. In the second line, we've written this in terms of Dirac spinors u, d, and e and the gamma matrices γ^μ to emphasise that, despite its chiral origins, this is the kind of vector-like current that we're used to in QED.

For the Z boson, we have a little more work to do. Some algebra reveals that the current takes the form

$$J^Z_\mu = \frac{1}{2}(\bar{u}_L\bar{\sigma}_\mu u_L - \bar{d}_L\bar{\sigma}_\mu d_L + \bar{\nu}_L\bar{\sigma}_\mu \nu_L - \bar{e}_L\bar{\sigma}_\mu e_L) - \sin^2\theta_W\, J^{\rm EM}_\mu\ . \quad (5.87)$$

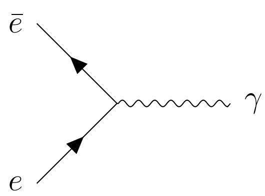
Finally, the currents for the W bosons can be read off immediately from (5.85); they are

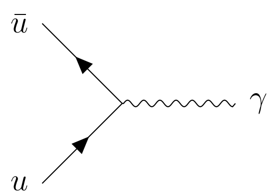
$$J^+_\mu = \bar{u}_L\bar{\sigma}_\mu d_L + \bar{\nu}_L\bar{\sigma}_\mu e_L \quad \text{and} \quad J^-_\mu = \bar{d}_L\bar{\sigma}_\mu u_L + \bar{e}_L\bar{\sigma}_\mu \nu_L\ . \quad (5.88)$$

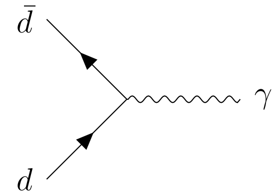
The currents for both the W and Z bosons are chiral, treating left-handed fermions differently from their right-handed counterparts.

5.3.2 Feynman Diagrams

From the interaction terms (5.85), we can read off the Feynman rules for the electroweak sector. We see from (5.86) that the photon couples in the usual way to the up and down quarks and to the electron, with coupling constant given by eq with q the charge. This gives rise to the kind of Feynman diagram that we met in our first course on Quantum Field Theory.

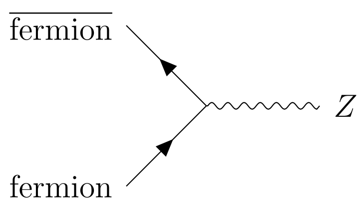






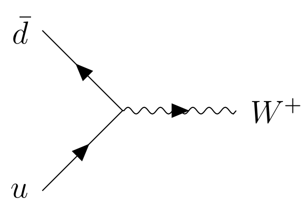
The photon couples to the up and down quarks and the electron. It doesn't couple to the neutrino because it's neutral.

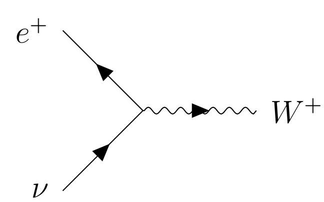
From (5.87), we see that there are similar diagrams involving the Z boson. But, in contrast to the photon, this couples to all low energy particles, including the neutrino. So we have diagrams of the form



where the fermion could be u, d, e or ν. This time, the coupling is more complicated: there is an overall factor of $e/\sin\theta_W\,\cos\theta_W$, with different coefficients depending on the fermion species. And more care is needed with the spinor indices because of the chiral nature of the coupling.

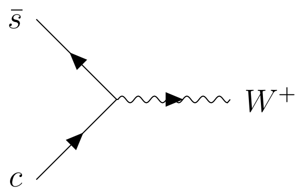
Finally, the W boson relates two different fermions. We have the Feynman diagrams:

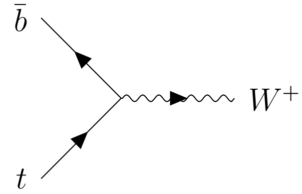




The two fermions in these diagrams have electric charges that differ by ± 1 to ensure that the overall electric charge is conserved at the vertex. We've included an arrow on the gauge boson propagator because it is now a complex spin 1 field. The arrow going the other way corresponds to the anti-particle W^-.

Here, we've only focussed on a single generation. There are similar diagrams where u, d, e and ν_e are replaced by their higher generational cousins. So, for example, there are additional W boson diagrams that connect the strange and charm quark, and the bottom and top quark:





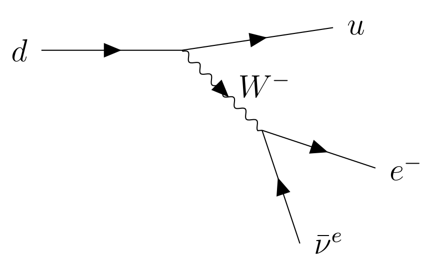
There are also diagrams with muons and taus replacing electrons. In fact, it turns out that there is an additional subtlety when considering these higher generations that we will turn to in Section 6.

5.3.3 A First Look at Weak Processes

Historically, the weak force was first observed in beta decay of nuclei. We can view this as a neutron decaying to a proton, electron and anti-neutrino

$$n \to p + e^- + \bar{\nu}^e \ . \tag{5.89}$$

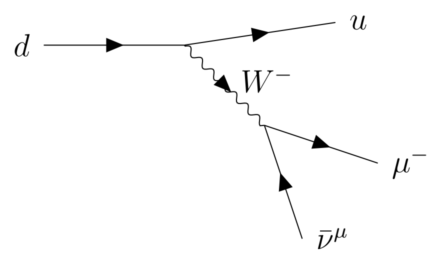
The possibility of such a process follows immediately from our discussion above. As we saw in Section 3, a neutron is a baryon with quark content *udd*. This decays to a proton with quark content *uud* through the tree level Feynman diagram



The lifetime of the neutron is about 10 minutes.

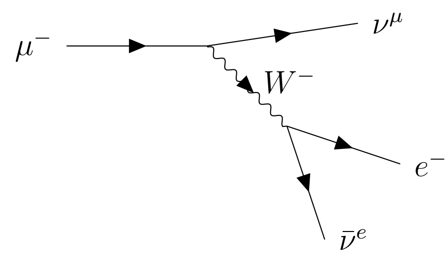
An obvious comment: the reason that down quarks decay into up quarks, rather than the other way around, is because the mass of the down quark is heavier than the masses of the decay products, $m_d > m_u + m_e + m_{\nu_e}$. As we've mentioned previously, we have no understanding of why the masses of fundamental particles are ordered in this way.

Neutrons are not the only victim of the weak force. A world without the weak force would be awash with pions which, as we saw in Section 3, are the lightest of the hadrons. The vast majority of the time (something like 99.99%) charged pion $\pi^- = d\bar{u}$ decays through the weak force to a muon and anti-neutrino. This occurs through a similar Feynman diagram to that responsible for beta decay, but with muons replacing electrons as the end products,



The resulting up quark then combines with the anti-up quark in the pion, and the two rapidly decay into photons. The lifetime of the charged pion is about 10^{-8} seconds.

The resulting muons don't live too long either. Their demise is also due to the weak force and they decay to electrons and neutrinos through the process



The lifetime of the muon is around 2×10^{-6} seconds. All other particles involving quarks and leptons from the second and third generation have the same fate, decaying through the weak force to the more familiar particles from the first generation.

5.3.4 4-Fermi Theory

Although the weak force is mediated by W and Z bosons, if we focus on processes that take place at low energies, $E \ll M_W,\ M_Z$, then it's possible to ignore these gauge bosons and write down interaction terms that describe the relevant physics directly.

There are a couple of (essentially equivalent) ways to remove the W and Z bosons while leaving behind the processes that they induce. The first, and most direct, way to see this is to start with the terms linear and quadratic in W bosons. (We'll ignore the Higgs field h in what follows but, crucially, keep its vev v.) We have

$$\begin{aligned}\mathcal{L}_{\text{weak}} = &-\frac{1}{2}(\partial_\mu W^+_\nu - \partial_\nu W^+_\mu)(\partial^\mu W^{-\,\nu} - \partial^\nu W^{-\,\mu}) \\ &+ \frac{g^2 v^2}{4} W^+_\mu W^{-\,\mu} - \frac{g}{\sqrt{2}}(W^+_\mu J^\mu_+ + W^-_\mu J^\mu_-)\ .\end{aligned} \tag{5.90}$$

At low energies, we can neglect the kinetic terms for the W bosons. We then proceed by completing the square in the remaining terms,

$$\mathcal{L}_{\rm weak} \approx \frac{g^2v^2}{4}\left(W^+_\mu - \frac{2\sqrt{2}}{gv^2}J_{-\,\mu}\right)\left(W^{-\,\mu} - \frac{2\sqrt{2}}{gv^2}J_+^\mu\right) - \frac{2}{v^2}J_{+\,\mu}J_-^\mu\ . \tag{5.91}$$

Performing the path integral over the W bosons effectively sets the first term to zero, leaving us just with the current-current interaction. We write this, for historic reasons, as

$$\mathcal{L}_{\rm weak} = -\frac{4G_F}{\sqrt{2}}\,J_{+\,\mu}\,J_-^\mu \tag{5.92}$$

with

$$G_F = \frac{1}{\sqrt{2}v^2} \approx 1.16\times 10^{-5}\ {\rm GeV}^{-2}\ . \tag{5.93}$$

Our final result (5.92) is a 4-fermion interaction. The coupling constant G_F is called the *Fermi coupling* and provides a direct measurement of the Higgs vev. It has dimensions $[G_F] = -2$ (because the fermion has dimension $3/2$ so the $J_\mu J^\mu$ term has dimension 6). This means that the four fermi term is irrelevant in the renormalisation group sense. It is, however, very relevant in the cosmic sense. For example, it is what makes the Sun shine.

There is a second way to arrive at the same result (5.92) using Feynman diagrams. In this approach, we start by examining the propagator for a massive vector field. In momentum space, it takes the form

$$D_{\mu\nu}(p) = \frac{i}{p^2-M^2}\left(-\eta_{\mu\nu} + \frac{p_\mu p_\nu}{M}\right)\ . \tag{5.94}$$

In the limit $E \ll M$, we ignore the momentum terms and get

$$D_{\mu\nu}(p) \approx \frac{i}{M^2}\eta_{\mu\nu} \quad\Longrightarrow\quad D_{\mu\nu}(x-y) = \frac{i}{M^2}\eta_{\mu\nu}\delta^4(x-y)\ . \tag{5.95}$$

In this limit, the propagator in position space becomes a delta-function, as shown, and the kind of couplings induced by the massive gauge boson, which are generally of the form $J^\mu(x)D_{\mu\nu}(x,y)J^\nu(y)$ collapse to the direct current-current interaction that we saw in (5.92).

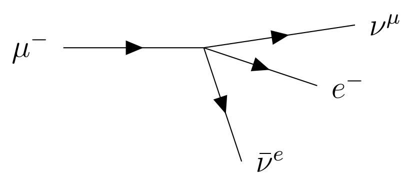
We can see what this means for, say, muon decay. If we ignore the quarks for now, but include both electron and muon contributions, then the W boson current (5.88) includes the term

$$J^+_\mu = \bar{\nu}^e_L \bar{\sigma}_\mu e_L + \bar{\nu}^\mu_L \bar{\sigma}_\mu \mu_L \ . \tag{5.96}$$

The 4-fermi terms then include

$$\mathcal{L}_{\text{weak}} \sim -\frac{4G_F}{\sqrt{2}} \, (\bar{\nu}^e_L \bar{\sigma}_\mu e_L)(\bar{\mu}_L \bar{\sigma}^\mu \nu^\mu_L) \ . \tag{5.97}$$

This gives rise directly to muon decay through the Feynman diagram



It's as if we've squinted and ignored the W boson that mediates the weak force.

These kinds of 4-fermion interactions were first written down by Fermi in 1933. His purpose was to describe beta decay, with the neutron coupled to the proton, electron and neutrino fields (the latter later realised to be an anti-neutrino). This was an important breakthrough in our understanding of particle physics because it changed the way we think about particles. In beta decay, a neutron decays into a proton and electron. But that doesn't mean that the neutron is *made* of a proton and electron! They're not sitting there inside the neutron all along, waiting to escape. Instead, the key idea of quantum field theory is that the four-fermion couplings allow one type of field to transmute into the others.

Second, there's some spin structure going on in (5.97) that Fermi was unaware of. This arises because the W boson couples only to left-handed fermions, not their right-handed counterparts. We can also write the resulting coupling in terms of Dirac spinors where we need a projection operator onto the left-handed part. The coupling (5.97) can then be written as

$$\mathcal{L}_{\text{weak}} \sim \frac{G_F}{\sqrt{2}} \left(\bar{\nu}^e \gamma_\mu (1+\gamma^5) e\right)\left(\bar{\mu}\gamma^\mu (1+\gamma^5)\nu^\mu\right) \ . \tag{5.98}$$

This is referred to as the *V-A theory*, because the coupling involves the difference between the vector current $\bar{\psi}\gamma^\mu\psi$ and the axial current $\bar{\psi}\gamma^5\gamma^\mu\psi$. (Admittedly, the expression "V-A" would probably have made more sense if I'd defined my γ^5 matrix with a different sign so that it appeared as $(1-\gamma^5)$ rather than $(1+\gamma^5)$ in the expressions above. Oh well.)

6 Flavour

The purpose in this section is to understand how the three different generations of the Standard Model fit into the story. We will focus on the quark fields, where this topic usually goes by the name of *flavour physics*. We will comment briefly on the leptons, but their full story will only be told in Section 7 when we discuss neutrino masses.

6.1 Diagonalising the Yukawa Interactions

Including three generations, the quark Yukawa terms read (5.22)

$$\mathcal{L}_{\rm Yuk} = -y^d_{ij}\,\bar{Q}^i_L H d^j_R - y^u_{ij}\,\bar{Q}^i_L \tilde{H} u^j_R + {\rm h.c.}\ . \tag{6.1}$$

Here the $i, j = 1, 2, 3$ indices label the generations. We can expand the fields out in terms of the more familiar quark names,

$$\begin{aligned} d^i_R &= \big\{ d_R\ ,\ s_R\ ,\ b_R\ \big\} \quad \text{and} \quad u^i_R = \big\{ u_R\ ,\ c_R\ ,\ t_R\ \big\} \\ Q^i_L &= \begin{pmatrix} u^i_L \\ d^i_L \end{pmatrix} = \left\{ \begin{pmatrix} u_L \\ d_L \end{pmatrix}\ ,\ \begin{pmatrix} c_L \\ s_L \end{pmatrix}\ ,\ \begin{pmatrix} t_L \\ b_L \end{pmatrix}\ \right\}\ . \end{aligned} \tag{6.2}$$

Now the Yukawa couplings y^d and y^u in (6.1) are each 3×3 matrices. Generally these coefficients can be complex, which means that we have $2\times3\times3 = 18$ complex parameters or, equivalently, 36 real parameters. That's a lot of parameters! The purpose of flavour physics is to understand what they mean and to put some order to them.

6.1.1 Counting Yukawa Parameters

Happily, many of these parameters are redundant. At this point, there are two ways to proceed. The first is to follow the restrictions imposed by gauge invariance. The second is to do something practical that helps comparison with experiment. For once, it turns out, these two requirements are rather different.

Let's first bow to the altar of gauge symmetry. The kinetic terms are (5.25)

$$\mathcal{L}_{\rm kin} = -i\sum_{i=1}^{3}\Big(\bar{Q}^i_L\bar{\sigma}^\mu\mathcal{D}_\mu Q^i_L + \bar{u}^i_R\sigma^\mu\mathcal{D}_\mu u^i_R + \bar{d}^i_R\sigma^\mu\mathcal{D}_\mu d^i_R\Big)\ . \tag{6.3}$$

We can always rotate the fermions among themselves, leaving these kinetic terms invariant, by acting with

$$Q^i_L \rightarrow V^i{}_j Q^j_L\ , \quad d^i_R \rightarrow (U^d)^i{}_j d^j_R\ , \quad u^i_R \rightarrow (U^u)^i{}_j u^j_R \tag{6.4}$$

with $V, U^u, U^d \in U(3)$. These transformations leave the kinetic terms invariant, but they change the Yukawa couplings which become

$$y^d \to V^\dagger y^d U^d \quad \text{and} \quad y^u \to V^\dagger y^u U^u \ . \tag{6.5}$$

Such field redefinitions don't change the physics. This means that we can use these rotations to diagonalise one of the Yukawa couplings – say y^u – but, because the same matrix $V \in U(3)$ appears in both the transformations of y^u and y^d, we cannot diagonalise both. The upshot is that if we insist on doing transformations (6.4) that respect the full gauge invariance of the Standard Model, then the mass terms for quarks will typically be non-diagonal.

Ultimately, we'll work with a different set of transformations that do not respect gauge invariance. But, before we do this, it's useful to do a little counting. We've already seen that the two Yukawa matrices y^d and y^u contain 36 real parameters. But we can act with $U(3)^3$ to rotate away some of these. We have $\dim U(3) = 9$, so naively we can remove $3 \times 9 = 27$ parameters. But, a closer inspection, shows that there's an overall $U(1) \subset U(3)^3$ that doesn't affect the Yukawa couplings in (6.5). This means that we can, in fact, eliminate 26 of the parameters in the Yukawa couplings by this method. We're left with

$$36 - 26 = 10 \tag{6.6}$$

physical parameters in y^u and y^d.

In fact, we can be a bit more precise than that. We can think of each of the elements of the Yukawa matrix as consisting of a real parameter, together with a complex phase, so that $y_{ij} = r_{ij} e^{i\theta_{ij}}$. So our original Yukawa matrices y^d and y^u each contain 9 real parameters and 9 complex phases.

How many of each of these are eliminated? Here's a slick argument. A real $N \times N$ unitary matrix $\mathcal{O}$ obeys $\mathcal{O}^T \mathcal{O} = \mathbb{1}$ which is the same thing as an orthogonal matrix. This suggests that, of the N^2 components of a unitary matrix, $\frac{1}{2}N(N-1)$ of them are "real parameters" and the remaining $\frac{1}{2}N(N+1)$ of them are "complex phases". So our $U(3)^3$ consists of 9 real parameters and 18 complex phases, with one complex phase corresponding to the overall $U(1)$ that doesn't affect the Yukawas. This means that, of the 10 physical parameters sitting inside y^d and y^u, we have

$$(2 \times 9) - 9 = 9 \text{ real parameters} \tag{6.7}$$

and

$$(2 \times 9) - (18 - 1) = 1 \text{ complex phase} \ . \tag{6.8}$$

Why is this distinction important? It's because a theory with non-vanishing complex phases violates CP symmetry. We'll look at this more closely in Section 6.4. For now, we note that if we took the Standard Model with $N = 1$ or $N = 2$ generations, then there's no possibility of writing down Yukawa matrices that violate CP. (You can do the same counting as above and see that there are no physical phases remaining after using the $U(N)^3$ symmetries.) The first time that CP violation becomes a possibility is with $N = 3$ and, moreover, it is a possibility that the Standard Model chooses to embrace. Presumably it is no coincidence that $N = 3$ is the minimal number of generations that allows for CP violation although the deeper significance of this remains something that we have yet to fully appreciate.

There is also a remarkable historical fact here. A counting similar to the one above was first done by Kobayashi and Maskawa in 1972 who argued that there must be a third generation of quarks to account for the observed CP violation in hadronic physics. This was before the discovery of the charm quark!

6.1.2 The Mass Eigenbasis

There's nothing wrong with the analysis above, but it doesn't jibe with how we usually do quantum field theory.

Typically, we start with terms in the Lagrangian that are quadratic in fields and make sure that they're diagonal. This is akin to working in the energy, or equivalently mass, eigenbasis of the free theory. We then add interaction terms which, as always in quantum mechanics, change the energy eigenstates. If the interaction terms are small, so that we can use perturbation theory, then this approach is the one that most clearly highlights the physics.

But, as we've seen, if we keep with gauge invariant fields then the transformation (6.5) is not sufficient to diagonalise both Yukawa matrices. We can achieve this only if we're willing to sacrifice gauge invariance and rotate the two components of Q_L independently, so

$$d^i_L \to (V^d)^i{}_j d^j_L \ , \quad u^i_L \to (V^u)^i{}_j u^j_L \ , \quad d^i_R \to (U^d)^i{}_j d^j_R \ , \quad u^i_R \to (U^u)^i{}_j u^j_R \tag{6.9}$$

with $V^u, V^d, U^u, U^d \in U(3)$. While this is necessary if we want to diagonalise both Yukawa matrices, it is only tenable because we have already spontaneously broken the $SU(2)$ gauge symmetry through the Higgs mechanism. The Yukawa couplings now transform independently as

$$y^d \to V^{d\,\dagger} y^d U^d \quad \text{and} \quad y^u \to V^{u\,\dagger} y^u U^u \ . \tag{6.10}$$

By a prudent choice of these unitary matrices, we can now diagonalise both Yukawa couplings

$$y^d = \text{diag}(y^d, y^s, y^b) \quad \text{and} \quad y^u = \text{diag}(y^u, y^c, y^t) \; . \tag{6.11}$$

These Yukawa couplings dictate the masses of the quarks, with

$$m^X = \frac{1}{\sqrt{2}} y^X v \tag{6.12}$$

now with X running over all quark fields, $X = d, u, s, c, b, t$. These diagonal components of the Yukawa matrices are such that they reproduce the quark masses that we met in Section 3,

$$\begin{array}{rccc}
\text{top}: & y^t \approx 1 & \Longrightarrow & m_t \approx 173 \text{ GeV} \\
\text{bottom}: & y^b \approx 2.5 \times 10^{-2} & \Longrightarrow & m_b \approx 4.2 \text{ GeV} \\
\text{charm}: & y^c \approx 7.5 \times 10^{-3} & \Longrightarrow & m_c \approx 1.3 \text{ GeV} \\
\text{strange}: & y^s \approx 5.5 \times 10^{-4} & \Longrightarrow & m_s \approx 93 \text{ MeV} \\
\text{up}: & y^u \approx 1.3 \times 10^{-5} & \Longrightarrow & m_u \approx 2 \text{ MeV} \\
\text{down}: & y^d \approx 2.7 \times 10^{-5} & \Longrightarrow & m_d \approx 5 \text{ MeV}
\end{array}$$

Although we've reduced the masses of the various quarks to dimensionless coupling constants y^X, we currently have no understanding of why the Yukawa couplings take these values. The Yukawa couplings span 5 orders of magnitude and we don't know why. In particular, the top Yukawa is apparently almost exactly one. Is this coincidence? We don't know. (I've not heard any convincing idea for it being anything other than a coincidence.)

Our counting in Section 6.1.1 told us to expect 10 physical parameters in the two Yukawa matrices. Yet now we've diagonalised the two Yukawa matrices to leave ourselves with just 6 masses. Which suggests that there are still 4 other parameters lurking somewhere. As we will see in Section 6.2, these have been pushed, like a bubble in wallpaper, to a different part of the theory.

6.1.3 A Brief Look at Leptons

So far, our attention has been solely on the quarks. We can ask: what's the analogous story for leptons? We decompose the left-handed leptons as (5.20)

$$L_L^i = \left\{ \begin{pmatrix} \nu_L^e \\ e_L \end{pmatrix} , \begin{pmatrix} \nu_L^\mu \\ \mu_L \end{pmatrix} , \begin{pmatrix} \nu_L^\tau \\ \tau_L \end{pmatrix} \right\} \; . \tag{6.13}$$

Their Yukawa couplings are given by

$$\mathcal{L}_{\text{Yuk}} = -y^e_{ij}\,\bar{L}^i_L H e^j_R - y^\nu_{ij}\,\bar{L}^i_L \tilde{H} \nu^j_R + \text{h.c.} \ . \tag{6.14}$$

However, as we mentioned previously, there remains a question mark about the existence of the right-handed neutrino. This is all tied up with how the neutrinos get a mass, a subject that we will discuss in Section 7. To avoid getting into this can of worms, lets for now assume that there is no right-handed neutrino, in which case the lepton Yukawa terms are just

$$\mathcal{L}_{\text{Yuk}} = -y^e_{ij}\,\bar{L}^i_L H e^j_R + \text{h.c.} \ . \tag{6.15}$$

Then we have a single 3×3 Yukawa matrix y^e and there is no obstacle to rotating the two fields, L_L and e_R, to ensure that this matrix is diagonal

$$y^e = \text{diag}(y^e, y^\mu, y^\tau) \ . \tag{6.16}$$

The values of these Yukawa couplings determine the masses of the electron, muon, and tau through the same formula (6.12) as the quarks. The experimentally measured values of these couplings are

$$\begin{aligned} \text{tau}: &\quad y^\tau \approx 1 \times 10^{-2} &\Longrightarrow &\quad m_\tau \approx 1.8 \text{ GeV} \\ \text{muon}: &\quad y^\mu \approx 6.1 \times 10^{-4} &\Longrightarrow &\quad m_\mu \approx 106 \text{ MeV} \\ \text{electron}: &\quad y^e \approx 2.9 \times 10^{-6} &\Longrightarrow &\quad m_e \approx 0.5 \text{ MeV} \ . \end{aligned}$$

We won't say any more about leptons in this section. Instead, we'll return to the quarks where the need to simultaneously diagonalise two Yukawa matrices implies something interesting. Having understood what happens for quarks, we'll then return to leptons in Section 7 and see how something similar plays out in the world of neutrinos.

6.2 The CKM Matrix

Although we've diagonalised the quark mass matrices, there's a price to pay. And this comes in the interactions with the gauge bosons. We computed these for a single generation in (5.85) where we saw that the interactions take the form

$$\mathcal{L}_{\text{kin}}\Big|_{\text{weak}} = -\frac{e}{\sqrt{2}\sin\theta_W}(W^+_\mu J^\mu_+ - W^-_\mu J^\mu_-) - \frac{e}{\sin\theta_W\ \cos\theta_W} Z^\mu J^Z_\mu - eA^\mu J^{\text{EM}}_\mu \tag{6.17}$$

with the various currents computed in (5.86), (5.87) and (5.88). To extend these results to multiple generations is easy: we simply sum over all generations. For our immediate purposes, we will ignore the coupling to leptons so the electromagnetic current (5.86) becomes

$$J_\mu^{\rm EM} = \sum_{i=1}^{3} \left(\frac{2}{3}(\bar{u}^i_L \bar{\sigma}_\mu u^i_L + \bar{u}^i_R \sigma^\mu u^i_R) - \frac{1}{3}(\bar{d}^i_L \bar{\sigma}_\mu d^i_L + \bar{d}^i_R \sigma^\mu d^i_R) \right) . \tag{6.18}$$

The coupling to the Z bosons (5.87) is

$$J_\mu^Z = \frac{1}{2} \sum_{i=1}^{3} \left(\bar{u}^i_L \bar{\sigma}_\mu u^i_L - \bar{d}^i_L \bar{\sigma}_\mu d^i_L \right) - \sin^2\theta_W \, J_\mu^{\rm EM} . \tag{6.19}$$

And, finally, the couplings to the W bosons (5.88) are

$$J_\mu^+ = \sum_{i=1}^{3} \bar{u}^i_L \bar{\sigma}_\mu d^i_L \quad \text{and} \quad J_\mu^- = \sum_{i=1}^{3} \bar{d}^i_L \bar{\sigma}_\mu u^i_L . \tag{6.20}$$

Each of these currents is diagonal in flavour, but this is *before* we do the rotation (6.9) needed to diagonalise the Yukawa matrices. What becomes of the currents after we rotate the quarks to go to the mass eigenbasis?

Neither the electromagnetic current $J_\mu^{\rm EM}$ nor the Z boson current J_μ^Z are affected by the change of basis (6.9). This is because the quarks in these currents always appear together with the corresponding anti-quark as $\bar{q}^i q^i$.

The novelty comes when we look at the W boson current. Here there are different kinds of quarks, $\bar{u}^i_L d^i_L$ and these rotate differently when we diagonalise the Yukawa matrices. This means that if we work in the mass eigenbasis, the coupling to the W boson takes the form

$$J_\mu^+ = \bar{u}^i_L \bar{\sigma}_\mu V_{ij} \, d^j_L \quad \text{and} \quad J_\mu^- = \bar{d}^i_L \bar{\sigma}_\mu V^\dagger_{ij} \, u^j_L . \tag{6.21}$$

where

$$V = (V^u)^\dagger V^d \tag{6.22}$$

captures the mismatch between the rotations of the left-handed up and down quarks. This matrix V is the *CKM matrix*, sometimes denoted as $V_{\rm CKM}$ and named after Cabibbo, Kobayashi and Maskawa. This is where the remaining parameters of the Yukawa couplings are hiding after we diagonalise them.

6.2.1 Two Generations and the Cabibbo Angle

Before we turn to the full CKM matrix, it's useful to look at what happens when we have just two generations. In this case the analogous matrix V is a 2×2 matrix. Moreover, as we can see from the form (6.22), the matrix is necessarily unitary. The most general unitary 2×2 matrix can be written as a rotation matrix, dressed with various complex phases

$$V_{2\times 2} = \begin{pmatrix} e^{i\delta_1}\cos\theta & e^{i\delta_2}\sin\theta \\ -e^{-i\delta_3}\sin\theta & e^{i\delta_4}\cos\theta \end{pmatrix} \tag{6.23}$$

where unitarity requires $\delta_1 - \delta_2 - \delta_3 + \delta_4 = 0$. Here we see the decomposition that we described in Section 6.1.1: the four parameters comprise of 3 complex phases and a single real angle θ.

However, we can eliminate all the complex phases in this case. This is because the diagonal mass terms are invariant under the $U(1)^4$ symmetry

$$d^i_{R,L} \rightarrow e^{i\alpha_i}\, d^i_{R,L} \quad \text{and} \quad u^i_{R,L} \rightarrow e^{i\beta_i} u^i_{R,L} \quad \text{with} \quad i = 1,2\ . \tag{6.24}$$

Of these, $U(1)^3$ acts on $V_{2\times 2}$, leaving the overall sum $\delta_1 - \delta_2 - \delta_3 + \delta_4$ unchanged. This means that the lone physical parameter in $V_{2\times 2}$ is the angle θ. This is known as the *Cabibbo angle* and we denote it $\theta = \theta_c$. We have

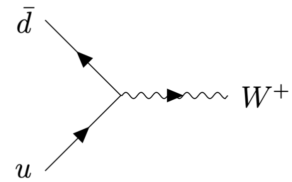
$$V_{2\times 2} = \begin{pmatrix} \cos\theta_c & \sin\theta_c \\ -\sin\theta_c & \cos\theta_c \end{pmatrix}\ . \tag{6.25}$$

To see the physical meaning of this, we can return to the W boson currents (6.21). For two generations, the quark labels are $d = (d, s)$ and $u = (u, c)$, so the current is

$$J^+_\mu = \cos\theta_c\,(\bar{u}_L\bar{\sigma}_\mu d_L + \bar{c}_L\bar{\sigma}_\mu s_L) + \sin\theta_c\,(\bar{u}_L\bar{\sigma}_\mu s_L - \bar{c}_L\bar{\sigma}_\mu d_L)\ . \tag{6.26}$$

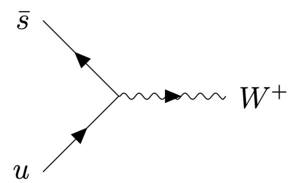
We see that we get two terms: the first, proportional to $\cos\theta_c$, relates quarks within the same generation: up to down, and charm to strange. The second term, proportional to $\sin\theta_c$, relates quarks within different generations: up to strange, and charm to down. This is what the additional parameters in the Yukawa matrices buy us.

This means that we have additional Feynman diagrams. The diagram that we met previously comes with a factor of $\cos\theta_c$,



$\sim\ \cos\theta_c$

But we also get a diagram that relates quarks in different generations,



$\sim\ \sin\theta_c$

This inter-generational mixing occurs only for interactions involving W bosons. They are referred to as *flavour changing currents*.

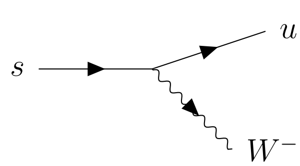
The value of the Cabibbo angle is, like all other things Yukawaesque, something that we cannot predict from first principles and have to go out and measure. It takes the value

$$\sin\theta_c \approx 0.22 \quad \Longrightarrow \quad \theta_c \approx \frac{\pi}{14} \approx 13^\circ\ . \tag{6.27}$$

We don't currently have any deeper explanation for this value.

This resolves an issue that we gracefully swept under the rug when describing weak decays in Section 5.3. How does the kaon decay?

Consider the kaon K^- whose quark content is $\bar{u}s$. If there was no way for the flavour to change, then there would be nowhere for the strange quark to go. It cannot decay into a charm quark because that is significantly heavier. But the quark mixing described above means that there is a Feynman diagram that allows the strange quark to decay to an up quark,



The resulting up quark can then annihilate with the $\bar{u}$ in the kaon, while the W^- can decay into, say, an electron and anti-neutrino in the usual way. This Feynman diagram comes with a factor of $\sin\theta_c$ which, in turn, means that the decay rate is suppressed by $\sin^2\theta_c \approx 0.05$. This results in an increased lifetime for mesons containing strange quarks.

6.2.2 Three Generations and the CKM Matrix

Now we can turn to the full CKM matrix (6.22). This is a unitary 3×3 matrix with the general form

$$V_{\text{CKM}} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} . \tag{6.28}$$

Each of these elements can, in principle, be complex and we will discuss the phases shortly. But for now we can give the experimentally measured absolute values, which are roughly

$$|V_{\text{CKM}}| = \begin{pmatrix} |V_{ud}| & |V_{us}| & |V_{ub}| \\ |V_{cd}| & |V_{cs}| & |V_{cb}| \\ |V_{td}| & |V_{ts}| & |V_{tb}| \end{pmatrix} \approx \begin{pmatrix} 0.97 & 0.22 & 0.004 \\ 0.22 & 0.97 & 0.04 \\ 0.009 & 0.04 & 0.999 \end{pmatrix} . \tag{6.29}$$

You can see the Cabibbo angle sitting there in $V_{us} \approx \sin\theta_c \approx 0.22$.

Just like we have no understanding of why the Cabibbo angle takes its particular value, nor do we have any good understanding of the CKM matrix. As you can see, it's not far from a diagonal matrix, with the Cabibbo terms V_{us} and V_{cd} the only ones that aren't completely tiny. We don't know why.

Not all the parameters in matrix (6.29) are independent. The CKM matrix is unitary and a general unitary matrix contains a total 9 parameters which decompose as 3 real angles and 6 phases. But, as in the 2×2 case, we can eliminate some of these because the diagonal mass terms are invariant under the $U(1)^6$ symmetry

$$d^i_{R,L} \rightarrow e^{i\alpha_i}\, d^i_{R,L} \quad \text{and} \quad u^i_{R,L} \rightarrow e^{i\beta_i} u^i_{R,L} . \tag{6.30}$$

Of these, $U(1)^5$ acts on the CKM matrix and can be used to set 5 of the phases to zero. The $U(1)$ symmetry that fails to act has α_i and β_i all equal and corresponds to the baryon number symmetry of the Standard Model. All of which means that we expect the CKM matrix to depend on four parameters, 3 real angles and one complex phase. This agrees with our counting in Section 6.1.1.

This prompts the question: how should we write the CKM matrix in terms of these four parameters? There's no right and wrong answer here: merely more or less convenient ways of doing things. One of the most standard choices is to take V_{ud}, V_{us}, V_{cb} and V_{tb} to be real and to write the CKM matrix in terms of three angle θ_{12}, θ_{13} and θ_{23}, together with a complex phase $e^{i\delta}$, constructed in a similar way to the Euler angles for rotating rigid bodies,

$$\begin{aligned} V_{\rm CKM} &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & c_{23} & s_{23} \\ 0 & -s_{23} & c_{23} \end{pmatrix} \begin{pmatrix} c_{13} & 0 & s_{13}e^{-i\delta} \\ 0 & 1 & 0 \\ -s_{13}e^{i\delta} & 0 & c_{13} \end{pmatrix} \begin{pmatrix} c_{12} & s_{12} & 0 \\ -s_{12} & c_{12} & 0 \\ 0 & 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} c_{12}\,c_{13} & s_{12}\,c_{13} & s_{13}e^{-i\delta} \\ -s_{12}\,c_{23} - c_{12}\,s_{23}\,s_{13}e^{i\delta} & c_{12}\,c_{23} - s_{12}\,s_{23}\,s_{13}e^{i\delta} & s_{23}\,c_{13} \\ s_{12}\,s_{23} - c_{12}\,c_{23}\,s_{13}e^{i\delta} & -c_{12}\,s_{23} - s_{12}\,c_{23}\,s_{13}e^{i\delta} & c_{23}\,c_{13} \end{pmatrix} . \end{aligned} \tag{6.31}$$

where we're using the convention

$$c_{ij} = \cos\theta_{ij} \quad \text{and} \quad s_{ij} = \sin\theta_{ij} \ . \tag{6.32}$$

Here $\theta_{12} = \theta_c$ is the Cabibbo angle. The angles are given in degrees by

$$\begin{aligned} \theta_{12} &= 13.02^\circ \pm 0.004^\circ \\ \theta_{13} &= 0.20^\circ \pm 0.02^\circ \\ \theta_{23} &= 2.56^\circ \pm 0.03^\circ \\ \delta &= 69^\circ \pm 5^\circ \ . \end{aligned} \tag{6.33}$$

We see that the complex phase δ is not at all small, but it appears in the elements of the CKM matrix multiplying $\sin\theta_{13}$ so its effects are tiny. We will see these effects in Section 6.4.

It's worth pausing to take in a bigger perspective here. In the first part of Section 5, we described how the matter content of the Standard Model interacts with the different forces. There we found a beautiful consistent picture – a perfect jigsaw – in which the interactions were largely forced upon us by the consistency requirements of anomaly cancellation. For a theoretical physicist, it is really the dream scenario. This contrasts starkly with the story of flavour. Even focussing solely on the quarks, we find that there are 6 Yukawa couplings that determine their mass, plus a further 4 entries of the CKM matrix that determine their mixing. And none of these parameters are fixed or understood at a deeper level.

Somewhat ironically, much of this complexity can be traced to the simplicity of the Higgs. Yang-Mills theories and Weyl fermions all come with subtleties that are responsible for the quantum consistency conditions. But the Higgs is a spin 0 particle and, as we observed earlier: scalars are basic. There are no consistency conditions beyond the requirements of Lorentz invariance and gauge invariance so the Higgs can do what it likes. This is what leads to the plethora of extra parameters that we've seen, and it is why the Higgs is simultaneously both the simplest and the most complicated field in the Standard Model.

Turning this on its head, the flavour sector of the Standard Model may well offer a unique opportunity. The structure of quark masses, together with the CKM matrix, surely contains clues for what lies beyond the Standard Model. Why the hierarchy of masses? Why these values of the CKM matrix? Hopefully one day we will find out.

6.2.3 The Wolfenstein Parameterisation

There is a way to write the CKM matrix that highlights the numerical values that the various elements take. This is motivated by the observation that the absolute values (6.29) seem to roughly follow the pattern

$$|V_{\rm CKM}| \sim \begin{pmatrix} 1 & \lambda & \lambda^3 \\ \lambda & 1 & \lambda^2 \\ \lambda^3 & \lambda^2 & 1 \end{pmatrix} \tag{6.34}$$

with $\lambda \approx 0.2$. The idea of the Wolfenstein parameterisation is that we take this as a starting point and then add corrections. We parameterise these corrections by one real number that we call A and one complex number that we write as $\rho - i\eta$, so that the overall number of parameters is the same as the CKM matrix. Then numbers A and $\rho - i\eta$ are all of order one. We then write

$$V_{\rm CKM} \approx \begin{pmatrix} 1-\lambda^2/2 & \lambda & A\lambda^3(\rho - i\eta) \\ -\lambda & 1-\lambda^2/2 & A\lambda^2 \\ A\lambda^3(1-\rho-i\eta) & -A\lambda^2 & 1 \end{pmatrix} . \tag{6.35}$$

You will recognise the upper-left 2×2 matrix as the Taylor expansion of $V_{2\times 2}$ given in (6.25), with $\lambda = \theta_c$.

The Wolfenstein parameterisation (6.35) is not unitary. It sacrifices that property of the CKM matrix to highlight some other numerical structure. Note, in particular, that only the far off-diagonal elements V_{ub} and V_{td} have an imaginary piece. This contrasts

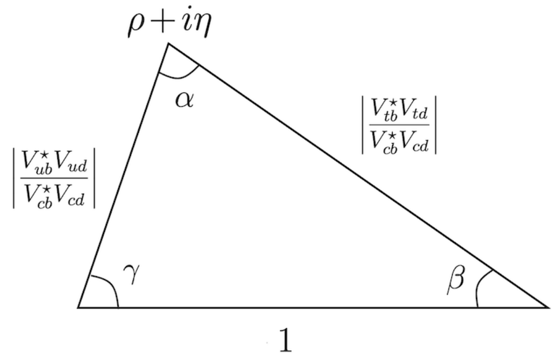


Figure 17. The unitarity triangle, plotted on the complex plane.

with the exact CKM matrix (6.31) where V_{cd}, V_{cs} and V_{ts} also have imaginary parts but you can check that these are one or two orders of magnitude smaller than $\text{Im}(V_{ub})$ and $\text{Im}(V_{td})$, which is why they are neglected in (6.35).

6.2.4 The Unitarity Triangle

The CKM matrix is unitary,

$$V_{\text{CKM}}^\dagger V_{\text{CKM}} = \mathbb{1} \ . \tag{6.36}$$

This means, in particular, that a given row of $V_{\text{CKM}}^\dagger$ is orthogonal to two of the three columns of V_{CKM}.

For example, if we contract the middle row of $V_{\text{CKM}}^\dagger$ with the first column of V_{CKM}, we have the requirement

$$\sum_{i=1}^{3} V_{is}^\star V_{id} = V_{us}^\star V_{ud} + V_{cs}^\star V_{cd} + V_{ts}^\star V_{td} = 0 \ . \tag{6.37}$$

If we look at this in the Wolfenstein parameterisation, then we see that the first two terms are of order λ while the final term is of order λ^5. This means that the equation essentially boils down to the requirement that $V_{us}^\star V_{ud} \approx V_{cs}^\star V_{cd}$.

We get something more interesting if we contract the bottom row of $V_{\text{CKM}}^\dagger$ with the first column of V_{CKM}. This reads

$$\sum_{i=1}^{3} V_{ib}^\star V_{id} = V_{ub}^\star V_{ud} + V_{cb}^\star V_{cd} + V_{tb}^\star V_{td} = 0 \ . \tag{6.38}$$

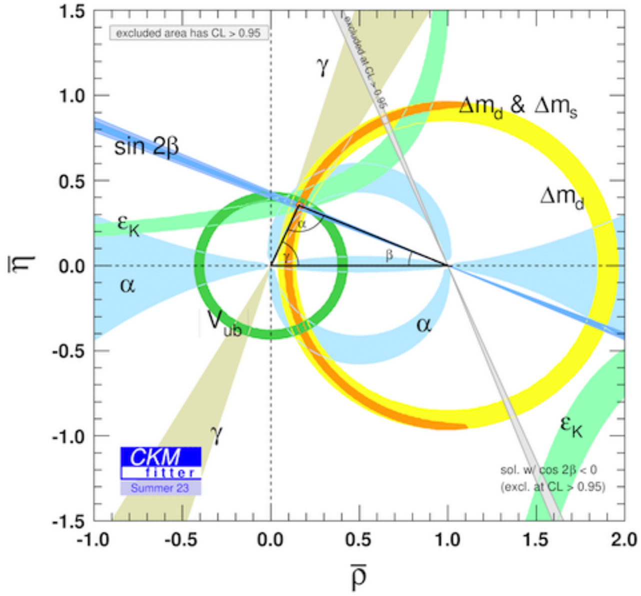


Figure 18. The experimental data, constraining the unitarity triangle. Taken from the CKMfitter website.

Now each of the terms has a comparable magnitude $\sim \lambda^3$, but they have different phases. But we can go out and measure each of the terms in this equation and check if they do, indeed, add up to zero. This gives us a very useful test on the whole framework of flavour, not to mention an opportunity to search for physics beyond the Standard model. So far, it is a test that the Standard Model has passed with flying colours.

To perform this test, it's traditional to divide by $V_{cb}^\star V_{cd}$ and write the constraint as

$$\frac{V_{ub}^\star V_{ud}}{V_{cb}^\star V_{cd}} + 1 + \frac{V_{tb}^\star V_{td}}{V_{cb}^\star V_{cd}} = 0 \ . \tag{6.39}$$

Each of the two non-trivial terms is a complex number whose magnitude is of order 1. We can then plot these numbers on the complex plane. You can check that, to leading order in λ, we have $V_{ub}^\star V_{ud}/V_{cb}^\star V_{cd} = -(\rho + i\eta)$. The result is called the *unitarity triangle* and is shown in Figure 17. The data from a multitude of experiments, constraining the corners of the triangle, is shown in Figure 18.

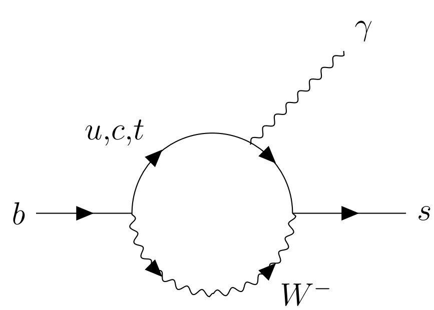
6.3 Flavour Changing Neutral Currents

When we diagonalise the mass matrices for quarks, neither the electromagnetic current (6.18) nor the Z boson current (6.19) are affected. It's only the W boson current that couples up-type and down-type quarks that gets hit by this diagonalisation, and that is where the CKM matrix sits.

This means that the tree level processes that change one generation of quarks with another always involve *charged* currents. So, for example, we can change a strange quark into an up quark by emitting a W boson. But we can't change a strange quark directly into a down quark which has the same charge. We phrase this as saying that there are no tree level *flavour changing neutral currents*, often abbreviated as *FCNC*.

That's not to say that flavour changing neutral currents don't exist. We can cook them up at loop level, and an example is given by the neutral kaon mixing that we will discuss in Section 6.4 where K^0 turns into the $\bar{K}^0$ by exchanging s and d quarks. But it does mean that these processes are suppressed because they can only come from loop diagrams.

In fact, the situation is even more interesting than that. The structure of the Standard Model is such that these one-loop contributions are further suppressed. A particularly simple example arises if we look at how a bottom quark might decay into a strange quark, with $b \to s\gamma$. The simplest Feynman diagrams take the form



As shown, we should sum over all up-like quarks running in the loop. But this means that the amplitude comes with factors of the CKM matrix,

$$\mathcal{M} \sim \sum_{i=1}^{3} V_{ib} V_{is}^{\star} = 0 \tag{6.40}$$

which vanishes by unitarity of the CKM matrix. This observation is known as the *GIM mechanism*, named after Glashow, Iliopoulos, and Maiani.

In fact, the cancellation isn't precise because the quarks running in the loop have different masses. This means that we actually get terms that are of the form

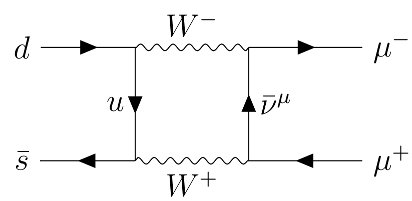
$$\mathcal{M} \sim \sum_{i=1}^{3} V_{ib} V_{is}^{\star} f(m_i) \tag{6.41}$$

for some function $f(m_i)$. These diagrams also contain a W boson running in the loop and, because $m_i \ll m_W$ for each of the u, c, and b quarks, it can be shown that this function takes the form $f(m_i) \sim m_i^2/m_W^2$.

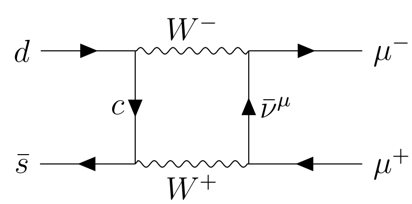
Remarkably, this kind of argument was first used by GIM to predict the existence of the charm quark in 1970, before its discovery in 1974. (This was also before the Standard Model had been fully constructed, and certainly before the importance of anomaly cancellation was realised.) The issue arose from looking at decays of the neutral kaon K^0 with quark content $d\bar{s}$ to a pair of muons.

$$K^0 \to \mu^+\mu^- \ . \tag{6.42}$$

This proceeds through the one loop diagram



The problem is that this diagram gives a contribution to $K^0 \to \mu^+\mu^-$ that is much greater than observed. The suggestion by GIM was to add an additional quark – the charm – that contributes with a similar diagram



Under the (obviously wrong!) assumption that the up and charm quark have similar masses, these two diagrams would cancel. This is because each is proportional to the appropriate CKM matrix elements which, with just two generations, can be written in terms of the Cabibbo angle. The resulting amplitude scales as

$$\mathcal{M} \sim V_{ud} V_{us}^{\star} + V_{cd} V_{cs}^{\star} = \cos\theta_c \, \sin\theta_c - \sin\theta_c \, \cos\theta_c = 0 \ . \tag{6.43}$$

This illustrates the general idea captured in (6.40). When you take into account the fact $m_u \neq m_c$, there is still partial cancellation but it is not complete. The amplitude scales as

$$\mathcal{M} \sim g^4 \frac{m_c^2}{m_W^2}\left(1 - \frac{m_u^2}{m_c^2}\right) \; . \tag{6.44}$$

It's that overall factor of $g^4 m_c^2/m_W^2$ that makes the decay rate to muons so small.

The lack of flavour changing neutral currents is special to the Standard Model and any attempt to introduce new physics that goes beyond the Standard Model will typically generate these currents. This means that experiments involving neutral currents provide an important class of constraints on what theories govern the next level of reality.

Here's an example. It's possible that flavour changing neutral currents could be generated by the Higgs field. But that doesn't happen in the Standard Model because the Higgs field couples, like its vev, to the mass matrix which, as we have seen, can be diagonalised for both up and down sectors. This means that we have, for example,

$$\mathcal{L}_{\text{Yuk}} = -y^d_{ij}(v+h)\bar{d}^i_L d^j_R \tag{6.45}$$

with a diagonal Yukawa matrix $y^d = \text{diag}(y^d, y^s, y^b)$. There is a similar term for the up sector.

Now suppose that we had a theory with two Higgs fields, H_1 and H_2. We'll assume (without any justification) that their vacuum expectation values align, so that $\langle H_1 \rangle = (0, v_1)$ and $\langle H_2 \rangle = (0, v_2)$. Then we should include two sets of Yukawa interactions that, for the down sector, take the form

$$\mathcal{L}_{\text{Yuk}} = y^1_{ij}(v_1 + h_1)\bar{d}^i_L d^j_R + y^2_{ij}(v_2 + h_2)\bar{d}^i_L d^j_R \; . \tag{6.46}$$

Now the fermion mass matrix is $M_{ij} = v_1 y^1_{ij} + v_2 y^2_{ij}$. We could rotate the quarks to ensure that this is diagonal, but the Higgs fields h_1 and h_2 will couple to the fermions through y^1_{ij} and y^2_{ij} respectively and there is no reason that these will be diagonal. This means that in a model with two Higgs fields, there will generically be flavour changing neutral currents at tree level, mediated by the two Higgses, in contradiction to what is observed in experiment. If you want to make a two-Higgs model fly (and many people do), then you need to find a way to suppress these currents.

6.4 CP Violation

The complex phase $e^{i\delta}$ in the CKM matrix (6.31) is important. This is because it is responsible for the laws of physics violating the symmetry CP. Said differently, because any relativistic quantum field theory is invariant under CPT, a non-vanishing phase δ means that the laws of physics are not invariant under time reversal.

We discussed the discrete symmetries of C, P and T in Section 1.4. There we saw that parity and charge conjugation both exchange left-handed and right-handed spinors. The electroweak sector of the Standard Model violates both parity and charge conjugation from the get go because, as a gauge chiral theory, the left- and right-handed fermions transform differently under the gauge symmetries. But the combination CP is more subtle.

We derived how CP acts on left-handed and right-handed Weyl spinors in (1.132). For fermions with real masses, we have

$$CP : \psi_L(t,\mathbf{x}) \mapsto \mp i\sigma^2\psi_L^\star(t,-\mathbf{x}) \quad \text{and} \quad CP : \psi_R(t,\mathbf{x}) \mapsto \pm i\sigma^2\psi_R^\star(t,-\mathbf{x}) \ . \tag{6.47}$$

From this, you can check that the fermion bilinear $\bar\psi_L\psi_R$ transforms under CP as

$$CP : \bar\psi_L\psi_R(t,\mathbf{x}) \mapsto \bar\psi_R\psi_L(t,-\mathbf{x}) \ . \tag{6.48}$$

A Yukawa coupling between two fermions and a scalar ϕ takes the form

$$\mathcal{L}_{\rm Yuk} = y\bar\psi_L\phi\psi_R + y^\star\bar\psi_R\phi^\dagger\psi_L \tag{6.49}$$

where the second term is what was hiding in the + h.c. in our previous expressions (5.74) and (6.1). The scalar gets mapped to its conjugate under CP, so these two terms get mapped into each other, with $CP : \bar\psi_L\phi\psi_R \mapsto \bar\psi_R\phi^\dagger\psi_L$. This means that the Yukawa terms (6.49) are invariant under CP only if the Yukawa coupling is real, so $y = y^\star$.

There's a quicker argument that gets us to the same conclusion. This is to note that T is an anti-unitary symmetry: it maps $i \mapsto -i$. Only theories with real parameters are invariant under time reversal.

From the structure of CKM matrix (6.31), we see that CP violation will only occur in processes that mix different generations. Moreover, as emphasised in the Wolfenstein parameterisation (6.35), CP violation will be strongest in processes that mix the first and third generations of quarks, even though this is the smallest element of the CKM matrix in magnitude.

6.4.1 How to Think of the Breaking of Time Reversal

The fact that the fundamental laws of physics are not invariant under time reversal is an extraordinarily big deal. And yet, when we get to see the details one can't help but be a little disappointed. It just boils down to a complex phase $e^{i\delta}$ in the CKM matrix that can't be removed by a field redefinition. Surely there's more to it than that!

The purpose of this section is to give some intuition for why such a complex phase results in the breaking of time reversal symmetry. We will do this by providing an analogy with the meaning of time-reversal in quantum mechanics.

Let's return to our Yukawa coupling matrices y^d_{ij} and y^u_{ij} in (6.1). We will consider the general case where we have $i, j = 1, \ldots, N$ generations rather than setting $N = 3$. Before we do any field redefinitions, each of these is an $N \times N$ complex matrix. Any complex matrix y can be written in terms of a matrix polar decomposition as

$$y = YU \ . \tag{6.50}$$

with U a unitary matrix and Y a Hermitian matrix, so $Y = Y^\dagger$. Because Y is Hermitian, it necessarily has real eigenvalues and these can always be taken to be non-negative. This is the matrix version of writing a complex number as $z = re^{i\theta}$. But, for each Yukawa coupling, the unitary matrix U can be absorbed into a redefinition of the right-handed quarks, as in (6.4). This means that we can always take the Yukawa matrices to be Hermitian. We will denote these two Hermitian Yukawa matrices as Y^u_{ij} and Y^u_{ij}.

One benefit of having Hermitian Yukawa matrices is that we can start to import some intuition from quantum mechanics. For example, we can consider conjugating the two matrices by a unitary matrix V,

$$Y^d \to V^\dagger Y^d V \quad \text{and} \quad Y^u \to V^\dagger Y^u V \ . \tag{6.51}$$

These are the remaining field redefinitions (6.5) that keep the matrices Hermitian. We know from quantum mechanics that it is possible to simultaneously diagonalise both Y^d and Y^u by such a transformation if and only if

$$[Y^d, Y^u] = 0 \ . \tag{6.52}$$

The fact that this condition isn't satisfied for the Yukawa matrices of the Standard Model is what leads to the CKM matrix. Said differently, the CKM matrix is a measure of the failure of Y^d and Y^u to commute.

There's also a less familiar question that we can ask: is it possible to find a unitary matrix V such that, by conjugation (6.51), we can make both Y^d and Y^u real? If this is possible, we will say that Y^d and Y^u are *mutually real*. First note that if Y^d and Y^u are simultaneously diagonalisable then they are necessarily mutually real. But the requirement that matrices are mutually real is weaker than the requirement that they commute.

Next we will show that if Y^d and Y^u are mutually real then the CKM matrix is real. (In fact, the converse also holds: a real CKM matrix implies that Y^d and Y^u are mutually real.) To see this, note that if $V^\dagger Y^d V$ and $V^\dagger Y^u V$ are both real then each can be diagaonalised by a (different) *orthogonal* real matrix, O^d and O^u:

$$(O^d)^T V^\dagger Y^d V O^d = \text{diag}(y^d, y^s, \ldots) \quad \text{and} \quad (O^u)^T V^\dagger Y^u V O^u = \text{diag}(y^u, y^c, \ldots) \tag{6.53}$$

Comparing to (6.10), we see that we can identify the unitary matrices V^d and V^u that diagonalise the Yukawa interactions as $V^d = VO^d$ and $V^u = VO^u$ so the CKM matrix is

$$V_{\text{CKM}} = (V^u)^\dagger V^d = (O^u)^T O^d \ . \tag{6.54}$$

This is now real as both O^u and O^d are real.

So far we've just phrased our previous results in a slightly different language. The Standard Model is not invariant under time reversal if the CKM matrix is not real. And this, in turn, holds if the Hermitian Yukawa matrices are not mutually real. Now we'd like to explain *why* this should result in breaking time reversal. We will do so by analogy with quantum mechanics.

A Quantum Mechanical Analogy

To this end, suppose that we have two $N \times N$ Hermitian matrices A and B that act on an N-dimensional Hilbert space. These will be analogous to our two Yukawa matrices Y^d and Y^u. What is the implication in quantum mechanics if A and B are mutually real? The answer, as we now explain, is related to time reversal invariance.

One particularly physical way to think of this is to take A to be the Hamiltonian of the system. We then measure B. Suppose that we find ourselves in one eigenstate $|b_i\rangle$ of B, evolve for some time under A, and then measure B again. The probability that we find ourselves in an eigenstate $|b_j\rangle$ is

$$\begin{aligned} P(i \to j; t) &= \left|\langle b_j|e^{-iAt}|b_i\rangle\right|^2 \\ &= \langle b_j|e^{-iAt}|b_i\rangle\langle b_i|e^{+iAt}|b_j\rangle \ . \end{aligned} \tag{6.55}$$

We can compare this to the same probability if we instead run time backwards

$$
\begin{aligned}
P(i \to j; -t) &= \left|\langle b_j|e^{+iAt}|b_i\rangle\right|^2 \\
&= \langle b_j|e^{+iAt}|b_i\rangle\langle b_i|e^{-iAt}|b_j\rangle \ . \qquad (6.56)
\end{aligned}
$$

First we see that

$$
P(i \to j; -t) = P(j \to i; +t) \ . \qquad (6.57)
$$

Now we can ask about time reversal invariance. When is the probability the same, regardless of whether we run backwards or forwards in time? In other words, when is $P(i \to j; t) = P(j \to i; t)$?

The answer is that these two probabilities are equal whenever A and B are mutually real or, equivalently, whenever the CKM-type matrix is real. First we introduce some notation. We introduce unitary matrices V_A and V_B that diagonalise A and B,

$$
V_A^\dagger A V_A = \text{diag}(a_1, \ldots, a_N) \quad \text{and} \quad V_B^\dagger B V_B = \text{diag}(b_1, \ldots, b_N) \ . \qquad (6.58)
$$

If we introduce the basis $|i\rangle$, then the eigenvectors of A are

$$
|a_i\rangle = (V_A)_{ij}|j\rangle \quad \Longrightarrow \quad A|a_i\rangle = a_i|a_i\rangle \qquad (6.59)
$$

and similar for B. If we're avoiding using subscripts, we will sometimes write this as $|a_i\rangle = V_A|i\rangle$. The eigenvectors of A and B are then related by

$$
|b_i\rangle = U_{ij}|a_j\rangle \quad \text{with} \quad U_{ij} = (V_B V_A^\dagger)_{ij} \ . \qquad (6.60)
$$

Notice that this isn't quite of the CKM matrix form (6.22); the CKM matrix is $V_{\text{CKM}} = V_B^\dagger V_A$ while here we have $U = V_B V_A^\dagger$. We've already shown that V_{CKM} is real if A and B are mutually real. It will turn out that the probability is time reversal invariant if we can pick phases for the bases $|a_i\rangle$ and $|b_i\rangle$ so that U is also real.

To show this, we will consider an anti-unitary time reversal operator Θ in our quantum mechanics. We will show that whenever A and B are mutually real, it's possible to construct a time reversal operator such that $[\Theta, A] = [\Theta, B] = 0$. We do this by showing that the eigenvectors $|a_i\rangle$ and $|b_i\rangle$, with suitably chosen phases, are also eigenvectors of Θ.

We start by taking the basis of states $|i\rangle$, with $i = 1, \ldots, N$, and introduce the anti-linear involution K defined by

$$K|i\rangle = |i\rangle \ . \tag{6.61}$$

If K were a linear operator, this equation would tell us that $K = \mathbb{1}$. But k is an anti-linear operator which means that, for any $\alpha \in \mathbb{C}$, we have

$$K(\alpha|i\rangle) = \alpha^\star|i\rangle \ . \tag{6.62}$$

Now we define the time reversal operator

$$\Theta = V_A K V_A^\dagger \ . \tag{6.63}$$

With this definition, it's straightforward to check that the eigenvectors of A, $|a_i\rangle$, are also eigenvectors of time reversal

$$\Theta|a_i\rangle = |a_i\rangle \ . \tag{6.64}$$

But, importantly, so too are the eigenvectors of B provided that A and B are mutually real. This follows by plugging in the various definitions,

$$\Theta|b_i\rangle = V_A K V_A^\dagger V_B|i\rangle = V_A(V_A^\dagger V_B)^\star K|i\rangle n = V_A V_A^\dagger V_B|i\rangle = |b_i\rangle \tag{6.65}$$

where, in the third equality, we've used the fact that the CKM-like matrix $V_A^\dagger V_B$ is real if A and B are mutually real.

But we can look at what this time reversal means for the matrix U defined in (6.60). We have

$$\Theta|b_i\rangle = \Theta U_{ij}|a_j\rangle = U_{ij}^\star|a_j\rangle = |b_i\rangle = U_{ij}|a_j\rangle \quad \Longrightarrow \quad U_{ij}^\star = U_{ij} \ . \tag{6.66}$$

Finally, we can now use this to prove that our forward probability (6.55) and backward probability (6.56) are equal, so that $P(i \to j; t) = P(j \to i; t)$. We could do this directly using the time reversal operator Θ, but it's a bit fiddly as we need to think about how anti-unitary operators act on the dual vectors $|b_i\rangle$. Instead, we can proceed in a more pedestrian fashion. We have

$$\begin{aligned} \langle b_j|e^{-iAt}|b_i\rangle &= \sum_k \langle a_k|U_{kj}^\star U_{ki} e^{-ia_k t}|a_k\rangle \\ &= \sum_k \langle a_k|U_{kj} U_{ki}^\star e^{-ia_k t}|a_k\rangle = \langle b_i|e^{-iAt}|b_j\rangle \end{aligned} \tag{6.67}$$

where, in the second line, we've used the fact that $U_{ij}^\star = U_{ij}$. This is exactly what we need to equate the probabilities in the forwards (6.55) and backwards (6.56) time directions.

This quantum mechanical story was designed to give some intuition for why having two mutually real Hermitian matrices – A and B above, or Y^d and Y^u in the Standard Model – implies time reversal symmetry. And why, conversely, the failure of these two matrices to be mutually real implies time reversal symmetry breaking. The analogy with the Standard Model isn't perfect but you could, for example, think of diagonalising Y^d so that this gives mass eigenstates, and then measuring flavour eigenstates of Y^u. Indeed, this way of thinking works better in the lepton sector where there is a similar issue that results in neutrino mixing, as explained in section 7.)

6.4.2 The Jarlskog Invariant

We can ask: how much does the CKM matrix violate CP or, equivalently, time reversal? Clearly the answer is "not much" but it would be nice to find a way to quantify this. There is a way that is independent of the choice of basis. This is known as the *Jarlskog invariant*.

To see this, it's useful to work with Hermitian Yukawa couplings Y^d and Y^u; this is always possible as explained above. Then we know that there can be no CP breaking whenever $[Y^d, Y^u] = 0$. This suggests that we look at the Hermitian matrix

$$C = [Y^u, Y^d] \tag{6.68}$$

as a way to measure CP breaking. We can individually diagonalise each of these Yukawa matrices by

$$\begin{aligned} (V^d)^\dagger Y^d V^d &= D^d := \text{diag}(y^d, y^s, y^b) \\ \text{and} \quad (V^u)^\dagger Y^u V^u &= D^u := \text{diag}(y^u, y^c, y^t) \ . \end{aligned} \tag{6.69}$$

The commutator then becomes

$$C = V^u [D^u, V_{\text{CKM}}\, D^d\, V_{\text{CKM}}^\dagger] V^{u\,\dagger} \ . \tag{6.70}$$

We would like to construct something that is invariant under the field redefinitions $Y^d \to V^\dagger Y^d V$ and $Y^u \to V^\dagger Y^u V$. The obvious way to do this is to take traces of powers of C. Clearly $\text{Tr}\, C = 0$ while $\text{Tr}\, C^2$ is a measure of the failure of Y^u and Y^d to commute or, in other words, a measure of the size of V_{CKM}. However, for a measure of CP violation, the relevant quantity is

$$\text{Tr}\, C^3 = 3 \det C \ . \tag{6.71}$$

It's straightforward to see why this is the appropriate measure of CP violation. From (6.70), the matrix C shares its eigenvalues with the matrix $[D^u, V_{\text{CKM}}\, D^d\, V_{\text{CKM}}^\dagger]$. But if $V_{\rm CKM}$ is real then this is an anti-symmetric matrix and so are pure imaginary and come in conjugate $\pm$ pairs. That means in particular that, for $N = 3$ generations, the matrix C must have a zero eigenvalue whenever $V_{\rm CKM}$ is real and hence $\det C = 0$.

We can see this through an explicit calculation: we have

$$\det C = -2iF^u F^d J \tag{6.72}$$

where

$$\begin{aligned} F^u &= (y^t - y^c)(y^t - y^u)(y^c - y^u) \\ \text{and} \quad F^d &= (y^b - y^s)(y^b - y^d)(y^s - y^d) \ . \end{aligned} \tag{6.73}$$

We see that these factors vanish if any of the quark masses of the same type are equal. That's because, in this case the CKM matrix degenerates to become analogous to the situation with just two flavours, but we know that there can be no CP violation in that case. For the situation where all quark masses differ, the relevant measure of CP violation lies in the remaining factor J which is given by

$$J = {\rm Im}\,(V_{ud}V^\star_{ub}V_{tb}V^\star_{td}) \ . \tag{6.74}$$

This is the *Jarlskog invariant.* Its measured value is

$$J = s_{12}\, s_{23}\, s_{13}\, c_{12}\, c_{23}\, c_{13}^2\, \sin\delta \approx 3 \times 10^{-5} \ . \tag{6.75}$$

The Jarlskog invariant depends on each of the mixing angles θ_{ij}. If any of them vanishes (or, indeed, if any of them equals $\pi/2$) then the situation effectively reduces to that of just two flavours where, as we have already seen, there is no CP violation. Conversely, you can show that the theoretical maximum value of the Jarlskog invariant is $J_{\rm max} = 1/6\sqrt{6} \approx 0.07$. The measured value of the Jarlskog invariant $J/J_{\rm max} \approx 4 \times 10^{-4}$ is telling us that CP violation in the quark sector of the Standard Model is *really* small. As we've mentioned before, this isn't because the complex phase δ is small: it's not. It's all those other angles that kill us. We can see this in the Wolfenstein parameterisation, which gives

$$J \approx \lambda^6 A^2 \eta \ . \tag{6.76}$$

CP violation is small because it's proportional to λ^6.

The Jarlskog invariant has a nice interpretation in terms of the unitarity triangle. The area of the triangle (6.38) (computed before normalising one of the sides to have length 1) is of order $\sim \lambda^6$. One can show that it is given by the Jarlskog invariant

$$\text{Area} = \frac{J}{2} \ . \tag{6.77}$$

In fact, this result is stronger. If one considers the area of the triangle formed by the (extremely squashed) triangle defined by the complex numbers in (6.37), that too obeys (6.77). Indeed, the areas of all such triangles are equal and given by $J/2$.

6.4.3 The Strong CP Problem Revisited

In Section 3.4, we described the theta term of QCD,

$$S_\theta = \frac{\theta g_s^2}{16\pi^2} \int d^4x \ \text{Tr}\, G_{\mu\nu}{}^\star G^{\mu\nu} \ . \tag{6.78}$$

This would provide a contribution to CP violation directly within the strong force except that, as far as we can tell, the theta angle takes the value $\theta = 0$. (Or, more precisely, $\theta < 10^{-10}$.) Understanding why $\theta = 0$ is known as the strong CP problem.

It's worth revisiting this now that we understand how CP is violated in the weak sector. In particular, this new perspective gives the strong CP problem extra bite.

The issue comes when we choose to remove various phases of the CKM matrix by shifting the phases of the up and down quarks in (6.30). As we saw in Section 4, the $U(1)$ symmetries in (6.30) have a mixed anomaly with the $SU(3)$ gauge group. This means that the phase rotations (6.30) are not entirely innocuous because they shift the QCD theta angle as described in Section 4.2.1.

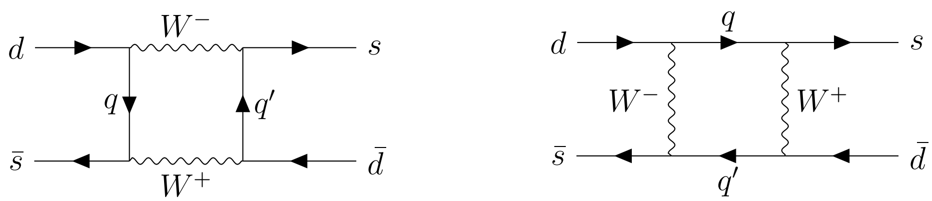
This suggests that the strong CP problem is tied up with the question of flavour and the CKM matrix. The fuller statement is that $\theta \approx 0$ when we remove all but one of the phases from the CKM matrix.

6.4.4 Neutral Kaons

How does CP violation manifest itself in our world? Although the imaginary part of the CKM matrix is largest in the V_{ub} and V_{td} components, the place where CP violation shows up most clearly is among kaons, for the simple reason that it's easy to produce a gazillion kaons and study them with precision.

Recall from Section 3 that the neutral kaon K^0 contains the quarks $d\bar{s}$. Its anti-particle $\bar{K}^0$ contains $s\bar{d}$. These mesons have mass $m_K \approx 498$ MeV.

The mesons K^0 and $\bar{K}^0$ are degenerate eigenstates of the strong interactions. (For example, they have well defined strangeness, which is a symmetry of QCD, but not of the full Standard Model.) However, the weak interactions can act to mix these two degenerate eigenstates. This happens through so-called *box diagrams* of the form



where the q and q' quarks in the diagrams can be either u, c or t. Each of these vertices comes with the corresponding CKM matrix element V_{dq} or $V^\star_{sq}$ and, as we've seen, some of these have imaginary parts, reflecting the fact that CP is broken. As we now explain, this has an interesting consequence for these kaons.

As usual in degenerate perturbation theory in quantum mechanics, we should figure out the new linear combinations of states that are energy eigenstates which, in the context of quantum field theory, is the same as a mass eigenstate.

To start, let's assume that CP is a good symmetry of the weak interactions. We will deduce the consequences of this and then see that these consequences are almost, but not quite, respected by nature, reflecting the fact that CP is almost, but not quite, a good symmetry.

If CP is a good symmetry of the weak force, then the mass eigenstates should be eigenstates of CP. But neither K^0 nor $\bar{K}^0$ are eigenstates of CP. To see this, first note that the kaon is a pseudoscalar meson (recall that it was a Goldstone boson from chiral symmetry breaking) and so, under parity, we have

$$P : |K^0\rangle \mapsto -|K_0\rangle \quad \text{and} \quad P : |\bar{K}^0\rangle \mapsto -|\bar{K}^0\rangle \ . \tag{6.79}$$

Meanwhile, under charge conjugation we have $C : d\bar{s} \mapsto \bar{d}s$ and so

$$C : |K^0\rangle \mapsto |\bar{K}^0\rangle \quad \text{and} \quad C : |\bar{K}^0\rangle \mapsto |K^0\rangle \ . \tag{6.80}$$

The upshot is that we can construct eigenstates under CP by taking

$$|K_1\rangle = \frac{1}{\sqrt{2}}(|K^0\rangle - |\bar{K}_0\rangle) \quad \text{and} \quad |K_2\rangle = \frac{1}{\sqrt{2}}(|K^0\rangle + |\bar{K}^0\rangle) \tag{6.81}$$

with

$$CP : |K_1\rangle \mapsto +|K_1\rangle \quad \text{and} \quad |K_2\rangle \mapsto -|K_2\rangle \ . \tag{6.82}$$

So we have two eigenstates of CP, $|K_1\rangle$ and $|K_2\rangle$, and if CP were a good symmetry then these would also be mass eigenstates. Let's now figure out what this means for the decay of kaons.

Kaons decay primarily to pions. The pions have mass $m_\pi \approx 140$ MeV which means that, in principle, a kaon could decay to either two pions or to three pions (because $140 \times 3 < 498$). Which of these happens is dictated by their CP quantum numbers.

Claim Two pion states have $CP = +1$.

Proof: There are actually two possible two pion decays: $\pi^0\pi^0$ and $\pi^+\pi^-$. We deal with each in turn.

The intrinsic parity of all pions is $P = -1$. (This was described in Section 3 and, as for the kaons, follows because they are Goldstone modes for chiral symmetry.) So the parity of a pair of pions is $P = (-1)^2 \times (-1)^L$ where L is the orbital angular momentum. But because the pions arise from the decay of a spin 0 particle, we must have $L = 0$ and hence $P = +1$.

That leaves us with charge conjugation. The neutral pion has quark content $\pi^0 = \frac{1}{\sqrt{2}}(u\bar{u} - d\bar{d})$ and so has $C = +1$. Meanwhile, the charged pions are exchanged under C. This means, in particular, that their positions are swapped and so charge conjugation acts in the same way as parity, meaning $C(\pi^+\pi^-) = P(\pi^+\pi^-) = (-1)^L$. But, as we've seen, $L = 0$ and so $\pi^+\pi^-$ also has $C = +1$.

Putting this together, we learn that the pair of pions has $CP = +1$. $\square$

Claim: The three pion states nearly always have $CP = -1$.

Proof: Again, we have two cases to consider: $\pi^0\pi^0\pi^0$ and $\pi^+\pi^-\pi^0$.

Each of these states has intrinsic parity $(-1)^3 = -1$, leaving us with the contribution from orbital angular momentum to worry about. Let's start with the $\pi^0\pi^0\pi^0$ state. We can think of the first two pions as having mutual angular momentum L_1 and the third as orbiting this pair with angular momentum L_2. The contribution to the parity of the state is then $(-1)^{L_1}(-1)^{L_2}$. We add angular momentum in the usual quantum mechanical way, $L_1 \oplus L_2 = |L_1 - L_2| + \ldots + |L_1 + L_2|$. But for this to include the required angular momentum $L = 0$ state, we must have $L_1 = L_2$ and so $(-1)^{L_1}(-1)^{L_2} = +1$. We learn that $\pi^0\pi^0\pi^0$ has parity $(-1)^3(-1)^{L_1}(-1)^{L_2} = -1$. It also has $C = +1$, and so $CP = -1$.

Things are a little more complicated for $\pi^+\pi^-\pi^0$. We again have total parity

$$P = (-1)^3(-1)^{L_1}(-1)^{L_2} = -1 \ . \tag{6.83}$$

The charge conjugation of π^0 is again $C = +1$, but the charge conjugation of the $\pi^+\pi^-$ pair is now $C(\pi^+\pi^-) = P(\pi^+\pi^-) = (-1)^{L_1}$ and this time there is no reason that L_1 should be even. This is why we've got the weasel words "nearly always" in the claim above. If the three pion state $\pi^+\pi^-\pi^0$ has $L_1 = 0$ then it does indeed have $CP = -1$ as claimed. But for $L_1 = +1$, the CP differs. Happily, this isn't an issue in practice because it costs extra kinetic energy for the pions to decay in the $L_1 = 1$ state but, with only $m_K - 3m_\pi \approx 80$ MeV to play with, these decay products with $L_1 \neq 0$ are strongly suppressed. □

The upshot of this argument is that, if CP is conserved, then the state $|K_1\rangle$ will decay to two pions, and the state $|K_2\rangle$ will decay to three pions. But there's a vast difference in the energy available for these decays. We have

$$m_K - 2m_\pi \approx 220 \text{ MeV} \quad \text{and} \quad m_K - 3m\pi \approx 80 \text{ MeV} \ . \tag{6.84}$$

This means that there's much more phase space available for the first decay than for the second and, correspondingly, we expect that the first decay will happen much faster than the second. Indeed, this is what is observed: the neutral kaons with mass $m_K \approx 498$ MeV have two different lifetimes, τ_{short} and τ_{long}, given by

$$\tau_{\text{short}} \approx 0.9 \times 10^{-10} \text{ s} \quad \text{and} \quad \tau_{\text{long}} \approx 0.5 \times 10^{-7} \text{ s} \ . \tag{6.85}$$

Putting all this together, we have the following conclusion: *if* CP is preserved, then we expect to identify the short-lived kaons with the $CP = +1$ eigenstates,

$$|K_{\text{short}}\rangle = |K_1\rangle = \frac{1}{\sqrt{2}}(|K^0\rangle - |\bar{K}^0\rangle) \ . \tag{6.86}$$

These will decay to two pions $K_S \to \pi\pi$ in time τ_{short}. Meanwhile, the long-lived kaons should correspond to the $CP = -1$ eigenstates,

$$|K_{\text{long}}\rangle = |K_2\rangle = \frac{1}{\sqrt{2}}(|K^0\rangle + |\bar{K}^0\rangle) \ . \tag{6.87}$$

These will decay to three pions $K_{\text{long}} \to \pi\pi\pi$ in a time τ_{long}.

So is this what's seen? Well, almost but not quite.

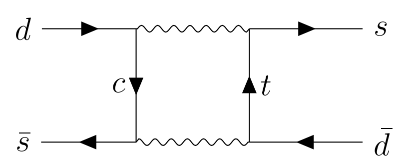
We can produce kaons through collisions $\pi^- + p \to \Lambda + K^0$. These kaons are a linear combination of CP even and odd eigenstates, $|K^0\rangle = \frac{1}{\sqrt{2}}(|K_1\rangle + |K_2\rangle)$. If we produce a beam of such kaons, then we should see them initially decay to two pions, and later decay to three pions. Indeed, that's what happens. Mostly.

Suppose that we wait for a time $\tau_{\rm short} \ll t \ll \tau_{\rm long}$, at which point we can be sure that the beam contains only $|K_{\rm long}\rangle$. We then look closely at the decay products. This is what Cronin and Fitch did in 1964. They observed 22700 kaon decays, of which 22655 decayed to three pions. But not all. There were 45 long-lived kaons that decayed to two pions. This tiny effect was the first evidence for CP violation. It arises because the long-lived energy eigenstates are *not* CP eigenstates. Instead, we have

$$\begin{aligned} |K_{\rm short}\rangle &= \frac{1}{\sqrt{1+|\epsilon|^2}}\big(|K_1\rangle + \epsilon|K_2\rangle\big) \\ |K_{\rm long}\rangle &= \frac{1}{\sqrt{1+|\epsilon|^2}}\big(|K_2\rangle + \epsilon|K_1\rangle\big) \ . \end{aligned} \tag{6.88}$$

Experimentally, $|\epsilon| \approx 2\times 10^{-3}$. This is the signature of CP violation in the neutral kaon system.

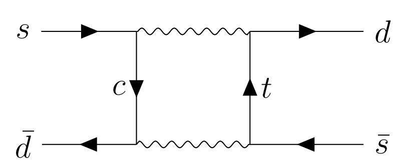
We can understand this from the box diagrams that we drew previously. We should sum over all different quarks running in the loop but, for simplicity, we will focus on the following diagram that mixes $K^0 \to \bar{K}^0$,



This diagram is proportional to the product of the CKM matrix elements,

$$\mathcal{M}(K \to \bar{K}) \sim V_{cd}V^\star_{cs}V_{td}V^\star_{ts} \ . \tag{6.89}$$

Meanwhile, the diagram that mixes $\bar{K}^0 \to K^0$ is



This diagram is proportional to

$$\mathcal{M}(\bar{K} \to K) \sim V^\star_{cd}V_{cs}V^\star_{td}V_{ts} = \mathcal{M}^\star(K \to \bar{K}) \ . \tag{6.90}$$

CP violation is reflected in the fact that the CKM matrix elements are not real, and hence $\mathcal{M}(K \to \bar{K}) \neq \mathcal{M}(\bar{K} \to K)$. The difference in the amplitude is

$$\mathcal{M}(K \to \bar{K}) - \mathcal{M}(\bar{K} \to K) \sim \text{Im}\big(V_{cd}V_{cs}^{\star}V_{td}V_{ts}^{\star}\big) \ . \tag{6.91}$$

The value of ϵ in (6.88) is set by this imaginary part, together with further contributions from other quarks running in the loop.

6.4.5 Wherefore CP Violation?

The CPT theorem tells us that CP violation is tantamount to a violation of time reversal. And that sounds interesting!

It's worth comparing the implications of parity violation and time reversal violation. At first glance, they seem very similar: one is a flip of spatial coordinates, $\mathbf{x} \to -\mathbf{x}$, the other a flip of time $t \to -t$. Yet, despite their similarities, the mathematical consequences of these two broken symmetries could not be more different.

The breaking of parity is sewn into the heart of the Standard Model which is a chiral gauge theory. As we've seen, the requirements of anomaly cancellation then put stringent constraints on the allowed interactions which pretty much fixes the gauge sector of the Standard Model.

This stands in sharp contrast to the theoretical consequences of time reversal violation, which shows up only as some complex phase in the CKM matrix. There are seemingly no deep mathematical consequences for theories that violate time reversal, no consistency requirements that we have to deal with. You just make a parameter complex and you're done. It's striking how little impact this has, not just on our daily lives, but on our deeper understanding of physics. It makes you wonder if there's something that we're missing!

There is, however, thought to be one very important implication of CP violation, albeit one that we don't fully understand. This follows from the fortunate observation that our universe contains lots of matter, but very little anti-matter. It is thought that this imbalance occurred naturally in the early universe, but for this to happen there have to be processes where matter and anti-matter behave differently. This, it turns out, requires CP violation.

It's not clear if the formation of matter over anti-matter can happen solely using the Standard Model (perhaps including some further CP violation that occurs in the lepton sector) or if it requires some new physics that lies beyond the Standard Model. This process, whatever causes it, goes by the name of *baryogenesis*.

7 Neutrinos

No one would accuse a neutrino of being gregarious. They interact less than a first year undergraduate mathematics student forced to sit next to their theoretical physics professor at a matriculation dinner (to give a weirdly specific yet shudderingly memorable analogy).

For example, in the time it takes you to read this sentence, around 100 trillion neutrinos will have passed through your body. Most of them came from the Sun, but a significant minority have a cosmic origin, and have been streaming through the universe, uninterrupted since the first few seconds after the Big Bang. Moreover, in contrast to photons, the number of neutrinos hitting you doesn't change appreciably as day turns into night. The neutrinos from the Sun will happily pass right through the Earth and out the other side. This is vividly demonstrated in the picture of the Sun at night shown in Figure 19.

There are two reasons why neutrinos are so intangible. The first is that they are the only particle to interact solely through the weak force. And, as we've seen, the weak force is weak. The second reason is that their mass is much much smaller than any other fermion which means that on the rare occasion that they do interact, they don't deliver much of a punch. The purpose of this section is to describe some properties of neutrinos in more detail.

7.1 Neutrino Masses

There is much that we don't know about neutrino masses. But we do know that the masses are not zero.

At the moment, we have no direct measurement of the mass of each neutrino. But we do have some precious information. First, we know that one neutrino must have a mass greater than

$$m_\nu \gtrsim 0.05 \text{ eV} . \tag{7.1}$$

Second, constraints from cosmology give us an upper bound on the sum of all neutrino masses. This comes from the imprint that neutrinos in the early universe leave on the cosmic microwave background radiation and on subsequent structure formation of galaxies (in particular, baryon acoustic oscillations – you can read more about this in the lecture notes on Cosmology.) This bound is

$$\sum_\nu m_\nu \lesssim 0.25 \text{ eV} . \tag{7.2}$$

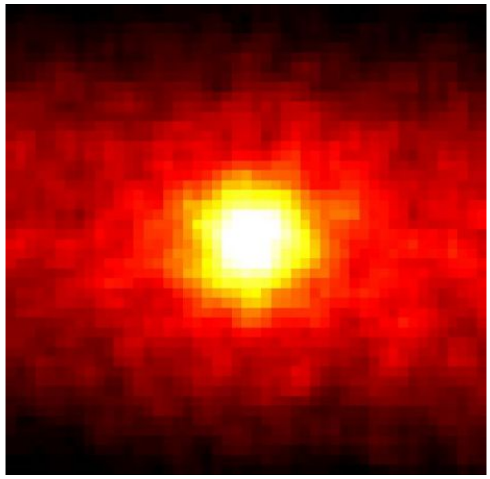


Figure 19. The Sun at night. This is a picture, taken by Super-Kamiokande, shows the neutrino flux coming from the Sun. The picture was taken at night, with the neutrinos passing through the Earth before hitting the detector.

In addition, we have information about the mass *differences* between neutrinos. We denote the mass of the neutrinos as m_1, m_2 and m_3. Much like for quarks, the mass eigenstates do not correspond to the flavour eigenstates ν_e, ν_μ and ν_τ and we will explain the relation more in the next section. We know that the mass splitting between two of the states is comparable to the overall mass of neutrinos,

$$|m_3^2 - m_2^2| = 2.5 \times 10^{-3} \text{ eV}^2 \ . \tag{7.3}$$

We've taken the magnitude of the difference on the left-hand side to hide the fact that we don't actually know which of m_3 and m_2 is heavier: we will describe this ambiguity further below. Then there is a much smaller mass splitting between of order

$$m_2^2 - m_1^2 \approx 7.4 \times 10^{-5} \text{ eV}^2 \ . \tag{7.4}$$

There are still a number of possibilities consistent with these bounds. It may even be, for example, that one neutrino is massless while others have mass ~ 0.1 eV or so. Still, our ignorance notwithstanding, a rough summary of the masses of all fermions is shown in Figure 20.

In the rest of this section, we will describe the basics of neutrino masses. We will learn how they can get a mass in the Standard Model and its extensions, and how we are able to determine the structure of masses described above.

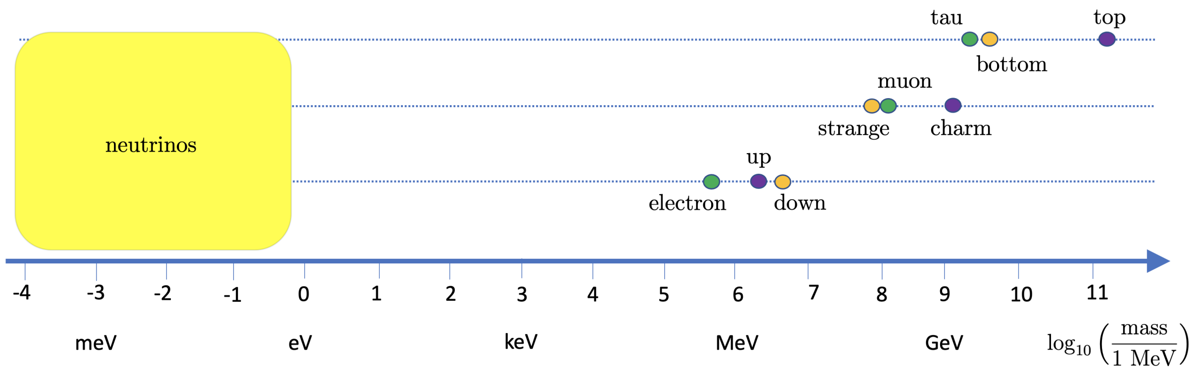


Figure 20. Fermion masses, arranged by generation. The charged leptons are green, the $-1/3$ quarks are orange, and the charge $+2/3$ quarks are purple. The neutrinos are way off to the left.

7.1.1 Dirac vs Majorana Masses

Even with our limited knowledge, it's clear that neutrinos aren't like the other particles. There are six orders of magnitude separating the mass of the top quark from the mass of the electron. Then there is a gap of another six orders of magnitude before we get to the neutrinos. The first question we should ask is: why?

We don't have a definitive answer to this question. But we do have a plausible answer. In what follows, I will sketch what appear to be the most reasonable ways in which neutrinos can get a mass. They are not the only ways: if you're willing to add new fields to the Standard Model, and then try to hide them from experiments, then you can cook up other possibilities. Ultimately, experiment must be our guide to figure out which is right.

The most obvious way to give neutrinos a mass is to add a right-handed neutrino ν_R to the Standard Model. Indeed, we already included this in Section 5 when describing the fields of the Standard Model, although we also raised a question mark about its existence. *If* we include a right-handed neutrino that is uncharged under the Standard Model gauge group, then it can participate in a Yukawa coupling. Restricting to a single generation for now, the lepton Yukawas are then (5.74),

$$\mathcal{L}_{\text{Yuk}} = -y^e\,\bar{L}_L H e_R - y^\nu\,\bar{L}_L \tilde{H} \nu_R + \text{h.c.}\ . \tag{7.5}$$

When the Higgs condenses, the neutrino gets a mass just like all other fermions, given by

$$m = \frac{y^\nu}{\sqrt{2}} v \ . \tag{7.6}$$

We refer to this as a *Dirac mass*.

There's nothing wrong with this explanation for neutrino masses. But it does raise a question of why the dimensionless Yukawa coupling is $y^\nu \sim 10^{-12}$. Of course, as we've repeatedly seen, we don't understand the values of any of the Yukawa couplings so perhaps this is just one more mystery to add to the list. Nonetheless, it's such a wildly small number that it feels like it's crying out for some explanation. And the good news is that there is a very natural explanation at hand.

Moreover, this explanation doesn't require us to do anything than follow our original philosophy when constructing the Standard Model. That is, given all the fields at our disposal, we should write down all possible relevant and marginal terms consistent with Lorentz invariance and gauge symmetry. And the addition of the right-handed neutrino allows for something new. This is the term

$$\mathcal{L}_{\text{Maj}} = \frac{1}{2} M \nu_R \nu_R + \text{h.c.} \ . \tag{7.7}$$

Here $M \in \mathbb{C}$. This is called a *Majorana mass*.

Suppose that we have both the Dirac mass m, as in (7.6), and the Majorana mass M, as in (7.7). What is the physical mass of the neutrinos? To answer this, we write the combined mass term as

$$\mathcal{L}_{\text{mass}} = \frac{1}{2} (\bar{\nu}_L, \nu_R) \begin{pmatrix} 0 & m \\ m & M \end{pmatrix} \begin{pmatrix} \bar{\nu}_L \\ \nu_R \end{pmatrix} + \text{h.c.} \ . \tag{7.8}$$

The physical masses are the eigenvalues of this matrix. We have

$$\text{mass} = \frac{1}{2} \left| M \pm \sqrt{M^2 + 4m^2} \right| \ . \tag{7.9}$$

What does this buy us? We know that the neutrinos have a mass in the eV range. One possibility is that both m and M are in this ballpark. But there's an alternative option, which is that the Majorana mass M is very large. If we take $M \gg m$, then the two masses above become

$$\text{mass} \approx M \quad \text{and} \quad \text{mass} \approx \frac{m^2}{M} \ . \tag{7.10}$$

The particle with mass $\approx M$ is mostly the right-handed neutrino, while the particle with mass $\approx m^2/M$ is approximately the left-handed neutrino. And, crucially, it's quite possible for the latter of these to be light, even if the Yukawa couplings are the same order of magnitude as those for electrons.

For example, if $y^\nu \approx 1$ (like the extraordinarily heavy top quark) then a Majorana mass of order 10^{13} GeV or so will get us in the ballpark of the observed masses. This is getting close to the realm of grand unified theories. Obviously, for smaller Yukawa couplings, the corresponding Majorana mass should be smaller. This suggests, somewhat counterintuitively, that the smallness of the neutrino mass might be because the right-handed neutrino gets a very large mass. This is known as the *seesaw mechanism*.

7.1.2 The Dimension 5 Operator

There's something a little unsettling about the seesaw mechanism. We introduced a right-handed neutrino to give both left- and right-handed particles a mass. But then we saw that the physical mass of one of these states M was extremely large, way beyond current experiments. Which suggests that it should be possible to describe the resulting physics without invoking it in the first place!

And, indeed there is. But it does require us to go beyond our original philosophy when constructing the Standard Model. We originally set ourselves the task of writing down all relevant and marginal terms consistent with Lorentz and gauge symmetries. We can incorporate neutrino masses without a right-handed neutrino if we also allow ourselves to include irrelevant operators.

As usual, operators in quantum field theory are classified by their dimension. Those with dimension $\Delta < 4$ are relevant, and those with dimension $\Delta = 4$ are (classically) marginal. There are an infinite number of irrelevant operators, but their importance can still be judged by how irrelevant they are. And, among them, there is a unique operator with dimension $\Delta = 5$. This is

$$\mathcal{L}_5 = \frac{\lambda}{M}(\bar{L}_L\tilde{H})(\bar{L}_L\tilde{H}) + \text{h.c.}\ . \tag{7.11}$$

This is sometimes called the *Weinberg operator* although Weinberg has so many things named after him in the Standard Model that I'm not sure it's helpful terminology. It has dimension 5 because it contains two fermions (each of dimension 3/2) and two scalars (each of dimension 1). Here λ is a dimensionless coupling and M is a mass scale. If we integrate out the massive right-handed neutrino, then we generate the coupling (7.11) with M the Majorana mass and $\lambda = (y^\nu)^2$. However, the operator (7.11) may be generated by something else that isn't associated to a right-handed neutrino.

We see that (7.11) captures the spirit of the seesaw mechanism: when the Higgs gets a vev v, the left-handed neutrino ν_L gets a Majorana mass $\sim \lambda v^2/M$. This retains the irony in which detecting a very small Majorana mass points towards physics at a very high energy scale.

7.1.3 Neutrinoless Double Beta Decay

Above, we've seen that there are two ways that a neutrino can get a mass: either a bog standard Dirac mass (7.6), or a Majorana mass (7.7) which, if large, is captured in the dimension 5 operator (7.11).

There is one important difference between these: the Majorana mass violates lepton number at tree level. This means that it might be possible to detect the neutrino Majorana mass by observing a process which explicitly violates lepton number.

You can't have a process that changes lepton number by just one because (in the absence of any other fermion getting involved) that would also violate $(-1)^F$ which is part of the Lorentz group. So, in searching for signals of lepton number violation, we are looking for processes that change L by two. The most clear cut process of this type is something called *neutrinoless double beta decay*, sometime referred to rather elliptically as $0\nu\beta\beta$.

Recall that beta decay is the process $n \to p + e^- + \bar{\nu}^e$. This increases the atomic number of an element by one. Double beta decay is what it sounds like: we have $2n \to 2p + 2e^- + 2\bar{\nu}^e$, increasing the atomic number of an element by two.

Double beta decay occurs, albeit rarely. It's most easy to observe in elements for which the normal single beta decay is forbidden. For example, ^{76}Ge (with atomic number 32) can't decay through single beta decay to ^{76}As (with atomic number 33) because the germanium nucleus is lighter than the arsenic nucleus. However, it is possible for germanium to decay to ^{76}Se (with atomic number 34) which happens to have a lighter nucleus. The decay process is

$$^{76}\text{Ge} \to {}^{76}\text{Se} + 2e^- + 2\bar{\nu}^e \ . \tag{7.12}$$

This decay has been observed with lifetime of around 10^{21} years. (That was a *very* long experiment.)

Ordinary double beta decay preserves lepton number. But if the neutrino has a Majorana mass, so lepton number is violated, then there is another option: this is neutrinoless double beta decay

$$^{76}\text{Ge} \to {}^{76}\text{Se} + 2e^- \ . \tag{7.13}$$

Despite many ongoing searches, no such decay process has been observed, either in germanium or the dozen or so other elements that exhibit ordinary double beta decay. Current bounds put the effective half-life of elements due to double beta decay at $> 10^{25}$ years or so. These put bounds on the mass of a neutrino coming from a dimension 5 operator of $m_\nu \lesssim 0.3$ eV.

7.1.4 The PMNS Matrix

The fact that we have three generations of fermions means that, as for quarks, there is a misalignment between the mass and flavour eigenstates of leptons. As we saw in Section 5, we label the three generations of leptons as (5.20),

$$L_L^i = \begin{pmatrix} \nu_L^i \\ e_L^i \end{pmatrix} = \left\{ \begin{pmatrix} \nu_L^e \\ e_L \end{pmatrix} , \begin{pmatrix} \nu_L^\mu \\ \mu_L \end{pmatrix} , \begin{pmatrix} \nu_L^\tau \\ \tau_L \end{pmatrix} \right\} . \tag{7.14}$$

These left-handed leptons appear in the charged currents that couple to the W bosons (5.88). If we omit the quarks terms, and focus only on the leptons, we have

$$J_\mu^+ = \bar{\nu}_L^i \bar{\sigma}_\mu e_L^i \quad \text{and} \quad J_\mu^- = \bar{e}_L^i \bar{\sigma}_\mu \nu_L^i . \tag{7.15}$$

As with the quarks, the leptons that appear here are *before* we diagonalise the mass matrices. In other words, the leptons that appear here are in the *flavour* basis.

If, however, we choose to work in the mass basis, which means that the mass terms are diagonal then, as with the quarks, we get a 3×3 unitary mixing matrix U appearing in the charged current which becomes

$$J_\mu^+ = \bar{\nu}_L^i \bar{\sigma}_\mu U_{ij}^\dagger e_L^j \quad \text{and} \quad J_\mu^- = \bar{e}_L^i \bar{\sigma}_\mu U_{ij} \nu_L^j . \tag{7.16}$$

This matrix U is known as the PMNS matrix, named after Pontecorvo, Maki, Nakagawa, and Sakata or simply the neutrino mixing matrix.

We learn that there are two natural bases that we can use: the mass basis in which the masses are diagonal, or the flavour basis in which the coupling to W bosons are diagonal. And these differ from each other. Correspondingly, there are two different linear combinations of fields.

What we usually refer to as the "electron neutrino", "muon neutrino", and "tau neutrino" are fields in the flavour basis. For example, beta decay happens by $n \to p + e^- + \bar{\nu}^e$ and that neutrino $\bar{\nu}^e$ is the one that couples to the W boson and electron, so it is $\bar{\nu}^e$ in the flavour eigenbasis. Which means that the neutrino that is emitted is *not* in a mass eigenstate!

It's useful to introduce some new notation to highlight what's going on. We will refer to the left-handed neutrinos in the flavour basis as ν_e and ν_μ and ν_τ. And we will refer to the neutrinos in the mass basis simply as ν_1 and ν_2 and ν_3. Each of these is a left-handed Weyl fermion, but we've suppressed the subscript L. The ν_i in (7.16) are in the mass basis and we see that these are related to the flavour basis by the PMNS matrix,

$$\begin{pmatrix} \nu_e \\ \nu_\mu \\ \nu_\tau \end{pmatrix} = \begin{pmatrix} U_{e1} & U_{e2} & U_{e3} \\ U_{\mu 1} & U_{\mu 2} & U_{\mu 3} \\ U_{\tau 1} & U_{\tau 2} & U_{\tau 3} \end{pmatrix} \begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \end{pmatrix} \ . \tag{7.17}$$

The PMNS matrix is to leptons what the CKM matrix is to quarks. Just as for the CKM matrix, we have no way to determine the values of U from first principles. Instead, we must measure these from experiment. The magnitude of each component is now known reasonably accurately: these are

$$\begin{pmatrix} |U_{e1}| & |U_{e2}| & |U_{e3}| \\ |U_{\mu 1}| & |U_{\mu 2}| & |U_{\mu 3}| \\ |U_{\tau 1}| & |U_{\tau 2}| & |U_{\tau 3}| \end{pmatrix} \approx \begin{pmatrix} 0.8 & 0.5 & 0.1 \\ 0.3 & 0.5 & 0.7 \\ 0.4 & 0.6 & 0.6 \end{pmatrix} \ . \tag{7.18}$$

Some values are known fairly well; others less well. There are, for example, error bars of ± 0.1 on $U_{\tau 2}$.

The first thing to note is that the PMNS matrix is strikingly different from the CKM matrix describing the mixing of quarks[10]. In the quark sector, the CKM matrix was close to being the unit matrix, with just small off-diagonal elements. This meant that there was close alignment between the masses and the weak force. But we see no such thing in the neutrino sector. The mixing is pretty much as big as it can be! The lepton sector is really nothing like the quark sector. We do not have an explanation for the structure of the PMNS matrix. Indeed, its form came as a surprise to theorists. Surely it is telling us something important. It's just we don't yet know what!

[10]Recall that $\begin{pmatrix} |V_{ud}| & |V_{us}| & |V_{ub}| \\ |V_{cd}| & |V_{cs}| & |V_{cb}| \\ |V_{td}| & |V_{ts}| & |V_{tb}| \end{pmatrix} \approx \begin{pmatrix} 0.97 & 0.22 & 0.004 \\ 0.22 & 0.97 & 0.04 \\ 0.009 & 0.04 & 0.999 \end{pmatrix}$. Note also that the indices of the CKM matrix and PMNS matrix are in the opposite order. For $V_{\rm CKM}$, the different rows are labelled by the up-type quarks, which is the first component of Q_L. For $U_{\rm PMNS}$, the rows are labelled by the charged lepton, which is the second component of L_L.

7.1.5 CP Violation in the Lepton Sector

As with the CKM matrix, CP violation is captured by the complex phases of the PMNS matrix. Here we must distinguish between neutrinos getting a purely Dirac mass and neutrinos getting a Majorana mass.

In the case where there are three right-handed neutrinos and each species of neutrino gets a Dirac mass, then the story is the same as for the CKM matrix: the neutrino mixing matrix has just a single phase.

But the counting is different if we have a Majorana mass. For this exercise, we will ignore the (unknown) mass of the right-handed neutrino and assume that the neutrino mass comes from the dimension 5 operator (7.11). With three generations, this takes the form

$$\mathcal{L}_5 = \frac{C_{ij}}{M}(\bar{L}_L^i \tilde{H})(\bar{L}_L^j \tilde{H}) \ . \tag{7.19}$$

Here C_{ij} is a complex *symmetric* 3×3 matrix, which means that it has 6 complex parameter or 12 real parameters. This means that in C_{ij} and the electron Yukawa y^e_{ij}, there are a total of $12 + 18 = 30$ real parameters. And we can eliminate some of these through $U(3)^2$ rotations acting on L^i_L and e^i_R. This leaves us with

$$30 - 2 \times 9 = 12 \tag{7.20}$$

physical parameters. That's two more than for the quark sector. Note that, in contrast to the quark sector, there's no overall $U(1)$ that leaves the parameters untouched: that's because of the Majorana mass.

As for quarks, we can also see how this decomposes into real mixing angles and complex phases. A $U(3)$ matrix has 3 real parameters and 6 complex phases, so the lepton sector with Majorana masses has

$$(6 + 9) - 2 \times 3 = 9 \text{ real parameters} \tag{7.21}$$

and

$$(6 + 9) - 2 \times 6 = 3 \text{ complex phases} \ . \tag{7.22}$$

We see that the total number of real parameters is the same as for the quarks: it decomposes into 6 masses for electrons and neutrinos, together with three angles which live inside the PMNS matrix. In contrast, with a Majorana mass there are two more complex phases lurking inside the PMNS matrix. The usual way to parameterise these is by embellishing the CKM matrix structure (6.31) with two additional phases,

$$U_{\rm PMNS} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & c_{23} & s_{23} \\ 0 & -s_{23} & c_{23} \end{pmatrix} \begin{pmatrix} c_{13} & 0 & s_{13}e^{-i\delta} \\ 0 & 1 & 0 \\ -s_{13}e^{i\delta} & 0 & c_{13} \end{pmatrix} \begin{pmatrix} c_{12} & s_{12} & 0 \\ -s_{12} & -s_{23} & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & e^{i\alpha_1} & 0 \\ 0 & 0 & e^{i\alpha_2} \end{pmatrix}$$

$$= \begin{pmatrix} c_{12}\,c_{13} & s_{12}\,c_{13} & s_{13}e^{-i\delta} \\ -s_{12}\,c_{23} - c_{12}\,s_{23}\,s_{13}e^{i\delta} & c_{12}\,c_{23} - s_{12}\,s_{23}\,s_{13}e^{i\delta} & s_{23}\,c_{13} \\ s_{12}\,s_{23} - c_{12}\,c_{23}\,s_{13}e^{i\delta} & -c_{12}\,s_{23} - s_{12}\,c_{23}\,s_{13}e^{i\delta} & c_{23}\,c_{13} \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & e^{i\alpha_1} & 0 \\ 0 & 0 & e^{i\alpha_2} \end{pmatrix} .$$

While the real angles θ_{ij} are measured with some precision, as shown in (7.18), the complex phases $e^{i\delta}$ and (if they exist) $e^{i\alpha_1}$ and $e^{i\alpha_2}$ remain unknown for neutrinos. This means that we don't currently know if CP violation is possible in the lepton sector of the Standard Model. We note, however, that because none of the mixing angles θ_{ij} are particularly small, there is the possibility that CP violation in the lepton sector is significantly larger than in the quark sector. Future experiments should decide this.

7.2 Neutrino Oscillations

So far we have described the different ways in which neutrinos can get a mass. But we haven't yet explained how we know that they have mass. After all, it's not like we can simply collect a bunch of neutrinos in a jar and weigh it. Instead, our information comes in a less direct manner.

We have met the key piece of physics already: the mass eigenstates of the neutrinos are misaligned with the flavour eigenstates. The two are related through the PMNS matrix (7.17).

Neutrinos are always created or observed in flavour eigenstates. For example, in beta decay we have

$$n \longrightarrow p + e^- + \bar{\nu}_e \tag{7.23}$$

and it's definitely an electron neutrino that is emitted. Relatedly, we can detect an electron neutrino through a neutrino capture process, $\nu^e + n \longrightarrow p + e^-$. For example, the earliest neutrino detection experiments used tanks filled with dry-cleaning fluid which was rich in chlorine and looked for electron neutrinos through the process

$$\nu_e + {}^{37}{\rm Cl} \longrightarrow {}^{37}{\rm Ar} + e^- \ . \tag{7.24}$$

Again, it's necessarily an electron neutrino that induces this process, not a neutrino of any other type.

However, as we have seen, the electron neutrino ν^e is *not* a mass eigenstate. In the language of quantum mechanics, this means that it's not an energy eigenstate. But we know from our first courses on quantum mechanics what happens when systems are placed in states that are not energy eigenstates: the state you sit in varies with time. And so it is with neutrinos: the flavour of neutrino oscillates over time.

Before we put some mathematical meat on these ideas, it's worth pointing out that neutrino mixing comes with a slightly different change of perspective compared to the entirely analogous quark mixing that we met in Section 6. When we talk about quarks, we usually think of mesons as energy eigenstates. The mixing then manifests itself as interactions allowing, say, a strange quark to decay to a up quark.

In contrast, in the world of leptons we can be confident that we have a particular flavour of neutrino to hand. The mixing then manifests itself as this flavour evolving, coherently, to a superposition of other flavours over time.

7.2.1 Oscillations with Two Generations

To see the basic physics, it's useful to restrict ourselves to the situation with just two flavours of neutrino. We'll take these to be the electron and muon neutrinos, related to mass eigenstates by the rotation matrix

$$\begin{pmatrix} \nu_e \\ \nu_\mu \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix} \ . \tag{7.25}$$

If the neutrinos have Majorana masses then there can be an additional complex phase in these relations. This will not affect neutrino oscillations and we won't consider it here.

We can think of the neutrinos as a 2-level system in quantum mechanics. Suppose that we start with an electron neutrino. Written in terms of energy eigenstates, this is

$$|\nu_e\rangle = \cos\theta|\nu_1\rangle + \sin\theta|\nu_2\rangle \ . \tag{7.26}$$

The neutrino ν_e is emitted with some energy E but, as we've seen, $|\nu_e\rangle$ isn't an energy eigenstate so we should view this as the average energy, $E = \cos^2\theta\, E_1 + \sin^2\theta\, E_2$, where

E_1 and E_2 are the energies of the states $|\nu_1\rangle$ and $|\nu_2\rangle$. Now, as we evolve in time, each of the energy eigenstates picks up a different phase,

$$|\nu_e(t)\rangle = e^{-iE_1 t}\cos\theta|\nu_1\rangle + e^{-iE_2 t}\sin\theta|\nu_2\rangle$$
$$= e^{-iE_1 t}\big(\cos\theta|\nu_1\rangle + e^{-i\Delta E\, t}\sin\theta|\nu_2\rangle\big) \tag{7.27}$$

where $\Delta E = E_2 - E_1$ is the energy difference between the states. Now we can convert back to the flavour eigenstates to get

$$|\nu_e(t)\rangle = e^{-iE_1 t}\Big(\big(\cos^2\theta + e^{-i\Delta E\, t}\sin^2\theta\big)|\nu_e\rangle - \cos\theta\,\sin\theta\big(1 - e^{-i\Delta E\, t}\big)|\nu_\mu\rangle\Big)\ . \tag{7.28}$$

This is a standard result in quantum mechanics, entirely analogous to, say, Rabi oscillations in atomic physics. We see that, as time evolves, we have a probability of the electron neutrino ν_e to convert to a muon neutrino ν_μ,

$$P(\nu_e \to \nu_\mu) = \sin^2(2\theta)\,\sin^2\left(\frac{\Delta E\, t}{2}\right)\ . \tag{7.29}$$

The fact that this probability depends on sine functions is telling us that the change of flavour is an oscillation, in the sense that it goes back and forth. At this point, we need an expression for the energy difference ΔE. For each of the mass eigenstates, we have the usual relativistic dispersion relation

$$E_i = \sqrt{\mathbf{p}_i^2 + m_i^2} \approx |\mathbf{p}_i| + \frac{m_i^2}{2|\mathbf{p}_i|} \tag{7.30}$$

where, in the second equality, we've used the fact that our neutrinos are ultra-relativistic with $|\mathbf{p}| \gg m$. We can think of the neutrinos as sitting in momentum eigenstates, so that $\mathbf{p}_1 = \mathbf{p}_2$. Further, we can replace the $\mathbf{p}$ in the denominator with the original energy E, giving

$$\Delta E = \frac{\Delta m^2}{2E} \tag{7.31}$$

with $\Delta m^2 = m_2^2 - m_1^2$. There's one final flourish: the neutrinos are travelling at very close to the speed of light and so, in time t, travel a distance $L = t$ (because, of course, $c = 1$). We can then write the probability for an electron neutrino to convert into a muon neutrino, depending on the distance it travels

$$P(\nu_e \to \nu_\mu) = \sin^2(2\theta)\,\sin^2\left(\frac{\Delta m^2}{4E}L\right)\ . \tag{7.32}$$

We can put some numbers in this to figure out what kind of length scales L we need to see neutrino oscillations. First, we should put factors of $\hbar$ and c back into the formula. On dimensional grounds, we should have

$$P(\nu_e \to \nu_\mu) = \sin^2(2\theta)\,\sin^2\left(\frac{\Delta m^2 c^4}{4E\hbar c}L\right)\ . \tag{7.33}$$

We have $\hbar = 6.5 \times 10^{-16}$ eV s. For mass differences Δmc^2 of order an eV (which, as we will see, is a little on the high side) and neutrino energies E measured in GeV (which, as we shall see, is also a little on the high side), the argument of the sine function is of order 1 for

$$L \sim 4\hbar c \times \frac{\text{GeV}}{(\text{eV})^2} \sim 1\ \text{km}\ . \tag{7.34}$$

That's a remarkably human length scale to emerge from fundamental physics! It sets the kind of scale over which neutrino experiments should take place. We will see examples below. Putting in the numbers, the probability is often written as

$$P(\nu_e \to \nu_\mu) \approx \sin^2(2\theta)\,\sin^2\left(1.27 \times \frac{\Delta m^2}{(\text{eV})^2}\frac{(\text{GeV})}{E}\frac{L}{(\text{km})}\right)\ . \tag{7.35}$$

This formula contains two fundamental parameters: the mixing angle θ and the difference in masses Δm^2. To see oscillations, both need to be non-zero. The formula also contains two parameters that can vary from one experiment to another: the energy E of the beam and the length travelled L. In principle, by varying E and L, and seeing how one kind of neutrino morphs into another, we can determine the mixing angle θ and mass difference Δm^2. As you can see from the formula above, to see oscillations it is best to tune $E/L \sim \Delta m^2$.

Oscillations with Three Flavours

Repeating this calculation with three species of neutrinos gives the probability for oscillation from one flavour species α to another β in terms of the PMNS matrix U,

$$P(\nu_\alpha \to \nu_\beta) = \left|U_{\alpha 1}U^\star_{\beta 1} + U_{\alpha 2}U^\star_{\beta 2}\,e^{-i\Delta m^2_{21}L/2E} + U_{\alpha 3}U^\star_{\beta 3}\,e^{-i\Delta m^2_{31}L/2E}\right|^2\ . \tag{7.36}$$

If we take a limit in which $\Delta m^2_{21}L \ll E$, then we have

$$P(\nu_\alpha \to \nu_\beta) = \left|U_{\alpha 1}U^\star_{\beta 1} + U_{\alpha 2}U^\star_{\beta 2} + U_{\alpha 3}U^\star_{\beta 3}\,e^{-i\Delta m^2_{31}L/2E}\right|^2\ . \tag{7.37}$$

But, because U is unitary, we have $U_{\alpha 1}U^\star_{\beta 1} + U_{\alpha 2}U^\star_{\beta 2} + U_{\alpha 3}U^\star_{\beta 3} = \delta_{\alpha\beta}$. For $\alpha \neq \beta$, we then have

$$P(\nu_\alpha \to \nu_\beta) = \left|U_{\alpha 3}U^\star_{\beta 3}\right|^2 \left|-1 + e^{i\Delta m^2_{31}L/2E}\right|^2\ . \tag{7.38}$$

This reproduces our two flavour result (7.35).

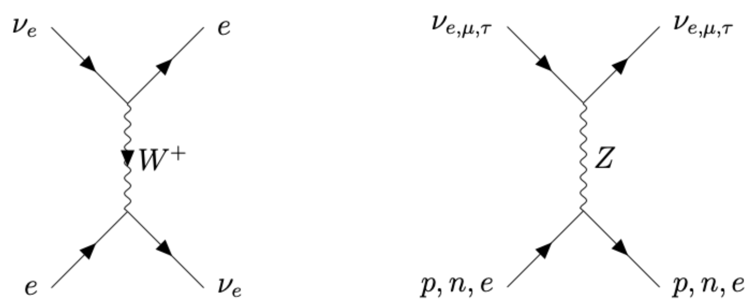


Figure 21. The scattering of electron neutrinos through a charged current, and any kind of neutrino through a neutral current.

7.2.2 Oscillations in Matter

There is a variation on the neutrino oscillation calculation that arises when neutrinos propagate through matter. This is both important and surprising.

The result is important because one source of neutrinos is the Sun, and the neutrinos that are created in the centre of the Sun have a way to travel before they emerge into empty space. And we would like to understand what happens to them on that journey. In addition, it is quite possible to detect neutrinos at night, after they have passed through the Earth and, again, we would like to understand if this last part of the journey has any noticeable effect.

The result is surprising because neutrinos are famously not impeded by things that sit in their way. Most happily pass straight through the Earth without being scattered. And yet, as we will see, the fact that they move in a density of matter does affect the oscillations. (There is also a second reason why the result is surprising which is to do with the orders of magnitude of energy involved and we will highlight this below.)

The effect that we care about arises from the elastic, forward scattering of neutrinos off a background of matter. This means that the neutrinos exchange neither energy nor momentum with the background matter. This process arises through the Feynman diagrams shown in Figure 21. All three types of neutrino can scatter off protons, neutrons and electrons through the exchange of a Z boson, while the electron neutrino can additionally scatter off electrons through the exchange of a W boson.

The neutral currents give the same contribution to all flavours of neutrinos while, for oscillations, we care about differences in neutrino energies. For this reason, we look only at the contribution from charged currents. We've already seen in Section 5 that, at low energies, this is captured by the 4-fermion current-current interaction (5.92) which, in the present context, we view as contribution to the Hamiltonian

$$\Delta H = 2\sqrt{2} G_F \, J_{+\,\mu} \, J_-^{\mu} \ . \tag{7.39}$$

Here, $G_F \approx 10^{-5}$ GeV^{-2} is the Fermi coupling. The currents $J_\mu^{\pm}$ were given in (5.88) and include the term

$$\begin{aligned} J_{+\,\mu} \, J_-^{\mu} &= (\bar{\nu}_L \bar{\sigma}_\mu e_L) \, (\bar{e}_L \bar{\sigma}^\mu \nu_L) + \ldots \\ &= (\bar{e}_L \bar{\sigma}_\mu e_L) \, (\bar{\nu}_L \bar{\sigma}^\mu \nu_L) + \ldots \end{aligned} \tag{7.40}$$

where, in the second line, we've done a Fierz shuffle to reorder the fermions. In the presence of matter, the $\mu = 0$ component of the vector $\bar{e}_L \bar{\sigma}^\mu e_L$ gets an expectation value

$$\langle \bar{e}_L \bar{\sigma}^\mu e_L \rangle = n \delta^{\mu 0} \tag{7.41}$$

where n is the background (number) density of electrons. This expectation value breaks Lorentz invariance, as a background density of matter must. It also breaks both CP and CPT as the background is made of normal matter, not anti-matter. (Recall that the CPT theorem is a statement about Lorentz invariant theories only.) The upshot is that we get a contribution to the Hamiltonian governing neutrinos that takes the form

$$\Delta H = V \, \bar{\nu}_L \bar{\sigma}^0 \nu_L \quad \text{where} \quad V = 2\sqrt{2} G_F n \ . \tag{7.42}$$

At this point, we see the next surprise. The extra term in the Hamiltonian H_c is quadratic in neutrinos and so, in that sense, looks like an additional contribution to the neutrino mass. The mass density of matter in the Sun is about $\rho \approx 1$ g cm^{-3} which gives $V \approx 10^{-12}$ eV. In the centre of the Earth, the density is an order of magnitude larger and, correspondingly, $V \approx 10^{-13}$ eV. Both of these are tiny compared to typical neutrino masses of 10^{-3} eV which naively suggests that this effect can't possibly be important for neutrino propagation.

But that intuition is wrong. And it's wrong because of the different index structure. That extra factor of $\bar{\sigma}^0$ in (7.42) makes all the difference: it is telling us that the background matter couples to neutrinos much like a background gauge field of the form $V^\mu = (V, 0, 0, 0)$. This means that the dispersion relation for neutrinos now takes the form

$$(p_\mu - V_\mu)(p^\mu - V^\mu) = m^2 \quad \Longrightarrow \quad (E - V)^2 = m^2 + \mathbf{p}^2 \ . \tag{7.43}$$

We're in a ultra-relativistic regime, with $E, p \gg m \gg V$, so we expand and drop the V^2 term to get the

$$E \approx p + \frac{m^2 + 2EV}{2p^2} + \dots \ . \tag{7.44}$$

We see that the relevant comparison is not m vs V but, instead, m^2 vs EV. And for energies in the MeV range, these can be comparable.

Our next task is to understand how this affects the oscillations. Recall that, in the vacuum, the neutrino Hamiltonian was diagonal in the mass basis. But now we've added an extra term that is diagonal in the flavour basis, contributing only to the electron neutrino. This means that we have some more matrix diagonalisation ahead of us.

To keep things simple, we'll stick to just two flavours of neutrino which we take to be ν_e and ν_μ. We'll again reduce things to a two-state quantum system. In the flavour basis, the vacuum Hamiltonian is given by

$$H = U \begin{pmatrix} E_1 & 0 \\ 0 & E_2 \end{pmatrix} U^\dagger \quad \text{with} \quad U = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \ . \tag{7.45}$$

We use the result (7.31) that gives the energy difference in terms of the mass difference, $E_2 - E_1 = \Delta m^2/2E$, to write

$$H = \frac{1}{2}(E_1 + E_2)\mathbb{1} + \frac{\Delta m^2}{4E} \begin{pmatrix} -\cos 2\theta & \sin 2\theta \\ \sin 2\theta & \cos 2\theta \end{pmatrix} \ . \tag{7.46}$$

The overall energy contribution $\frac{1}{2}(E_1 + E_2)\mathbb{1}$ is unimportant for our needs and we drop it in what follows. This is the vacuum Hamiltonian. Now we want to include the effects of matter which, as we have seen, give a new contribution

$$H + \Delta H = \frac{\Delta m^2}{4E} \begin{pmatrix} -\cos 2\theta & \sin 2\theta \\ \sin 2\theta & \cos 2\theta \end{pmatrix} + \begin{pmatrix} V & 0 \\ 0 & 0 \end{pmatrix} \ . \tag{7.47}$$

We need to extract the new eigenvalues and eigenvectors of this matrix. If we call these eigenvalues λ_1 and λ_2 then the effective mass splitting in the presence of matter is $\Delta m_m^2 = 2E(\lambda_2 - \lambda_1)$. A short calculation shows that

$$\Delta m_m^2 = \sqrt{(\Delta m^2 \cos 2\theta - 2EV)^2 + (\Delta m^2 \sin 2\theta)^2} \ . \tag{7.48}$$

Meanwhile, we also want to know the effective mixing angle θ_m. This comes from computing the eigenvectors of the new Hamiltonian which take the form $(\cos\theta_m, -\sin\theta_m)$ and $(\sin\theta_m, \cos\theta_m)$. The result is most simply expressed using a double angle formula as

$$\tan 2\theta_m = \frac{\sin 2\theta}{\cos 2\theta - 2EV/\Delta m^2} \ . \tag{7.49}$$

The probability for oscillation from one species to the other is then given by our previous expression (7.33) with Δm^2 and θ replaced by Δm_m^2 and θ_m. This probability is maximised when

$$\cos 2\theta = \frac{2EV}{\Delta m^2} \quad \Longrightarrow \quad \theta_m = \frac{\pi}{4} \ . \tag{7.50}$$

For anti-neutrinos, we replace V with $-V$ in the expressions above. This means that when mixing is maximal for neutrinos, with $\cos 2\theta = 2EV/\Delta m^2$, it is not maximal for anti-neutrinos.

Briefly, the MSW Effect

You might think that it's rather unlikely that we will hit the resonance condition (7.50) for maximal mixing. However, as neutrinos propagate outwards from the centre of the Sun, they experience a changing matter density. This means that we should think of the parameter V in our 2-state quantum system as being time-dependent. It may well be that, at some point on its journey, a given neutrino experiences a point where the effective mixing is maximal. In this way, large mixing can be generated even though the fundamental mixing angles may be small. This is known as the *MSW effect*.

We saw in the lectures on Topics in Quantum Mechanics that there are two limits in which it is straightforward to analyse systems with time-dependent parameters. When the time dependent is slow (in a suitable sense), we can use the adiabatic approximation. This is appropriate in the interior of the Sun. When the time dependence is fast, we can use the sudden approximation. This is appropriate when the neutrinos exit the Sun or when they enter the Earth. Both of these effects are important when understanding the observed oscillations in solar neutrinos.

7.2.3 Neutrino Detection Experiments

Nature provides two different sources of neutrinos that allow us to see oscillations. In what follows, we provide some very brief sketches of the experiments that revealed oscillations in each of these sources. In recent years, these results have been confirmed by looking at terrestrial neutrinos, created in reactors and accelerators.

Solar Neutrinos

Most neutrinos in the Sun are created in a reaction that turns hydrogen into helium,

$$4p \to {}^4\text{He} + 2e^+ + 2\nu_e + 2\gamma \ . \tag{7.51}$$

This produces neutrinos with energy $E \lesssim 400$ keV. There are also further reactions, notably those involving ^{7}Be and ^{8}Be that produce significantly fewer neutrinos, but at energy up to 10 MeV. It is now thought that we have a reasonably good understanding of the neutrinos at various energy scales produced by the Sun. A number of experiments show very cleanly that what leaves the Sun is rather different from what reaches Earth.

- The first set of experiments use neutrino capture,

  $$\nu_e + n \to p + e^- \ . \tag{7.52}$$

  Clearly, this only works for electron neutrinos. This was first done in the late 1960s, useing tanks of chlorine with the reaction

  $$\nu_e + {}^{37}\text{Cl} \ \longrightarrow \ {}^{37}\text{Ar} + e^- \ . \tag{7.53}$$

  The resulting argon atoms were then counted and used as a proxy for the original neutrino. The incoming neutrinos require an energy of $E > 800$ keV to achieve this heat, which means that this is detecting the neutrinos produced in the rarer neutrino processes. The observed solar neutrinos are a factor of 3 smaller than expected.

  This experiment can be repeated with the chlorine replaced by gallium,

  $$\nu_e + {}^{71}\text{Ga} \ \longrightarrow \ {}^{71}\text{Ge} + e^- \ . \tag{7.54}$$

  Now the threshold is lower, needing only energies of $E \approx 200$ keV, meaning that many more of the Sun's neutrinos can partake. Indeed, the number of events seen is significantly higher, but still with a shortfall of about 40% compared to the theoretical prediction. This shows that the oscillations are energy-dependent, as predicted.

- It is possible to see neutrinos of any type by looking at the scattering process

  $$\nu_\alpha + e^- \to \nu_\alpha + e^- \ . \tag{7.55}$$

  As shown in Figure 21, all neutrinos scatter by exchanging Z bosons, while the electron neutrinos have an additional contribution coming from exchanging a W boson.

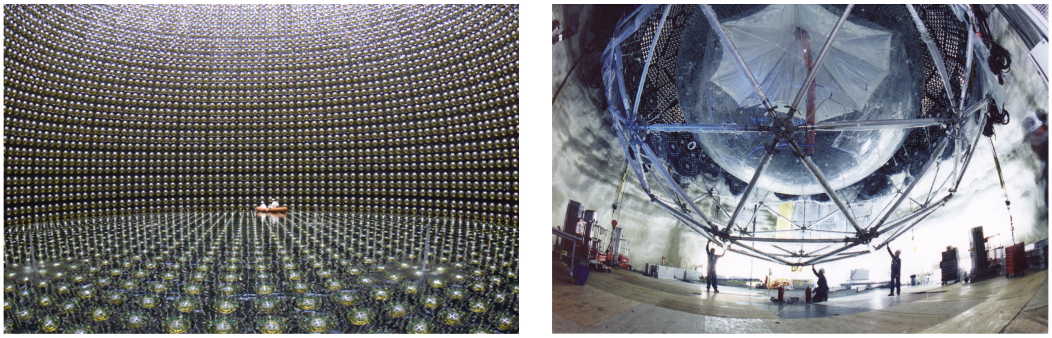


Figure 22. Neutrino detectors tend to look like the lair of a James Bond villain. On the left is a boat cleaning the Super-Kamiokande photosensors as the tank slowly fills up. On the right is the SNO tank, filled with heavy water.

Typically, the neutrinos are scattered off electrons which sit in a large tank of water and detected by the resulting Cerenkov radiation. This, for example, is how the super-Kamiokande experiment in Japan works. The neutrinos must have an energy threshold of $E \approx 8$ MeV and so, as with the chlorine experiments, is sensitive only to the rarer beryllium neutrinos. This time there is a shortfall of around 50%.

These experiments have the advantage that they reveal the direction of the incoming neutrino, and show clearly that the neutrinos are indeed coming from the Sun. In addition, the neutrinos are measured in real time which means that it's possible to detect differences between day, when the neutrinos come directly from the Sun, and night, when the neutrinos must first pass through the Earth before reaching the detector. (We will explain below why such a difference is expected.)

- The state of the art in neutrino detection is offered by the Sudbury neutrino observatory (SNO). This has a tank filled with heavy water, D_20, where the hydrogen is replaced by deuterium D. It doesn't take much to split the deuterium nucleus apart; just 2 MeV of energy is enough. Moreover, neutrinos can knock apart a deuterium nucleus in two different ways. A weak interaction involving an intermediate W boson does the job through a neutrino capture process analogous to those that occur in chlorine or gallium,

$$\nu_e + D \;\rightarrow\; p + p + e^- \; . \tag{7.56}$$

Only electron neutrinos contribute to such processes. However, the neutrinos can also split the deuterium through a weak interaction involving a Z boson,

$$\nu + D \;\rightarrow\; n + p + \nu \; . \tag{7.57}$$

This time there is no charged lepton created, meaning that all three kinds of neutrinos, ν_e, ν_μ and ν_τ contribute.

In addition, SNO measured neutrino scattering events of the form $\nu + e^- \rightarrow \nu + e^-$ where, again, the electron neutrinos have an additional scattering mode through the W boson. The upshot is that SNO was able to see everything – electron, muon and tau neutrinos. And once you see everything, nothing is missing. The end result agreed perfectly with theoretical expectations of the nuclear reactions inside the Sun. The electron neutrinos missed by previous experiments had transmuted into muon and tau neutrinos, incontrovertible evidence for neutrino oscillations.

Atmospheric Neutrinos

The story of missing neutrinos is repeated when we look elsewhere. Cosmic rays, mostly in the form of protons or helium nuclei, are constantly bombarding the Earth. When they hit the atmosphere they create a constant stream of $\pi^\pm$ pions. These pions decay to muons

$$\pi^+ \;\longrightarrow \mu^+ + \nu_\mu \quad \text{and} \quad \pi^- \;\longrightarrow \mu^- + \bar{\nu}_\mu$$

and the muons then quickly decay to electrons,

$$\mu^+ \;\longrightarrow e^+ + \nu_e + \bar{\nu}_\mu \quad \text{and} \quad \mu^- \;\longrightarrow e^- + \bar{\nu}_e + \nu_\mu$$

The resulting *atmospheric neutrinos* have significantly higher energies than solar neutrinos; often around a GeV or higher. Given the decay processes described above, each collision should result in two muon neutrinos (strictly one ν_μ, one $\bar{\nu}_\mu$) for every electron neutrino. The question is: can we find them?

The answer, given by Super-Kamiokande, is interesting and shown in Figure 23. These show plots of the neutrino flux (on the vertical axis) against the angle at which the neutrinos come into the detector (on the horizontal axis). An angle $\cos\theta = 1$, on the far right, means that the neutrinos come directly down. An angle $\cos\theta = -1$, on the far left, means that neutrinos come up, through the Earth.

The data on the left two boxes is for electron neutrinos, both for low-energy events (shown in the top box) and high-energy events (in the bottom box). The red line is the theoretical expectation; the black dots the observed flux. We see that the agreement between experiment and theory works well.

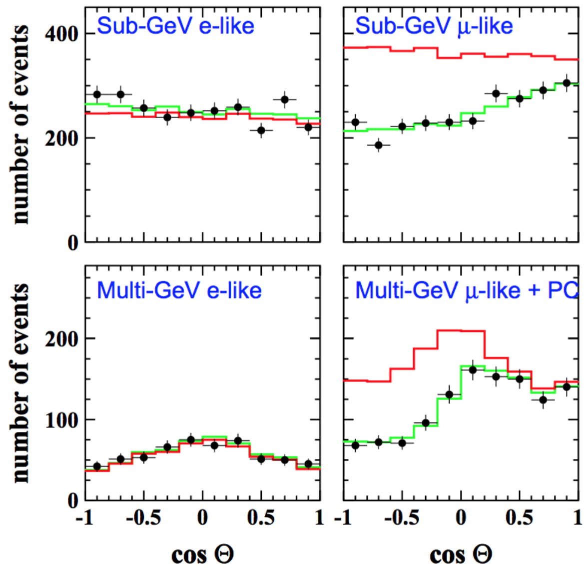


Figure 23. The observed flux of electron neutrinos (on the left) and muon neutrinos (on the right). The top boxes show low-energy neutrinos; the lower boxes high-energy neutrinos. The red line is the theoretical expectation without neutrino oscillations, and the black boxes the data.

The story is more interesting for muon neutrinos, shown in the two boxes on the right. The number of neutrinos coming straight down agrees perfectly with what we expect, but there's a clear deficit for those that come up through the Earth. Why?

For any other particle, you might think that the Earth is simply getting in the way. But neutrinos pass right through the Earth without any difficulty. (Remember the picture of the Sun at night in Figure 19.) Besides: theorists aren't stupid and had taken the presence of the Earth into account when computing the red line! Instead, the key point is that the muon neutrinos have travelled further, and so had more opportunity to convert into other neutrinos, in this case tau.

Importantly, the atmospheric neutrinos clearly show us that neutrino oscillations depend on the length L that neutrinos travel. For those neutrinos that come straight down, we have $L \approx 15$ km and no oscillations are seen. Meanwhile, for those that come up through the Earth we have $L \approx 13000$ km and ν_e is unaffected, while $\nu_\mu \to \nu_\tau$.

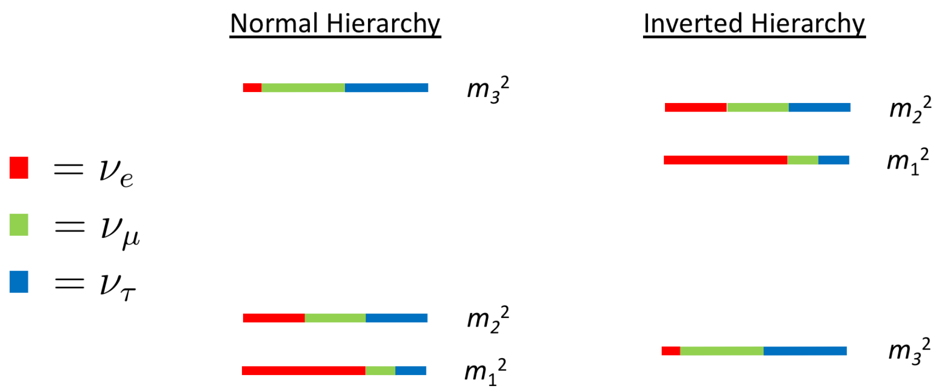


Figure 24. A colour coded description of the possible ordering of neutrino masses.

Neutrino Mass Differences

The experiments sketched above, together with similar terrestrial experiments, are how we determine the precious information about the fundamental parameters in the Standard Model. These tell us the values of the mixing angles that lie within the PMNS matrix (7.17) which, roughly speaking, translate into the following statements about the mass eigenstates: ν_1, ν_2 and ν_3

- ν_1 acts like an electron neutrino two thirds of the time, and as a muon or tau neutrino the other third.
- ν_2 acts like any one of the three neutrinos one third of the time.
- ν_3 acts like a tau neutrino 45% of the time and like a muon neutrino 45% of the time. The remaining 10%, it acts like an electron neutrino.

We also get information about mass differences. The eigenstate ν_1 is known to be lighter than ν_2 and the squares of their masses differ by

$$m_2^2 - m_1^2 \approx 7.4 \times 10^{-5} \text{ eV}^2$$

The resulting difference in their masses is of order $\sim 10^{-2}$ eV, an order of magnitude smaller than the biggest mass. We also know the difference between the masses of ν_3 and ν_2 but, crucially, we don't yet know which one is heavier! We have

$$m_3^2 - m_2^2 = \pm\, 2.5 \times 10^{-3} \text{ eV}^2$$

Of course, if we could measure the mass difference between m_1 and m_3 then we would be able to resolve this $\pm$ ambiguity. As it stands, we just don't know the order of the masses.

The two possibilities are shown in Figure 24. Given the pattern seen in all other fermions, one might expect that the electron neutrino ν_e would be the lightest. Since the ν_e has the biggest overlap with ν_1, this would mean that ν_1 is lightest. This is referred to as the *normal hierarchy*. But, as we've seen, very little about the neutrinos follows our expectation. So another possibility is that ν_3, which contains very little of the electron neutrino, is the lightest. This is called the *inverted hierarchy*. The latest evidence from cosmological observations of the CMB and structure formation give an improved bound on $\sum_i m_i$ and point towards the normal hierarchy.