Preprint typeset in JHEP style - HYPER VERSION 2018

Gauge Theory

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/gaugetheory.html
d.tong@damtp.cam.ac.uk

Contents

For Mum.

Acknowledgements, Caveats and Apologies

The subject of quantum gauge dynamics is a rather mathematical one. These lectures make no pretence at mathematical rigour. I have tried to put the physics front and centre and introduced the relevant mathematics only when necessary. This might not be everyone's cup of tea. But it's mine. Also, I am not an ornithologist. I have no idea whether birds prefer to perch on wires at the end of the day or during their mid-morning brunch. The opening paragraph should be read with poetic licence. The subsequent 400 pages should be more reliable.

My thanks to Pietro Benetti Genolini for comments and to Cory Fletcher for spotting many typos. I am supported by the STFC, by a Royal Society Wolfson Merit award, and by a Simons Investigator award.

Pre-Requisites

These are advanced lectures on quantum field theory. They assume that you are comfortable with the basics of canonical quantisation and, most importantly, path integral techniques. You can find an introduction to the former in my introductory lectures on Quantum Field Theory. Many of the ideas covered in these lectures have their genesis in statistical physics and, in particular, Wilson's development of the renormalisation group; these were covered in the lectures on Statistical Field Theory.

Recommended Books and Resources

Much of the material covered in these lectures was discovered in a golden period of quantum field theory, dating from the mid 1970s and early 1980s, and underlies large swathes of current research. Some of this material can be found in the usual quantum field theory textbooks, but often they tend to peter out just as the fun gets going. Here are some books and resources which cover some relevant topics:

- John Preskill, Lectures on Quantum Field Theory

Preskill's beautiful and comprehensive lectures on quantum field theory are the closest to this course and, in places, offer substantially more detail. Unfortunately they are available only in hand-written form, which means it can take some time to search for the topic you're interested in.

They can be downloaded here: http://www.theory.caltech.edu/~preskill/notes.html

- Sidney Coleman, *Aspects of Symmetry*

Despite their age, Coleman's Erice lectures still sparkle. They cover only a small subset of the material we'll need – solitons, instantons and large N are highlights – but do so with such charm that they shouldn't be missed.

- Alexander Polyakov, *Gauge Fields and Strings*

Polyakov is one of the masters of the path integral, whose pioneering work over the decades did much to cement our current understanding of quantum field theory. His book is not easy going, but rewards anyone who persists.

- Gerard 't Hooft, *Under the Spell of the Gauge Principle*

During the 1970s, 't Hooft wrote a series of papers, each of which changed the way we think about quantum field theory. His name is attached to so many things in these lectures that it can, at times, get confusing. (How do 't Hooft anomalies affect 't Hooft lines in the 't Hooft limit?) This book is a collection of preprints, prefaced by some brief remarks. Still, the originals are well worth the read.

- Yitzhak Frishman and Cobi Sonnenschein, *Non-Perturbative Field Theory: From Two Dimensional Conformal Field Theory to QCD in Four Dimensions*

The goal of this book is similar to these lectures but the itinerary is run in reverse, starting in two dimensions and building up to four.

- Eduardo Fradkin, *Field Theories in Condensed Matter Physics*

- Shankar, *Quantum Field Theory and Condensed Matter*

Both of these books discuss quantum field theory in condensed matter physics. Much of the material is restricted to field theories in $d = 1+1$ and $d = 2+1$ dimensions, and so useful for Sections 7 and 8. But the general approach to understanding the phase structure and behaviour of field theories should resonate.

Lecture notes on various topics discussed in these lectures can be downloaded from the course webpage.

0. Introduction

Towards the end of the day, as feathers droop and hearts flutter from too much flapping, it is not unusual to find flocks of birds resting on high voltage wires. For someone unacquainted with the gauge principle, this may seem like a dangerous act. But birds know better. There is no absolute sense in which the voltage of the wire is high. It is only high in comparison to the Earth.

Of the many fillets and random facts that we are fed in high school science classes, the story of the birds is perhaps the deepest. Most other ideas from our early physics lessons look increasingly antiquated as we gain a deeper understanding of the Universe. The concept of "force", for example, is very 17th century. Yet the curious fact that the electrostatic potential does not matter, only the potential difference, blossoms into the gauge symmetry which underlies the Maxwell equations, the Standard Model and, in the guise of diffeomorphism invariance, general relativity.

Gauge symmetry is, in many ways, an odd foundation on which to build our best theories of physics. It is not a property of Nature, but rather a property of how we choose to describe Nature. Gauge symmetry is, at heart, a redundancy in our description of the world. Yet it is a redundancy that has enormous utility and brings a subtlety and richness to those theories that enjoy it.

This course is about the quantum dynamics of gauge theories. It is here that the utility of gauge invariance is clearest. At the perturbative level, the redundancy allows us to make manifest the properties of quantum field theories, such as unitarity, locality, and Lorentz invariance, that we feel are vital for any fundamental theory of physics but which teeter on the verge of incompatibility. If we try to remove the redundancy by fixing some specific gauge, some of these properties will be brought into focus, while others will retreat into murk. By retaining the redundancy, we can flit between descriptions as is our want, keeping whichever property we most cherish in clear sight.

The purpose of this course is not so much to convince you that gauge theories are useful, but rather to explore their riches. Even at the classical level they have much to offer. Gauge theories are, like general relativity, founded in geometry. They are not associated only to the geometry of spacetime, but to a less intuitive and more general mathematical construct known as a fibre bundle. This brings something new to the table. While most interesting applications of general relativity are restricted to ripples of the curved, but topologically flat, spacetime in which we live, gauge fields are more supple: they can twist and wind in novel ways, bringing the subject of topology firmly into the realm of physics. This will be a dominant theme throughout these lectures. It is a theme that becomes particularly subtle when we include fermions in the mix, and see how they intertwine with the gauge fields.

However, the gauge theoretic fun really starts when we fully immerse ourselves in the quantum world. The vast majority of gauge theories are strongly coupled quantum field theories, where the usual perturbative techniques are insufficient to answer many questions of interest. Despite many decades of work, our understanding of this area remains rather primitive. Yet this is where the most interesting phenomena occur. In particle physics, the strong coupling dynamics of quantum field theory causes quarks and gluons to bind into protons, neutrons and other particles. In condensed matter physics, it causes electrons, which are indivisible particles, to fractionalise in high magnetic fields. There are even tantalising hints that such dynamics may be responsible for the emergence of space and time itself from more fundamental underlying degrees of freedom. The focus of these lectures is not on any particular phenomenon (although confinement in QCD will be something of a pre-occupation). Rather we will try to explain some of the ways in which we can make progress, primitive as it may be, in understanding gauge fields when interactions become strong, and quantum fluctuations wild.

1. Topics in Electromagnetism

We start these lectures by reviewing some topics in Maxwell theory. As we will see, there are some beautiful topological surprises hiding in electromagnetism that are not usually covered in our first undergraduate lectures. These topics will then follow us through these lectures as we explore other examples of gauge theories.

1.1 Magnetic Monopoles

A *magnetic monopole* is an object which emits a radial magnetic field of the form

$$\mathbf{B} = \frac{g\hat{\mathbf{r}}}{4\pi r^2} \quad \Rightarrow \quad \int d\mathbf{S}\cdot\mathbf{B} = g \tag{1.1}$$

Here g is called the *magnetic charge*.

We learn as undergraduates that magnetic monopoles don't exist. First, and most importantly, they have never been observed. Second, there's a law of physics which insists that they can't exist. This is the Maxwell equation

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

Third, this particular Maxwell equation would appear to be non-negotiable. This is because it follows from the definition of the magnetic field in terms of the gauge potential

$$\mathbf{B} = \nabla\times\mathbf{A} \quad \Rightarrow \quad \nabla\cdot\mathbf{B} = 0$$

Yet the gauge potential $\mathbf{A}$ is indispensable in theoretical physics. It is needed whenever we describe the quantum physics of particles moving in magnetic fields. Underlying this statement is the fact that the gauge potential is needed in the classical Hamiltonian treatment. Moreover, there are more subtle phenomena such as the Aharonov-Bohm effect which tell us that there is further, non-local information stored in the gauge potentials. (The Aharonov-Bohm effect was covered in the lectures on Applications of Quantum Mechanics.) All of this points to the fact that we would be wasting our time discussing magnetic monopoles.

Happily, there is a glorious loophole in all of these arguments, first discovered by Dirac, and magnetic monopoles play a crucial role in our understanding of the more subtle effects in gauge theories. The essence of this loophole is that there is an ambiguity in how we define the gauge potentials. In this section, we will see how we can exploit this.

1.1.1 Dirac Quantisation

It turns out that not any magnetic charge g is compatible with quantum mechanics. Since this will be important, we will present several different arguments for the allowed values of g.

We start the simplest, and most physical of these arguments. For this we need to know a fact from quantum mechanics. Suppose that we take a particle which carries electric charge e. We adiabatically transport it along some closed path C in the background of some gauge potential $\mathbf{A}(\mathbf{x},t)$. Then, upon returning to its initial starting position, the wavefunction of the particle picks up a phase

$$\psi \to e^{ie\alpha/\hbar}\psi \quad \text{with} \quad \alpha = \oint_C \mathbf{A}\cdot d\mathbf{x} \tag{1.2}$$

There are different ways to see this, but the simplest is from the path integral approach to quantum mechanics, where the action for a point particle includes the term $\int dt\, e\dot{\mathbf{x}}\cdot\mathbf{A}$; this directly gives the phase above.

The phase of the wavefunction is not an observable quantity in quantum mechanics. However, the phase in (1.2) is really a *phase difference*. We could, for example, place a particle in a superposition of two states, one of which stays still while the other travels around the loop C. The subsequent interference will depend on the phase $e^{ie\alpha/\hbar}$. Indeed, this is the essence of the Aharonov-Bohm effect.

Let's now see what this has to do with magnetic monopoles. We place our electric particle, with charge e, in the background of a magnetic monopole with magnetic charge g. We keep the magnetic monopole fixed, and let the electric particle undergo some journey along a path C. We will ask only that the path C avoid the origin where the magnetic monopole is sitting. This is shown in the left-hand panel of the figure. Upon returning, the particle picks up a phase $e^{ie\alpha/\hbar}$ with

$$\alpha = \oint_C \mathbf{A}\cdot d\mathbf{x} = \int_S d\mathbf{S}\cdot\mathbf{B}$$

where, as shown in the figure, S is the area enclosed by C. Using the fact that $\int_{\mathbf{S}^2} d\mathbf{S}\cdot\mathbf{B} = g$, if the surface S makes a solid angle Ω, this phase can be written as

$$\alpha = \frac{\Omega g}{4\pi}$$

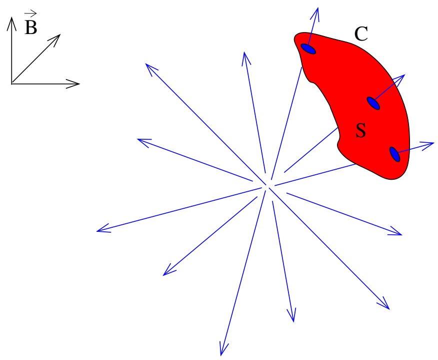


Figure 1: Integrating over S...

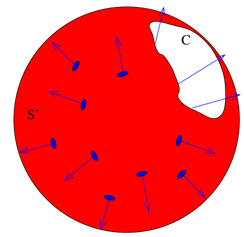


Figure 2: ...or over S'.

However, there's an ambiguity in this computation. Instead of integrating over S, it is equally valid to calculate the phase by integrating over S', shown in the right-hand panel of the figure. The solid angle formed by S' is $\Omega' = 4\pi - \Omega$. The phase is then given by

$$\alpha' = -\frac{(4\pi - \Omega)g}{4\pi}$$

where the overall minus sign comes because the surface S' has the opposite orientation to S. As we mentioned above, the phase shift that we get in these calculations is observable: we can't tolerate different answers from different calculations. This means that we must have $e^{ie\alpha/\hbar} = e^{ie\alpha'/\hbar}$. This gives the condition

$$eg = 2\pi\hbar n \quad \text{ with } n \in \mathbf{Z} \tag{1.3}$$

This is the famous Dirac quantisation condition. The smallest such magnetic charge is also referred to as the *quantum of flux*, $\Phi_0 = 2\pi\hbar/e$.

Above we worked with a single particle of charge e. Obviously, the same argument holds for any other particle of charge e'. There are two possibilities. The first is that all particles carry charge that is an integer multiple of some smallest unit. In this case, it's sufficient to impose the Dirac quantisation condition (1.3) where e is the smallest unit of charge. For example, in our world we should take e to be the electron charge. (You might want to insist that monopoles carry a larger magnetic charge so that they are consistent with quarks which have one third the electron charge. However, it turns out this isn't necessary if the monopoles also carry colour magnetic charge.)

The second possibility is that the particles carry electric charges which are irrational multiples of each other. For example, there may be a particle with charge e and another particle with charge $\sqrt{2}e$. In this case, no magnetic monopoles are allowed.

It's sometimes said that the existence of a magnetic monopole would imply the quantisation of electric charges. This, however, has it slightly backwards. (It also misses the point that we have a beautiful explanation of the quantisation of charges from anomaly cancellation in the Standard Model; we will tell this story in Section 3.4.4.) Instead, the key distinction is the choice of Abelian gauge group. A $U(1)$ gauge group has only integer electric charges and admits magnetic monopoles. In contrast, a gauge group $\mathbf{R}$ can have any irrational charges, but the price you pay is that there are no longer monopoles.

Above we looked at an electrically charged particle moving in the background of a magnetically charged particle. It is simple to generalise the discussion to particles that carry both electric and magnetic charges. These are called *dyons*. For two dyons, with charges (e_1, g_1) and (e_2, g_2), the generalisation of the Dirac quantisation condition requires

$$e_1 g_2 - e_2 g_1 \in 2\pi\hbar\mathbf{Z} \tag{1.4}$$

This is sometimes called the *Dirac-Zwanziger* condition.

1.1.2 A Patchwork of Gauge Fields

The discussion above shows how quantum mechanics constrains the allowed values of magnetic charge. It did not, however, address the main obstacle to constructing a magnetic monopole out of gauge fields $\mathbf{A}$ when the condition $\mathbf{B} = \nabla \times \mathbf{A}$ would seem to explicitly forbid such objects.

Let's see how to do this. Our goal is to write down a configuration of gauge fields which give rise to the magnetic field (1.1) of a monopole which we will place at the origin. We will need to be careful about what we want such a gauge field to look like.

The first point is that we won't insist that the gauge field is well defined at the origin. After all, the gauge fields arising from an electron are not well defined at the position of an electron and it would be churlish to require more from a monopole. This fact gives us our first bit of leeway, because now we need to write down gauge fields on $\mathbf{R}^3\backslash\{0\}$, as opposed to $\mathbf{R}^3$ and the space with a point cut out enjoys some non-trivial topology that we will make use of.

Now consider the following gauge connection, written in spherical polar coordinates

$$A^N_\phi = \frac{g}{4\pi r}\frac{1-\cos\theta}{\sin\theta} \tag{1.5}$$

The resulting magnetic field is

$$\mathbf{B} = \nabla \times \mathbf{A} = \frac{1}{r\sin\theta}\frac{\partial}{\partial\theta}(A^N_\phi\ \sin\theta)\,\hat{\mathbf{r}} - \frac{1}{r}\frac{\partial}{\partial r}(rA^N_\phi)\hat{\boldsymbol{\theta}}$$

Substituting in (1.5) gives

$$\mathbf{B} = \frac{g\hat{\mathbf{r}}}{4\pi r^2} \tag{1.6}$$

In other words, this gauge field results in the magnetic monopole. But how is this possible? Didn't we learn as undergraduates that if we can write $\mathbf{B} = \nabla \times \mathbf{A}$ then $\int d\mathbf{S} \cdot \mathbf{B} = 0$? How does the gauge potential (1.5) manage to avoid this conclusion?

The answer is that $\mathbf{A}^N$ in (1.5) is actually a singular gauge connection. It's not just singular at the origin, where we've agreed this is allowed, but it is singular along an entire half-line that extends from the origin to infinity. This is due to the $1/\sin\theta$ term which diverges at $\theta = 0$ and $\theta = \pi$. However, the numerator $1 - \cos\theta$ has a zero when $\theta = 0$ and the gauge connection is fine there. But the singularity along the half-line $\theta = \pi$ remains. The upshot is that this gauge connection is not acceptable along the line of the south pole, but is fine elsewhere. This is what the superscript N is there to remind us: this gauge connection is fine as long as we keep north.

Now consider a different gauge connection

$$A^S_\phi = -\frac{g}{4\pi r}\frac{1+\cos\theta}{\sin\theta} \tag{1.7}$$

This again gives rise to the magnetic field (1.6). This time it is well behaved at $\theta = \pi$, but singular at the north pole $\theta = 0$. The superscript S is there to remind us that this connection is fine as long as we keep south.

At this point, we make use of the ambiguity in the gauge connection. We are going to take $\mathbf{A}^N$ in the northern hemisphere and $\mathbf{A}^S$ in the southern hemisphere. This is allowed because the two gauge potentials are the same up to a gauge transformation, $\mathbf{A} \to \mathbf{A} + \nabla\omega$. Recalling the expression for $\nabla\omega$ in spherical polars, we find that for $\theta \neq 0, \pi$, we can indeed relate A^N_ϕ and A^S_ϕ by a gauge transformation,

$$A^N_\phi = A^S_\phi + \frac{1}{r\sin\theta}\,\partial_\phi\omega \quad \text{where} \;\; \omega = \frac{g\phi}{2\pi} \tag{1.8}$$

However, there's still a question remaining: is this gauge transformation allowed? The problem is that the function ω is not single valued: $\omega(\phi = 2\pi) = \omega(\phi = 0) + g$. Should this concern us?

To answer this, we need to think more carefully about what we require from a gauge transformation. This is where the charged matter comes in. In quantum mechanics, the gauge transformation acts on the wavefunction of the particle as

$$\psi \to e^{ie\omega/\hbar}\psi$$

In quantum field theory, we have the same transformation but now with ψ interpreted as the field. We will not require that the gauge transformation ω is single-valued, but only that the wavefunction ψ is single-valued. This holds for the gauge transformation (1.8) provided that we have

$$eg = 2\pi\hbar n \quad \text{with } n \in \mathbf{Z}$$

This, of course, is the Dirac quantisation condition (1.3).

Mathematically, this is a construction of a topologically non-trivial $U(1)$ bundle over the $\mathbf{S}^2$ surrounding the origin. In this context, the integer n is called the first Chern number.

1.1.3 Monopoles and Angular Momentum

Here we provide yet another derivation of the Dirac quantisation condition, this time due to Saha. The key idea is that the quantisation of magnetic charge actually follows from the more familiar quantisation of angular momentum. The twist is that, in the presence of a magnetic monopole, angular momentum isn't quite what you thought.

Let's start with some simple classical mechanics. The equation of motion for a particle of mass m and charge e and position $\mathbf{r}$, moving in a magnetic field $\mathbf{B}$, is the familiar Lorentz force law

$$\frac{d\mathbf{p}}{dt} = e\,\dot{\mathbf{r}} \times \mathbf{B}$$

with $\mathbf{p} = m\dot{\mathbf{r}}$ the mechanical momentum. If you remember the Hamiltonian formalism for a particle in a magnetic field, you might recall that $\mathbf{p}$ is not the canonical momentum, a fact which is hiding in the background in what follows. Now let's consider this equation in the background of a magnetic monopole, with

$$\mathbf{B} = \frac{g}{4\pi}\frac{\mathbf{r}}{r^3}$$

The monopole has rotational symmetry so we would expect that the angular momentum, $\mathbf{r} \times \mathbf{p}$, is conserved. Let's check:

$$\frac{d(\mathbf{r} \times \mathbf{p})}{dt} = \dot{\mathbf{r}} \times \mathbf{p} + \mathbf{r} \times \dot{\mathbf{p}} = \mathbf{r} \times \dot{\mathbf{p}} = e\mathbf{r} \times (\dot{\mathbf{r}} \times \mathbf{B})$$

$$= \frac{eg}{4\pi r^3}\,\mathbf{r}\times(\dot{\mathbf{r}}\times\mathbf{r}) = \frac{eg}{4\pi}\left(\frac{\dot{\mathbf{r}}}{r} - \frac{\dot{r}\mathbf{r}}{r^2}\right)$$
$$= \frac{d}{dt}\left(\frac{eg}{4\pi}\hat{\mathbf{r}}\right)$$

We see that in the presence of a magnetic monopole, the naive angular momentum $\mathbf{r}\times\mathbf{p}$ is not conserved! However, we can easily write down a modified angular momentum that is conserved, namely[1]

$$\mathbf{L} = \mathbf{r}\times\mathbf{p} - \frac{eg}{4\pi}\hat{\mathbf{r}}$$

The extra term can be thought of as the angular momentum stored in $\mathbf{E}\times\mathbf{B}$. The surprise is that the particle has angular momentum even if it doesn't move!

Before we move on, there's a nice and quick corollary that we can draw from this. The angular momentum vector $\mathbf{L}$ does not change with time. But the angle that the particle makes with this vector is

$$\mathbf{L}\cdot\hat{\mathbf{r}} = -\frac{eg}{4\pi} = \text{constant}$$

This means that the particle moves on a cone, with axis $\mathbf{L}$ and angle $\cos\theta = -eg/4\pi L$.

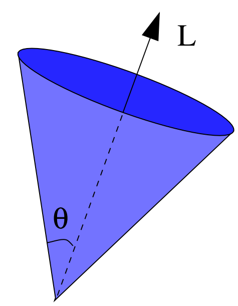


Figure 3:

So far, our discussion has been classical. Now we invoke some simple quantum mechanics: the angular momentum should be quantised. In particular, the angular momentum in the z-direction should be $L_z \in \frac{1}{2}\hbar\mathbf{Z}$. Using the result above, we have

$$\frac{eg}{4\pi} = \frac{1}{2}\hbar n \quad \Rightarrow \quad eg = 2\pi\hbar n \qquad \text{with } n\in\mathbf{Z}$$

Once again, we find the Dirac quantisation condition.

On Bosons and Fermions

There is an interesting factor of 2 buried in the discussion above. Consider a minimal Dirac monopole, with $g = 2\pi\hbar/e$. In the background of this monopole, we will throw in a particle of spin $\mathbf{S}$. The total angular momentum $\mathbf{J}$ is then

$$\mathbf{J} = \mathbf{L} + \mathbf{S} = \mathbf{r}\times\mathbf{p} + \mathbf{S} - \frac{1}{2}\hat{\mathbf{r}} \tag{1.9}$$

The key observation is that the final term, due to the monopole, shifts the total angular momentum by $1/2$. That means, in the presence of a monopole, bosons have half-integer angular momentum while fermions have integer angular momentum! We'll not need this curious fact for most of these lectures, but it will return in Section 8.6 when we discuss some surprising dualities in $d = 2+1$ quantum field theories.

[1]We also noticed this in the lecture notes on Classical Dynamics; see Section 4.3.2.

1.2 The Theta Term

In relativistic notation, the Maxwell action for electromagnetism takes a wonderfully compact form,

$$S_{\text{Maxwell}} = \frac{1}{\mu_0}\int d^4x\ -\frac{1}{4}F^{\mu\nu}F_{\mu\nu} = \int d^4x\ \left(\frac{\epsilon_0}{2}\,\mathbf{E}^2 - \frac{1}{2\mu_0}\mathbf{B}^2\right) \tag{1.10}$$

Here $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ and $E_i = cF_{0i}$ and $F_{ij} = -\epsilon_{ijk}B_k$.

One reason that the Maxwell action is so simple is that there is very little else we can write down that is both gauge invariant and Lorentz invariant. There are terms of order $\sim F^4$ and higher, which give rise to non-linear electrodynamics, but these will always be suppressed by some high mass scale and are unimportant at low-energies.

There is, however, one other term that we can add to the Maxwell action that, at first glance, would seem to be of equal importance. A second glance then shows that it is completely unimportant and it's on the third glance that we see the role it plays. This is the *theta term*.

We start by defining the dual tensor

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

$^\star F^{\mu\nu}$ takes the same form as the original electromagnetic tensor $F_{\mu\nu}$, but with $\mathbf{E}/c \to -\mathbf{B}$ and $\mathbf{B} \to +\mathbf{E}/c$. The theta term is then given by

$$S_\theta = \frac{\theta e^2}{4\pi^2\hbar}\int d^4x\ \frac{1}{4}{}^\star F^{\mu\nu}F_{\mu\nu} = -\frac{\theta e^2}{4\pi^2\hbar c}\int d^4x\ \mathbf{E}\cdot\mathbf{B} \tag{1.11}$$

where θ is a parameter. The morass of constants which accompany it ensure, among other things, that θ is dimensionless; we will have more to say about this in Section 1.2.4. Like the original Maxwell term, the theta term is quadratic in electric and magnetic fields. However, it is simple to check that the theta term can be written as a total derivative,

$$S_\theta = \frac{\theta e^2}{8\pi^2\hbar}\int d^4x\ \partial_\mu(\epsilon^{\mu\nu\rho\sigma}A_\nu\,\partial_\rho A_\sigma) \tag{1.12}$$

We say that the theta term is *topological*. It depends only on boundary information. Another way of saying this is that we don't need to use the spacetime metric to define the theta term; we instead use the volume form $\epsilon^{\mu\nu\rho\sigma}$. The upshot is that the theta term does not change the equations of motion and, it would seem, can have little effect on the physics.

As we will now see, this latter conclusion is a little rushed. There are a number of situations in which the theta term does lead to interesting physics. These situations often involve subtle interplay between quantum mechanics and topology.

Axion Electrodynamics

We start by looking at situations where θ affects the dynamics classically. This occurs when θ is not constant, but instead varies in space and, possibly, time: $\theta = \theta(\mathbf{x}, t)$. In general, the action governing the electric and magnetic field is given by

$$S = \int d^4x \, \left(-\frac{1}{4} F^{\mu\nu} F_{\mu\nu} + \frac{e^2}{16\pi^2 \hbar} \, \theta(\mathbf{x}, t) {}^\star F^{\mu\nu} F_{\mu\nu} \right)$$

The equations of motion from this action read

$$\nabla \cdot \mathbf{E} = \frac{\alpha c}{\pi} \nabla\theta \cdot \mathbf{B} \quad \text{and} \quad \frac{1}{c^2} \frac{\partial \mathbf{E}}{\partial t} + \nabla \times \mathbf{B} = \frac{\alpha}{\pi c} \left(\dot{\theta} \mathbf{B} + \nabla\theta \times \mathbf{E} \right) \tag{1.13}$$

where

$$\alpha = \frac{1}{4\pi\epsilon_0} \frac{e^2}{\hbar c}$$

is the dimensionless fine structure constant. It takes the approximate value $\alpha \approx 1/137$. The deformed Maxwell equations are sometimes referred to as the equations of *axion electrodynamics*. The name is slightly misleading; an axion is what you get if you promote θ to a new dynamical field. Here we're considering it to be some fixed background. They are accompanied by the usual Bianchi identities, $\partial_\mu {}^\star F^{\mu\nu} = 0$, which remain unchanged

$$\nabla \cdot \mathbf{B} = 0 \quad \text{and} \quad \frac{\partial \mathbf{B}}{\partial t} + \nabla \times \mathbf{E} = 0$$

The equations (1.13) carry much – although not all – of the new physics. The first tells us that in regions of space where θ varies, a magnetic field $\mathbf{B}$ acts like an electric charge density. The second tells us that the combination $(\dot{\theta}\mathbf{B} + \nabla\theta \times \mathbf{E})$ acts like a current density.

1.2.1 The Topological Insulator

There are a fascinating class of materials, known as *topological insulators*, whose dynamics is characterised by the fact that $\theta = \pi$. (We'll see what's special about the value $\theta = \pi$ in Section 1.2.4.) Examples include the Bismuth compounds Bi_2Se_3 and Bi_2Te_3.

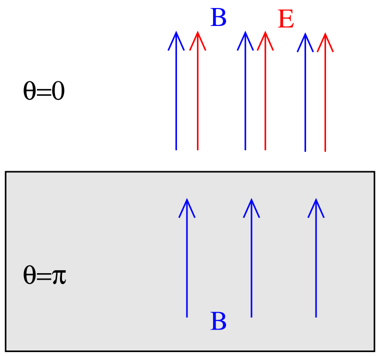


Figure 4: Applying a magnetic field

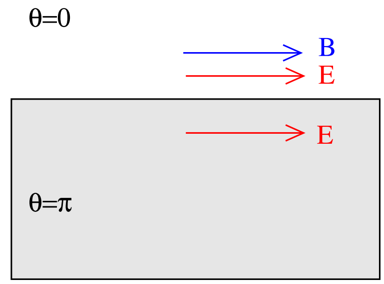


Figure 5: Applying an electric field

Consider a topological insulator, with $\theta = \pi$, filling (most of) the lower-half plane, $z < -\epsilon$. We fill (most of) the upper-half plane, $z > \epsilon$, with the vacuum which has $\theta = 0$. In the intermediate region $z \in [-\epsilon, \epsilon]$ we have $\partial_z \theta \neq 0$.

Let's first shine a magnetic field $B_z = B$ on this interface from below, as shown in the left-hand panel of the figure. The first equation in (1.13) tells us that there is an effective accumulation of charge density, $\rho = \alpha c (\partial_z \theta) B / \pi$. The surface charge per unit area is given by

$$\sigma = \int_{-\epsilon}^{\epsilon} dz \ \rho = \alpha c B$$

This surface charge will give rise to an electric field outside the topological insulator. We learn that the boundary of a topological insulator has rather striking properties: it takes a magnetic field inside and generates an electric field outside!

Alternatively, we can turn on an electric field which lies tangential to the interface, say $E_y = E$. This is shown in the right-hand panel of the figure. The second equation in (1.13) tells us that, in the regime where $\partial_z \theta \neq 0$, the electric field acts as a surface current $\mathbf{K}$, lying within the interface, perpendicular to $\mathbf{E}$,

$$K_x = \alpha \epsilon_0 c E_y \tag{1.14}$$

This, in turn, then generates a magnetic field outside the topological insulator, parallel to $\mathbf{E}$. This, again, is shown in the right-hand panel of the figure.

The creation of a two-dimensional current which lies perpendicular to an applied electric field is called the *Hall effect*. The coefficient of proportionality is known as the *Hall conductivity* and there is a long and beautiful story about how it takes certain very special values which are rational multiples of $e^2/2\pi\hbar$. (More details can be found in the lecture notes on the Quantum Hall Effect.) In the present example (1.14), the Hall conductivity is

$$\sigma_{xy} = \frac{1}{2}\frac{e^2}{2\pi\hbar}$$

This is usually abbreviated to say that the interface of the topological insulator has Hall conductivity 1/2.

The general phenomenon in which electric fields induce magnetic fields and vice versa goes by the name of the *topological magneto-electric effect*.

Continuity Conditions

There's a slightly different, but equivalent way of describing the physics above. It doesn't tell us anything new, but it does make contact with the language we previously used to describe electrodynamics in materials[2]. We introduce the electric displacement

$$\mathbf{D} = \epsilon_0\left(\mathbf{E} - \frac{\alpha c\theta}{\pi}\mathbf{B}\right)$$

Comparing to the usual expression for $\mathbf{D}$, we see that, in a topological insulator, a magnetic field $\mathbf{B}$ acts like polarisation. When θ varies, we have a varying polarisation, resulting in bound charge. This is what we saw in the topological insulator interface above. Similarly, we define the magnetising field

$$\mathbf{H} = \frac{1}{\mu_0}\left(\mathbf{B} + \frac{\alpha}{\pi c}\theta\mathbf{E}\right)$$

We see in a topological insulator, $\mathbf{E}$ acts like magnetisation. When θ varies, we get a varying magnetisation which results in bound currents.

With these definitions, the equations of axion electrodynamics (1.13) take the usual form of the Maxwell equations in matter

$$\nabla\cdot\mathbf{D} = 0 \quad \text{and} \quad \nabla\times\mathbf{H} - \frac{\partial\mathbf{D}}{\partial t} = 0$$

Now we can use the standard arguments (involving Gaussian pillboxes and line integrals) that tell us $\mathbf{B}$ perpendicular to a surface and $\mathbf{E}$ tangential to a surface are necessarily continuous. This means that if we introduce the normal vector to the surface $\hat{\mathbf{n}}$, then

$$\hat{\mathbf{n}}\cdot\Delta\mathbf{B} = 0 \quad \text{and} \quad \hat{\mathbf{n}}\times\Delta\mathbf{E} = 0 \tag{1.15}$$

[2]See Section 7 of the lectures on Electromagnetism.

For a usual dielectric, $\mathbf{D}$ perpendicular to a surface and $\mathbf{H}$ parallel to a surface are both discontinuous, with the discontinuity given by the surface charge and current respectively. Here, we've absorbed the θ-induced surface charges and currents into the definition of $\mathbf{D}$ and $\mathbf{H}$. If there are no further, external charges we have

$$\hat{\mathbf{n}} \cdot \Delta \mathbf{D} = 0 \quad \text{and} \quad \hat{\mathbf{n}} \times \Delta \mathbf{H} = 0 \tag{1.16}$$

It is simple to check that this condition reproduces the topological magneto-electric results that we described above.

1.2.2 A Mirage Monopole

Let's continue to explore the physics of interface between the vacuum (filling $z > 0$) and a topological insulator (filling $z < 0$). Here's a fun game to play: take an electric charge q and place it in the vacuum at point $\mathbf{x} = (0, 0, d)$, a distance d above a topological insulator. What do the resulting electric and magnetic fields look like?

We can answer this using the continuity conditions described above, together with the idea of an image charge. (We met the image charge in the Electromagnetism lecture notes when discussing metals. One can also use the same tricks to describe the electric field in the presence of a dielectric, which is closer in spirit to the calculation here.) As always with the method of images, we need a flash of insight to write down an ansatz. (Or, equivalently, someone to tell us the answer.) However, if we find a solution that works then general results about the uniqueness of boundary-value problems ensure that this is the unique condition.

In the present case, the answer is quite cute: we will see that if we sit in the vacuum $z > 0$, the electric and magnetic field lines are those due to the original particle at $\mathbf{x} = (0, 0, d)$, together with a mirror dyon sitting at $\mathbf{x} = (0, 0, -d)$ with electric and magnetic charges (q', g). Meanwhile, if we sit in the topological insulator, $z < 0$, the electric and magnetic field lines are those due to the original particle, now superposed with those arising from a mirror dyon with charges $(q', -g)$, also sitting at $\mathbf{x} = (0, 0, d)$. Note that in both cases, the dyon is a mirage: it sits outside of the region we have access to. If we try to reach it by crossing the boundary, it switches to the other side!

To see that this is the correct answer (and to compute q' and g), we work with scalar potentials. It's familiar to use the electrostatic equation $\nabla \times \mathbf{E} = 0$ to write $\mathbf{E} = -\nabla \phi$. The electric potential in the two regions is

$$\phi = \frac{1}{4\pi\epsilon_0} \left(\frac{q}{\sqrt{x^2 + y^2 + (z-d)^2}} + \frac{q'}{\sqrt{x^2 + y^2 + (z+d)^2}} \right) \qquad z > 0$$

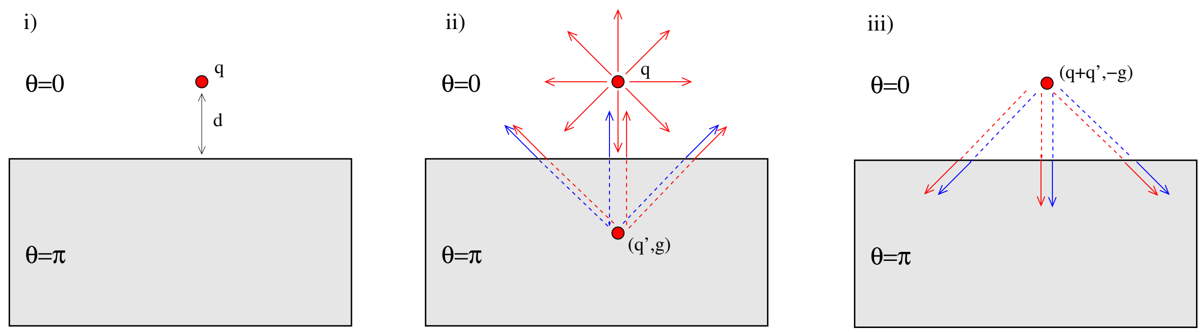


Figure 6: i) An electric charge q placed near a topological insulator. ii) The resulting electric and magnetic field lines as seen outside. iii) The field lines as seen inside.

and

$$\phi = \frac{1}{4\pi\epsilon_0}\frac{q+q'}{\sqrt{x^2+y^2+(z-d)^2}} \qquad\qquad z<0$$

Note that $E_y = -\partial_y\phi$ and $E_x = -\partial_x\phi$ are both continuous at the interface $z=0$, as required by (1.15). In contrast, E_z will be discontinuous; we'll look at this shortly.

For the magnetic field, this is one of the few occasions where it's useful to work with the magnetic scalar potential. This means that we use the fact that $\nabla\times\mathbf{B}=0$ to write $\mathbf{B} = -\nabla\Omega$. (Recall the warning from earlier lectures: unlike the electric scalar ϕ, there is nothing fundamental about Ω; it is merely a useful computational trick.) We then have

$$\Omega = \frac{1}{4\pi}\frac{g}{\sqrt{x^2+y^2+(z+d)^2}} \qquad\qquad z>0$$

and

$$\Omega = -\frac{1}{4\pi}\frac{g}{\sqrt{x^2+y^2+(z-d)^2}} \qquad\qquad z<0$$

Note that $B_z = -\partial_z\Omega$ is continuous across the plane $z=0$, as required by the condition (1.15).

Let's now look at the (dis)continuity conditions (1.16). From the expressions above, we have

$$D_z\Big|_{z=0^+} = \epsilon_0 E_z\Big|_{z=0^+} = -\frac{q-q'}{4\pi}\frac{d}{(x^2+y^2)^{3/2}}$$

and

$$D_z\Big|_{z=0^-} = \epsilon_0\Big(E_z + \alpha c B_z\Big)\Big|_{z=0^-} = -\frac{q+q'+\alpha c\epsilon_0 g}{4\pi}\frac{d}{(x^2+y^2)^{3/2}}$$

Equating these tells us that the magnetic charge on the image dyon is

$$g = \frac{2q'}{\alpha c\epsilon_0} \tag{1.17}$$

Similarly, the magnetic field tangent to the interface is

$$H_x\Big|_{z=0^+} = \frac{1}{\mu_0}B_x\Big|_{z=0^+} = \frac{g}{4\pi\mu_0}\frac{d}{(x^2+y^2)^{3/2}}$$

and

$$H_x\Big|_{z=0^-} = \frac{1}{\mu_0}\left(B_x - \frac{\alpha}{c}E_x\right)\Big|_{z=0^-} = -\frac{g-(q+q')\alpha/c\epsilon_0}{4\pi\mu_0}\frac{d}{(x^2+y^2)^{3/2}}$$

which gives us

$$g = \frac{(q+q')\alpha}{2c\epsilon_0} \tag{1.18}$$

Happily we have found a solution both (1.15) and (1.16) can be satisfied across the boundary. Uniqueness means that this must be the correct solution. As we have seen, it involve mirage dyons sitting beyond our reach. From (1.17) and (1.18), we learn that the electric and magnetic charges carried by these dyons are given by

$$q' = -\frac{\alpha^2}{4+\alpha^2}q \quad \text{and} \quad g = \frac{2\alpha}{(4+\alpha^2)c\epsilon_0}q$$

The monopoles and dyons that arise in this way are mirages. Experimentally, we're in the slightly unusual situation where we can see mirage monopoles, but not real monopoles!

1.2.3 The Witten Effect

There is also an interesting story to tell about genuine magnetic monopoles. As we now show, the effect of the θ term is to endow the magnetic monopole with an electric charge. This is known as the *Witten effect*.

It's simplest to frame the set-up by first taking a magnetic monopole with magnetic charge g and placing it inside a vacuum, with $\theta = 0$. We then surround this with a medium that has $\theta \neq 0$ as shown in the figure. We know what happen from our discussion above. When the magnetic field crosses the interface where θ changes, it will induce an electric charge. This charge follows from the first equation in (1.13). From inside the medium when $\theta \neq 0$, it looks as if the monopole has electric charge

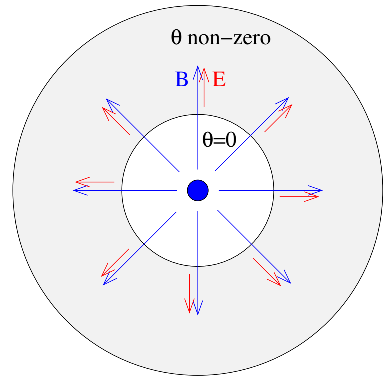


Figure 7:

$$q = -\alpha c \epsilon_0 \frac{\theta g}{\pi} = -\frac{e^2}{4\pi\hbar}\frac{\theta g}{\pi} \tag{1.19}$$

Note, however, that this result is independent of the size of the interior region where $\theta = 0$. We could shrink this region down until it is infinitesimally small, and we still find that the monopole has charge q. The correct interpretation of this is that when $\theta \neq 0$, a monopole is, in fact, a dyon: it carries electric charge (1.19).

When the monopole carries the minimum allowed magnetic charge, its electric charge is given by

$$g = \frac{2\pi\hbar}{e} \quad \Rightarrow \quad q = \frac{e\theta}{2\pi}$$

In particular, if we place a magnetic monopole inside a topological insulator, it turns into a dyon which carries half the charge of the electron.

Note that if we take $\theta = 2\pi$ then the electric charge of the monopole coincides with that of the electron; in this case, we can construct a neutral monopole by considering a bound state of the dyon + positron. However, when θ is not a mutliple of 2π, all monopoles necessarily carry electric charge.

One might wonder why we had to introduce the region with $\theta = 0$ at all. What happens if we simply insist that we place the monopole directly in a system with $\theta \neq 0$? You would again discover the Witten effect, but now you have to be careful about the boundary conditions you can place on the gauge field at the origin. We won't describe this here. We will, however, give a slightly different derivation. Consider, once again first, placing a monopole in a medium with $\theta = 0$. This time we will very slowly increase θ. (Don't ask me how...I don't know! We just imagine it's possible.) The second equation in (1.13) contains a $\dot{\theta}$ term which tells us that this will be accompanied by a time-varying electric field which lies parallel to $\mathbf{B}$. At the end of this process, the final electric field will be

$$\mathbf{E} = \frac{-\alpha c}{\pi} \int \dot{\theta} dt \, \mathbf{B} = -\frac{\alpha c \theta}{\pi} \mathbf{B}$$

Once again, we learn that the monopole carries an electric charge given by (1.19).

We'll see various other manifestations of the Witten effect as these lectures progress including, in Section 2.8.2, for monopoles in non-Abelian gauge theories.

1.2.4 Why θ is Periodic

In classical axion electrodynamics, θ can take any value. Indeed, as we have seen, it is only spatial and temporal variations of θ that play a role. However, in the quantum theory θ is a periodic variable: it lies in the range

$$\theta \in [0, 2\pi)$$

This is the real reason why θ was accompanied by that mess of other constants multiplying the action; it is to ensure that the periodicity is something natural.

The periodicity of θ in electrodynamics is actually fairly subtle. It hinges on the topology of the $U(1)$ gauge fields. We'll see that, after imposing appropriate boundary conditions, S_θ can only take values of the form

$$S_\theta = \hbar \theta N \quad \text{with} \quad N \in \mathbf{Z} \tag{1.20}$$

This means that the theta angle contributes to the partition function as

$$\exp\left(\frac{iS_\theta}{\hbar}\right) = e^{iN\theta}$$

The factor of i here is all important. In Minkowski signature, the action always sits with a factor of i. However, one of the special things about the theta term is that it has only a single time derivative in the integrand, a fact which can be traced to the appearance of the $\epsilon^{\mu\nu\rho\sigma}$ anti-symmetric tensor. This means that the factor of i persists even in Euclidean signature. Since N is an integer, we see that the value of θ in the partition function is only important modulo 2π.

So our task is to show that, when evaluated on any field configuration, S_θ must take the form (1.20). The essence of the argument follows from the fact that the theta term is a total derivative (1.12), which shows us that the value of S_θ depends only on the boundary condition. To exploit the topology lurking in the $U(1)$ gauge field, we will work on a compact Euclidean spacetime which we take to be $\mathbf{T}^4$. We'll take each of the circles in the torus to have radii R.

We'll make life even easier for ourselves by restricting to the special case $\mathbf{E} = (0,0,E)$ and $\mathbf{B} = (0,0,B)$ with E and B constant. The integral that we're interested in is

$$\int_{\mathbf{T}^4} d^4x\ EB = \int_{\mathbf{T}^2} dx^0 dx^3 E \int_{\mathbf{T}^2} dx^1 dx^2\ B \tag{1.21}$$

This still looks like it can take any value we like. But we need to recall that E and B are not the fundamental fields; these are the gauge fields A_μ. And these must be well defined on the underlying torus. As we'll now show, this puts restrictions on the allowed values of E and B.

First, we need the following result: when a direction of space, say x^1, is periodic with radius R, then the constant part of the corresponding gauge field (also known as the *zero mode*) also becomes periodic with radius

$$A_1 \equiv A_1 + \frac{\hbar}{eR} \tag{1.22}$$

This arises because the presence of a circle allows us to do something interesting with gauge transformations $A_1 \to A_1 + \partial_1\omega$. As in Section 1.1.2, we do not insist that $\omega(x)$ is single valued. Instead, we require only that $e^{ie\omega/\hbar}$ is single-valued, since this is what acts on the wavefunction. This allows us to perform gauge transformations that wind around the circle, such as

$$\omega = \frac{x^1\hbar}{eR}$$

These are sometimes called *large* gauge transformations, a name which reflects the fact that they cannot be continuously deformed to the identity. Under such a gauge transformation, we see that

$$A_1 \to A_1 + \partial_1\omega = A_1 + \frac{\hbar}{eR}$$

But field configurations that are related by a gauge transformation are to be viewed as physically equivalent. We learn that the constant part of the gauge field is periodically identified as (1.22) as claimed.

Now let's see how this fact restricts the allowed values of the integral (1.21). The magnetic field is written as

$$B = \partial_1 A_2 - \partial_2 A_1$$

We can work in a gauge where $A_1 = 0$, so that $B = \partial_1 A_2$. If we want a uniform, constant B then we need to write $A_2 = Bx^1$. This isn't single valued. However, that needn't be a problem because, as we've seen above, A_2 is actually a periodic variable with periodicity $\hbar/eR$. This means that we're perfectly at liberty to write $A_2 = Bx_1$, but only if this has the correct period (1.22). This holds provided

$$B = \frac{\hbar n}{2\pi e R^2} \quad \text{with} \quad n \in \mathbf{Z} \qquad \Rightarrow \qquad \int_{\mathbf{T}^2} dx^1 dx^2 \; B = \frac{2\pi\hbar n}{e} \tag{1.23}$$

Note that this is the same as the condition on $\int d\mathbf{S}\cdot\mathbf{B} = g$ that we derived from the Dirac quantisation condition (1.3). Indeed, the derivation above relies on the same kind of arguments that we used when discussing magnetic monopoles.

We can now apply exactly the same argument to the electric field,

$$\frac{E}{c} = \partial_0 A_3 - \partial_3 A_0$$

Let's work in a gauge with $A_0 = 0$, so that $E/c = \partial_0 A_3$. We can write $A_3 = (E/c)x^0$, which is compatible with the periodicity of A_3 only when $E/c = \hbar n'/2\pi e R^2$ for some $n' \in \mathbf{Z}$. We find

$$\int_{\mathbf{T}^2} dx^0 dx^3 \, E = \frac{2\pi c\hbar n'}{e} \tag{1.24}$$

Before we go on, let me point out something that may be confusing. You may have thought that the relevant equation for $\mathbf{E}$ is Gauss' law which, given the quantisation of charge, states that $\int d\mathbf{S}\cdot\mathbf{E} = en'$ for some $n' \in \mathbf{Z}$. But that's not what we computed in (1.24) because $\mathbf{E} = (0,0,E)$ lies parallel to the side of the torus, not perpendicular. Instead, both (1.23) and (1.24) are best thought of as integrating the 2-form $F_{\mu\nu}$ over the appropriate $\mathbf{T}^2$. For the magnetic field, this coincides with $\int d\mathbf{S}\cdot\mathbf{B}$ which measures the magnetic charge enclosed in the manifold. It does not, however, coincide with $\int d\mathbf{S}\cdot\mathbf{E}$ which measures the electric charge.

Armed with (1.23) and (1.24), we see that, at least for this specific example,

$$\int_{\mathbf{T}^4} d^4x \; \mathbf{E}\cdot\mathbf{B} = \frac{4\pi^2\hbar^2 c N}{e^2} \qquad \Rightarrow \qquad S_\theta = \hbar\theta N \quad \text{with} \quad N = nn' \in \mathbf{Z}$$

which is our promised result (1.20)

The above explanation was rather laboured. It's pretty straightforward to generalise it to non-constant $\mathbf{E}$ and $\mathbf{B}$ fields. If you're mathematically inclined, it is the statement that the second Chern number of a $U(1)$ bundle is integer valued and, as we have seen above, is actually equal to the product of two first Chern numbers. Finally note that, although we took Euclidean spacetime to be a torus T^4, the end result does not depend on the volume of the torus which is set by R. Nonetheless, the introduction of the torus was crucial in our argument: we needed the circles of T^4 to exploit the fact that $\Pi_1(U(1)) \cong \mathbf{Z}$. We will see another derivation of this when we come to discuss the anomaly in Section 3.3.1.

1.2.5 Parity, Time-Reversal and $\theta = \pi$

The theta term does not preserve the same symmetries as the Maxwell term. It is, of course, gauge invariant and Lorentz invariant. But it is not invariant under certain discrete symmetries.

The discrete symmetries of interest are parity $\mathcal{P}$ and time reversal invariance $\mathcal{T}$. Parity acts by flipping all directions of space

$$\mathcal{P} : \mathbf{x} \mapsto -\mathbf{x} \tag{1.25}$$

(At least this is true in any odd number of spatial dimensions; in an even number of spatial dimensions, this is simply a rotation.) Meanwhile, as the name suggests, time reversal flips the direction of time

$$\mathcal{T} : \ t \to -t$$

We would like to understand how these act on the electric and magnetic fields. This follows from looking at the Lorentz force law,

$$m\ddot{\mathbf{x}} = e\left(\mathbf{E} + \dot{\mathbf{x}} \cdot \mathbf{B}\right)$$

This equation is invariant under neither parity, nor time reversal. However it can be made invariant if we simultaneously act on both $\mathbf{E}$ and $\mathbf{B}$ as

$$\mathcal{P} : \mathbf{E}(\mathbf{x}, t) \mapsto -\mathbf{E}(-\mathbf{x}, t) \quad \text{ and } \quad \mathcal{P} : \mathbf{B}(\mathbf{x}, t) \mapsto \mathbf{B}(-\mathbf{x}, t)$$

and

$$\mathcal{T} : \mathbf{E}(\mathbf{x}, t) \mapsto \mathbf{E}(\mathbf{x}, -t) \quad \text{ and } \quad \mathcal{T} : \mathbf{B}(\mathbf{x}, t) \mapsto -\mathbf{B}(\mathbf{x}, -t)$$

We say that $\mathbf{E}$ is odd under parity and even under time reversal; $\mathbf{B}$ is even under parity and odd under time reversal.

As an aside, note that a high energy theorist usually refers to $\mathcal{CP}$ rather than $\mathcal{T}$. Here $\mathcal{C}$ is charge conjugation which acts as $\mathcal{C} : \mathbf{E} \mapsto -\mathbf{E}$ and $\mathcal{C} : \mathbf{B} \mapsto -\mathbf{B}$, with the consequence that $\mathcal{CP} : \mathbf{E} \mapsto \mathbf{E}$ and $\mathcal{CP} : \mathbf{B} \mapsto -\mathbf{B}$, rather like $\mathcal{T}$. However, there is a difference between the two symmetries: $\mathcal{CP}$ is unitary, while $\mathcal{T}$ is anti-unitary.

This means that, in general, the theta term breaks both parity and time-reversal invariance. We say that $\theta \mapsto -\theta$ under $\mathcal{P}$ and $\mathcal{T}$. There are two exceptions. One of these is obvious: when $\theta = 0$, the theory is invariant under these discrete symmetries. However, when $\theta = \pi$ the theory is also invariant. This is because, as we have seen θ is periodic so $\theta = \pi$ is the same as $\theta = -\pi$.

This observation also gives some hint as to why the topological insulator has $\theta = \pi$. These are materials which are defined to be time-reversal invariant. As we have seen, there are two possibilities for the dynamics of such materials. (In fancy language, they are said to have a $\mathbf{Z}_2$ classification.) Most materials are boring and have $\theta = 0$. But some materials have a band structure which is twisted in a particular way. This results in $\theta = \pi$.

1.3 Further Reading

Anyone who has spent even the briefest time looking into the history of physics will have learned one thing: it's complicated. It's vastly more complicated than the air-brushed version we're fed as students. A fairly decent summary is: everyone was confused. Breakthroughs are made by accident, or for the wrong reason, or lie dormant until long after they are rediscovered by someone else. Mis-steps later turn out to be brilliant moves. Ideas held sacred by one generation are viewed as distractions by the next.

Things become even harder when attempting to assign attribution. The scientific literature alone does not tell the full story. It misses the conference coffee conversations, the petty rivalries, the manoeuvering for glory. It misses the fact that, for most of the time, everyone was confused. Gell-Mann, who perhaps did more than anyone to lay the groundwork for particle physics and quantum field theory, captures this in an uncharacteristically rambling manner [47]

> My whole life was like that and I think many people's lives are like that. If you generalised my errors and the things that I got wrong and the things that I didn't follow up properly and the things that I saw and I did not believe in and the things I saw and did not write up. Almost always there was some error of level. That is I knew that a certain thing was right and felt that it was right and it was contradictory to what I was doing and I could not get used to the idea that some things you have to answer late: you just put them off, but you answer some of the things now and...

These lectures are concerned with the theoretical structure of gauge theories. It is a subject whose history is inextricably bound with experimental discoveries in particle physics and the development of the Standard Model. Our understanding of gauge theory took place slowly, over many decades, and involved many hundreds, if not thousands, of physicists.

Each chapter of these lectures ends with a section in which I offer a broad brush account of this history. It is flawed. In places, given the choice between accuracy or a good story, I have erred towards a good story. I have, however, included references to the original literature. More usefully for students, I have also included references to reviews where a number of topics are treated in much greater detail.

Gauge Symmetry

These lectures are about gauge symmetry. Although the use of a gauge choice was commonplace among classical physicists, it was viewed as a trick for finding solutions to the equations of electromagnetism. It took a surprisingly long time for physicists to appreciate the idea of gauge invariance as an important principle in its own right. Fock was the first to realise, in 1926, that the action of gauge symmetry is intricately tied to the phase of the wavefunction in quantum mechanics [62]. The credit for viewing gauge symmetry (or "eichinvarianz") as a desirable property of our theories of Nature is usually attributed to Weyl [203] although, as with many stories in the history of physics, his motivation now seems somewhat misplaced as he tried to prematurely develop a unified theory of gravity and electromagnetism [204]. (His approach survives in the Weyl invariance enjoyed by the worldsheet in string theory.) More historical background on the long road to the gauge principle can be found in [116, 150].

Monopoles

Debrett's style guide for physics papers includes the golden rule: one idea per paper. Many authors flaunt this, but few flaunt it in as spectacular a fashion as Dirac. His 1931 paper "Quantised Singularities in the Electromagnetic Field" [44] is primarily about the possibility of magnetic monopoles obeying the quantisation condition that now bears his name. But the paper starts by reflecting on the negative energy states predicted by the Dirac equation which, he is convinced, cannot be protons as he originally suggested. Instead, he argues, the negative energy states must correspond to novel particles, equal in mass to the electron but with positive charge.

It seems that Dirac held anti-matter and magnetic monopoles, both predictions made within a few pages of each other, on similar footing [55]. He returned to the subject of monopoles only once, in 1948, elaborating on the concept of the "Dirac string" [45]. But the spectacular experimental discovery of anti-matter in 1932, followed by a long, fruitless wait in the search for monopoles, left Dirac disillusioned. Fifty years after his first paper, and seemingly unconvinced by theoretical arguments (like "monopoles are heavy"), he wrote in a letter to Salam [46],

> "I am inclined now to believe that monopoles do not exist. So many years have gone by without any encouragement from the experimental side."

In the intervening years, the experimental situation has not improved. But monopoles now sit at the heart of our understanding of quantum field theory. This story took some decades to unfold and only really came to fruition with the discovery, by 't Hooft [99] and Polyakov [158], of solitons carrying magnetic charge in non-Abelian gauge theories; this will be described in Section 2.8.

As we saw in these lectures, the angular momentum of a particle-monopole pair has an extra anomalous term. This fact was noted long ago by Poincaré [155] in the charmingly titled short story "Remarques sur une expérience de M. Birkeland". In 1936, the Indian physicist Meghnad Saha showed that the quantum version of this observation provides a re-derivation of the Dirac quantisation [171]. The paper is an ambitious, but flawed, attempt to explain the mass of the neutron in terms of a monopole-anti-monopole bound state, and the argument for which it is now remembered is dealt with in a couple of brief sentences. The angular momentum derivation was later rediscovered by H. Wilson [213], prompting an "I did it first" response from Saha [172]. The implication for the spin-statistics of monopoles was pointed out in [92] and [112]; a more modern take can be found in [137].

The extension of the Dirac quantisation condition to dyons was made by Zwanziger in 1968 [235], while the idea of patching gauge fields is due to Wu and Yang [231].

There are many good reviews on magnetic monopoles. More details on the material discussed in this section can be found in the review by Preskill [164] or the book by Shnir [183]. More references are given in the next section when we discuss 't Hooft-Polyakov monopoles.

Topological Insulators

The story of topological insulators started in the study of band structures, and the ways in which they can twist. The first examples are the TKNN invariant for the integer quantum Hall effect [194], and the work of Haldane on Chern insulators [88]. Both of these were described in the lectures on the quantum Hall effect [193]. For this work, Thouless and Haldane were awarded the 2016 Nobel prize.

The possibility that a topologically twisted band structure could exist in 3d materials was realised only in 2006. In July of that year, three groups posted papers on the arXiv [139, 170, 66], and in November of that year, Fu and Kane predicted the existence of this phase in a number of real materials [67]. This was quickly confirmed in experiments [109].

The effective field theory of a topological insulator, in terms of electrodynamics with $\theta = \pi$, was introduced by Qi, Hughes and Zhang in [165]. This took the subject away from its lattice underpinnings, and into the realm of quantum field theory. Indeed, Wilczek had already discussed a number of properties of electrodynamics in the presence of a theta angle [212], including the Witten effect [216]. The existence of the mirror monopole was shown in [166], and a number of further related effects were discussed in [54].

More details of topological insulators can be found in the reviews [91, 167, 17].

2. Yang-Mills Theory

Pure electromagnetism is a free theory of a massless spin 1 field. We can ask: is it possible to construct an interacting theory of spin 1 fields? The answer is yes, and the resulting theory is known as Yang-Mills. The purpose of this section is to introduce this theory and some of its properties.

As we will see, Yang-Mills is an astonishingly rich and subtle theory. It is built upon the mathematical structure of Lie groups. These Lie groups have interesting topology which ensures that, even at the classical (or, perhaps more honestly, semi-classical) level, Yang-Mills exhibits an unusual intricacy. We will describe these features in Sections 2.2 and 2.3 where we introduce the theta angle and instantons.

However, the fun really gets going when we fully embrace $\hbar$ and appreciate that Yang-Mills is a strongly coupled quantum field theory, whose low-energy dynamics looks nothing at all like the classical theory. Our understanding of quantum Yang-Mills is far from complete, but we will describe some of the key ideas from Section 2.4 onwards.

A common theme in physics is that Nature enjoys the rich and subtle: the most beautiful theories tend to be the most relevant. Yang-Mills is no exception. It is the theory that underlies the Standard Model of particle physics, describing both the weak and the strong forces. Much of our focus, and much of the terminology, in this section has its roots in QCD, the theory of the strong force.

For most of this section we will be content to study pure Yang-Mills, without any additional matter. Only in Sections 2.7 and 2.8 will we start to explore how coupling matter fields to the theory changes its dynamics. We'll then continue our study of Yang-Mills coupled to matter in Section 3 where we discuss anomalies, and in Section 5 where we discuss chiral symmetry breaking.

2.1 Introducing Yang-Mills

Yang-Mills theory rests on the idea of a Lie group. The basics of Lie groups and Lie algebras were covered in the Part 3 lectures on *Symmetries and Particle Physics*. We start by introducing our conventions. A compact Lie group G has an underlying Lie algebra $\mathfrak{g}$, whose generators T^a satisfy

$$[T^a, T^b] = i f^{abc} T^c \tag{2.1}$$

Here $a, b, c = 1, \ldots, \dim G$ and f^{abc} are the fully anti-symmetric structure constants. The factor of i on the right-hand side is taken to ensure that the generators are Hermitian: $(T^a)^\dagger = T^a$.

Much of our discussion will hold for general compact, simple Lie group G. Recall that there is a finite classification of these objects. The possible options for the group G, together with the dimension of G and the dimension of the fundamental (or minimal) representation F, are given by

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

where 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 F rather than the rank of the Lie algebra $\mathfrak{g}$.)

Although we will present results for general G, when we want to specialise, or give examples, we will frequently turn to $G = SU(N)$. We will also consider $G = U(1)$, in which case Yang-Mills theory reduces to Maxwell theory.

We will need to normalise our Lie algebra generators. We require that the generators in the fundamental (i.e. minimal) representation F satisfy

$$\operatorname{tr} T^a T^b = \frac{1}{2}\delta^{ab} \tag{2.2}$$

In what follows, we use T^a to refer to the fundamental representation, and will refer to generators in other representations R as $T^a(R)$. Note that, having fixed the normalisation (2.2) in the fundamental representation, other $T^a(R)$ will have different normalisations. We will discuss this in more detail in Section 2.5 where we'll extract some physics from the relevant group theory.

For each generator of the algebra, we introduce a gauge field A^a_μ. These are then packaged into the Lie-algebra valued gauge potential

$$A_\mu = A^a_\mu T^a \tag{2.3}$$

This is a rather abstract object, taking values in a Lie algebra. For $G = SU(N)$, a more down to earth perspective is to view A_μ simply as a traceless $N \times N$ Hermitian matrix.

We will refer to the fields A^a_μ collectively as *gluons*, in deference to the fact that the strong nuclear force is described by $G = SU(3)$ Yang-Mills theory. From the gauge potential, we construct the Lie-algebra valued field strength

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu - i[A_\mu, A_\nu] \tag{2.4}$$

Since this is valued in the Lie algebra, we could also expand it as $F_{\mu\nu} = F^a_{\mu\nu} T^a$. In more mathematical terminology, A_μ is called a *connection* and the field strength $F_{\mu\nu}$ is referred to as the *curvature*. We'll see what exactly the connection connects in Section 2.1.3.

Although we won't look at dynamical matter fields until later in this section, it will prove useful to briefly introduce relevant conventions here. Matter fields live in some representation R of the gauge group G. This means that they sit in some vector ψ of dimension $\dim R$. Much of our focus will be on matter fields in the fundamental representation of $G = SU(N)$, in which case ψ is an N-dimensional complex vector. The matter fields couple to the gauge fields through a *covariant derivative*, defined by

$$\mathcal{D}_\mu \psi = \partial_\mu \psi - i A_\mu \psi \tag{2.5}$$

However, the algebra $\mathfrak{g}$ has many different representations R. For each such representation, we have generators $T^a(R)$ which we can think of as square matrices of dimension $\dim R$. Dressed with all their indices, they take the form

$$T^a(R)^i{}_j \quad i,j = 1,\ldots,\dim R \ \ ; \ a = 1\ldots,\dim G$$

For each of these representations, we can package the gauge fields into a Lie algebra valued object $A^a_\mu T^a(R)^i{}_j$. We can then couple matter in the representation R by generalising the covariant derivative from the fundamental representation to

$$\mathcal{D}_\mu \psi^i = \partial_\mu \psi^i - i A^a_\mu T^a(R)^i{}_j \psi^j \qquad i,j = 1,\ldots,\dim R \tag{2.6}$$

Each of these representations offers a different way of packaging the fields A^a_μ into Lie-algebra valued objects A_μ. As we mentioned above, we will mostly focus on $G = SU(N)$: in this case, we usually take T^a in the fundamental representation, in which case A_μ is simply an $N \times N$ traceless Hermitian matrix.

Aside from the fundamental, there is one other representation that will frequently arise: this is the adjoint, for which $\dim R = \dim G$. We could think of these fields as forming a vector ϕ^a, with $a = 1, \ldots, \dim G$, and then use the form of the covariant derivative (2.6). In fact, it turns out to be more useful to package adjoint valued matter fields into a Lie-algebra valued object, $\phi = \phi^a T^a$. In this language the covariant derivative can be written as

$$\mathcal{D}_\mu \phi = \partial_\mu \phi - i[A_\mu, \phi] \tag{2.7}$$

The field strength can be constructed from the commutator of covariant derivatives. It's not hard to check that

$$[\mathcal{D}_\mu, \mathcal{D}_\nu]\psi = -iF_{\mu\nu}\psi$$

The same kind of calculation shows that if ϕ is in the adjoint representation,

$$[\mathcal{D}_\mu, \mathcal{D}_\nu]\phi = -i[F_{\mu\nu}, \phi]$$

where the right-hand-side is to be thought of as the action of $F_{\mu\nu}$ on fields in the adjoint representation. More generally, we write $[\mathcal{D}_\mu, \mathcal{D}_\nu] = -iF_{\mu\nu}$, with the understanding that the right-hand-side acts on fields according to their representation.

2.1.1 The Action

The dynamics of Yang-Mills is determined by an action principle. We work in natural units, with $\hbar = c = 1$ and take the action

$$S_{\text{YM}} = -\frac{1}{2g^2} \int d^4x \ \text{tr}\, F^{\mu\nu} F_{\mu\nu} \tag{2.8}$$

where g^2 is the Yang-Mills coupling. (It's often called the "coupling constant" but, as we will see in Section 2.4, there is nothing constant about it so I will try to refrain from this language.)

If we compare to the Maxwell action (1.10), we see that there is a factor of $1/2$ outside the action, rather than a factor of $1/4$; this is accounted for by the further factor of $1/2$ that appears in the normalisation of the trace (2.2). There is also the extra factor of $1/g^2$ that we will explain below.

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

$$\mathcal{D}_\mu F^{\mu\nu} = 0 \tag{2.9}$$

where, because $F_{\mu\nu}$ is Lie-algebra valued, the definition (2.7) of the covariant derivative is the appropriate one.

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

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

and noting that this obeys the identity

$$\mathcal{D}_\mu {}^\star F^{\mu\nu} = 0 \tag{2.10}$$

The equations (2.9) and (2.10) are the non-Abelian generalisations of the Maxwell equations. They differ only in commutator terms, both those inside $\mathcal{D}_\mu$ and those inside $F_{\mu\nu}$. Even in the classical theory, this is a big difference as the resulting equations are non-linear. This means that the Yang-Mills fields interact with themselves.

Note that we need to introduce the gauge potentials A_μ in order to write down the Yang-Mills equations of motion. This is in contrast to Maxwell theory where the Maxwell equations can be expressed purely in terms of $\mathbf{E}$ and $\mathbf{B}$ and we introduce gauge fields, at least classically, merely as a device to solve them.

A Rescaling

Usually in quantum field theory, the coupling constants multiply the interaction terms in the Lagrangian; these are terms which are higher order than quadratic, leading to non-linear terms in the equations of motion.

However, in the Yang-Mills action, all terms appear with fixed coefficients determined by the definition of the field strength (2.4). Instead, we've chosen to write the (inverse) coupling as multiplying the entire action. This difference can be accounted for by a trivial rescaling. We define

$$\tilde{A}_\mu = \frac{1}{g}A_\mu \quad \text{and} \quad \tilde{F}_{\mu\nu} = \partial_\mu \tilde{A}_\nu - \partial_\nu \tilde{A}_\mu - ig[\tilde{A}_\mu, \tilde{A}_\nu]$$

Then, in terms of this rescaled field, the Yang-Mills action is

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

In the second version of the action, the coupling constant is buried inside the definition of the field strength, where it multiplies the non-linear terms in the equation of motion as expected.

In what follows, we will use the normalisation (2.8). This is the more useful choice in the quantum theory, where $S_{\rm YM}$ sits exponentiated in the partition function. One way to see this is to note that g^2 sits in the same place as $\hbar$ in the partition function. This already suggests that $g^2 \to 0$ will be a classical limit. Heuristically you should think that, for g^2 small, one pays 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.

Based on this, you might think that we can just set g^2 to be small and a classical analysis of the equations of motion (2.9) and (2.10) will be a good starting point to understand the quantum theory. As we will see in Section 2.4, it turns out that this is not an option; instead, the theory is much more subtle and interesting.

2.1.2 Gauge Symmetry

The action (2.8) has a very large symmetry group. These come from spacetime-dependent functions of the Lie group G,

$$\Omega(x) \in G$$

The set of all such transformations is known as the *gauge group*. Sometimes we will 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

$$A_\mu \to \Omega(x) A_\mu \Omega^{-1}(x) + i\Omega(x)\partial_\mu \Omega^{-1}(x) \tag{2.11}$$

A short calculation shows that this induces the action on the field strength

$$F_{\mu\nu} \to \Omega(x)\, F_{\mu\nu}\, \Omega^{-1}(x) \tag{2.12}$$

The Yang-Mills action is then invariant by virtues of the trace in (2.8).

In the case that $G = U(1)$, the transformations above reduce to the familiar gauge transformations of electromagnetism. In this case we can write $\Omega = e^{i\omega}$ and the transformation of the gauge field becomes $A_\mu \to A_\mu + \partial_\mu \omega$.

Gauge symmetry is poorly named. It is not a symmetry of the system in the sense that it takes one physical state to a different physical state. Instead, it is a redundancy in our description of the system. This is familiar from electromagnetism and remains true in Yang-Mills theory.

There are a number of ways to see why we should interpret the gauge symmetry as a redundancy of the system. Roughly speaking, all of them boil down to the statement that the theory fails to make sense unless we identify states related by gauge transformations. This can be seen classically where the equations of motion (2.9) and (2.10) do not uniquely specify the evolution of A_μ, but only its equivalence class subject to the identification (2.11). In the quantum theory, the gauge symmetry is needed to remove various pathologies which arise, such as the presence of negative norm states in the Hilbert space. A more precise explanation for the redundancy comes from appreciating that Yang-Mills theory is a constrained system which should be analysed as such using the technology of Dirac brackets; we will not do this here.

Our best theories of Nature are electromagnetism, Yang-Mills and general relativity. Each is based on an underlying gauge symmetry. Indeed, the idea of gauge symmetry is clearly something deep. Yet it is, at heart, nothing more than an ambiguity in the language we chose to present the physics? Why should Nature revel in such ambiguity?

There are two reasons why it's advantageous to describe Nature in terms of a redundant set of variables. First, although gauge symmetry means that our presentation of the physics is redundant, it appears to be by far the most concise presentation. For example, we will shortly describe the gauge invariant observables of Yang-Mills theory; they are called "Wilson lines" and can be derived from the gauge potentials A_μ. Yet presenting a configuration of the Yang-Mills field in terms of a complete set of Wilson lines would require vastly more information than specifying the four matrix-valued fields A_μ.

The second reason is that the redundant gauge field allow us to describe the dynamics of the theory in a way that makes manifest various properties of the theory that we hold dear, such as Lorentz invariance and locality and, in the quantum theory, unitarity. This is true even in Maxwell theory: the photon has two polarisation states. Yet try writing down a field which describes the photon that has only two indices and which transforms nicely under the $SO(3,1)$ Lorentz group; its not possible. Instead we introduce a field with four indices – A_μ – and then use the gauge symmetry to kill two of the resulting states. The same kind of arguments also apply to the Yang-Mills field, where there are now two physical degrees of freedom associated to each generator T^a.

The redundancy inherent in the gauge symmetry means that only gauge independent quantities should be considered physical. These are the things that do not depend on our underlying choice of description. In general relativity, we would call such objects "coordinate independent", and it's not a bad metaphor to have in mind for Yang-Mills.

It's worth pointing out that in Yang-Mills theory, the "electric field" $E_i = F_{0i}$ and the "magnetic field" $B_i = -\frac{1}{2}\epsilon_{ijk}F_{jk}$ are not gauge invariant as they transform as (2.12). This, of course, is in contrast to electromagnetism where electric and magnetic fields are physical objects. Instead, if we want to construct gauge invariant quantities we should work with traces such as $\mathrm{tr}\, F_{\mu\nu}F_{\rho\sigma}$ or the Wilson lines that we will describe below. (Note that, for simple gauge groups such as $SU(N)$, the trace of a single field strength vanishes: $\mathrm{tr}\, F_{\mu\nu} = 0$.)

Before we proceed, it's useful to think about infinitesimal gauge transformations. To leading order, gauge transformations which are everywhere close to the identity can be written as

$$\Omega(x) \approx 1 + i\omega^a(x)T^a + \ldots$$

The infinitesimal change of the gauge field from (2.11) becomes

$$\delta A_\mu = \partial_\mu \omega - i[A_\mu, \omega] \equiv \mathcal{D}_\mu \omega$$

where $\omega = \omega^a T^a$. Similarly, the infinitesimal change of the field strength is

$$\delta F_{\mu\nu} = i[\omega, F_{\mu\nu}]$$

Importantly, however, there are classes of gauge transformations which cannot be deformed so that they are everywhere close to the identity. We will study these in Section 2.2.

2.1.3 Wilson Lines and Wilson Loops

It is a maxim in physics, one that leads to much rapture, that "gravity is geometry". But the same is equally true of all the forces of Nature since gauge theory is rooted in geometry. In the language of mathematics, gauge theory is an example of a fibre bundle, and the gauge field A_μ is referred to as a *connection*.

We met the idea of connections in general relativity. There, the Levi-Civita connection $\Gamma^\rho_{\mu\nu}$ tells us how to parallel transport vectors around a manifold. The Yang-Mills connection A_μ plays the same role, but now for the appropriate "electric charge". First we need to explain what this appropriate charge is.

Throughout this section, we will consider a fixed background Yang-Mills fields $A_\mu(x)$. In this background, we place a test particle. The test particle is going to be under our control: we're holding it and we get to choose how it moves and where it goes. But the test particle will carry an internal degree of freedom – this is the "electric charge" – and the evolution of this internal degree of freedom is determined by the background Yang-Mills field.

This internal degree of freedom sits in some representation R of the Lie group G. To start with, we will think of the particle as carrying a complex vector, w, of fixed length,

$$w_i \quad i = 1, \ldots \dim R \qquad \text{such that } w^\dagger w = \text{constant}$$

In analogy with QCD, we will refer to the electrically charged particles as *quarks*, and to w_i as the *colour* degree of freedom. The w_i is sometimes called *chromoelectric charge*.

As the particle moves around the manifold, the connection A_μ (or, to dress it with all its indices, $(A_\mu)^i_{\ j} = A^a_\mu (T^a)^i_{\ j}$) tells this vector w how to rotate. In Maxwell theory, this "parallel transport" is nothing more than the Aharonov-Bohm effect that we discussed in Section 1.1. Upon being transported around a closed loop C, a particle returns with a phase given by $\exp\left(i \oint_C A\right)$. We'd like to write down the generalisation of this formula for non-Abelian gauge theory. For a particle moving with worldline $x^\mu(\tau)$, the rotation of the internal vector w is governed by the parallel transport equation

$$i\frac{dw}{d\tau} = \frac{dx^\mu}{d\tau} A_\mu(x) w \tag{2.13}$$

The factor of i ensues that, with A_μ Hermitian, the length of the vector $w^\dagger w$ remains constant. Suppose that the particle moves along a curve C, starting at $x^\mu_i = x^\mu(\tau_i)$ and finishing at $x^\mu_f = x^\mu(\tau_f)$. Then the rotation of the vector depends on both the starting and end points, as well as the path between them,

$$w(\tau_f) = U[x_i, x_f; C] w(\tau_i)$$

where

$$U[x_i, x_f; C] = \mathcal{P} \exp\left(i \int_{\tau_i}^{\tau_f} d\tau \ \frac{dx^\mu}{d\tau} A_\mu(x(\tau)) \right) = \mathcal{P} \exp\left(i \int_{x_i}^{x_f} A \right) \tag{2.14}$$

where $\mathcal{P}$ stands for *path ordering*. It means that when expanding the exponential, we order the matrices $A_\mu(x(\tau))$ so that those at earlier times are placed to the left. (We met this notation previously in the lectures on quantum field theory when discussing Dyson's formula and you can find more explanation there.) The object $U[x_i, x_f; C]$ is referred to as the *Wilson line*. Under a gauge transformation $\Omega(x)$, it changes as

$$U[x_i, x_f; C] \to \Omega(x_i) U[x_i, x_f; C] \, \Omega^\dagger(x_f)$$

If we take the particle on a closed path C, this object tells us how the vector w differs from its starting value. In mathematics, this notion is called *holonomy*. In this case, we can form a gauge invariant object known as the *Wilson loop*,

$$W[C] = \mathrm{tr}\, \mathcal{P} \exp\left(i \oint_C A \right) \tag{2.15}$$

The Wilson loop $W[C]$ depends on the representation R of the gauge field, and its value along the path C. This will play an important role in Section 2.5 when we describe ways to test for confinement.

Quantising the Colour Degree of Freedom

Above we viewed the colour degree of freedom as a vector w. This is a very classical perspective. It is better to think of each quark as carrying a finite dimensional Hilbert space $\mathcal{H}_{\rm quark}$, of dimension $\dim \mathcal{H}_{\rm quark} = \dim R$.

Here we will explain how to accomplish this. This will provide yet another perspective on the Wilson loop. What follows also offers an opportunity to explain a basic aspect of quantum mechanics which is often overlooked when we first meet the subject. The question is the following: what classical system gives rise to a finite dimensional quantum Hilbert space? Even the simplest classical systems that we meet as undergraduates, such as the harmonic oscillator, give rise to an infinite dimensional Hilbert space. Instead, the much simpler finite dimensional systems, such as the spin of the electron, are typically introduced as having no classical analog. Here we'll see that there is an underlying classical system and that it's rather simple.

We'll stick with a $G = SU(N)$ gauge theory. We consider a single test particle and attach to it a complex vector w, but this time we will insist that w has dimension N. We will restrict its length to be

$$w^\dagger w = \kappa \tag{2.16}$$

The action which reproduces the equation of motion (2.13) is

$$S_w = \int d\tau \; i w^\dagger \frac{dw}{d\tau} + \lambda(w^\dagger w - \kappa) + w^\dagger A(x(\tau)) w \tag{2.17}$$

where λ is a Lagrange multiplier to impose the constraint (2.16), and where $A = A_\mu \, dx^\mu/d\tau$ is to be thought of as a fixed background gauge field $A_\mu(x)$ which varies in time in some fixed way as the particle moves along the path $x^\mu(\tau)$.

Perhaps surprisingly, the action (2.17) has a $U(1)$ worldline gauge symmetry. This acts as

$$w \to e^{i\alpha} w \quad \text{and} \quad \lambda \to \lambda + \dot{\alpha}$$

for any $\alpha(\tau)$. Physically, this gauge symmetry means that we should identify vectors which differ only by a phase: $w_\gamma \sim e^{i\alpha} w_\gamma$. Since we already have the constraint (2.16), this means that the vectors parameterise the projective space $\mathbf{S}^{2N-1}/U(1) \cong \mathbf{CP}^{N-1}$.

Importantly, our action is first order in time derivatives rather than second order. This means that the momentum conjugate to w is $iw^\dagger$ and, correspondingly, $\mathbf{CP}^{N-1}$ is the phase space of the system rather than the configuration space. This, it turns out, is the key to getting a finite dimensional Hilbert space: you should quantise a system with a finite volume phase space. Indeed, this fits nicely with the old-fashioned Bohr-Sommerfeld view of quantisation in which one takes the phase space and assigns a quantum state to each region of extent $\sim \hbar$. A finite volume then gives a finite number of states.

We can see this in a more straightforward way doing canonical quantisation. The unconstrained variables w_i obey the commutation relations

$$[w_i, w_j^\dagger] = \delta_{ij} \tag{2.18}$$

But we recognise these as the commutation relations of creation and annihilation operators. We define a "ground state" $|0\rangle$ such that $w_i|0\rangle = 0$ for all $i = 1, \ldots, N$. A general state in the Hilbert space then takes the form

$$|i_1, \ldots, i_n\rangle = w_{i_1}^\dagger \ldots w_{i_n}^\dagger |0\rangle \tag{2.19}$$

However, we also need to take into account the constraint (2.16). Note that this now arises as the equation of motion for the worldline gauge field λ. As such, it is analogous to Gauss' law when quantising Maxwell theory and we should impose it as a constraint that defines the physical Hilbert space. There is an ordering ambiguity in defining this constraint in the quantum theory: we chose to work with the normal ordered constraint

$$(w_i^\dagger w_i - \kappa)|\text{phys}\rangle = 0$$

This tells us that the physical spectrum of the theory has precisely κ excitations. In this way, we restrict from the infinite dimensional Hilbert space (2.19) to a finite dimensional subspace. However, clearly this restriction only makes sense if we take

$$\kappa \in \mathbf{Z}^+ \tag{2.20}$$

This is interesting. We have an example where a parameter in an action can only take integer values. We will see many further examples as these lectures progress. In the present context, the quantisation of κ means that the $\mathbf{CP}^{N-1}$ phase space of the system has a quantised volume. Again, this sits nicely with the Bohr-Sommerfeld interpretation of dividing the phase space up into parcels.

For each choice of κ, the Hilbert space inherits an action under the $SU(N)$ symmetry. For example:

- $\kappa = 0$: The Hilbert space consists of a single state, $|0\rangle$. This is equivalent to putting a particle in the trivial representation of the gauge group.
- $\kappa = 1$: The Hilbert space consists of N states, $w_i^\dagger|0\rangle$. This describes a particle transforming in the fundamental representation of the $SU(N)$ gauge group.
- $\kappa = 2$: The Hilbert space consists of $\frac{1}{2}N(N+1)$ states, $w_i^\dagger w_j^\dagger|0\rangle$, transforming in the symmetric representation of the gauge group.

In this way, we can build any symmetric representation of $SU(N)$. If we were to treat the degrees of freedom w_i as Grassmann variables, and so replace the commutators in (2.18) with anti-commutators, $\{w_i, w_j^\dagger\} = \delta_{ij}$, then it's easy to convince yourself that we would end up with particles in the anti-symmetric representations of $SU(N)$.

The Path Integral over the Colour Degrees of Freedom

We can also study the quantum mechanical action (2.17) using the path integral. Here we fix the background gauge field A_μ and integrate only over the colour degrees of freedom $w(\tau)$ and the Lagrange multiplier $\lambda(\tau)$.

First, we ask: how can we see the quantisation condition of κ (2.20) in the path integral? There is a rather lovely topological argument for this, one which will be repeated a number of times in subsequent chapters. The first thing to note is that the term $\kappa\lambda$ in the Lagrangian transforms as a total derivative under the gauge symmetry. Naively we might think that we can just ignore this. However, we shouldn't be quite so quick as there are situations where this term is non-vanishing.

Suppose that we think of the worldline of the system, parameterised by $\tau \in \mathbf{S}^1$ rather than $\mathbf{R}$. Then we can consider gauge transformations $\alpha(\tau)$ in which α winds around the circle, so that $\int d\tau\ \dot{\alpha} = 2\pi n$ for some $n \in \mathbf{Z}$. The action (2.17) would then change as

$$S_w \to S_w + 2\pi\kappa n$$

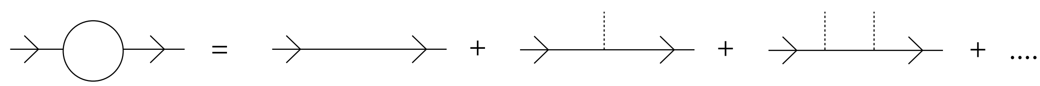
under a gauge transformation which seems bad. However, in the quantum theory it's not the action S_w that we have to worry about but e^{iS_w} because this is what appears in the path integral. And e^{iS_w} is gauge invariant provided that $\kappa \in \mathbf{Z}$.

It is not difficult to explicitly compute the path integral. For convenience, we'll set $\kappa = 1$, so we're looking at objects in the $\mathbf{N}$ representation of $SU(N)$. It's not hard to see that the path integral over λ causes the partition function to vanish unless we put in two insertions of w. We should therefore compute

$$Z_w[A] := \int \mathcal{D}\lambda\mathcal{D}w\mathcal{D}w^\dagger \; e^{iS_w(w,\lambda;A)} w_i(\tau=\infty)w_i^\dagger(\tau=-\infty)$$

The insertion at $\tau = -\infty$ can be thought of as placing the particle in some particular internal state. The partition function measures the amplitude that it remains in that state at $\tau = +\infty$

We next perform the path integral over w and $w^\dagger$. This is tantamount to summing a series of diagrams like this:



where the straight lines are propagators for w_i which are simply $\theta(\tau_1 - \tau_2)\delta_{ij}$, while the dotted lines represent insertions of the gauge fields A. It's straightforward to sum these. The final result is something familiar:

$$Z_w[A] = \operatorname{tr} \mathcal{P} \exp\left(i\int d\tau \; A(\tau)\right) \tag{2.21}$$

This, of course, is the Wilson loop $W[C]$. We see that we get a slightly different perspective on the Wilson loop: it arises by integrating out the colour degrees of freedom of the quark test particle.

2.2 The Theta Term

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

$$S_{\rm YM} = -\frac{1}{2g^2}\int d^4x \; \operatorname{tr} F^{\mu\nu}F_{\mu\nu}$$

There is, however, one further term that we can add which is Lorentz invariant, gauge invariant and quadratic in field strengths. This is the theta term,

$$S_\theta = \frac{\theta}{16\pi^2}\int d^4x \; \operatorname{tr} {}^\star F^{\mu\nu}F_{\mu\nu} \tag{2.22}$$

where ${}^\star F^{\mu\nu} = \frac{1}{2}\epsilon^{\mu\nu\rho\sigma}F_{\rho\sigma}$. Clearly, this is analogous to the theta term that we met in Maxwell theory in Section 1.2. Note, however, that the canonical normalisation of the

Yang-Mills theta term differs by a factor of $\frac{1}{2}$ from the Maxwell term (a fact which is a little hidden in this notation because it's buried in the definition of the trace (2.2)). We'll understand why this is the case below. (A spoiller: it's because the periodicity of the Maxwell theta term arises from the first Chern number, $c_1(A)^2$ while the periodicity of the non-Abelian theta-term arises from the second Chern number $c_2(A)$.)

The non-Abelian theta term shares a number of properties with its Abelian counterpart. In particular,

- The theta term is a total derivative. It can be written as

$$S_\theta = \frac{\theta}{8\pi^2}\int d^4x\ \partial_\mu K^\mu \tag{2.23}$$

  where

$$K^\mu = \epsilon^{\mu\nu\rho\sigma}\mathrm{tr}\left(A_\nu\partial_\rho A_\sigma - \frac{2i}{3}A_\nu A_\rho A_\sigma\right) \tag{2.24}$$

  This means that, as in the Maxwell case, the theta term does not change the classical equations of motion.

- θ is an angular variable. For simple gauge groups, it sits in the range

$$\theta \in [0, 2\pi)$$

  This follows because the total derivative (2.23) counts the winding number of a gauge configuration known as the Pontryagin number such that, evaluated on any configuration, $S_\theta = \theta n$ with $n \in \mathbf{Z}$. This is similar in spirit to the kind of argument we saw in Section 1.2.4 for the $U(1)$ theta angle, although the details differ because non-Abelian gauge groups have a different topology from their Abelian cousins. We will explain this in the rest of this section and, from a slightly different perspective, in Section 2.3.

  There can, however, be subtleties associated to discrete identifications in the gauge group in which case the range of θ should be extended. We'll discuss this in more detail in Section 2.6.

In Section 1.2, we mostly focussed on situations where θ varies in space. This kind of "topological insulator" physics also applies in the non-Abelian case. However, as we mentioned above, the topology of non-Abelian gauge groups is somewhat more complicated. This, it turns out, affects the spectrum of states in the Yang-Mills theory even when θ is constant. The purpose of this section is to explore this physics.

2.2.1 Canonical Quantisation of Yang-Mills

Ultimately, we want to see how the θ term affects the quantisation of Yang-Mills. But we can see the essence of the issue already in the classical theory where, as we will now show, the θ term results in a shift to the canonical momentum. The full Lagrangian is

$$\mathcal{L} = -\frac{1}{2g^2}\mathrm{tr}\, F^{\mu\nu}F_{\mu\nu} + \frac{\theta}{16\pi^2}\mathrm{tr}\, {}^\star F^{\mu\nu}F_{\mu\nu} \tag{2.25}$$

To start, we make use of the gauge redundancy to set

$$A_0 = 0$$

With this ansatz, the Lagrangian becomes

$$\mathcal{L} = \frac{1}{g^2}\mathrm{tr}\left(\dot{\mathbf{A}}^2 - \mathbf{B}^2\right) + \frac{\theta}{4\pi^2}\,\mathrm{tr}\,\dot{\mathbf{A}}\cdot\mathbf{B} \tag{2.26}$$

Here $B_i = -\frac{1}{2}\epsilon_{ijk}F_{jk}$ is the non-Abelian magnetic field (sometimes called the *chromo-magnetic* field). Meanwhile, the non-Abelian electric field is $E_i = \dot{A}_i$. I've chosen not to use the electric field notation in (2.26) as the $\dot{\mathbf{A}}$ terms highlight the canonical structure. Note that the θ term is linear in time derivatives; this is reminiscent of the effect of a magnetic field in Newtonian particle mechanics and we will see some similarities below.

The Lagrangian (2.26) is not quite equivalent to (2.25); it should be supplemented by the equation of motion for A_0. In analogy with electromagnetism, we refer to this as Gauss' law. It is

$$\mathcal{D}_i E_i = 0 \tag{2.27}$$

This is a constraint which should be imposed on all physical field configurations.

The momentum conjugate to $\mathbf{A}$ is

$$\boldsymbol{\pi} = \frac{\partial\mathcal{L}}{\partial\dot{\mathbf{A}}} = \frac{1}{g^2}\mathbf{E} + \frac{\theta}{8\pi^2}\mathbf{B}$$

From this we can build the Hamiltonian

$$\mathcal{H} = \frac{1}{g^2}\mathrm{tr}\left(\mathbf{E}^2 + \mathbf{B}^2\right) \tag{2.28}$$

We see that, when written in terms of the electric field $\mathbf{E}$, neither the constraint (2.27) nor the Hamiltonian (2.28) depend on θ; all of the dependence is buried in the Poisson bracket structure. Indeed, when written in terms of the canonical momentum $\boldsymbol{\pi}$, the constraint becomes

$$\mathcal{D}_i \pi_i = 0$$

where the would-be extra term $\mathcal{D}_i B_i = 0$ by virtue of the Bianchi identity (2.10). Meanwhile the Hamiltonian becomes

$$\mathcal{H} = g^2 \mathrm{tr}\left(\boldsymbol{\pi} - \frac{\theta}{8\pi^2}\mathbf{B}\right)^2 + \frac{1}{g^2}\mathrm{tr}\,\mathbf{B}^2$$

It is this θ-dependent shift in the canonical momentum which affects the quantum theory.

Building the Hilbert Space

Let's first recall how we construct the physical Hilbert space of Maxwell theory where, for now, we set $\theta = 0$. For this Abelian theory, Gauss' law (2.27) is linear in $\mathbf{A}$ and it is equivalent to $\nabla \cdot \mathbf{A} = 0$. This makes it simple to solve: the constraint kills the longitudinal photon mode, leaving us with two, physical transverse modes. We can then proceed to build the Hilbert space describing just these physical degrees of freedom. This was the story we learned in our first course on Quantum Field Theory.

In contrast, things aren't so simple in Yang-Mills theory. Now the Gauss' law (2.27) is non-linear and it's not so straightforward to solve the constraint to isolate only the physical degrees of freedom. Instead, we proceed as follows. We start by constructing an auxiliary Hilbert space built from all spatial gauge fields: we call these states $|\mathbf{A}(\mathbf{x},t)\rangle$. The physical Hilbert space is then defined as those states $|\mathrm{phys}\rangle$ which obey

$$\mathcal{D}_i E_i \, |\mathrm{phys}\rangle = 0 \tag{2.29}$$

Note that we do not set $\mathcal{D}_i E_i = 0$ as an operator equation; this would not be compatible with the commutation relations of the theory. Instead, we use it to define the physical states.

There is an alternative way to think about the constraint (2.29). After we've picked $A_0 = 0$ gauge, we still have further time-independent gauge transformations of the form

$$\mathbf{A} \to \Omega \mathbf{A} \Omega^{-1} + i\Omega \boldsymbol{\nabla} \Omega^{-1}$$

Among these are *global* gauge transformations which, in the limit $\mathbf{x} \to \infty$, asymptote to $\Omega \to$ constant $\neq 1$. These are sometimes referred to as *large* gauge transformations.

They should be thought of as global, physical symmetries rather than redundancies. A similar interpretation holds in Maxwell theory where the corresponding conserved quantity is electric charge. In the present case, we have a conserved charge for each generator of the gauge group. The form of the charge follows from Noether's theorem and, for the gauge transformation $\Omega = e^{i\omega}$, is given by

$$\begin{aligned} Q(\omega) &= \int d^3x\ \mathrm{tr}\left(\boldsymbol{\pi}\cdot\delta\mathbf{A}\right) \\ &= \frac{1}{g^2}\int d^3x\ \mathrm{tr}\left(E_i + \frac{\theta g^2}{8\pi^2}B_i\right)\mathcal{D}_i\omega \\ &= -\frac{1}{g^2}\int d^3x\ \mathrm{tr}\left(\mathcal{D}_iE_i\,\omega\right) \end{aligned} \tag{2.30}$$

where we've used the fact that $\mathcal{D}_iB_i = 0$. This is telling us that the Gauss' law $G^a = (\mathcal{D}_iE_i)^a$ plays the role of the generator of the gauge symmetry. The constraint (2.29) is the statement that we are sitting in the gauge singlet sector of the Hilbert space where, for all ω, $Q(\omega) = 0$.

2.2.2 The Wavefunction and the Chern-Simons Functional

It's rare in quantum field theory that we need to resort to the old-fashioned Schrödinger representation of the wavefunction. But we will find it useful here. We will think of the states in the auxiliary Hilbert space as wavefunctions of the form $\Psi(\mathbf{A})$. (Strictly speaking, these are wavefunctionals because the argument $\mathbf{A}(\mathbf{x})$ is itself a function.)

In this language, the canonical momentum π^i is, as usual in quantum mechanics, $\pi^i = -i\delta/\delta A_i$. The Gauss' law constraint then becomes

$$\mathcal{D}_i\left(-i\frac{\delta\Psi}{\delta A_i}\right) = 0 \tag{2.31}$$

Meanwhile, the Schrödinger equation is

$$\mathcal{H}\Psi = g^2\mathrm{tr}\left(-i\frac{\delta}{\delta\mathbf{A}} - \frac{\theta}{8\pi^2}\mathbf{B}\right)^2\Psi + \frac{1}{g^2}\mathrm{tr}\,\mathbf{B}^2\Psi = E\Psi \tag{2.32}$$

This is now in a form that should be vaguely familiar from our first course in quantum mechanics, albeit with an infinite number of degrees of freedom. All we have to do is solve these equations. That, you may not be surprised to hear, is easier said than done.

We can, however, try to see the effect of the θ term. Suppose that we find a physical, energy eigenstate — call it $\Psi_0(\mathbf{A})$ — that solves both (2.31), as well as the Schrödinger equation (2.32) with $\theta = 0$. That is,

$$-g^2 \mathrm{tr}\, \frac{\delta^2 \Psi_0}{\delta \mathbf{A}} + \frac{1}{g^2} \mathrm{tr}\, \mathbf{B}^2 \Psi_0 = E \Psi_0 \tag{2.33}$$

Now consider the following state

$$\Psi(\mathbf{A}) = e^{i\theta\, W[\mathbf{A}]}\, \Psi_0(\mathbf{A}) \tag{2.34}$$

where $W(\mathbf{A})$ is given by

$$W(\mathbf{A}) = \frac{1}{8\pi^2} \int d^3x\ \epsilon^{ijk}\, \mathrm{tr}\left(F_{ij} A_k + \frac{2i}{3} A_i A_j A_k \right) \tag{2.35}$$

This is known as the *Chern-Simons functional*. It has a number of beautiful and subtle properties, some of which we will see below, some of which we will explore in Section 8. It also plays an important role in the theory of the Quantum Hall Effect. Note that we've already seen the expression (2.35) before: when we wrote the θ term as a total derivative (2.24), the temporal component was $K^0 = 4\pi^2 W$.

For now, the key property of $W(\mathbf{A})$ that we will need is

$$\frac{\delta W(\mathbf{A})}{\delta A_i} = \frac{1}{8\pi^2} \epsilon^{ijk} F_{jk} = \frac{1}{4\pi^2} B_i$$

which gives us the following relation,

$$-i \frac{\delta \Psi(\mathbf{A})}{\delta A_i} = -i e^{i\theta W[\mathbf{A}]}\, \frac{\delta \Psi_0(\mathbf{A})}{\delta A_i} + \frac{\theta}{4\pi^2} B_i\, \Psi(\mathbf{A})$$

This ensures that Ψ satisfies the Gauss law constraint (2.31). (To see this, you need to convince yourself that the $\mathcal{D}_i$ in (2.31) acts only on $\delta\Psi_0/\delta A_i$ in the first term above and on B_i in the second and then remember that $\mathcal{D}_i B_i = 0$ by the Bianchi identity.) Moreover, if Ψ_0 obeys the Schrödinger equation (2.33), then Ψ will obey the Schrödinger equation (2.32) with general θ.

The above would seem to show that if we can construct a physical state Ψ_0 with energy E when $\theta = 0$ then we can dress this with the Chern-Simons functional $e^{i\theta W(\mathbf{A})}$ to construct a state Ψ which has the same energy E when $\theta \neq 0$. In other words, the physical spectrum of the theory appears to be independent of θ. In fact, this conclusion is wrong! The spectrum *does* depend on θ. To understand the reason behind this, we have to look more closely at the Chern-Simons functional (2.35).

Is the Chern-Simons Functional Gauge Invariant?

The Chern-Simons functional $W[\mathbf{A}]$ is not obviously gauge invariant. In fact, not only is it not obviously gauge invariant, it turns out that it's not actually gauge invariant! But, as we now explain, it fails to be gauge invariant in an interesting way.

Let's see what happens. In $A_0 = 0$ gauge, we can still act with time-independent gauge transformations $\Omega(\mathbf{x}) \in G$, under which

$$\mathbf{A} \rightarrow \Omega \mathbf{A} \Omega^{-1} + i\Omega \boldsymbol{\nabla} \Omega^{-1}$$

The spatial components of the field strength then changes as $F_{ij} \rightarrow \Omega F_{ij} \Omega^{-1}$. It is not difficult to check that the Chern-Simons functional (2.35) transforms as

$$W[\mathbf{A}] \rightarrow W[\mathbf{A}] + \frac{1}{4\pi^2} \int d^3x \, \left\{ i\epsilon^{ijk}\partial_j \text{tr}\,(\partial_i \Omega\, \Omega^{-1} A_k) - \frac{1}{3}\epsilon^{ijk} \text{tr}\,(\Omega^{-1}\partial_i \Omega\, \Omega^{-1}\partial_j \Omega\, \Omega^{-1}\partial_k \Omega) \right\}$$

The first term is a total derivative. It has an interesting role to play on manifolds with boundaries but will not concern us here. Instead, our interest lies in the second term. This is novel to non-Abelian gauge theories and has a beautiful interpretation.

To understand this interpretation, we need to understand something about the topology of non-Abelian gauge transformations. As we now explain, these gauge transformations fall into different classes.

We've already met the first classification of gauge transformations. Those with $\Omega \neq 1$ at spatial infinity, $\mathbf{S}^2_\infty \cong \partial \mathbf{R}^3$, are to be thought of as global symmetries. The remaining gauge symmetries have $\Omega = 1$ on $\mathbf{S}^2_\infty$. These are the ones that we are interested in here.

Insisting that $\Omega \rightarrow 1$ at $\mathbf{S}^2_\infty$ is equivalent to working on spatial $\mathbf{S}^3$ rather than $\mathbf{R}^3$. Each gauge transformation with this property then defines a map,

$$\Omega(\mathbf{x}) : \mathbf{S}^3 \mapsto G$$

Such maps fall into disjoint classes. This arises 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. We'll meet this kind of idea a lot throughout these lectures. Such maps are characterised by homotopy theory. In general, we will be interested in the different classes of maps from spheres $\mathbf{S}^n$ into some space X. Two maps are said to be homotopic if they can be continuously deformed into each other.

The homotopically distinct maps are classified by the group $\Pi_n(X)$. For us, the relevant formula is

$$\Pi_3(G) = \mathbf{Z}$$

for all simple, compact Lie groups G. In words, this means that the winding of gauge transformations is classified by an integer n. This statement is intuitive for $G = SU(2)$ since $SU(2) \cong \mathbf{S}^3$, so the homotopy group counts the winding of maps from $\mathbf{S}^3 \mapsto \mathbf{S}^3$. For higher dimensional G, it turns out that it's sufficient to pick an $SU(2)$ subgroup of G and consider maps which wind within that. It turns out that these maps cannot be unwound within the larger G. Moreover, all topologically non-trivial maps within G can be deformed to lie within an $SU(2)$ subgroup. It can be shown that this winding is computed by,

$$n(\Omega) = \frac{1}{24\pi^2} \int_{\mathbf{S}^3} d^3S \; \epsilon^{ijk} \mathrm{tr} \left(\Omega^{-1}\partial_i \Omega \, \Omega^{-1}\partial_j \Omega \, \Omega^{-1}\partial_k \Omega\right) \tag{2.36}$$

We claim that this expression always spits out an integer $n(\Omega) \in \mathbf{Z}$. This integer characterises the gauge transformation. It's simple to check that $n(\Omega_1\Omega_2) = n(\Omega_1) + n(\Omega_2)$.

An Example: $SU(2)$

We won't prove that the expression (2.36) is an integer which counts the winding. We will, however, give a simple example which illustrates the basic idea. We pick gauge group $G = SU(2)$. This is particularly straightforward because, as a manifold, $SU(2) \cong \mathbf{S}^3$ and it seems eminently plausible that $\Pi_3(\mathbf{S}^3) \cong \mathbf{Z}$.

In this case, it is not difficult to give an explicit mapping which has winding number n. Consider the radially symmetric gauge transformation

$$\Omega_n(\mathbf{x}) = \exp\left(if(r)\sigma_i \hat{x}^i\right) = \cos(f) + i\sin(f)\,\sigma^i \cdot \hat{x}^i \tag{2.37}$$

where $\omega(r)$ is some monotonic function such that

$$f(r) = \begin{cases} 0 & r = 0 \\ 2\pi n & r = \infty \end{cases}$$

Evaluating the integrand in (2.36), we have

$$\frac{1}{24\pi^2}\epsilon^{ijk}\mathrm{tr}\left(\Omega_n^{-1}\partial_i\Omega_n \, \Omega_n^{-1}\partial_j\Omega_n \, \Omega_n^{-1}\partial_k\Omega_n\right) = \frac{1}{2\pi^2}\frac{df}{dr}\frac{\sin^2 f}{r^2}$$

and so the winding is

$$n(\Omega_n) = \frac{2}{\pi}\int_0^\infty dr\ f'(r)\sin^2 f = \frac{1}{\pi}\left[f - \frac{1}{2}\sin(2f)\right]_0^\infty = n$$

For more general non-Abelian gauge groups G, one can always embed the winding $\Omega_n(\mathbf{x})$ into an $SU(2)$ subgroup. It turns out that it is not possible to unwind this by moving in the larger G. Moreover, the converse also holds: given any non-trivial winding $\Omega(\mathbf{x})$ in G, one can always deform $\Omega(\mathbf{x})$ until it sits entirely within an $SU(2)$ subgroup.

The Chern-Simons Functional is not Gauge Invariant!

We now see the relevance of these topologically non-trivial gauge transformations. Dropping the boundary term, the transformation of the Chern-Simons functional is

$$W[\mathbf{A}] \to W[\mathbf{A}] + n$$

We learn that the Chern-Simons functional is not quite gauge invariant. But it only changes under topologically non-trivial gauge transformations, where it shifts by an integer.

What does this mean for our wavefunctions? We will require that our wavefunctions are gauge invariant, so that $\Psi(\mathbf{A}') = \Psi(\mathbf{A})$ with $\mathbf{A}' = \Omega\mathbf{A}\Omega^{-1} + i\Omega\boldsymbol{\nabla}\Omega^{-1}$. Now, however, we see the problem with our dressing argument. Suppose that we find a wavefunction $\Psi_0(\mathbf{A})$ which is a state when $\theta = 0$ and is gauge invariant. Then the dressed wavefunction

$$\Psi(\mathbf{A}) = e^{i\theta W[\mathbf{A}]}\,\Psi_0(\mathbf{A}) \tag{2.38}$$

will indeed solve the Schrödinger equation for general θ. But it is not gauge invariant: instead it transforms as $\Psi(\mathbf{A}') = e^{i\theta n}\Psi(\mathbf{A})$.

This then, is the way that the θ angle shows up in the states. We do require that $\Psi(\mathbf{A})$ is gauge invariant which means that it's not enough to simply dress the $\theta = 0$ wavefunctions $\Psi_0(\mathbf{A})$ with the Chern-Simons functional $e^{i\theta W[\mathbf{A}]}$. Instead, if we want to go down this path, we must solve the $\theta = 0$ Schrödinger equation with the requirement that $\Psi_0(\mathbf{A}') = e^{-i\theta n}\Psi_0(\mathbf{A})$, so that this cancels the additional phase coming from the dressing factor so that $\Psi(\mathbf{A})$ is gauge invariant.

There is one last point: the value of θ only arises in the phase $e^{i\theta n}$ with $n \in \mathbf{Z}$. This, is the origin of the statement that θ is periodic mod 2π. We take $\theta \in [0, 2\pi)$.

We have understood that the spectrum does depend on θ. But we have not understood *how* the spectrum depends on θ. That is much harder. We will not have anything to say here, but will return to this a number of times in these lectures, both in Section 2.3 where we discuss instantons and in Section 6 when we discuss the large N expansion.

2.2.3 Analogies From Quantum Mechanics

There's an analogy that exhibits some (but not all) of the ideas above in a much simpler setting. Consider a particle of unit charge, restricted to move on a circle of radius R. Through the middle of the circle we thread a magnetic flux Φ. Because the particle sits away from the magnetic field, its classical motion is unaffected by the flux. Nonetheless, the quantum spectrum does depend on the flux and this arises for reasons very similar to those described above.

Let's recall how this works. The Hamiltonian for the particle is

$$H = \frac{1}{2m}\left(-i\frac{\partial}{\partial x} + \frac{\Phi}{2\pi R}\right)^2$$

We can now follow our previous train of logic. Suppose that we found a state Ψ_0 which is an eigenstate of the Hamiltonian when $\Phi = 0$. We might think that we could then just write down the new state $\Psi = e^{-i\Phi x/2\pi R}\Psi_0$ which is an eigenstate of the Hamiltonian for non-zero Φ. However, as in the Yang-Mills case above, this is too quick. For our particle on a circle, it's not large gauge transformations that we have to worry about; instead, it's simply the requirement that the wavefunction is single valued. The dressing factor $e^{i\Phi x/2\pi R}$ is only single valued if Φ is a multiple of 2π.

Of course, the particle moving on a circle is much simpler than Yang-Mills. Indeed, there is no difficulty in just solving it explicitly. The single-valued wavefunctions have the property that they are actually independent of Φ. (There is no reason to believe that this property also holds for Yang-Mills.) They are

$$\Psi = \frac{1}{\sqrt{2\pi R}}\, e^{inx/R} \qquad n \in \mathbf{Z}$$

These solve the Schrödinger equation $H\Psi = E\Psi$ with energy

$$E = \frac{1}{2mR^2}\left(n + \frac{\Phi}{2\pi}\right)^2 \qquad n \in \mathbf{Z} \tag{2.39}$$

We see that the spectrum of the theory does depend on the flux Φ, even though the particle never goes near the region with magnetic field. Moreover, as far as the particle is concerned, the flux Φ is a periodic variable, with periodicity 2π. In particular, if Φ is an integer multiple of 2π, then the spectrum of the theory is unaffected by the flux.

The Theta Angle as a "Hidden" Parameter

There is an alternative way to view the problem of the particle moving on a circle. We explain this here before returning to Yang-Mills where we offer the same viewpoint. This new way of looking at things starts with a question: why should we insist that the wavefunction is single-valued? After all, we only measure probability $|\Psi|^2$, which cares nothing for the phase. Does this mean that it's consistent to work with wavefunctions that are not single-valued around the circle?

The answer to this question is "yes". Let's see how it works. Consider the Hamiltonian for a free particle on a circle of radius R,

$$H = -\frac{1}{2m}\frac{\partial^2}{\partial x^2} \tag{2.40}$$

In this way of looking at things, the Hamiltonian contains no trace of the flux. Instead, it will arise from the boundary conditions that we place on the wavefunction. We will not require that the wavefunction is single valued, but instead that it comes back to itself up to some specified phase Φ, so that

$$\Psi(x + 2\pi R) = e^{i\Phi}\,\Psi(x)$$

The eigenstates of (2.40) with this requirement are

$$\Psi = \frac{1}{\sqrt{2\pi R}}\,e^{i(n+\Phi/2\pi)x/R} \quad n \in \mathbf{Z}$$

The energy of these states is again given by (2.39). We learn that allowing for more general wavefunctions doesn't give any new physics. Instead, it allows for a different perspective on the same physics, in which the presence of the flux does not appear in the Hamiltonian, but instead is shifted to the boundary conditions imposed on the wavefunction. In this framework, the phase Φ is sometimes said to be a "hidden" parameter because you don't see it directly in the Hamiltonian.

We can now ask this same question for Yang-Mills. We'll start with Yang-Mills theory in the absence of a θ term and will see how we can recover the states with $\theta \neq 0$. Here, the analog question is whether the wavefunction $\Psi_0(\mathbf{A})$ should really be gauge invariant, or whether we can suffer an additional phase under a gauge transformation. The phase that the wavefunction picks up should be consistent with the group structure of gauge transformations: this means that we are looking for a one-dimensional representation (the phase) of the group of gauge transformations.

Topologically trivial gauge transformations (which have $n(\Omega) = 0$) can be continuously connected to the identity. For these, there's no way to build a non-trivial phase factor consistent with the group structure: it must be the case that $\Psi_0(\mathbf{A}') = \Psi_0(\mathbf{A})$ whenever $\mathbf{A}' = \Omega\mathbf{A}\Omega^{-1} + i\Omega\boldsymbol{\nabla}\Omega^{-1}$ with $n(\Omega) = 0$.

However, things are different for the topologically non-trivial gauge transformations. As we've seen above, these are labelled by their winding $n(\Omega) \in \mathbf{Z}$. One could require that, under these topologically non-trivial gauge transformations, the wavefunction changes as

$$\Psi_0(\mathbf{A}') = e^{-i\theta n}\, \Psi_0(\mathbf{A}) \tag{2.41}$$

for some choice of $\theta \in [0, 2\pi)$. This is consistent with consecutive gauge transformations because $n(\Omega_1\Omega_2) = n(\Omega_1) + n(\Omega_2)$. In this way, we introduce an angle θ into the definition of the theory through the boundary conditions on wavefunctions.

It should be clear that the discussion above is just another way of stating our earlier results. Given a wavefunction which transforms as (2.41), we can always dress it with a Chern-Simons functional as in (2.38) to construct a single-valued wavefunction. These are just two different paths that lead to the same conclusion. We've highlighted the "hidden" interpretation here in part because it is often the way the θ angle is introduced in the literature. Moreover, as we will see in more detail in Section 2.3, it is closer in spirit to the way the θ angle appears in semi-classical tunnelling calculations.

Another Analogy: Bloch Waves

There's another analogy which is often wheeled out to explain how θ affects the states. This analogy has some utility, but it also has some flaws. I'll try to highlight both below.

So far our discussion of the θ angle has been for all states in the Hilbert space. For this analogy, we will focus on the ground state. Moreover, we will work "semi-classically", which really means "classically" but where we use the language of wavefunctions. I should stress that this approximation is *not* valid: as we will see in Section 2.4, Yang-Mills theory is strongly coupled quantum theory, and the true ground state will bear no resemblance to the classical ground state. The purpose of what follows is merely to highlight the basic structure of the Hilbert space.

With these caveats out the way, let's proceed. The classical ground states of Yang-Mills are pure gauge configurations. This means that they take the form

$$\mathbf{A} = iV\boldsymbol{\nabla}V^{-1} \tag{2.42}$$

for some $V(\mathbf{x}) \in G$. But, as we've seen above, such configurations are labelled by the integer $n(V)$. This is a slightly different role for the winding: now it is labelling the zero energy states in the theory, as opposed to gauge transformations. At the semi-classical level, the configurations (2.42) map into quantum states. Since the classical configurations are labelled by an integer $n(V)$, this should carry over to the quantum Hilbert space. We call the corresponding ground states $|n\rangle$ with $n \in \mathbf{Z}$.

If we were to stop here, we might be tempted to conclude that Yang-Mills has multiple ground states, $|n\rangle$. But this would be too hasty. All of these ground states are connected by gauge transformations. But the gauge transformations themselves must have non-trivial topology. Specifically, if Ω is a gauge transformation with $n(\Omega) = n'$ then $\Omega|n\rangle = |n + n'\rangle$.

The true ground state, like all states in the Hilbert space, should obey (2.41). For our states, this reads

$$\Omega|\Psi\rangle = e^{i\theta n'}|\Psi\rangle$$

This means that the physical ground state of the system is a coherent sum over all the states $|n\rangle$. It takes the form

$$|\theta\rangle = \sum_n e^{i\theta n}|n\rangle \tag{2.43}$$

This is the semi-classical approximation to the ground state of Yang-Mills theory. These states are sometimes referred to as *theta vacua*. Once again, I stress that the semi-classical approximation is a rubbish approximation in this case! This is not close to the true ground state of Yang-Mills.

Now to the analogy, which comes from condensed matter physics. Consider a particle moving in a one-dimensional periodic potential

$$V(x) = V(x + a)$$

Classically there are an infinite number of ground states corresponding to the minima of the potential. We describe these states as $|n\rangle$ with $n \in \mathbf{Z}$. However, we know that these aren't the true ground states of the Hamiltonian. These are given by Bloch's theorem which states that all eigenstates have the form

$$|k\rangle = \sum_n e^{ikan}|n\rangle \tag{2.44}$$

for some $k \in [-\pi/a, \pi/a)$ called the lattice momentum. Clearly there is a parallel between (2.43) and (2.44). In some sense, the θ angle plays a role in Yang-Mills similar to the combination ka for a particle in a periodic potential. This similarity can be traced to the underlying group theory structure. In both cases there is a $\mathbf{Z}$ group action on the states. For the particle in a lattice, this group is generated by the translation operator; for Yang-Mills it is generated by the topologically non-trivial gauge transformation with $n(\Omega) = 1$.

There is, however, an important difference between these two situations. For the particle in a potential, all the states $|k\rangle$ lie in the Hilbert space. Indeed, the spectrum famously forms a band labelled by k. In contrast, in Yang-Mills theory there is only a single state: each theory has a specific θ which picks out one state from the band. This can be traced to the different interpretation of the group generators. The translation operator for a particle is a genuine symmetry, moving one physical state to another. In contrast, the topologically non-trivial gauge transformation Ω is, like all gauge transformations, a redundancy: it relates physically identical states, albeit up to a phase.

2.3 Instantons

We have argued that the theta angle is an important parameter in Yang-Mills, changing the spectrum and correlation functions of the theory. This is in contrast to electromagnetism where θ only plays a role in the presence of boundaries (such as topological insulators) or magnetic monopoles. It is natural to ask: how do we see this from the path integral?

To answer this question, recall that the theta term is a total derivative

$$S_\theta = \frac{\theta}{16\pi^2} \int d^4x \; \mathrm{tr}\, {}^\star F^{\mu\nu} F_{\mu\nu} = \frac{\theta}{8\pi^2} \int d^4x \; \partial_\mu K^\mu$$

where

$$K^\mu = \epsilon^{\mu\nu\rho\sigma} \mathrm{tr} \left(A_\nu \partial_\rho A_\sigma - \frac{2i}{3} A_\nu A_\rho A_\sigma \right)$$

This means that if a field configuration is to have a non-vanishing value of S_θ, then it must have something interesting going on at infinity.

At this point, we do something important: we Wick rotate so that we work in Euclidean spacetime $\mathbf{R}^4$. We will explain the physical significance of this in Section 2.3.2. Configurations that have finite action S_{YM} must asymptote to pure gauge,

$$A_\mu \to i\Omega \partial_\mu \Omega^{-1} \quad \text{as } x \to \infty \tag{2.45}$$

with $\Omega \in G$. This means that finite action, Euclidean field configurations involve a map

$$\Omega(x) : \mathbf{S}^3_\infty \mapsto G$$

But we have met such maps before: they are characterised by the homotopy group $\Pi_3(G) = \mathbf{Z}$. Plugging this asymptotic ansatz (2.45) into the action S_θ, we have

$$S_\theta = \theta\nu \tag{2.46}$$

where $\nu \in \mathbf{Z}$ is an integer that tells us the number of times that $\Omega(x)$ winds around the asymptotic $\mathbf{S}^3_\infty$,

$$\nu(\Omega) = \frac{1}{24\pi^2} \int_{\mathbf{S}^3_\infty} d^3S\ \epsilon^{ijk} \mathrm{tr}\,(\Omega\partial_i\Omega^{-1})(\Omega\partial_j\Omega^{-1})(\Omega\partial_k\Omega^{-1}) \tag{2.47}$$

This is the same winding number that we met previously in (2.36).

This discussion is mathematically identical to the classification of non-trivial gauge transformations in Section 2.2.2. However, the physical setting is somewhat different. Here we are talking about maps from the boundary of (Euclidean) spacetime $\mathbf{S}^3_\infty$, while in Section 2.2.2 we were talking about maps from a spatial slice, $\mathbf{R}^3$, suitably compactified to become $\mathbf{S}^3$. We will see the relationship between these in Section 2.3.2.

2.3.1 The Self-Dual Yang-Mills Equations

Among the class of field configurations with non-vanishing winding ν there are some that are special: these solve the classical equations of motion,

$$\mathcal{D}_\mu F^{\mu\nu} = 0 \tag{2.48}$$

There is a cute way of finding solutions to this equation. The Yang-Mills action is

$$S_{YM} = \frac{1}{2g^2} \int d^4x\ \mathrm{tr}\, F_{\mu\nu} F^{\mu\nu}$$

Note that in Euclidean space, the action comes with a + sign. This is to be contrasted with the Minkowski space action (2.8) which comes with a minus sign. We can write this as

$$S_{YM} = \frac{1}{4g^2} \int d^4x\ \mathrm{tr}\,(F_{\mu\nu} \mp {}^\star F_{\mu\nu})^2 \pm \frac{1}{2g^2} \int d^4x\ \mathrm{tr}\, F_{\mu\nu} {}^\star F^{\mu\nu} \geq \frac{8\pi^2}{g^2}|\nu|$$

where, in the last line, we've used the result (2.46). We learn that in the sector with winding ν, the Yang-Mills action is bounded by $8\pi^2|\nu|/g^2$. The action is minimised when the bound is saturated. This occurs when

$$F_{\mu\nu} = \pm^\star F_{\mu\nu} \tag{2.49}$$

These are the (anti) self-dual Yang-Mills equations. The argument above shows that solutions to these first order equations necessarily minimise the action in a given topological sector and so must solve the equations of motion (2.48). In fact, it's straightforward to see that this is the case since it follows immediately from the Bianchi identity $\mathcal{D}_\mu{}^\star F^{\mu\nu} = 0$. The kind of "completing the square" trick that we used above, where we bound the action by a topological invariant, is known as the Bogomolnyi bound. We'll see it a number of times in these lectures.

Solutions to the (anti) self-dual Yang-Mills equations (2.49) are known as *instantons*. This is because, as we will see below, the action density is localised at both a point in space and at an instant in (admittedly, Euclidean) time. They contribute to the path integral with a characteristic factor

$$e^{-S_{\text{instanton}}} = e^{-8\pi^2|\nu|/g^2} e^{i\theta\nu} \tag{2.50}$$

Note that the Yang-Mills contribution is real because we've Wick rotated to Euclidean space. However, the contribution from the theta term remains complex even after Wick rotation. This is typical behaviour for such topological terms that sit in the action with epsilon symbols.

A Single Instanton in $SU(2)$

We will focus on gauge group $G = SU(2)$ and solve the self-dual equations $F_{\mu\nu} = {}^\star F_{\mu\nu}$ with winding number $\nu = 1$. As we've seen, asymptotically the gauge field must be pure gauge, and so takes the form $A_\mu \to i\Omega\partial_\mu\Omega^{-1}$. An example of a map $\Omega(x) \in SU(2)$ with winding $\nu = 1$ is given by

$$\Omega(x) = \frac{x_\mu \sigma^\mu}{\sqrt{x^2}} \quad \text{where } \sigma^\mu = (1, -i\vec{\sigma})$$

with this choice, the asymptotic form of the gauge field is given by[3]

$$A_\mu \;\to\; i\Omega\partial_\mu\Omega^{-1} = \frac{1}{x^2}\eta^i_{\mu\nu}x^\nu\sigma^i \quad \text{as } x \to \infty$$

[3]In the lecture notes on Solitons, the instanton solution was presented in singular gauge, where it takes a similar, but noticeably different form.

Here the $\eta^i_{\mu\nu}$ 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

$$\eta^1_{\mu\nu}=\begin{pmatrix}0&1&0&0\\-1&0&0&0\\0&0&0&1\\0&0&-1&0\end{pmatrix}\ ,\ \ \eta^2_{\mu\nu}=\begin{pmatrix}0&0&1&0\\0&0&0&-1\\-1&0&0&0\\0&1&0&0\end{pmatrix}\ ,\ \ \eta^3_{\mu\nu}=\begin{pmatrix}0&0&0&1\\0&0&1&0\\0&-1&0&0\\-1&0&0&0\end{pmatrix}$$

These matrices are self-dual: they obey $\frac{1}{2}\epsilon_{\mu\nu\rho\sigma}\eta^i_{\rho\sigma}=\eta^i_{\mu\nu}$. This will prove important. (Note that we're not being careful about indices up vs down as we are in Euclidean space with no troublesome minus signs.) The full gauge potential should now be of the form $A_\mu = if(x)\Omega\partial_\mu\Omega^{-1}$ for some function $f(x)\to 1$ as $x\to\infty$. The right choice turns out to be $f(x)=x^2/(x^2+\rho^2)$ where ρ is a parameter whose role will be clarified shortly. We then have the gauge field

$$A_\mu = \frac{1}{x^2+\rho^2}\,\eta^i_{\mu\nu}x^\nu\sigma^i \tag{2.51}$$

You can check that the associated field strength is

$$F_{\mu\nu} = -\frac{2\rho^2}{(x^2+\rho^2)^2}\eta^i_{\mu\nu}\sigma^i$$

This inherits its self-duality from the 't Hooft matrices and therefore solves the Yang-Mills equations of motion.

The instanton solution (2.51) is not unique. By acting on this solution with various symmetries, we can easily generate more solutions. The most general solution with winding $\nu=1$ depends on 8 parameters which, in this context, are referred to as *collective coordinates*. Each of them has a simple explanation:

- The instanton solution above is localised at the origin. But we can always generate a new solution localised at any point $X\in\mathbf{R}^4$ simply by replacing $x^\mu\to x^\mu - X^\mu$ in (2.51). This gives 4 collective coordinates.

- We've kept one parameter ρ explicit in the solution (2.51). This is the scale size of the instanton, an interpretation which is clear from looking at the field strength which is localised in a ball of radius ρ. The existence of this collective coordinate reflects the fact that the classical Yang-Mills theory is scale invariant: if a solution exists with one size, it should exist with any size. This property is broken in the quantum theory by the running of the coupling constant, and this has implications for instantons that we will describe below.

- The final three collective coordinates arise from the global part of the gauge group. These are gauge transformations which do not die off asymptotically, and correspond to three physical symmetries of the theory, rather than redundancies. For our purposes, we can consider a constant $V \in SU(2)$, and act as $A_\mu \to VA_\mu V^{-1}$. These determine the internal orientation of the instanton within the gauge group.

Before we proceed, we pause to mention that it is straightforward to write down a corresponding anti-self-dual instanton with winding $\nu = -1$. We simply replace the 't Hooft matrices with their anti-self dual counterparts,

$$\bar{\eta}^1_{\mu\nu} = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{pmatrix} \ , \ \bar{\eta}^2_{\mu\nu} = \begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \ , \ \bar{\eta}^3_{\mu\nu} = \begin{pmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

They obey $\frac{1}{2}\epsilon_{\mu\nu\rho\sigma}\eta^i_{\rho\sigma} = -\eta^i_{\mu\nu}$, and one can use these to build a gauge potential (2.51) with $\nu = -1$. These too form an irreducible representation of $su(2)$, and obey $[\eta^i, \bar{\eta}^j] = 0$. The fact that we can find two commuting $su(2)$ algebras hiding in a 4×4 matrix reflects the fact that $Spin(4) \cong SU(2) \times SU(2)$ and, correspondingly, the Lie algebras are $so(4) = su(2) \oplus su(2)$.

General Instanton Solutions

To get an instanton solution in $SU(N)$, we could take the $SU(2)$ solution (2.51) and simply embed it in the upper left-hand corner of an $N \times N$ matrix. We can then rotate this into other embeddings by acting with $SU(N)$, modulo the stabilizer which leaves the configuration untouched. This leaves us with the action

$$\frac{SU(N)}{S[U(N-2) \times U(2)]}$$

where the $U(N-2)$ hits the lower-right-hand corner and doesn't see our solution, while the $U(2)$ is included in the denominator because it acts like V in the original solution (2.51) and we don't want to over count. The notation $S[U(p) \times U(q)]$ means that we lose the overall central $U(1) \subset U(p) \times U(q)$. The coset space above has dimension $4N - 8$. This means that the solution in which (2.51) is embedded into $SU(N)$ comes with $4N$ collective coordinate. This is the most general $\nu = 1$ instanton solution in $SU(N)$.

What about solutions with higher ν? There is a beautiful story here. It turns out that such solutions exist and have $4N|\nu|$ collective coordinates. Among these

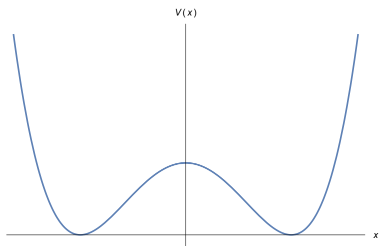


Figure 8: The double well

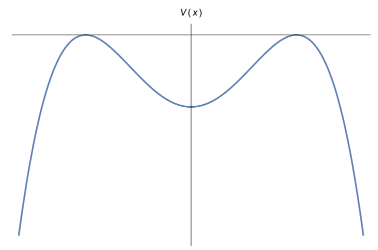


Figure 9: The upside down double well

solutions are configurations which look like ν well separated instantons, each with $4N$ collective coordinates describing their position, scale size and orientation. However, as the instantons overlap this interpretation breaks down.

Remarkably, there is a procedure to generate all solutions for general ν. It turns out that one can reduce the non-linear partial differential equations (2.49) to a straightforward algebraic equation. This is known as the ADHM construction and is possible due to some deep integrable properties of the self-dual Yang-Mills equations. You can read more about this construction (from the perspective of D-branes and string theory) in the lectures on Solitons.

2.3.2 Tunnelling: Another Quantum Mechanics Analogy

We've found solutions in Euclidean spacetime that contribute to the theta dependence in the path integral. But why Euclidean rather than Lorentzian spacetime? The answer is that solutions to the Euclidean equations of motion describe quantum tunnelling.

This is best illustrated by a simple quantum mechanical example. Consider the double well potential shown in the left-hand figure. Clearly there are two classical ground states, corresponding to the two minima. But we know that a quantum particle sitting in one minimum can happily tunnel through to the other. The end result is that the quantum theory has just a single ground state.

How can we see this behaviour in the path integral? There are no classical solutions to the equations of motion which take us from one minimum to the other. However, things are rather different in Euclidean time. We define

$$\tau = it$$

After this Wick rotation, the action

$$S[x(t)] = \int dt\; \frac{m}{2}\left(\frac{dx}{dt}\right)^2 - V(x)$$

turns into the Euclidean action

$$S_E[x(\tau)] = -iS = \int d\tau \; \frac{m}{2}\left(\frac{dx}{d\tau}\right)^2 + V(x)$$

We see that the Wick rotation has the effect of inverting the potential: $V(x) \to -V(x)$. In Euclidean time, the classical ground states correspond to the maxima of the inverted potential. But now there is a perfectly good solution to the equations of motion, in which we roll from one maximum to the other. We come to a rather surprising conclusion: quantum tunnelling can be viewed as classical motion in imaginary time!

As an example, consider the quartic potential

$$V(x) = \lambda(x^2 - a^2)^2 \tag{2.52}$$

which has minima at $x = \pm a$. Then a solution to the equations of motion which interpolates between the two ground states in Euclidean time is given by

$$\bar{x}(\tau) = a \tanh\left(\frac{\omega}{2}(\tau - \tau_0)\right) \tag{2.53}$$

with $\omega^2 = 8\lambda a^2/m$. This solution is the instanton for quantum mechanics in the double well potential. There is also an anti-instanton solution that interpolates from $x = +a$ to $x = -a$. The (anti)-instanton solution is localised in a region $1/\omega$ in imaginary time. In this case, there is just a single collective coordinate, τ_0, whose existence follows from time translational invariance of the quantum mechanics.

Returning to Yang-Mills, we now seek a similar tunnelling interpretation for the instanton solutions. In the semi-classical approximation, the instantons tunnel between the $|n\rangle$ vacua that we described in Section 2.2.3. Recall that the semi-classical vacuum is defined by $A_i = iV\partial_i V^{-1}$ on a spatial slice $\mathbf{R}^3$, which we subsequently compactify to $\mathbf{S}^3$. The vacuum $|n\rangle$ is associated to maps $V(\mathbf{x}) : \mathbf{S}^3 \mapsto G$ with winding n, defined in (2.36).

We noted previously that the construction of the vacua $|n\rangle$ in terms of winding relies on topological arguments which are similar to those which underlie the existence of instantons. To see the connection, we can take the definition of the instanton winding (2.47) and deform the integration region from the asymptotic $\mathbf{S}^3_\infty = \partial\mathbf{R}^4$ to the two asympotic three spheres $\mathbf{S}^3_\pm$ which we think of as the compactified $\mathbf{R}^3_\pm$ spatial slices ar $t = \pm\infty$. We can then compare the instanton winding (2.47) to the definition of the vacuum states (2.36), to write

$$\nu(U) = n_+(U) - n_-(U)$$

We learn that the Yang-Mills instanton describes tunnelling between the two semi-classical vacua, $|n_-\rangle \;\to\; |n_+\rangle = |n_- + \nu\rangle$, as shown in the figure.

2.3.3 Instanton Contributions to the Path Integral

Given an instanton solution, our next task is to calculate something. The idea is to use the instanton as the starting point for a semi-classical evaluation of the path integral.

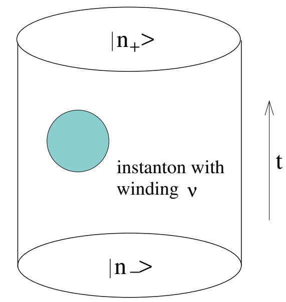


Figure 10:

We can first illustrate this in our quantum mechanics analogy, where we would like to compute the amplitude to tunnel from one classical ground state $|x=-a\rangle$ to the other $|x=+a\rangle$ over some time T.

$$\langle a|e^{-HT}|-a\rangle = \mathcal{N}\int_{x(0)=-a}^{x(T)=+a} \mathcal{D}x(\tau)\ e^{-S_E[x(\tau)]}$$

with $\mathcal{N}$ a normalisation constant that we shall do our best to avoid calculating. There is a general strategy for computing instanton contributions to path integrals which we sketch here. This strategy will be useful in later sections (such as Section 7.2 and 8.3 where we discuss instantons in 2d and 3d gauge theories respectively). However, we'll see that we run into some difficulties when applying these ideas to Yang-Mills theories in $d=3+1$ dimensions.

Given an instanton solution $\bar{x}(\tau)$, like (2.53), we write the general $x(\tau)$ as

$$x(\tau) = \bar{x}(\tau) + \delta x(\tau)$$

and expand the Euclidean action as

$$S_E[x(\tau)] = S_{\text{instanton}} + \int d\tau\ \delta x\, \Delta \delta x + \mathcal{O}(\delta x^3) \tag{2.54}$$

Here $S_{\text{instanton}} = S_E[\bar{x}(\tau)]$. There are no terms that are linear in δx because $\bar{x}(\tau)$ solves the equations of motion. The expansion of the action to quadratic order gives the differential operator Δ. The semi-classical approach is valid if the higher order terms give sub-leading corrections to the path integral. For our quantum mechanics double well potential, one can check that this holds provided $\lambda \ll 1$ in (2.52). For Yang-Mills, this requirement will ultimately make us think twice about the semi-classical expansion.

Substituting the expansion (2.54) into the path integral, we're left with the usual Gaussian integral. It's tempting to write

$$\begin{aligned}\int_{x(0)=-a}^{x(T)=+a} \mathcal{D}x(\tau)\ e^{-S_E[x(\tau)]} &= e^{-S_{\text{instanton}}}\int_{\delta x(0)=0}^{\delta x(T)=0} \mathcal{D}\delta x(\tau)\ e^{-\delta x\Delta\delta x + \mathcal{O}(\delta x^3)} \\ &\approx \frac{e^{-S_{\text{instanton}}}}{\det^{1/2}\Delta}\end{aligned}$$

This, however, is a little too quick. The problem comes because the operator Δ has a zero eigenvalue which makes the answer diverge. A zero eigenvalue of Δ occurs if there are any deformations of the solution $\bar{x}(\tau)$ which do not change the action. But we know that such deformations do indeed exist since the instanton solutions are never unique: they depend on collective coordinates. In our quantum mechanics example, there is a just a single collective coordinate, called τ_0 in (2.53), which means that the deformation $\delta x = \partial\bar{x}/\partial\tau_0$ is a *zero mode*: it is annihilated by Δ.

To deal with this, we need to postpone the integration over any zero mode. These can then be replaced by an integration over the associated collective coordinate. For our quantum mechanics example, we have

$$\int_{x(0)=-a}^{x(T)=+a} \mathcal{D}x(\tau)\; e^{-S_E[x(\tau)]} \approx \int_0^T d\tau_0\; J\,\frac{e^{-S_{\rm instanton}}}{\det'^{1/2}\Delta}$$

Here J is the Jacobian factor that comes from changing the integration variable from the zero mode to the collective coordinate. We will not calculate it here. Meanwhile the notation $\det'$ means that we omit the zero eigenvalue of Δ when computing the determinant. The upshot is that a single instanton gives a saddle point contribution to the tunnelling amplitude,

$$\langle a|e^{-HT}|-a\rangle \approx KT\,e^{-S_{\rm instanton}} \quad \text{with } K = \frac{\mathcal{N}J}{\det'^{1/2}\Delta}$$

Note that we've packaged all the things that we couldn't be bothered to calculate into a single constant, K.

The result above gives the contribution from a single instanton to the tunnelling amplitude. But, it turns out, this is not the dominant contribution. That, instead, comes from summing over many such tunnelling events.

Consider a configuration consisting of a string of instantons and anti-instantons. Each instanton must be followed by an anti-instanton and vice versa. This configuration does not satisfy the equation of motion. However, if the (anti) instantons are well separated, with a spacing $\gg 1/\omega$, then the configuration very nearly satisfies the equations of motion; it fails only by exponentially suppressed terms. We refer to this as a dilute gas of instantons.

As above, we should integrate over the positions of the instantons and anti-instantons. Because each of these is sandwiched between two others, this leads to the integration

$$\int_0^T dt_1 \int_{t_1}^T dt_2 \ldots \int_{t_{n-1}}^T dt_n = \frac{T^n}{n!}$$

where we're neglecting the thickness $1/\omega$ of each instanton.

A configuration consisting of n instantons and anti-instantons is more highly suppressed since its action is approximately $nS_{\text{instanton}}$. But, as we now see, these contributions dominate because of entropic factors: there are many more of them. Summing over all such possibilities, we have

$$\langle a|e^{-HT}|-a\rangle \sim \sum_{n \text{ odd}} \frac{1}{n!}(KTe^{-S_{\text{instanton}}})^n = \sinh\left(KTe^{-S_{\text{instanton}}}\right)$$

where we restrict the sum to n odd to ensure that we end up in a different classical ground state from where we started. We haven't made any effort to normalise this amplitude, but we can compare it to the amplitude to propagate from the state $|-a\rangle$ back to $|-a\rangle$,

$$\langle -a|e^{-HT}|-a\rangle \sim \sum_{n \text{ even}} \frac{1}{n!}(KTe^{-S_{\text{instanton}}})^n = \cosh\left(KTe^{-S_{\text{instanton}}}\right)$$

In the long time limit $T \to \infty$, we see that we lose information about where we started, and we're equally likely to find ourselves in either of the ground states $|a\rangle$ or $|-a\rangle$. If we were more careful about the overall normalisation, we can also use this argument to compute the energy splitting between the ground state and the first excited state.

As an aside, you may notice that the calculation above is identical to the argument for why there are no phase transitions in one dimensional thermal systems given in the lectures on Statistical Field Theory.

Back to Yang-Mills Instantons

Now we can try to apply these same ideas to Yang-Mills instantons. Unfortunately, things do not work out as nicely as we might have hoped. We would like to approximate the Yang-Mills path integral

$$Z = \int \mathcal{D}A \; e^{-S_{YM}+iS_\theta}$$

by the contribution from the instanton saddle point. There are the usual issues related to gauge fixing, but these do not add anything new to our story so we neglect them here and focus only on the aspects directly related to instantons. (We'll be more careful about gauge fixing in Section 2.4.2 when we discuss the beta function.)

Let's start by again considering the contribution from a single instanton. The story proceeds as for the quantum mechanics example until we come to discuss the collective coordinates. For the instanton in quantum mechanics, there was just a single collective coordinate τ_0. For our Yang-Mills instanton in $SU(2)$, there are eight. Four of these are associated to translations in Euclidean spacetime; these play the same role as τ_0 and integrating over them gives a factor of the Euclidean spacetime volume VT, with V the 3d spatial volume. Three of the collective coordinates arise from the global part of the gauge symmetry and can be happily integrated over. But this leaves us with the scale size ρ. This too should be singled out from the path integral and integrated over. We find ourselves with an integral of the form,

$$Z \approx \int_0^\infty d\rho\ K(\rho)\, VT\, e^{-8\pi^2/g^2} e^{i\theta}$$

where, as before, $K(\rho)$ includes contributions from the Jacobians and the one-loop determinant. Now, however, it is a function of the instanton scale size ρ and so we should do the hard work of calculating it.

We won't do this hard work, in part because the calculation is rather involved and in part because, as we advertised above, the end result doesn't offer quantitative insights into the behaviour of Yang-Mills. It turns out that $K(\rho)$ causes the integral to diverge at large ρ. This raises two concerns. First, it is difficult to justify the dilute instanton gas approximation if it is dominated by instantons of arbitrarily large size which are surely overlapping. Second, and more pressing, it is difficult to justify the saddle point expansion at all. This is because, as we describe in some detail in the next section, the gauge coupling in Yang-Mills runs; it is small at high energy but becomes large at low energies. This means that any semi-classical approximation, such as instantons, is valid for describing short distance processes but breaks down at large distances. The fact that our attempt to compute the partition function is dominated by instantons of large size is really telling us that the whole semi-classical strategy has broken down. Instead, we're going to have to face up to the fact that Yang-Mills is a strongly coupled quantum field theory.

It's a little disappointing that we can't push the instanton programme further in Yang-Mills. However, it's not all doom and gloom and we won't quite leave instantons behind in these lectures. There are situations where instantons are the leading contribution to certain processes. We will see one such example in Section 3.3.2 in the context of the anomaly, although for more impressive examples one has to look to supersymmetric field theories which are under greater control and beyond the scope of these lectures.

2.4 The Flow to Strong Coupling

Our discussion in the previous sections has focussed on the classical (or, at the very least, semi-classical) approach to Yang-Mills. Such a description gives good intuition for the physics when a theory is weakly coupled, but often fails miserably at strong coupling. The next question we should ask is whether Yang-Mills theory is weakly or strongly coupled.

We have chosen a scaling in which the coupling g^2 sits in front of the action

$$S_{\mathrm{YM}} = \frac{1}{2g^2}\int d^4x \; \mathrm{tr}\, F^{\mu\nu}F_{\mu\nu} \tag{2.55}$$

The quantum theory is defined, in the framework of path integrals, by summing over all field configurations weighted, with $e^{iS_{YM}}$ in Minkowski space or $e^{-S_{YM}}$ in Euclidean space. When g^2 is small, the Euclidean action has a deep minimum on the solutions to the classical equations of motion, and these dominate the path integral. In this case, the classical field configurations provide a good starting point for a saddle point analysis. (In Minkowski space, the action is a stationary point rather than a minimum on classical solutions but, once again, these dominate the path integral.) In contrast, when g^2 is large, many field configurations contribute to the path integral. In this case, we sometimes talk about quantum fluctuations being large. Now the quantum state will look nothing like the solutions to the classical equations of motion.

All of this would seem to suggest that life is easy when g^2 is small, and harder when g^2 is large. However, things are not quite so simple. This is because the effective value of g^2 differs depending on the length scale on which you look: we write $g^2 = g^2(\mu)$, where μ is an appropriate energy scale, or inverse length scale. Note that this is quite a radical departure from the the classical picture where any constants you put in the action remain constant. In quantum field theory, these constants are more wilful: they take the values they want to, rather than the values we give them.

We computed the running of the gauge coupling g^2 at one-loop in our previous course on *Advanced Quantum Field Theory*. (We will review this computation in Section 2.4.2 below.) The upshot is that the coupling constant depends on the scale μ as

$$\frac{1}{g^2(\mu)} = \frac{1}{g_0^2} - \frac{11}{3}\frac{C(\mathrm{adj})}{(4\pi)^2}\log\frac{\Lambda_{UV}^2}{\mu^2} \tag{2.56}$$

where g_0^2 is the coupling constant evaluated at the cut-off scale Λ_{UV}.

Here $C(\mathrm{adj})$ is a group theoretic factor. Recall that we have fixed a normalisation of the Lie algebra generators in the fundamental representation to be (2.2),

$$\mathrm{tr}\left[T^aT^b\right] = \frac{1}{2}\delta^{ab} \tag{2.57}$$

Having pinned down the normalisation in one representation, the other representations R will have different normalisations,

$$\mathrm{tr}\,\left[T^a(R)T^b(R)\right] = I(R)\,\delta^{ab}$$

The coefficient $I(R)$ is called the *Dynkin index* of the representation R. The convention (2.57) means that $I(F) = \frac{1}{2}$. The group theoretic factor appearing in the beta function is simply the Dynkin index in the adjoint representation,

$$C(\mathrm{adj}) = I(\mathrm{adj})$$

It is also known as the *quadratic Casimir*, which is why it is denoted by a different letter. For the various simple, compact Lie groups it is given by

G	$SU(N)$	$SO(N)$	$Sp(N)$	E_6	E_7	E_8	F_4	G_2
$C(\mathrm{adj})$	N	$\frac{1}{2}N-1$	$N+1$	2	3/2	1/2	3/2	2

Note that the adjoint representation of E_8 is the minimal representation; hence the appearance of $C(\mathrm{adj}) = I(F) = \frac{1}{2}$.

The running of the gauge coupling (2.56) is often expressed in terms of the beta function

$$\beta(g) \equiv \mu\frac{dg}{d\mu} = \beta_0 g^3 \quad \text{with} \quad \beta_0 = -\frac{11}{3}\frac{C(\mathrm{adj})}{(4\pi)^2} \tag{2.58}$$

The minus sign in (2.56) or, equivalently, in (2.58), is all important. It tells us that the gauge coupling gets stronger as we flow to longer length scales. In contrast, it is weaker at short distance scales. This phenomena is called *asymptotic freedom*.

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. In particular, our discussion of instantons in Section 2.3 is valid at short distance scales. However, it becomes much harder to understand what is going on at large distances where the coupling gets strong. Indeed, the beta function (2.58) is valid only when $g^2(\mu) \ll 1$. 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 $g^2(\mu)$ diverge? This happens at a finite energy

$$\Lambda_{QCD} = \Lambda_{UV}\, e^{1/2\beta_0 g_0^2} \tag{2.59}$$

For historical reasons, we refer to this as the "QCD scale", reflecting its importance in the strong force. Alternatively, we can write Λ_{QCD} in terms of any scale μ,

$$\Lambda_{QCD} = \mu\, e^{1/2\beta_0 g^2(\mu)}$$

and $d\Lambda_{QCD}/d\mu = 0$. For this reason, it is sometimes referred to as the RG-invariant scale.

Asymptotic freedom means that $\beta_0 < 0$. This ensures that if $g_0 \ll 1$, so that the theory is weakly coupled at the cut-off, then $\Lambda_{QCD} \ll \Lambda_{UV}$. This is interesting. Yang-Mills theory naturally generates a scale Λ_{QCD} which is exponentially lower than the cut-off Λ_{UV} of the theory. Theoretical physicists spend a lot of time worrying about "naturalness" which, at heart, is the question of how Nature generates different length scales. The logarithmic running of the coupling exhibited by Yang-Mills theory provides a beautiful mechanism to do this. As we will see moving forwards, all the interesting physics in Yang-Mills occurs at energies of order Λ_{QCD}.

Viewed naively, there's something very surprising about the emergence of the scale Λ_{QCD}. This is because classical Yang-Mills 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 (2.59) 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 this cut-off which evolves to the physical scale Λ_{QCD}.

The question we would like to ask is: what does Yang-Mills theory look like at low energies, comparable to Λ_{QCD}? This is a difficult question to answer, and our current understanding comes primarily from experiment and numerical work, with intuition built from different analytic approaches. The answer is rather startling: Yang-Mills theory does not describe massless particles. Instead, the gluons bind together to form massive particles known as *glueballs*. These particles have a mass that is of the order of Λ_{QCD}, but figuring out the exact spectrum remains challenging. We sometimes say that the theory is *gapped*, meaning that there is a gap in the energy spectrum between the ground state, which we can take to have $E = 0$, and the first excited state with energy $E = Mc^2$, where M is the mass of the lightest glueball.

Proving the mass gap for Yang-Mills is one of the most important and difficult open problems in mathematical physics. In these lectures we will restrict ourselves to building some intuition for Yang-Mills theory, and understanding some of the consequences of the mass gap. In later sections, will also see how the situation changes when we couple Yang-Mills to dynamical matter fields.

Before we proceed, I should mention a rather subtle and poorly understood caveat. We have argued in Sections 2.2 and 2.3 that the dynamics of Yang-Mills theory also depends on the theta parameter and we can ask: how does θ affect the spectrum? We have only a cursory understanding of this. It is thought that, for nearly all gauge groups, Yang-Mills remains gapped for all values of θ. However, something interesting happens at $\theta = \pi$. Recall from Section 1.2.5 that $\theta = \pi$ is special because it preserves time-reversal invariance, more commonly known in particle physics as $\mathcal{CP}$. For most gauge groups, it is thought that the dynamics spontaneously breaks time reversal invariance at $\theta = \pi$, so that Yang-Mills has two degenerate ground states. We will give an argument for this in Section 3.6 using discrete anomalies, and another in Section 6.2.5 when we discuss the large N expansion. However, there is speculation that the behaviour of Yang-Mills is rather different for gauge group $G = SU(2)$ and that, while gapped for all $\theta \neq \pi$, this theory actually becomes gapless at $\theta = \pi$, where it is conjectured to be described by a free $U(1)$ photon. We will have nothing to say about this in these lectures.

2.4.1 Anti-Screening and Paramagnetism

The computations of the 1-loop beta functions are rather involved. It's useful to have a more down-to-earth picture in mind to build some understanding for what's going on. There is nice intuitive analogy that comes 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

$$\mathbf{D} = \epsilon \mathbf{E}$$

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

$$\frac{1}{e^2(\mu)} = \frac{1}{e_0^2} + \frac{1}{12\pi^2}\log\left(\frac{\Lambda_{UV}^2}{\mu^2}\right)$$

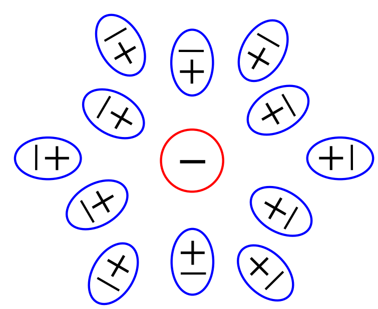


Figure 11:

This tells us that the charge of the electron gets effectively smaller as we look at larger distance scales. This can be 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 unusual. 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

$$\mathbf{B} = \mu \mathbf{H}$$

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 and one finds that the paramagnetism wins.

In the next section we will compute the usual one-loop beta-function. We present the computation in such a way that it makes clear the distinction between the diamagnetic and paramagnetic contributions. Viewed in this light, asymptotic freedom can be traced to the paramagnetic contribution from the gluon spins.

2.4.2 Computing the Beta Function

In this section, we will sketch the derivation of the beta function (2.58). We're going to use an approach known as the *background field method*. We work in Euclidean space and decompose the gauge field as

$$A_\mu = \bar{A}_\mu + \delta A_\mu$$

We will think of $\bar{A}_\mu$ as the low-energy, slowly moving part of the field. It is known as the *background field*. Meanwhile, δA_μ describes the high-energy, short-wavelength modes whose effect we would like to understand. The field strength becomes

$$F_{\mu\nu} = \bar{F}_{\mu\nu} + \bar{\mathcal{D}}_\mu \delta A_\nu - \bar{\mathcal{D}}_\nu \delta A_\mu - i[\delta A_\mu, \delta A_\nu]$$

where $\bar{\mathcal{D}}_\mu = \partial_\mu - i[\bar{A}_\mu, \cdot]$ is the covariant derivative with respect to the background field $\bar{A}_\mu$. From this, we can write the action (2.55) as

$$\begin{aligned} S_{YM} = \frac{1}{g^2}\int d^4x \ \mathrm{tr}\Big[\frac{1}{2}\bar{F}_{\mu\nu}\bar{F}^{\mu\nu} &+ 2\bar{F}^{\mu\nu}\bar{\mathcal{D}}_\mu\delta A_\nu \\ &+ \bar{\mathcal{D}}^\mu\delta A^\nu\,\bar{\mathcal{D}}_\mu\delta A_\nu - \bar{\mathcal{D}}^\mu\delta A^\nu\,\bar{\mathcal{D}}_\nu\delta A_\mu - i\bar{F}^{\mu\nu}[\delta A_\mu,\delta A_\nu] \\ &- 2i\bar{\mathcal{D}}^\mu\delta A^\nu[\delta A_\mu,\delta A_\nu] - \frac{1}{2}[\delta A^\mu,\delta A^\nu][\delta A_\mu,\delta A_\nu]\Big] \end{aligned} \tag{2.60}$$

where we've ordered the terms in the action depending on the number of δA's. Note that the middle line is quadratic in δA.

Gauge Fixing and Ghosts

Our plan is to integrate over the fluctuations δA_μ in the path integral, leaving ourselves with an effective action for the background field $\bar{A}_\mu$. To do this, we must first deal with the gauge symmetry. While the action of the gauge symmetry on A_μ is clear, there is no unique decomposition of the gauge transformation between on $\bar{A}_\mu$ and δA_μ. However, the calculation is simplest if we load the full gauge transformation into δA_μ, so

$$\delta_{\text{gauge}}\bar{A}_\mu = 0 \quad \text{and} \quad \delta_{\text{gauge}}(\delta A_\mu) = \bar{\mathcal{D}}_\mu\omega - i[\delta A_\mu, \omega]$$

where, for this section alone, we've changed our notation for infinitesimal gauge transformations so as not to confuse them with the fluctuating field δA_μ. With this choice, δA_μ transforms as any other adjoint field.

As usual, field configurations related by a gauge symmetry should be viewed as physically equivalent. This is necessary in the present context because the kinetic terms for δA_μ are not invertible. For this reason, we first need a way to fix the gauge.

We do this using the Faddeev-Popov procedure that we saw in the lectures on *Advanced Quantum Field Theory*. We choose to work in the gauge

$$G(\bar{A};\delta A) = \bar{\mathcal{D}}^\mu \delta A_\mu = 0 \tag{2.61}$$

Note that this gauge fixing condition depends on our choice of background field. This is the advantage of this method; we will find that the gauge invariance of $\bar{A}_\mu$ is retained throughout the calculation.

We add to our action the gauge-fixing term

$$S_{gf} = \frac{1}{g^2}\int d^4x \ \mathrm{tr}\,(\bar{\mathcal{D}}^\mu \delta A_\mu)^2 \tag{2.62}$$

The choice of overall coefficient of the gauge fixing term is arbitrary. But nice things happen if we make the choice above. To see why, let's focus on the $\bar{\mathcal{D}}^\mu \delta A^\nu \bar{\mathcal{D}}_\nu \delta A_\mu$ term in (2.60). Integrating by parts, we have

$$\begin{aligned}\int d^4x \ \mathrm{tr}\,\bar{\mathcal{D}}^\mu \delta A^\nu \bar{\mathcal{D}}_\nu \delta A_\mu &= -\int d^4x \ \mathrm{tr}\,\delta A_\nu \bar{\mathcal{D}}^\mu \bar{\mathcal{D}}^\nu \delta A_\mu \\ &= -\int d^4x \ \ \mathrm{tr}\,\delta A_\nu \Big([\bar{\mathcal{D}}^\mu, \bar{\mathcal{D}}^\nu] + \bar{\mathcal{D}}^\nu \bar{\mathcal{D}}^\mu\Big)\delta A_\mu \\ &= \int d^4x \ \mathrm{tr}\,\Big[(\bar{\mathcal{D}}^\mu \delta A_\mu)^2 + i\delta A_\nu [\bar{F}^{\mu\nu}, \delta A_\mu]\Big]\end{aligned}$$

The first of these terms is then cancelled by the gauge fixing term (2.62), leaving us with

$$\begin{aligned}S_{YM} + S_{gf} = \frac{1}{g^2}\int d^4x \ \mathrm{tr}\Big[&\frac{1}{2}\bar{F}_{\mu\nu}\bar{F}^{\mu\nu} + 2\bar{F}^{\mu\nu}\bar{\mathcal{D}}_\mu \delta A_\nu \\ &+\bar{\mathcal{D}}^\mu \delta A^\nu \,\bar{\mathcal{D}}_\mu \delta A_\nu - 2i\bar{F}^{\mu\nu}[\delta A_\mu, \delta A_\nu] \\ &-2i\bar{\mathcal{D}}^\mu \delta A^\nu [\delta A_\mu, \delta A_\nu] - \frac{1}{2}[\delta A^\mu, \delta A^\nu][\delta A_\mu, \delta A_\nu]\Big]\end{aligned}$$

and we're left with just two terms that are quadratic in δA. We'll return to these shortly.

The next step of the Faddeev-Popov procedure is to implement the gauge fixing condition (2.61) as a delta-function constraint in the path integral. We denote the gauge transformed fields as $\bar{A}^\omega_\mu = \bar{A}_\mu$ and $\delta A^\omega_\mu = \delta A_\mu + \bar{\mathcal{D}}\omega - i[\delta A_\mu, \omega]$. We then use the identity

$$\int \mathcal{D}\omega \ \delta(G(\bar{A}^\omega, \delta A^\omega)) \ \det\left(\frac{\partial G(\bar{A}^\omega, \delta A^\omega)}{\partial \omega}\right) = 1$$

The determinant can be rewritten through the introduction of adjoint-valued ghost fields c. For the gauge fixing condition (2.61), we have

$$\det\left(\frac{\partial G(\bar{A},\delta A^\omega)}{\partial\omega}\right)=\int\mathcal{D}c\,\mathcal{D}c^\dagger\,\exp\left(-\frac{1}{g^2}\int d^4x\ \mathrm{tr}\Big[-c^\dagger\bar{\mathcal{D}}^2c+ic^\dagger[\bar{\mathcal{D}}^\mu\delta A_\mu,c]\Big]\right)$$

where we've chosen to include an overall factor of $1/g^2$ in the ghost action purely as a convenience; it doesn't effect subsequent calculations. The usual Faddeev-Popov story tells us that the integration $\int\mathcal{D}\omega$ now decouples, resulting in an unimportant overall constant. We're left with an action that includes both the fluctuating gauge field δA_μ and the ghost field c, $S=S_{YM}+S_{gf}+S_{\rm ghost}$,

$$\begin{aligned}S=\frac{1}{g^2}\int d^4x\ \mathrm{tr}\Big[&\frac{1}{2}\bar{F}_{\mu\nu}\bar{F}^{\mu\nu}+2\bar{F}^{\mu\nu}\bar{\mathcal{D}}_\mu\delta A_\nu\\&+\bar{\mathcal{D}}^\mu\delta A^\nu\,\bar{\mathcal{D}}_\mu\delta A_\nu-2i\bar{F}^{\mu\nu}[\delta A_\mu,\delta A_\nu]+\bar{\mathcal{D}}_\mu c^\dagger\bar{\mathcal{D}}^\mu c\\&-2i\bar{\mathcal{D}}^\mu\delta A^\nu[\delta A_\mu,\delta A_\nu]-\frac{1}{2}[\delta A^\mu,\delta A^\nu][\delta A_\mu,\delta A_\nu]+ic^\dagger[\bar{\mathcal{D}}^\mu\delta A_\mu,c]\Big]\end{aligned}$$

As previously, we have arranged the terms so that the middle line is quadratic in fluctuating fields, while the final line is cubic and higher.

One-Loop Determinants

Our strategy now is to integrate out the fluctuating fields, δA_μ and c, to determine their effect on the dynamics of the background field $\bar{A}_\mu$.

$$e^{-S_{\rm eff}[\bar{A}]}=\int\mathcal{D}\delta A\,\mathcal{D}c\,\mathcal{D}c^\dagger\ e^{-S[\bar{A},\delta A,c]}$$

Things are simplest if we take our background field to obey the classical equations of motion, $\bar{\mathcal{D}}_\mu\bar{F}^{\mu\nu}$, which ensures that the term linear in δA_μ in the action disappears. Furthermore, at one loop it will suffice to ignore the terms cubic and quadratic in fluctuating fields that sit on the final line of the action above. We're then left just with Gaussian integrations, and these are easy to do,

$$e^{-S_{\rm eff}[\bar{A}]}=\det{}^{-1/2}\Delta_{\rm gauge}\ \det{}^{+1}\Delta_{\rm ghost}\ e^{-\frac{1}{2g^2}\int d^4x\ \mathrm{tr}\,\bar{F}_{\mu\nu}\bar{F}^{\mu\nu}}$$

where the quadratic fluctuation operators can be read off from the action and are given by

$$\Delta^{\mu\nu}_{\rm gauge}=-\bar{\mathcal{D}}^2\delta^{\mu\nu}+2i[\bar{F}^{\mu\nu}\,\cdot]\quad\text{and}\quad\Delta_{\rm ghost}=-\bar{\mathcal{D}}^2$$

where the $\bar{F}^{\mu\nu}$ should be thought of as an operator acting on objects in the adjoint representation. This extra term, $\bar{F}_{\mu\nu}$, arising from the gauge fields can be traced to the fact that they are spin 1 excitations. As we will see below, this contributes the paramagnetic part to the beta function and, ultimately, is responsible for the famous minus sign that leads to anti-screening.

Taking logs of both sides, the effective action is given by

$$S_{\text{eff}}[\bar{A}] = \frac{1}{2g^2} \int d^4x \ \text{tr} \, \bar{F}_{\mu\nu} \bar{F}^{\mu\nu} + \frac{1}{2} \text{Tr} \log \Delta_{\text{gauge}} - \text{Tr} \log \Delta_{\text{ghost}} \tag{2.63}$$

where the Tr means the trace over group, Lorentz and momentum indices (as opposed to tr which is over only gauge group indices). We need to figure out how to compute the contributions from these quadratic fluctuation operators.

<u>The Ghost Contribution</u>

The contribution from the ghost fields are simplest because it has the least structure. We write

$$\Delta_{\text{ghost}} = -\partial^2 + \Delta_1 + \Delta_2$$

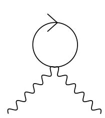
where the subscripts keep track of how many $\bar{A}_\mu$ terms each operator has,

$$\Delta_1 = i\partial^\mu \bar{A}_\mu + i\bar{A}_\mu \partial^\mu \quad \text{and} \quad \Delta_2 = [\bar{A}^\mu, [\bar{A}_\mu, \cdot]]$$

where, again these operators act on objects in the adjoint representation. This will prove important to get the right normalisation factor. We then have

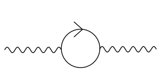
$$\begin{aligned} \text{Tr} \log \Delta_{\text{ghost}} &= \text{Tr} \log \left(-\partial^2 + \Delta_1 + \Delta_2 \right) \\ &= \text{Tr} \log(-\partial^2) + \text{Tr} \log \left(1 + (-\partial^2)^{-1}(\Delta_1 + \Delta_2) \right) \\ &= \text{Tr} \log(-\partial^2) + \text{Tr} \left((-\partial^2)^{-1}(\Delta_1 + \Delta_2) \right) - \frac{1}{2} \text{Tr} \left((-\partial^2)^{-1}(\Delta_1 + \Delta_2) \right)^2 + \ldots \end{aligned}$$

The first term is just an overall constant. We can ignore it. In the second term, $\text{Tr}\, \Delta_1$ includes the trace over gauge indices and vanishes because $\text{tr}\, \bar{A}_\mu = 0$. This is just the statement that there is no gauge invariant contribution to the kinetic term linear in $\bar{A}_\mu$. So the first terms that we need to worry about are the quadratic terms.



$$= \text{Tr}\left((-\partial^2)^{-1} \Delta_2 \right) = \int \frac{d^4k}{(2\pi)^4} \text{tr}_{\text{adj}} [\bar{A}_\mu(k) \bar{A}_\nu(-k)] \int \frac{d^4p}{(2\pi)^4} \frac{\delta^{\mu\nu}}{p^2}$$

where we've also included a graphical reminder of where these terms come from in a more traditional Feynman diagram approach. We also have



$$= -\frac{1}{2} \text{Tr}\left((-\partial^2)^{-1} \Delta_1 (-\partial^2)^{-1} \Delta_1 \right) = \frac{1}{2} \int \frac{d^4k}{(2\pi)^4} \text{tr}_{\text{adj}} [\bar{A}_\mu(k) \bar{A}_\nu(-k)] \times f_{\mu\nu}(k)$$

with

$$f_{\mu\nu}(k) = \int \frac{d^4p}{(2\pi)^4} \frac{(2p+k)^\mu (2p+k)^\nu}{p^2(p+k)^2}$$

Note that the trace over group indices should be taken with A_μ acting on adjoint valued objects, as opposed to our convention in (2.3) where it naturally acts on fundamental objects.

We would like to massage these into the form of the Yang Mills action. In momentum space, the quadratic part of the Yang-Mills action reads

$$\begin{aligned} S_{\text{quad}} &= \frac{1}{g^2} \int d^4x \; \text{tr} \left(\partial_\mu \bar{A}_\nu \partial^\mu \bar{A}^\nu - \partial_\mu \bar{A}^\nu \partial_\nu \bar{A}^\mu \right) \\ &= \frac{1}{g^2} \int \frac{d^4k}{(2\pi)^4} \; \text{tr} \left[\bar{A}_\mu(k) \bar{A}_\nu(-k) \right] \left(k^\mu k^\nu - k^2 \delta^{\mu\nu} \right) \end{aligned}$$

There are a couple of issues that we need to deal with. First, the Yang-Mills action is written in terms of fundamental generators which, as in (2.57), are normalised as $\text{tr}\, T^a T^b = \frac{1}{2}\delta^{ab}$. Meanwhile, the trace in the one-loop contributions is in the adjoint representation, and is given by

$$\text{tr}_{\text{adj}} T^a T^b = C(\text{adj})\, \delta^{ab}$$

Second, we must perform the integral over the loop momentum p. This, of course, diverges. These are the kind of integrals that were covered in previous QFT courses. We implement a UV cut-off Λ_{UV} to get

$$-\text{Tr} \log \Delta_{\text{ghost}} = -\frac{C(\text{adj})}{3(4\pi)^2} \int \frac{d^4k}{(2\pi)^4} \; \text{tr} \left[\bar{A}_\mu(k) \bar{A}_\nu(-k) \right] \left(k^\mu k^\nu - k^2 \delta^{\mu\nu} \right) \log \left(\frac{\Lambda_{UV}^2}{k^2} \right)$$

This is our first contribution to the logarithmic running of the coupling that we advertised in (2.56).

Above we focussed purely on the quadratic terms. Expanding the Yang-Mills action also gives us cubic and quartic terms and, for consistency, we should check that they too receive the same corrections. Indeed they do. In fact, this is guaranteed to work because of the manifest gauge invariance $\delta_{\text{gauge}} \bar{A}_\mu = \bar{\mathcal{D}}_\mu \omega$.

The Gauge Contribution

Next up is the contribution $\frac{1}{2}\text{Tr}\,\log\Delta_{\text{gauge}}$, where

$$\Delta^{\mu\nu}_{\text{gauge}} = \Delta_{\text{ghost}}\delta^{\mu\nu} + 2i[\bar{F}^{\mu\nu},\cdot]$$

We see that part of the calculation involves Δ_{ghost}, and so it gives the same answer as above. The only difference is the spin indices $\delta^{\mu\nu}$ which give an extra factor of 4 after taking the trace. This means that

$$\text{Tr}\,\log\Delta_{\text{gauge}} = 4\text{Tr}\,\log\Delta_{\text{ghost}} + \bar{F}_{\mu\nu}\text{ terms}$$

On rotational grounds, there is no term linear in $\bar{F}_{\mu\nu}$. This means that the first term comes from expanding out $\log\Delta_{\text{gauge}}$ to quadratic order and focussing on the $\bar{F}^2_{\mu\nu}$ terms,

$$\begin{aligned}\bar{F}_{\mu\nu}\text{ terms} &= -\frac{1}{2}(2i)^2\,\text{Tr}\,\left((-\partial^2)^{-1}[\bar{F}_{\mu\nu},[(-\partial^2)^{-1}\bar{F}^{\mu\nu},\cdot]]\right)\\ &= -\frac{1}{2}\int\frac{d^4k}{(2\pi)^4}\text{tr}_{\text{adj}}[\bar{A}_\mu(k)\bar{A}_\nu(-k)]\int\frac{d^4p}{(2\pi)^4}\,\frac{-4(k^\rho\delta^{\mu\sigma}-k^\sigma\delta^{\mu\rho})(k_\sigma\delta^\nu_\rho-k_\rho\delta^\nu_\sigma))}{p^2(p+k)^2}\end{aligned}$$

Once again, we have a divergent integral to compute. This time we get

$$\bar{F}_{\mu\nu}\text{ terms} = -\frac{8C(\text{adj})}{(4\pi)^2}\int\frac{d^4k}{(2\pi)^4}\ \text{tr}\left[\bar{A}_\mu(k)\bar{A}_\nu(-k)\right](k^\mu k^\nu - k^2\delta^{\mu\nu})\log\left(\frac{\Lambda^2_{UV}}{k^2}\right)$$

The sum then gives the contribution to the effective action,

$$\frac{1}{2}\text{Tr}\,\log\Delta_{\text{gauge}} = \frac{1}{2}\left[\frac{4}{3}-8\right]\frac{C(\text{adj})}{(4\pi)^2}\int\frac{d^4k}{(2\pi)^4}\ \text{tr}\left[\bar{A}_\mu(k)\bar{A}_\nu(-k)\right](k^\mu k^\nu - k^2\delta^{\mu\nu})\log\left(\frac{\Lambda^2_{UV}}{k^2}\right)$$

Here the 4/3 is the diamagnetic contribution. In fact, it's overkill since it neglects the gauge redundancy. This is subtracted by including the contribution from the ghost fields. Together, these give rise to a positive beta function. In contrast, the -8 term is the paramagnetic piece, and can be traced to the spin 1 nature of the gauge field. This is where the overall minus sign comes from.

The coefficient of the kinetic terms is precisely the gauge coupling $1/g^2$. Combining both gauge and ghost contributions, and identifying the momentum k of the background field as the relevant scale μ, we have

$$\begin{aligned}\frac{1}{g^2(\mu)} &= \frac{1}{g^2} + \frac{C(\text{adj})}{(4\pi)^2}\left[-\frac{1}{3}+\frac{1}{2}\left(\frac{4}{3}-8\right)\right]\log\left(\frac{\Lambda^2_{UV}}{\mu^2}\right)\\ &= \frac{1}{g^2} - \frac{11}{3}\frac{C(\text{adj})}{(4\pi)^2}\log\left(\frac{\Lambda^2_{UV}}{\mu^2}\right)\end{aligned}$$

This is in agreement with the advertised result (2.58). As explained previously, the overall minus sign here is important. Indeed, it was worth a Nobel prize.

2.5 Electric Probes

When we first studied Maxwell's theory of Electromagnetism, one of the most basic questions we asked was: what's the force between two charged particles? In these calculations, the charged particles are sources which we've inserted by hand; we're using them as a probe of the theory, to see how the electromagnetic fields respond in their presence. In this section we will develop the tools that will allow us to ask similar questions about non-Abelian gauge theories.

2.5.1 Coulomb vs Confining

We start by building up some expectation from the classical physics. Asymptotic freedom means that these classical results will be valid when the particles are close by, separated by distances $\ll 1/\Lambda_{QCD}$, but are unlikely to hold when they are far separated. Nonetheless, it will be useful to understand the theory in this regime, if only because it highlights just how surprising the long distance, quantum behaviour actually is.

In electromagnetism, two particles of equal and opposite charges $\pm e$, separated by a distance r, experience an attractive Coulomb force. This can be described in terms of the potential energy $V(r)$,

$$V(r) = -\frac{e^2}{4\pi r}$$

In the framework of QED, we can reproduce this from the the tree-level exchange of a single photon, as shown in the figure. We did this in first course on Quantum Field Theory.

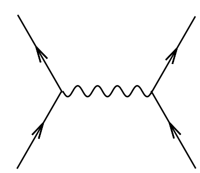


Figure 12:

Here we do the same calculation in $SU(N)$ Yang-Mills theory. We refer to the charged particles as *quarks*. For now, we'll take these particles to sit in the fundamental representation of $SU(N)$, although the methods we use here easily generalise to arbitrary gauge groups and representations. Each quark and anti-quark carries a colour index, $i = 1, \ldots, N$. Moreover, when they exchange a gluon, this colour index can change. The tree-level diagram takes the same form, but with a gluon exchanged instead of a photon. It gives

$$V(r) = \frac{g^2}{4\pi r} T^a_{ki} T^{\star\, a}_{lj} \tag{2.64}$$

But we've still got those colour indices to deal with, i, j for ingoing, and k, l for outgoing. We should think of $T^a T^{\star\, a}$ as an $N^2 \times N^2$ matrix, acting on the N^2 different ingoing colour states. These different N^2 states then split into different irreducible representations. For our quark and anti-quark, we have

$$\mathbf{N} \otimes \bar{\mathbf{N}} = \mathbf{1} \oplus \text{adj} \tag{2.65}$$

where the adjoint representation has dimension $N^2 - 1$. The matrix $T^a T^{\star\, a}$ will then have two different eigenvalues, one for each of these representations. This will lead to two different coefficients for the forces.

An Aside on Group Theory

We need a way to compute the eigenvalues of $T^a T^{\star\, a}$ in these two different representations. In fact, we've met this kind of problem before; it's the same kind of issue that arose in our lectures on Applications of Quantum Mechanics when we treated the spin-orbit coupling $\mathbf{L}\cdot\mathbf{S}$ of an atom. In that case we wrote $\mathbf{J} = \mathbf{L} + \mathbf{S}$ and used the identity $\mathbf{L}\cdot\mathbf{S} = \frac{1}{2}(\mathbf{J}^2 - \mathbf{L}^2 + \mathbf{S}^2) = \frac{1}{2}(j(j+1) - l(l+1) - s(s+1))$.

We can repeat this trick for any group G. Consider two representations R_1 and R_2 and the associated generators $T^a(R_1)$ and $T^a(R_2)$. We construct a new operator

$$S^a(R) = T^a(R_1)\otimes \mathbf{1} \ + \ \mathbf{1}\otimes T^a(R_2)$$

We then have

$$T^a(R_1)\otimes T^a(R_2) = \frac{1}{2}\left[S^a(R)S^a(R) - T^a(R_1)T^a(R_1)\otimes\mathbf{1} - \mathbf{1}\otimes T^a(R_2)T^a(R_2)\right]$$

But it is simple to show that $T^a(R)T^a(R)$ commutes with all elements of the group and so is proportional to the identity,

$$T^a(R)T^a(R) = C(R)\,\mathbf{1} \tag{2.66}$$

where $C(R)$ is known as the quadratic Casimir, a number which characterises the representation R. In our discussion of beta functions in Section 2.4, we encountered the Dynkin index, which is the coefficient of the trace normalisation

$$\text{tr}\, T^a(R)T^b(R) = I(R)\delta^{ab}$$

The two are related by

$$I(R)\dim(G) = C(R)\dim(R)$$

where $\dim(G)$ is the dimension of the group and $\dim(R)$ is the dimension of the representation. Note that this is consistent with our earlier claim that $I(\text{adj}) = C(\text{adj})$. For $G = SU(N)$, the fundamental and adjoint representations have

$$C(\mathbf{N}) = C(\bar{\mathbf{N}}) = \frac{N^2-1}{2N} \quad \text{and} \quad C(\text{adj}) = N$$

while the symmetric $\square\square$ and anti-symmetric $\begin{smallmatrix}\square\\\square\end{smallmatrix}$ representations have

$$C\left(\square\square\right) = \frac{(N-1)(N+2)}{N} \quad \text{and} \quad C\left(\begin{smallmatrix}\square\\\square\end{smallmatrix}\right) = \frac{(N-2)(N+1)}{N}$$

Non-Abelian Coulomb Force

Let's now apply this to the force between quarks. The group theory machinations above tell us that the operator $T^a(R_1)T^a(R_2)$ decomposes into a block diagonal matrix, with entries labelled by the irreducible representations $R \subset R_1 \otimes R_2$ and given by

$$T^a(R_1)T^a(R_2)\Big|_R = \frac{1}{2}\left[C(R) - C(R_1) - C(R_2)\right]$$

The quark and anti-quark can sit in two different irreducible representations: the singlet and the adjoint (2.65). For the singlet, we have

$$\frac{1}{2}\left[C(\mathbf{1}) - C(\mathbf{N}) - C(\bar{\mathbf{N}})\right] = -\frac{N^2-1}{2N}$$

The minus sign ensures that the force between the quark and anti-quark in the singlet channel is attractive. This is what we would have expected from our classical intuition. However, when the quarks sit in the adjoint channel, we have

$$\frac{1}{2}\left[C(\text{adj}) - C(\mathbf{N}) - C(\bar{\mathbf{N}})\right] = \frac{1}{2N}$$

Perhaps surprisingly, this is a repulsive force.

The group theory analysis above makes it simple to compute the classical force between quarks in any representation. Suppose, for example, we have two quarks, both in the fundamental representation. They decompose as

$$\mathbf{N} \otimes \mathbf{N} = \square\square \oplus \begin{smallmatrix}\square\\\square\end{smallmatrix}$$

where $\dim(\square\square) = \frac{1}{2}N(N+1)$ and $\dim(\begin{smallmatrix}\square\\\square\end{smallmatrix}) = \frac{1}{2}N(N-1)$. We then have

$$\frac{1}{2}\left[C\left(\square\square\right) - C(\mathbf{N}) - C(\mathbf{N})\right] = \frac{N-1}{2N}$$

and

$$\frac{1}{2}\left[C\left(\begin{smallmatrix}\square\\\square\end{smallmatrix}\right) - C(\mathbf{N}) - C(\mathbf{N})\right] = -\frac{N+1}{2N}$$

and the force is repulsive between quarks in the symmetric channel, but attractive in the anti-symmetric channel.

We see that, even classically, Yang-Mills theory provides a somewhat richer structure to the forces between particles. However, at the classical level, Yang-Mills retains the familiar $1/r$ fall-off from Maxwell theory. This is the signature of a force due to the exchange of massless particles in $d = 3+1$ dimensions, whether photons or gravitons or, in this case, gluons. As we now explain, at the quantum level things are very different.

The Confining Force

In the previous section, 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 $\bar{\mathbf{N}}$ representations respectively. The energy between the two turns out to grow linearly with distance

$$V(r) = \sigma r \tag{2.67}$$

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 scale in the game.

For two quarks, 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 between 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*: an individual quark cannot survive on its own, but is forced to enjoy the company of friends.

There is a possibility for confusion in the the claim that only singlet states survive in a confining gauge theory. In any gauge theory, one should only talk about gauge invariant states and a single quark is not a gauge invariant object. However, we can render the quark gauge invariant by attaching a Wilson line (2.14) which stretches from the position of the quark to infinity. When we blithely talk about a single quark, we should really be thinking of this composite object. This is not directly related to the issue of confinement. Indeed, the statements above hold equally well for electrons in QED: these too are only gauge invariant when attached to a Wilson line. Instead the issue of confinement is a dynamical statement, rather than a kinematical one. Confinement means that the quark + Wilson line costs infinite energy in Yang-Mills, while the electron + Wilson line (suitably regulated) costs finite energy in QED.

There are situations where it's not possible to form a singlet from a pair of particles, but it is possible if enough particles are added. The baryon provides a good example, in which N quarks, each in the fundamental representation of $SU(N)$, combine to form a singlet $\mathcal{B} = \epsilon^{i_1 \ldots i_N} q_{i_1} \ldots q_{i_N}$. These too are finite energy states.

Confinement in Yang-Mills is, like the mass gap, a challenging problem. There is no analytic demonstration of this phenomenon. Instead, we will focus on building some intuition for why this might occur and understanding the right language to describe it.

2.5.2 An Analogy: Flux Lines in a Superconductor

There is a simple system which provides a useful analogy for confinement. This is a superconductor.

One of the wonders of the superconducting vacuum is its ability to expel magnetic fields. If you attempt to pass a magnetic field through a superconductor, it resists. This is known as the *Meissner effect.* If you insist, by cranking up the magnetic field, the superconductor will relent, but it will not do so uniformly. Instead, the magnetic field will form string-like filaments known as vortices.

We can model this using the Abelian Higgs model. This is a $U(1)$ gauge field, coupled to a complex scalar

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

with $\mathcal{D}_\mu \phi = \partial_\mu \phi - iA_\mu \phi$. (As an aside: in an actual superconductor, the complex scalar field describes the Cooper pair of electrons, and should have a non-relativistic kinetic term rather than the relativistic kinetic terms we use here.)

In the vacuum, the scalar has an expectation value, $\langle |\phi| \rangle = v$, spontaneously breaking the $U(1)$ gauge symmetry and giving the photon a mass, $m_\gamma^2 = 2e^2v^2$. This is, of course, the Higgs mechanism. In this vacuum, the scalar also has a mass given by $m_\phi^2 = 4\lambda v^2$.

Let's start by seeing how this explains the Meissner effect. We'll look for time independent solutions, with $A_0 = 0$ and a magnetic field $B^i = -\frac{1}{2}\epsilon^{ijk}F_{jk}$. If we assume that the Higgs field doesn't deviate from $\phi = v$ then the equation of motion for the gauge field is

$$\nabla \times \mathbf{B} = -m_\gamma^2 \mathbf{A} \quad \Rightarrow \quad \nabla^2 \mathbf{B} = m_\gamma^2 \mathbf{B}$$

This is known as the *London equation.* It tells us that magnetic fields are exponentially damped in the Higgs phase, with solutions of the form $\mathbf{B}(x) = \mathbf{B}_0\, e^{-m_\gamma x}$. In the context of superconductors, the length scale $L = 1/m_\gamma$ is known as the *penetration depth*. Later another length scale, $\xi \sim 1/m_\phi$, will also be important; this is called the *correlation length*.

Of course, the assumption that $\phi = v$ is not justified: ϕ is a dynamical field and is determined by its equation of motion. This is where we will find the vortices. We decompose the complex scalar as

$$\phi = \rho e^{i\alpha}$$

All finite energy, classical configurations must have $\rho \to v$ as $x \to \infty$. But the phase α is arbitrary. This opens up an interesting topological possibility. Consider a classical configuration which is invariant in the x^3 direction, but is localised in the (x^1, x^2) plane. The translational invariance x^3 reflects the fact that we will be constructing an infinite string solution, aligned along x^3. We parameterise the plane by radial coordinates $x^1 + ix^2 = re^{i\theta}$. Then all configurations whose energy is finite when integrated over the (x^1, x^2) plane involve a map

$$\alpha(\theta) : \mathbf{S}^1_\infty \mapsto \mathbf{S}^1 \tag{2.68}$$

These maps fall into disjoint classes, labelled by the number of times that α winds as we move around the asymptotic circle $\mathbf{S}^1_\infty$. This is the same kind of idea that we met when discussing theta vacua and instantons in Sections 2.2 and 2.3. In that case we were dealing with the homotopy group $\Pi_3(\mathbf{S}^3)$; here we have a simpler situation, with maps of the form (2.68) classified by

$$\Pi_1(\mathbf{S}^1) = \mathbf{Z}$$

In this case, it is simple to write down an expression for the integer $n \in \mathbf{Z}$ which classifies the map. It is the winding number,

$$n = \frac{1}{2\pi} \int_{\mathbf{S}^1_\infty} d\theta \, \frac{\partial \alpha}{\partial \theta} \ \in \ \mathbf{Z} \tag{2.69}$$

In this way, the space of field configurations decompose into sectors, labelled by $n \in \mathbf{Z}$. The vacuum sits in the sector $n = 0$. A particularly simple way to find classical solutions is to minimize the energy in a sector $n \neq 0$. These solutions, which are stabilised by their winding at infinity, are often referred to as *topological solitons*. In the present context, these solitons will be the vortices that we are looking for.

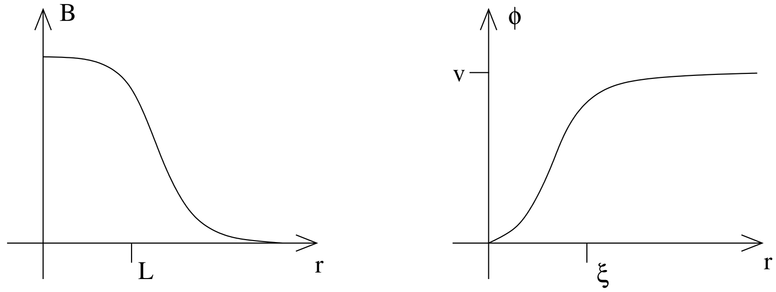


Figure 13: The profile for the magnetic field and Higgs field in a vortex.

We'll consider radially symmetric scalar profiles of the form

$$\phi(r,\theta) = \rho(r)e^{in\theta} \tag{2.70}$$

We will first see why any configuration with $n \neq 0$ necessarily comes with a magnetic field. Because our configurations are invariant under x^3 translations, they will always have a linearly diverging energy corresponding to the fact that we have an infinite string. But the energy density in the (x^1, x^2) plane should integrate to a finite number. We denote the energy per unit length of the vortex string by σ. The kinetic term for the scalar gives a contribution to the energy that includes

$$\sigma \sim \int dr d\theta \; r \left| \left(\frac{1}{r}\frac{\partial}{\partial\theta} - iA_\theta \right) \phi \right|^2 = \int dr d\theta \; r \left| \frac{in\rho}{r} - iA_\theta \rho \right|^2$$

If we try to set $A_\theta = 0$, the energy has a logarithmic divergence from the integral over the (x^1, x^2) plane. To compensate we must turn on $A_\theta \to n/r$ as $r \to \infty$. But this means that the configuration (2.70) is accompanied by a magnetic flux

$$\Phi = \int d^2x \; B_3 = \oint d\theta \; rA_\theta = 2\pi n \tag{2.71}$$

We see that the flux is quantised. This is the same quantisation condition that we saw for magnetic monopoles in Section 1.1 (albeit with a rescaled convention for the gauge field because we chose to put the coupling e^2 in front of the action). Note, however, that here we haven't invoked any quantum mechanics; in the Higgs phase, the quantisation of flux happens for topological reasons, rather than quantum reasons.

So far we have talked about configurations with winding, but not yet discussed whether they are solutions to the equations of motion. It is not hard to find solutions for a single vortex with $n = 1$ (or, equivalently, an anti-vortex with $n = -1$). We write an ansatz for the gauge field as $A_\theta = f(r)/r$ and require $f(r) \to 1$ as $r \to \infty$. The equations of motion then reduce to ordinary differential equations for $\rho(r)$ and $f(r)$. Although no analytic solutions are known, it is simple to solve them numerically. These solutions are often referred to as *Nielsen-Olesen vortices*.

Here we will build some intuition for what these look like without doing any hard work. The key feature is that ϕ winds asymptotically, as in (2.70), which means that by the time we get to the origin it has something of an identity crisis and does not know which way to point. The only way in which the configuration can remain smooth is if $\phi = 0$ at the origin. But it costs energy for ϕ to deviate from the vacuum, so it must do so over as small a scale as possible. This scale is $\xi \sim 1/m_\phi$.

Similarly, we know that the flux (2.71) must be non-zero. It is energetically preferable for this flux to sit at the origin, since this is where the Higgs field vanishes. This flux spreads over a region associated to the penetration length $L \sim 1/m_\gamma$. The resulting profiles for the Higgs and magnetic fields are sketched in the figures.

Type I, Type II and Bogomonlyi

Before we explain why these vortices provide a good analogy for confinement, we first make a small aside. As described above, there are two length scales at play in the vortex solutions. The Higgs field drops to zero over a region of size $\sim \xi$ while the magnetic field is spread over a region of size $\sim L$.

The ratio of these two scales determines the force between two parallel vortices. For far separated vortices, the force is exponentially suppressed, reflecting the fact that the theory is gapped. As they come closer, either their magnetic flux will begin to overlap (if $L > \xi$), or their scalar profiles will begin to overlap (if $\xi > L$). The magnetic flux is repulsive, while the scalar field is attractive. Based on this distinction, superconductors are divided into two classes:

Type I: $\xi > L$. In this case, the overlap of the scalar profiles of vortices provide the dominant, attractive force. If one applies a uniform magnetic field to a superconductor, it turns into one big vortex. But a big vortex is effectively the same as turning the system back into the normal phase. This means that the superconductor resists an applied magnetic field until it reaches a critical value, at which point the system exits the Higgs phase. This means that no vortices are seen in Type I superconductors.

Type II: $\xi < L$. Now the magnetic flux of the vortices overlap as they approach, resulting in a repulsive force. This means that when a uniform magnetic field is applied to a Type II superconductor, it will form many vortices, each of which wants to be as

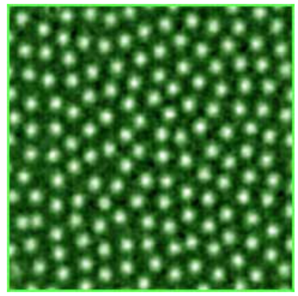


Figure 14: An Abrikosov lattice in a Type II superconductor.

far from the others as possible. The result is a periodic array of vortices known as an *Abrikosov lattice*. An example is shown in the figure[4].

At the boundary between Type I and Type II superconductors, the heuristic arguments above suggest that there are no forces between vortices. Mathematically, something rather pretty happens at this point. We have $m_\gamma^2 = m_\phi^2$ or, equivalently, $\lambda = e^2/2$. At this special value, we can write the tension of the vortex string as the sum of squares,

$$\begin{aligned}\sigma &= \int d^2x\ \frac{1}{2e^2}B_3^2 + \sum_{i=1,2}|\mathcal{D}_i\phi|^2 + \frac{e^2}{2}(|\phi|^2 - v^2)^2 \\ &= \int d^2x\ \ |\mathcal{D}_1\phi - i\mathcal{D}_2\phi|^2 + i\mathcal{D}_1\phi^\dagger\mathcal{D}_2\phi - i\mathcal{D}_2\phi^\dagger\mathcal{D}_1\phi \\ &\qquad\qquad + \frac{1}{2e^2}\left(B_3 + e^2(|\phi|^2 - v^2)\right)^2 - B_3(|\phi|^2 - v^2) \\ &= \int d^2x\ \ |\mathcal{D}_1\phi - i\mathcal{D}_2\phi|^2 - i\phi^\dagger[\mathcal{D}_1,\mathcal{D}_2]\phi + \frac{1}{2e^2}\left(B_3^2 + e^2(|\phi|^2 - v^2)\right)^2 - B_3(|\phi|^2 - v^2) \\ &= \int d^2x\ \ |\mathcal{D}_1\phi + i\mathcal{D}_2\phi|^2 + \frac{1}{2e^2}\left(B_3 + e^2(|\phi|^2 - v^2)\right)^2 + v^2 B_3\end{aligned}$$

where, in going to the last line, we used the fact that $[\mathcal{D}_1,\mathcal{D}_2] = -iF_{12} = +iB_3$. This "completing the square" trick is the same kind of Bogomolnyi argument that we used in Section 2.3 when discussing instantons. Since the two squares are necessarily positive, the energy can be bounded by

$$\mathcal{E} \geq \int d^2x\ v^2 B_3 = 2\pi v^2 n$$

where we have related the flux to the winding using (2.71). This is nice. In a sector

[4] This picture is taken from P. Goa et al, Supercond. Sci. Technol. 14, 729 (2001).

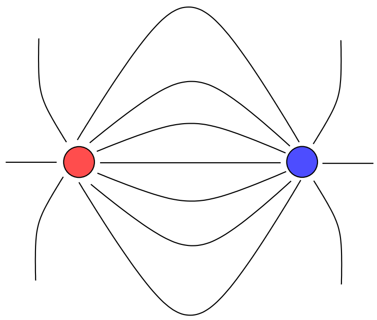


Figure 15: The flux lines for a monopole and anti-monopole in vacuum.

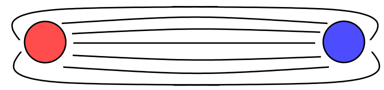


Figure 16: The same flux lines in a superconductor.

with winding $n > 0$, there is a minimum energy bound. Moreover, we can saturate this bound by requiring that the quantities in the squares vanish,

$$\mathcal{D}_1\phi = i\mathcal{D}_2\phi \quad \text{and} \quad B_3 = -e^2(|\phi|^2 - v^2) \tag{2.72}$$

These are the Bogomolnyi vortex equations. For $n < 0$, one can play a similar game with some minus signs shuffled around to derive Bogomolnyi equations for anti-vortices.

The vortex equations (2.72) have a number of remarkable properties. In particular, it can be shown that the general solution has $2n$ parameters which, at least for far separated vortices, can be thought of as the position of n vortices on the plane. Physically, this arises because there is no force between the vortices. You can read more about this in the lecture notes on Solitons.

The Confinement of Monopoles

So far we've reviewed some basic physics of the Higgs phase of electromagnetism. But what does this have to do with confinement? To see the connection, we need to think about what would happen if we place a Dirac monopole inside a superconductor.

To get some grounding, let's first consider a monopole and anti-monopole in vacuum. Their magnetic field lines spread out in a pattern that is familiar from the games we played with iron filings and magnets when we were kids. This is sketched in the left-hand figure. These field lines result in a Coulomb-like force between the two particles, $V(r) \sim 1/r$.

Now what happens when we place these particles inside a superconductor? The magnetic flux lines can no longer spread out, but instead must form collimated tubes. This is sketched in the right-hand figure. This tube of flux is the vortex that we described above. As we have seen, happily the magnetic flux carried by a single vortex

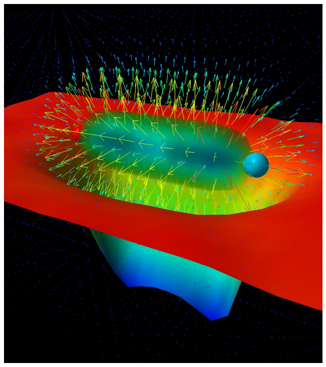


Figure 17: A simulation of a separated quark anti-quark pair in QCD.

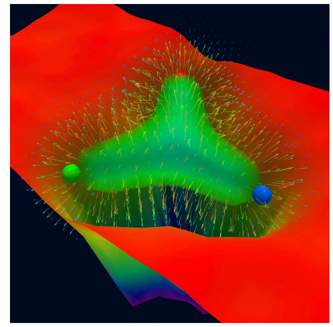


Figure 18: A simulation a separated baryon state in QCD.

coincides with the magnetic flux emitted by a single Dirac monopole. The energy cost in separating the monopole and anti-monopole by a distance r is now

$$V(r) = \sigma r$$

where σ is the energy per unit length of the vortex string. In other words, inside a superconductor, magnetic monopoles are confined!

What lesson for Yang-Mills can we take away from this? First, it seems very plausible that the confinement of quarks in Yang-Mills is again due to the emergence of flux lines, this time (chromo)electric rather than magnetic flux lines. However, in contrast to the Abelian Higgs model, the Yang-Mills flux tube is not expected to arise as a semi-classical solution of the Yang-Mills equations. Instead, the flux tube should emerge in the strongly coupled quantum theory where one sums over many field configurations. Indeed, such flux tubes are seen in lattice simulations where they provide dominant contributions to the path integral. An example is shown in the figure[5].

It is less obvious how these flux tubes form between N well separated quarks which form a baryon. Simulations suggest that the flux tubes emitted by each quark can join together at an N-string vertex. The picture for a well separated baryon in QCD, with $G = SU(3)$ gauge group, is shown in the figure.

We might also wish to take away another lesson from the superconducting story. In the Abelian Higgs model, the electrically charged field ϕ condenses, resulting in the confinement of monopoles. Duality then suggests that to confine electrically charged objects, such as quarks, we should look to condense magnetic monopoles. This idea

[5]These simulations were created by Derek Leinweber. You can find a host of beautiful QCD animations on his webpage.

smells plausible, but there has been scant progress in making it more rigorous in the context of Yang-Mills theory. (For what it's worth, the idea can be shown to work in certain supersymmetric theories.) Nonetheless, it encourages us to look for magnetic objects in non-Abelian gauge theories. We will describe these in Sections 2.6 and 2.8.

Regge Trajectories

The idea that quark anti-quark pairs are held together by flux tubes has experimental support. Here we'll provide a rather simplistic model of this set up. Ignoring the overall translational motion, the energy of two, massless relativistic quarks, joined together by a string, is given by

$$E = p + \sigma r$$

with $p = p_1 - p_2$ the relative momentum. We'll embrace the spirit of Bohr, and require that the angular momentum is quantised: $J = pr \in \mathbf{Z}$. We can then write the energy as

$$E = \frac{J}{r} + \sigma r$$

For a fixed J, this is minimized at $r = \sqrt{J/\sigma}$, which gives us the relationship between the energy and angular momentum of the states,

$$E^2 \sim \sigma J$$

We can now compare this to the data for hadrons. A plot of the mass^2 vs spin is known as a Chew-Frautschi plot. It is shown on the right for light vector mesons[6]. We see that families of mesons and their resonances do indeed sit on nice straight lines, referred to as *Regge trajectories*. The slope of the lines is determined by the QCD string tension, which turns out to be around $\sigma \sim 1.2\ GeV^2$. Perhaps more surprisingly, the data also reveals nice straight Regge trajectories in the baryon sector.

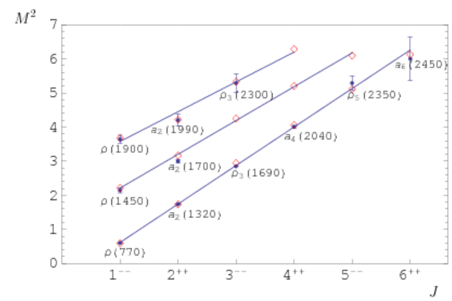


Figure 19:

2.5.3 Wilson Loops Revisited

Above we identified two different possible phases of Yang-Mills theory: the Coulomb phase and the confining phase. The difference between them lies in the forces experienced by two well-separated probe particles.

[6]This plot was taken from the paper by D. Ebert, R. Faustov and V. Galkin, arXiv:0903.5183.

- Coulomb: $V(r) \sim 1/r$
- Confining: $V(r) \sim r$

To this, we could add a third possibility that occurs when the gauge field is Higgsed, so that electric charges are completely screened. In this case we have

- Higgs: $V(r) \sim$ constant

We'll discuss this phase more in Section 2.7.3.

Usually in a quantum field theory (or in a statistical field theory) we identify the phase by computing the expectation value of some order parameter. The question that we would like to ask here is: what is the order parameter for confinement?

To answer this, we can rephrase our earlier discussion in terms of the path integral. To orient ourselves, let's first return to Maxwell theory. If we want to compute the path integral in the presence of an electrically charged probe particle, we simply introduce the particle by its associated current J^μ, which now acts as a source. We then add to the action the term $A_\mu J^\mu$. Moreover, for a probe particle which moves along a worldline C, the current J is a delta-function localised on C. We then compute the partition function with the insertion $e^{i\oint_C A}$,

$$\left\langle \exp\left(i\oint_C A\right)\right\rangle = \int \mathcal{D}A\ \exp\left(i\oint_C A\right) e^{iS_{\rm Maxwell}} \tag{2.73}$$

where we're being a little sloppy on the right-hand-side, omitting both gauge fixing terms and the normalisation factor coming from the denominator.

In Yang-Mills, there is a similar story. The only difference is that we can't just stipulate a fixed current J^μ because the term $A_\mu J^\mu$ is not gauge invariant. Instead, we must introduce some internal colour degrees of freedom for the quark, as we described previously in Section 2.1.3. As we saw, integrating over these colour degrees of freedom leaves us with the Wilson loop $W[C]$, which we take in the fundamental representation

$$W[C] = {\rm tr}\,\mathcal{P}\exp\left(i\oint_C A\right)$$

Performing the further path integral over the gauge fields A leaves us with the expectation value of this Wilson loop

$$\Big\langle W[C]\Big\rangle = \int \mathcal{D}A\ {\rm tr}\,\mathcal{P}\exp\left(i\oint_C A\right) e^{iS_{YM}} \tag{2.74}$$

Now consider the specific closed loop C shown in the figure. We again take this to sit in the fundamental representation. It has the interpretation that we create a quark anti-quark pair, separated by a distance r, at some time in the past. These then propagate forward for time T, before they annihilate back to the vacuum.

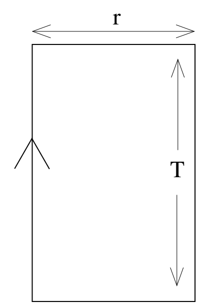


Figure 20:

What behaviour would we expect from the expectation value $\langle W[C]\rangle$? We'll work in Euclidean space. Recall from our earlier lectures on quantum field theory that, for long times, the path integral projects the system onto the lowest energy state. Before the quarks appear, and after they've gone, this is the ground state of the system which we can take to have energy zero. (Actually, you can take it to have any energy you like; its contribution will disappear from our analysis when we divide by the normalisation factor that is missing on the right-hand-side of (2.73) and (2.74).) However, in the presence of the sources, the ground state of the system has energy $V(r)$. This means that we expect the Euclidean path integral to give

$$\lim_{r,T\to\infty}\Big\langle W[C]\Big\rangle \sim e^{-V(r)T}$$

This now gives us a way to test for the existence of the confining phase directly in Yang-Mills theory. If the theory lies in the confining phase, we should find

$$\lim_{r,T\to\infty}\Big\langle W[C]\Big\rangle \sim e^{-\sigma A[C]} \tag{2.75}$$

where $A[C]$ is the area of the the loop C. This is known as the *area law criterion* for confinement. We won't be able to prove that Wilson loops in Yang-Mills exhibit an area law, although we'll offer an attempt in Section 4.2 when we discuss the strong coupling expansion of lattice gauge theory. We will have more success in Section 7 and 8 when we demonstrate confinement in lower dimensional gauge theories.

If a theory does not lie in the confining phase, we get different behaviour for the Wilson loop. For example, we could add scalar fields which condense and completely break the gauge symmetry. This is the *Higgs phase*, and we will discuss it in more detail in Section 2.7 where we first introduce dynamical matter fields. In the Higgs phase, we have

$$\lim_{r,T\to\infty}\Big\langle W[C]\Big\rangle \sim e^{-\mu L}$$

where $L = 2(r+T)$ is the perimeter of the loop and μ is some mass scale associated to the energy in the fields that screen the particle. This kind of perimeter law is characteristic of the screening phase of a theory.

Wilson Loops as Operators

There is a slightly different perspective on Wilson loops that will also prove useful: we can view them as operators on the Hilbert space of states. Since we are now dealing with Hilbert spaces and states, it's important that we are back in Lorentzian signature.

In quantum field theory, states are defined as living on a spacelike slice of the system. For this reason, we should first rotate our Wilson loop so that C is a spacelike, closed curve, sitting at a fixed point in time. The interpretation of the operator $W[C]$ is that it adds to the state a loop of electric flux along C. To see this, we can again revert to the canonical formalism that we introduced in Section 2.2. The electric field is $E^i = -i\delta/\delta A_i(x)$, so we have

$$E^i W[C] = \operatorname{tr}\mathcal{P}\left(\left[\frac{\delta}{\delta A_i(x)}\oint_C A\right] W[C]\right)$$

which indeed has support only on C.

The expectation value $\langle W[C]\rangle$ is now interpreted as the amplitude for a loop of electric flux $W[C]|0\rangle$ to annihilate to the vacuum $\langle 0|$. In the confining phase, this is unlikely because the flux tube is locally stable. The flux tube can, of course, shrink over time and disappear, but that's not what $\langle W[C]\rangle$ is measuring. Instead, it's looking for the amplitude that the flux tube instantaneously disappears. This can happen only through a tunnelling effect which, in Euclidean space, involves a string stretched across the flux tube acting. This Euclidean action of this string is proportional to its area, again giving $\langle W[C]\rangle \sim e^{-\sigma A}$ with $A[C]$ the minimal area bounding the curve.

In contrast, in the Higgs phase the string is locally unstable. Each part of the string can split into pieces and dissolve away. This is still unlikely: after all, it has to happen at all parts of the string simultaneously. Nonetheless, it is more likely than the corresponding process in the confining phase, and this is reflected in the perimeter law $\langle W[C]\rangle \sim e^{-\mu L}$.

2.6 Magnetic Probes

Much of our modern understanding of gauge theories comes from the interplay between electric and magnetic degrees of freedom. In the previous section we explored how Yang-Mills fields respond to electric probes. In this section, we will ask how they respond to magnetic probes.

A warning: the material in this section is a little more advanced than what we covered until now and won't be required for much of what follows. (An exception is Section 3.6 which discusses discrete anomalies and builds on the machinery we develop here.) In particular, sections 2.7 and 2.8 can both be read without reference to this section.

2.6.1 't Hooft Lines

Our first task is to understand how to construct an operator that corresponds to the insertion of a magnetic monopole. These are referred to as *'t Hooft lines*. For electric probes, we could build the corresponding Wilson line out of local fields A_μ. But there are no such fields that couple to magnetic charges. This means that we need to find a different way to describe the magnetic probes.

We will achieve this by insisting that the fields of the theory have a prescribed singular behaviour on a given locus which, in our case, will be a line C in spacetime. Because such operators disrupt the other fields in the theory, they are sometimes referred to as *disorder operators*.

't Hooft Lines in Electromagnetism

To illustrate this idea, we first describe 't Hooft lines in $U(1)$ electromagnetism. We have already encountered magnetic monopoles in Section 1.1. Suppose that a monopole of charge m traces out a worldline C in $\mathbf{R}^{3,1}$. (We referred to magnetic charge as g in Section 1.1, but this is now reserved for the Yang-Mills coupling so we have to change notation.) For any $\mathbf{S}^2$ that surrounds C, we then have

$$\int_{\mathbf{S}^2} \mathbf{B}\cdot d\mathbf{S} = m \tag{2.76}$$

We normalise the $U(1)$ gauge field to have integer electric charges. As explained in Section 1.1, the requirement that the monopole is compatible with these charges gives the Dirac quantisation condition (1.3), which now reads

$$e^{im} = 1 \quad \Rightarrow \quad m \in 2\pi\mathbf{Z} \tag{2.77}$$

For the magnetic field to carry flux (2.76), we must impose singular boundary conditions on the gauge field. As an example, suppose that we take the line C to sit at the spatial origin $\mathbf{x}=0$ and extend in the temporal direction t. Then, as explained in Section 1.1 we can cover the $\mathbf{S}^2$ by two charts. Working in polar coordinates with $A_r = 0$ gauge, in the northern hemisphere, we take the gauge field to have the singular behaviour

$$A_\phi \to \frac{m(1-\cos\theta)}{2r\,\sin\theta} \quad \text{as} \quad r\to 0$$

There is a similar condition (1.7) in the southern hemisphere, related by a gauge transformation.

We now define the 't Hooft line $T[C]$ by requiring that we take the path integral only over fields subject to the requirement that they satisfy (2.76) on C. This is a rather unusual definition of an "operator" in quantum field theory. Nonetheless, despite its unfamiliarity, we can – at least in principle – use it to compute correlation functions of $T[C]$ with other, more traditional operators.

't Hooft Lines in Yang-Mills

What's the analogous object in Yang-Mills theory with gauge group G. To explain the generalisation of Dirac quantisation to an arbitrary, semi-simple Lie group we need to invoke a little bit of Lie algebra-ology that was covered in the *Symmetries and Particles* course.

We work with a Lie algebra $\mathfrak{g}$. We denote the Cartan sub-algebra as $\mathbf{H} \subset \mathfrak{g}$. Recall that this is a set of r mutually commuting generators, where r is the rank of the Lie algebra. Throughout the rest of this section, bold (and not silly gothic) font will denote an r-dimensional vector.

We again define a 't Hooft line for a timelike curve C sitting at the origin. We will require that the magnetic field B^i, $i = 1, 2, 3$, takes the form

$$B^i \to \frac{x^i}{4\pi r^3} Q(x) \quad \text{as } r \to 0$$

where $Q(x)$ is a Lie algebra valued object which specifies the magnetic charge of the 't Hooft line. Spherical symmetry requires that $Q(x)$ be covariantly constant. We can again cover the $\mathbf{S}^2$ with two charts, and in each pick $Q(x)$ to be a constant which, by a suitable gauge transformation, we take to sit in the Cartan subalgebra. We write

$$Q = \mathbf{m} \cdot \mathbf{H}$$

for some r-dimensional vector $\mathbf{m}$ which determines the magnetic charge. We can think of this as r Dirac monopoles, embedded in the Cartan subalgebra.

The requirement that the 't Hooft lines are consistent in the presence of Wilson lines gives the generalised Dirac quantisation condition,

$$\exp\left(i\mathbf{m} \cdot \mathbf{H}\right) = 1 \tag{2.78}$$

The twist is that this must hold for all representations of the Lie algebra. To see why this requirement affects the allowed magnetic charges, consider the case of $G = SU(2)$.

We can pick a $U(1) \subset SU(2)$ in which we embed a Dirac monopole of charge m. The W-bosons have electric charge $q = \pm 1$ and are consistent with a 't Hooft line of charge $m = 2\pi$. However, our 't Hooft line should also be consistent with the insertion of a Wilson line in the fundamental representation, and this carries charge $q = \pm 1/2$. This means that, for $G = SU(2)$, the 't Hooft line must carry $m = 4\pi$, twice the charge of the simplest Dirac monopole.

To extend this to a general group and representation, we need the concept of *weights*. Given a d dimensional representation, $|\mu_a\rangle$ with $a = 1, \ldots, d$ of $\mathfrak{g}$, we may introduce a set of weights, which are the eigenvalues

$$\mathbf{H}|\mu_a\rangle = \boldsymbol{\mu}_a|\mu_a\rangle \tag{2.79}$$

All such weights span the *weight lattice* $\Lambda_w(\mathfrak{g})$.

The weights of the adjoint representation are special and are referred to as *roots*. Recall that these roots $\boldsymbol{\alpha}$ can be used to label the other generators of the Lie algebra, which are denoted as $E_{\boldsymbol{\alpha}}$. In the adjoint representation, the eigenvalue condition (2.79) becomes the commutation relation $[\mathbf{H}, E_{\boldsymbol{\alpha}}] = \boldsymbol{\alpha} E_{\boldsymbol{\alpha}}$. Importantly, the roots also span a lattice

$$\Lambda_{\text{root}}(\mathfrak{g}) \subset \Lambda_w(\mathfrak{g})$$

The weights and roots have the property that

$$\frac{\boldsymbol{\alpha} \cdot \boldsymbol{\mu}}{\boldsymbol{\alpha}^2} \in \frac{1}{2}\mathbf{Z}$$

for all $\boldsymbol{\mu} \in \Lambda_w(\mathfrak{g})$ and $\boldsymbol{\alpha} \in \Lambda_{\text{root}}(\mathfrak{g})$. This is exactly what we need to solve the Dirac quantisation condition (2.78), which becomes $\mathbf{m} \cdot \boldsymbol{\mu} \in 2\pi\mathbf{Z}$ for all $\boldsymbol{\mu} \in \Lambda_w(\mathfrak{g})$. We define the *co-root*

$$\boldsymbol{\alpha}^\vee = \frac{2\boldsymbol{\alpha}}{\boldsymbol{\alpha}^2}$$

These co-roots also span a lattice, which we call $\Lambda_{\text{co-root}}(\mathfrak{g})$. Clearly, we have $\boldsymbol{\alpha}^\vee \cdot \boldsymbol{\mu} \in \mathbf{Z}$ for all $\boldsymbol{\alpha}^\vee \in \Lambda_{\text{co-root}}(\mathfrak{g})$ and $\boldsymbol{\mu} \in \Lambda_w(\mathfrak{g})$. If the magnetic charge vector sits in the co-root lattice, then the Dirac quantisation condition is obeyed. More generally, it turns out that for simply connected groups we have

$$\mathbf{m} \in 2\pi\, \Lambda_{\text{co-root}}(\mathfrak{g}) \tag{2.80}$$

This is sometimes referred to as the Goddard-Nuyts-Olive (or GNO) quantisation condition. We will look at the possible magnetic charges for non-simply connected groups shortly.

There is one last part of this story. The co-root lattice can be viewed as the root lattice for a Lie algebra $\mathfrak{g}^\vee$, so that $\Lambda_{\text{co-root}}(\mathfrak{g}) = \Lambda_{\text{root}}(\mathfrak{g}^\vee)$ For simply laced algebras (these are the ADE series, and so includes $su(N)$), all roots have the same length and are normalised to $\boldsymbol{\alpha}^2 = 2$. In this case, the roots and co-roots are the same and $\mathfrak{g}^\vee = \mathfrak{g}$. For non-simply laced groups, the long and short roots get exchanged. This means that, for example, $so(2N+1)^\vee = sp(N)$ and $sp(N)^\vee = so(2n+1)$.

2.6.2 $SU(N)$ vs $SU(N)/\mathbf{Z}_N$

There seems to be something of an imbalance between the Wilson line operators and the 't Hooft line operators. Of course, these electric and magnetic probes are defined in rather different ways, but that's not our concern. Instead, it's slightly disconcerting that there are more Wilson line operators than 't Hooft line operators. This is because Wilson line operators are labelled by representations R which, in turn, are associated to elements of the weight lattice $\Lambda_w(\mathfrak{g})$. In contrast, 't Hooft lines are labelled by elements of $\Lambda_{\text{root}}(\mathfrak{g}^\vee)$ which is a subset of $\Lambda_w(\mathfrak{g}^\vee)$. Roughly speaking, this means that Wilson lines can sit in any representation, including the fundamental, while 't Hooft lines can only sit in representations that arise from tensor products of the adjoint. Why?

To better understand the allowed magnetic probes, we need to look more closely at the global topology of the gauge group. We will focus on pure Yang-Mills with $G = SU(N)$. Because the gauge bosons live in the adjoint representation, they are blind to any transformation which sits in the centre $\mathbf{Z}_N \subset SU(N)$,

$$\mathbf{Z}_N = \left\{ e^{2\pi i k N},\ k = 0, 1, \ldots, N-1 \right\}$$

The gauge bosons do not transform under this centre $\mathbf{Z}_N$ subgroup. In the older literature, it is sometimes claimed that the correct gauge group of Yang-Mills is actually $SU(N)/\mathbf{Z}_N$. But this is a bit too fast. In fact, the right way to proceed is to understand that there are two different Yang-Mills theories, defined by the choice of gauge group

$$G = SU(N) \quad \text{or} \quad G = SU(N)/\mathbf{Z}_N$$

Indeed, more generally we have a different theory with gauge group $G = SU(N)/\mathbf{Z}_p$ for any $\mathbf{Z}_p$ subgroup of $\mathbf{Z}_N$. The difference between these theories is rather subtle. We can't distinguish them by looking at the action, since this depends only on the shared $su(N)$ Lie algebra. Moreover, this means that the correlation functions of all local operators are the same in the two theories so you don't get to tell the difference by doing any local experiments. Nonetheless, different they are. The first place this shows up is in the kinds of operators that we can use to probe the theory.

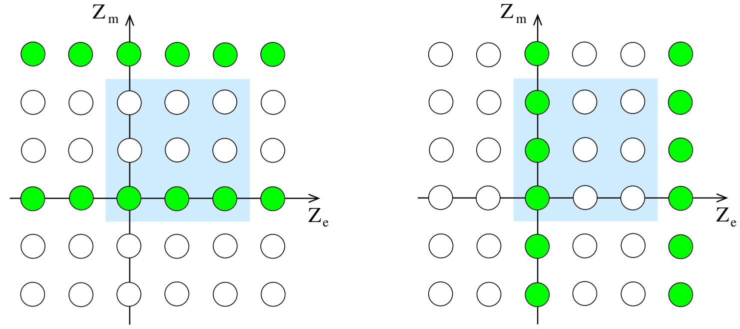


Figure 21: The figure on the left shows the allowed Wilson and 't Hooft lines (in green) for the gauge group $SU(3)$. The figure on the right shows the allowed lines for gauge group $SU(3)/\mathbf{Z}_3$.

Let's start with the Wilson lines. As we saw in Section 2.5, these are labelled by a representation of the group. The representations of $G = SU(N)/\mathbf{Z}_N$ are a subset of those of $G = SU(N)$; any representation that transforms non-trivially under $\mathbf{Z}_N$ is prohibited. This limits the allowed Wilson lines. In particular, the theory with $G = SU(N)/\mathbf{Z}_N$ does not admit the Wilson line in the fundamental representation, but Wilson lines in the adjoint representation are allowed. Similarly, the theory with gauge group $G = SU(N)/\mathbf{Z}_N$ cannot be coupled to fundamental matter; it can be coupled to adjoint matter.

This has a nice description in terms of the lattices that we introduced. For $G = SU(N)$, the representations are labelled by the weight lattice $\Lambda_w(\mathfrak{g})$. (The precise statement is that there is a one-to-one correspondence between representations and $\Lambda_w(\mathfrak{g})/W$ where W is the Weyl group.) However, for $G = SU(N)/\mathbf{Z}_N$, the representations are labelled by the root lattice $\Lambda_{\rm root}(\mathfrak{g})$. Indeed, the difference between the weight and root lattice for $\mathfrak{g} = su(N)$ is precisely the centre,

$$\Lambda_w(\mathfrak{g})/\Lambda_{\rm root}(\mathfrak{g}) = \mathbf{Z}_N$$

Now we come to the 't Hooft lines. When we introduced 't Hooft lines in the previous section, we were implicitly working with the universal cover of the gauge group, so that all possible Wilson lines were allowed. The requirement that magnetic charges are compatible with all representations and, in particular, the fundamental representation, resulted in the GNO condition (2.80) in which 't Hooft lines are labelled by $\Lambda_{\rm root}(\mathfrak{g})$. But what if we work with $G = SU(N)/\mathbf{Z}_N$? Now we have fewer Wilson lines, and so the demands of Dirac quantisation are less onerous. Correspondingly, in this theory the 't Hooft lines are labelled by $\Lambda_w(\mathfrak{g})$.

We can summarise the situation by labelling any line operator by a pair of integers

$$(z^e, z^m) \in \mathbf{Z}_N^e \times \mathbf{Z}_N^m \tag{2.81}$$

These describe how a given line operator transforms under the electric and magnetic centres of the group. If we have two line operators, labelled by (z^e, z^m) and (z'^e, z'^m) then Dirac quantisation requires $z^e z'^m - z^m z'^e = 0 \bmod N$. Note the similarity with the quantisation condition on dyons (1.4) that we met earlier.

For gauge group $G = SU(N)$, the line operators are labelled by $(z^e, 0)$ with $z^e = 0, \ldots, N-1$. Note that this doesn't mean that there are no magnetically charged 't Hooft lines: just that these lines sit in the root lattice and so have $z^m = 0 \bmod N$.

In contrast, for $G = SU(N)/\mathbf{Z}_N$ the line operators are labelled by $(0, z^m)$ with $z^m = 0, \ldots, N-1$. This time the Wilson lines must transform trivially under the centre of the group, so $z^e = 0 \bmod N$. The resulting line operators for $G = SU(3)$ and $G = SU(3)/\mathbf{Z}_3$ are shown in Figure 21. Yang-Mills with $G = SU(N)$ has more Wilson lines; Yang-Mills with $G = SU(N)/\mathbf{Z}_N$ has more 't Hooft lines.

There is a slightly more sophisticated way of describing these different line operators using the idea of generalised symmetries. We postpone this discussion until Section 3.6 where we will find an application in discrete anomalies.

The Theta Angle and the Witten Effect

The Witten effect gives rise to an interesting interplay between 't Hooft lines and the theta angle of Yang-Mills. Recall from Section 1.2.3, that a Dirac monopole of charge m in Maxwell theory picks up an electric charge proportional to the θ angle, given by

$$q = \frac{\theta m}{2\pi}$$

This analysis carries over to 't Hooft lines in both Maxwell and Yang-Mills theories. In the latter case, a shift of $\theta \to \theta + 2\pi$ changes the electric charge carried by a line operator,

$$\theta \to \theta + 2\pi \quad \Rightarrow \quad (z^e, z^m) \to (z^e + z^m, z^m)$$

For $G = SU(N)$, this maps the spectrum of line operators back to itself. However, for $G = SU(N)/\mathbf{Z}_N$ there is something of a surprise, because after a shift by 2π, the spectrum of line operators changes. This is shown in Figure 22 for $G = SU(3)/\mathbf{Z}_3$. We learn that the theory is *not* invariant under a shift of $\theta \to \theta + 2\pi$. Instead, to return to

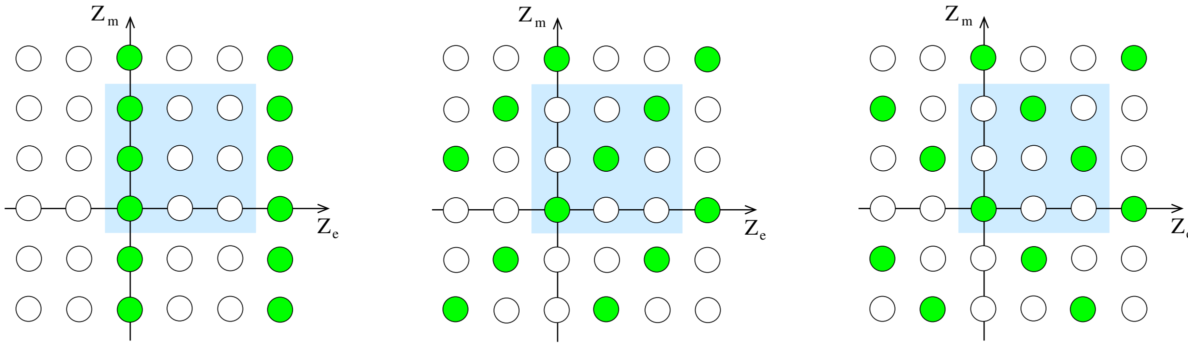


Figure 22: The spectrum of dyonic line operators in gauge group $SU(3)/\mathbf{Z}_3$, shown for $\theta = 0$ (on the left), $\theta = 2\pi$ (in the middle) and $\theta = 4\pi$ (on the right).

our original theory, with the same line operators, we must send $\theta \rightarrow \theta + 2\pi N$. In other words,

$$G = SU(N) \text{ has } \theta \in [0, 2\pi) \quad , \quad G = SU(N)/\mathbf{Z}_N \text{ has } \theta \in [0, 2\pi N)$$

We'll explore some consequences of this in Section 3.6 when we discuss anomalies in discrete symmetries.

One of the arguments we gave in Section 2.2 for the periodicity $\theta \in [0, 2\pi)$ was the appropriate quantisation of the topological charge $\int d^4x \; \mathrm{tr} \, {}^\star F^{\mu\nu} F_{\mu\nu}$. Instantons provide solutions to the equations of motion with non-vanishing topological charge. For Yang-Mills with $G = SU(N)/\mathbf{Z}_N$, the enlarged range of θ suggests that there might be "fractional instantons", configurations that carry $1/N^{\text{th}}$ the charge of an instanton. In fact, there are no such non-singular configurations on $\mathbf{R}^4$. But these fractional instantons do arise on manifolds with non-trivial topology. For example, if we take Euclidean spacetime to be $\mathbf{T}^4$, we can impose twisted boundary conditions in which, upon going around any circle, gauge fields come back to themselves up to a gauge transformation which lies in the centre $\mathbf{Z}_N$. Such boundary conditions are allowed for gauge group $G = SU(N)/\mathbf{Z}_N$, but not for $G = SU(N)$. One can show that these classes of configurations carry the requisite fractional topological charge.

't Hooft Lines as Order Parameters

One of the primary motivations for introducing line operators is to find order parameters that will distinguish between different phases of the theory. When $G = SU(N)$ we have the full complement of Wilson lines. As we saw in Section 2.5, an area law for the fundamental Wilson loop signals that the theory lies in the confining phase, which is the expected behaviour for pure Yang-Mills. If we also add scalar fields to the theory, these could condense so that we sit in the Higgs phase; in this case the Wilson loop exhibits a perimeter law.

If the gauge group is $G = SU(N)/\mathbf{Z}_N$, we no longer have the fundamental Wilson line at our disposal. Instead, we have the fundamental 't Hooft line with $z^m = 1$, and this now acts as our order parameter. Since the local dynamics is independent of the global topology of the gauge group, pure Yang-Mills theory is again expected to confine. But, as in our discussion of superconductors in Section 2.5.2, the confinement of electric charge is equivalent to the screening of magnetic charge. This means that the signature of electric confinement is now a perimeter law for the 't Hooft line.

We can also consider $G = SU(N)/\mathbf{Z}_N$ Yang-Mills in the Higgs phase. The theory does not admit scalar fields in the fundamental representation, so we introduce adjoint scalars which subsequently condense. A single adjoint scalar will break the gauge group to its maximal torus, $U(1)^{N-1}$, but with two misaligned adjoint Higgs fields we can break the gauge symmetry completely. This is the Higgs phase. As described in Section 2.5.2, the Higgs phase can be thought of as confinement of magnetic charges. Correspondingly, the 't Hooft line now exhibits an area law.

That's All Well and Good, but...

The difference between Yang-Mills with $G = SU(N)$ and $G = SU(N)/\mathbf{Z}_N$ seems rather formal. As we mentioned above, all correlation functions of local operators in the two theories coincide, which means that any local experiment that we can perform will agree. The theories only differ in the kinds of non-local probes that we can introduce. You might wonder whether this is some pointless intellectual exercise.

If we consider Yang-Mills on flat $\mathbf{R}^{3,1}$, then there is some justification in ignoring these subtleties: the physics of the two theories is the same, and we're just changing the way we choose to describe it. However, even in this case these subtleties will help us say something non-trivial about the dynamics as we will see in Section 3.6 when we discuss discrete anomalies.

The real differences between the two theories arise when we study them on background manifolds with non-trivial topology. Here the two theories can have genuinely different dynamics. Perhaps the most straightforward case arises for Yang-Mills coupled to a single, massless adjoint Weyl fermion. This theory turns out to have supersymmetry and goes by the name of $\mathcal{N} = 1$ super Yang-Mills. Although supersymmetry is beyond the scope of these lectures, it turns out that it provides enough of a handle for us to make quantitative statements about their dynamics. If we consider these theories on spacetime $\mathbf{R}^{2,1} \times \mathbf{S}^1$, the low energy dynamics, specifically the number of ground states, does depend on the global topology of the gauge group.

2.6.3 What is the Gauge Group of the Standard Model?

We all know the answer to the question in the heading. The gauge group of the Standard Model is

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

Or is it?

The fermions in a single generation sit in the following representations of G,

$$\begin{array}{rrlcl} \text{Leptons:} & l_L: & (\mathbf{2},\mathbf{1})_{-3} & \Rightarrow & (z_2^e, z_3^e)_Y = (1,0)_{-3} \\ & e_R: & (\mathbf{1},\mathbf{1})_{-6} & \Rightarrow & (z_2^e, z_3^e)_Y = (0,0)_{-6} \\ \text{Quarks:} & q_L: & (\mathbf{2},\mathbf{3})_{+1} & \Rightarrow & (z_2^e, z_3^e)_Y = (1,1)_{+1} \\ & u_R: & (\mathbf{1},\mathbf{3})_{+4} & \Rightarrow & (z_2^e, z_3^e)_Y = (0,1)_{+4} \\ & d_R: & (\mathbf{1},\mathbf{3})_{-2} & \Rightarrow & (z_2^e, z_3^e)_Y = (0,1)_{-2} \end{array}$$

where the subscript denotes $U(1)_Y$ hypercharge Y, normalised so that $Y \in \mathbf{Z}$. We could add to this the right-handed neutrino ν_R which is a gauge singlet. In the table above, we have also written the charges z_2^e and z_3^e under the $\mathbf{Z}_2 \times \mathbf{Z}_3$ centre of $SU(2) \times SU(3)$. Finally, the Higgs boson sits in the representation $(\mathbf{2},\mathbf{1})_3 \;\Rightarrow\; (z_2^e, z_3^e)_Y = (1,0)_3$.

Each of these representations has the property that

$$Y = 3z_2^e - 2z_3^e \mod 6$$

This means that there is a $\mathbf{Z}_6$ subgroup of $G = U(1)_Y \times SU(2) \times SU(3)$ under which all the fields are invariant: we must simultaneously act with the $\mathbf{Z}_6 = \mathbf{Z}_2 \times \mathbf{Z}_3$ centre of $SU(2) \times SU(3)$, together with a $\mathbf{Z}_6 \subset U(1)_Y$. Because nothing transforms under this $\mathbf{Z}_6$ subgroup, you can sometimes read in the literature that the true gauge group of the Standard Model is

$$G = \frac{U(1)_Y \times SU(2) \times SU(3)}{\Gamma} \tag{2.82}$$

where $\Gamma = \mathbf{Z}_6$. But this is also too fast. The correct statement is that there is a fourfold ambiguity in the gauge group of the Standard Model: it takes the form (2.82), where Γ is a subgroup of $\mathbf{Z}_6$, i.e.

$$\Gamma = 1,\ \mathbf{Z}_2,\ \mathbf{Z}_3,\ \text{or } \mathbf{Z}_6$$

We note in passing that we can embed the Standard Model in a grand unified group, such as $SU(5)$ or $Spin(10)$, only if $\Gamma = \mathbf{Z}_6$.

As we mentioned above, the choice of Γ does not affect any local correlation functions and, in particular, does not affect physics at the LHC. Nonetheless, each choice of Γ defines a different theory and, in principle, the distinction could have observable consequences. One place that the difference in Γ shows up is in the magnetic sector. Previously we discussed the allowed 't Hooft lines. However, there is a folk theorem that when a quantum field theory is coupled to gravity then any allowed electric or magnetic charge has a realisation as a physical state. In other words, particles (or groups of particles) should exist with each of the allowed electric and magnetic charges. We'll see in Section 2.8 how magnetic monopoles can arise as dynamical particles in a non-Abelian gauge theory.

The arguments for this are far from rigorous and, for magnetic charges, boil down to the fact that an attempt to define an infinitely thin 't Hooft line in a theory coupled to gravity will result in a black hole. If we now let this black hole evaporate, and insist that there are no remnants, then it should spit out a particle with the desired magnetic charge.

So what magnetic monopoles are allowed for each choice of Γ? First, let's recall how electromagnetism arises from the Standard Model. The electromagnetic charge q of any particle is related to the hypercharge Y and the $SU(2)$ charge τ^3 by

$$q = -\frac{Y}{6} + \tau^3$$

This gives us the familiar electric charges: for the electron $q = -1$; for the up quark $q = +2/3$; and for the down quark $q = -1/3$.

We denote the magnetic charge under $U(1)_Y$ as m_Y. As we explained in Section 2.5.2, when a Higgs field condenses, many of the magnetically charged states are confined. In the Standard Model, those that survive must have

$$\frac{6m_Y}{2\pi} = z_2^m \mod 2$$

The magnetic charge under $U(1)_Y$ and $SU(2)$ then conspires so that these states are blind to the Higgs field. For such states, the resulting magnetic charge under electromagnetism is

$$m = 6m_Y$$

Now we're in a position to see the how the global structure of the gauge group affects the allowed monopole charge. Suppose that we take $\Gamma = 1$. Here, the monopoles must obey the Dirac quantisation condition with respect to each gauge group individually. This means that $m_Y \in 2\pi\mathbf{Z}$, and so the magnetic charge of any particle is quantised as $m \in 12\pi\mathbf{Z}$. This is six times greater than the magnetic charge envisaged by Dirac. Of course, Dirac only knew about the existence of the electron with charge $q = 1$. The quarks, together with the structure of the electroweak force, impose a more stringent constraint.

In contrast, if $\Gamma = \mathbf{Z}_6$, more magnetic charges are allowed. This is entirely analogous to the situation that we saw in the previous section. The Dirac quantisation condition now imposes a single constraint on the combined gauge charges from each factor of the gauge group,

$$3z_2^e z_2^m + 2z_3^e z_3^m - \frac{6Y m_Y}{2\pi} \in 6\mathbf{Z}$$

But this gives us more flexibility. Now we are allowed a magnetic monopole with $m_Y = \frac{1}{6} \times 2\pi$ provided that it also carries a magnetic charge under the other groups, $z_2^m = 1$ and $z_3^m = 1$. In other words, the Standard Model with $\Gamma = \mathbf{Z}_6$ admits the kind of magnetic monopole that Dirac would have expected, with $m = 2\pi$. Of course, this obeys Dirac quantisation with respect to the electron. But it also obeys Dirac quantisation with respect to the fractionally charged quarks because it carries a compensating non-Abelian magnetic charge.

2.7 Dynamical Matter

Until now, we have (mostly) focussed on pure Yang-Mills, without any additional, dynamical matter fields. It's time to remedy this. We will consider coupling either scalar fields, ϕ, or Dirac spinors ψ to Yang-Mills.

Each matter field must transform in a representation R of the gauge group G. In the Lagrangian, the information about our chosen representation is often buried in the covariant derivative, which reads

$$\mathcal{D}_\mu = \partial_\mu - iA_\mu^a T^a(R)$$

where $T^a(R)$ are the generators of the Lie algebra in the representation R. For scalar fields, the action is

$$S_{\text{scalar}} = \int d^4x\ \mathcal{D}_\mu\phi^\dagger\mathcal{D}^\mu\phi - V(\phi)$$

where $V(\phi)$ can include both mass terms and ϕ^4 interactions. For spinors, the action is

$$S_{\text{fermion}} = \int d^4x\ i\bar{\psi}\not{\mathcal{D}}\psi - m\bar{\psi}\psi$$

If we have both scalars and fermions then we can also include Yukawa interactions between them.

Our ultimate goal is to understand the physics described by non-Abelian gauge theories coupled to matter. What is the spectrum of excitations of these theories? How do these excitations interact with other? How does the system respond to various probes and sources? In this section, we will start to explore this physics.

2.7.1 The Beta Function Revisited

The first question we will ask is: how does the presence of these matter degrees of freedom affect the running of the gauge coupling $g^2(\mu)$? This is simplest to answer for massless scalars and fermions. Suppose that we have N_s scalars in a representation R_s and N_f Dirac fermions in a representation R_f. The 1-loop running of the gauge coupling is

$$\frac{1}{g^2(\mu)} = \frac{1}{g_0^2} - \frac{1}{(4\pi)^2}\left[\frac{11}{3}I(\text{adj}) - \frac{1}{3}N_s I(R_s) - \frac{4}{3}N_f I(R_f)\right]\log\left(\frac{\Lambda_{UV}^2}{\mu^2}\right) \tag{2.83}$$

This generalises the Yang-Mills beta function (2.56). Recall that the Dynkin indices $I(R)$ are group theoretic factors defined by the trace normalisations,

$$\text{tr}\, T^a(R)T^b(R) = I(R)\delta^{ab}$$

and we are working in the convention in which $I(F) = \frac{1}{2}$ for the fundamental (or minimal) representation of any group.

When a field has mass m, it contributes to the running of the coupling only at scales $\mu > m$, and decouples when $\mu < m$. There is a smooth crossover from one behaviour to the other at scales $\mu \sim m$, but the details of this will not be needed in these lectures.

Here we will briefly sketch the derivation of the running of the coupling, following Section 2.4.2. We will then look at some of the consequences of this result.

The Beta Function for Scalars

If we integrate out a massless, complex scalar field, we get a contribution to the effective action for the gauge field given by

$$S_{\text{eff}}[A] = \frac{1}{2g^2}\int d^4x\ \text{tr}F_{\mu\nu}F^{\mu\nu} + \text{Tr}\,\log(-\mathcal{D}^2)$$

But this is something we've computed before, since it is the same as the ghost contribution to the effective action. The only differences are that we get a plus sign instead of a minus sign, because our scalars are the sensible kind that obey spin statistics, and that we pick up the relevant trace coefficient $I(R)$, as opposed to $I(\text{adj})$ for the ghosts. We can then immediately import our results from Section 2.4.2 to get the scalar contribution in (2.83).

The Beta Function for Fermions

If we integrate out a massless Dirac fermion, we get a contribution to the effective action for the gauge field given by

$$S_{\text{eff}}[A] = \frac{1}{2g^2}\int d^4x\ \text{tr}F_{\mu\nu}F^{\mu\nu} - \log\det(i\not{\mathcal{D}})$$

To compute the determinant, it's useful to expand as

$$\begin{aligned}\det(i\not{\mathcal{D}}) &= \det{}^{1/2}(-\gamma^\mu\gamma^\nu\mathcal{D}_\mu\mathcal{D}_\nu)\\ &= \det{}^{1/2}\Big(\frac{1}{2}\{\gamma^\mu,\gamma^\nu\}\mathcal{D}_\mu\mathcal{D}_\nu - \frac{1}{2}[\gamma^\mu,\gamma^\nu]\mathcal{D}_\mu\mathcal{D}_\nu\Big)\\ &= \det{}^{1/2}\Big(-\mathcal{D}^2 + \frac{i}{4}[\gamma^\mu,\gamma^\nu]F_{\mu\nu}\Big)\end{aligned}$$

where, to go to the final line, we have used both the Clifford algebra $\{\gamma^\mu,\gamma^\nu\} = 2\delta^{\mu\nu}$, as well as the fact that $[\mathcal{D}_\mu,\mathcal{D}_\nu] = -iF_{\mu\nu}$. The contribution to the effective action is then

$$\begin{aligned}-\log\det(i\not{\mathcal{D}}) &= -\frac{1}{2}\text{Tr}\ \log\Big(-\mathcal{D}^2\mathbf{1}_4 + \frac{i}{4}[\gamma^\mu,\gamma^\nu]F_{\mu\nu}\Big)\\ &= -2\text{Tr}\log(-\mathcal{D}^2) + [\gamma^\mu,\gamma^\nu]F_{\mu\nu}\ \text{terms}\end{aligned}$$

Here the $\frac{1}{2}$ has changed into a 2 after tracing over the spinor indices. We're left having to compute the contribution from the $[\gamma^\mu,\gamma^\nu]F_{\mu\nu}$ terms. This is very similar in spirit to the extra term that we had to compute for the gauge fluctuations in Section 2.4.2. However, the difference in spin structure means that it differs from the gauge contribution by a factor of 1/2. The upshot is that we have

$$-\log\det(i\not{\mathcal{D}}) = -\frac{1}{2}\left[\frac{4}{3}-4\right]\frac{T(R)}{(4\pi)^2}\int\frac{d^4k}{(2\pi)^4}\ \text{tr}\left[\bar{A}_\mu(k)\bar{A}_\nu(-k)\right](k^\mu k^\nu - k^2\delta^{\mu\nu})\log\left(\frac{\Lambda_{UV}^2}{k^2}\right)$$

which gives the fermionic contribution to the running of the gauge coupling in (2.83). Note that, once again, contributions from the extra spin term (the -4) overwhelm the contribution from the kinetic term (the $+4/3$). But, because we are dealing with fermions, there is an overall minus sign. This means that fermions, like scalars, give a positive contribution to the beta function.

2.7.2 The Infra-Red Phases of QCD-like Theories

We will start by ignoring the scalars and considering non-Abelian gauge theories coupled to fermions. In many ways, this is the most subtle and interesting class of quantum field theories and we will devote Sections 3 and 5 to elucidating some of their properties. Here we start by giving a brief tour of what is expected from these theories.

Obviously, there are many gauge groups and representations that we could pick. We will restrict ourselves to gauge group $SU(N_c)$, where N_c is referred to as the number of *colours*. We will couple to this gauge field N_f Dirac fermions, each transforming in the fundamental representation of the gauge group. Here N_f is referred to as the number of *flavours*. We will further take the fermions to be massless, although we will comment briefly on what happens as they are given masses. This class of theories will be sufficient to exhibit many of the interesting phenomena that we care about. Moreover, this class of theories boasts QCD as one of its members (admittedly you should relax the massless nature of the quarks just a little bit.)

At one-loop, the running of the gauge coupling can be read off from (2.83)

$$\frac{1}{g^2(\mu)} = \frac{1}{g_0^2} - \frac{1}{(4\pi)^2}\left[\frac{11N_c}{3} - \frac{2N_f}{3}\right]\log\left(\frac{\Lambda_{UV}^2}{\mu^2}\right) \tag{2.84}$$

These theories exhibit different dynamics depending on the ratio N_f/N_c.

The Infra-Red Free Phase

Life is simplest when $N_f > 11N_c/2$. In this case, the contribution to the beta function from the matter overwhelms the contribution from the gauge bosons, and the coupling g^2 becomes weaker as we flow towards the infra-red. Such theories are said to be *infra-red free*. This means that, for once, we can trust the classical description at low energies, where we have weakly coupled massless gauge bosons and fermions.

The force between external, probe electric charges takes the form

$$V_{\text{electric}}(r) \sim \frac{1}{r\,\log(r\Lambda_{UV})}$$

which is Coulombesque, but dressed with the extra log term which comes from the running of the gauge coupling. This is the same kind of behaviour that we would get in (massless) QED. Meanwhile, the potential between two external magnetic charges takes the form

$$V_{\text{magnetic}}(r) \sim \frac{\log(r\Lambda_{UV})}{r}$$

The log in the numerator reflects the fact that magnetic charges experience a force proportional to $1/g^2$ rather than g^2.

When $N_f = 11N_c/2$, the one-loop beta function vanishes. To see the fate of the theory, we must turn to the two-loop beta function which we discuss below. It will turn out that the theory is again infra-red free.

These theories are ill-defined in the UV, where there is a Landau pole. However, it's quite possible that theories of these types arise as the low-energy limit of other theories.

The Conformal Window

Next, consider N_f just below $11N_c/2$. To understand the behaviour of the theory, we can look at the two-loop contribution to the beta function,

$$\beta(g) = \mu\frac{dg}{d\mu} = \beta_0 g^3 + \beta_1 g^5 + \ldots$$

with the one-loop beta function extracted from (2.84)

$$\beta_0 = \frac{1}{(4\pi)^2}\left(-\frac{11N_c}{3} + \frac{2N_f}{3}\right)$$

We won't compute the two-loop beta function here, but just state the result:

$$\beta_1 = \frac{1}{(16\pi^2)^2}\left(-\frac{34N_c^2}{3} + \frac{N_f(N_c^2-1)}{N_c} + \frac{10N_fN_c}{3}\right)$$

Note that $\beta_1 > 0$ as long as the number of flavours sits in the range $N_f > 34N_c^3/(13N_c^2 - 3)$. But $\beta_0 < 0$ provided $N_f < 11N_c/2$ and so we can play the one-loop beta function against the two-loop beta function, to find a non-trivial fixed point of the RG flow, at which $\beta(g_\star) = 0$. This is given by

$$g_\star^2 = -\frac{\beta_0}{\beta_1}$$

Importantly, for $N_f/N_c = 11/2 - \epsilon$, with ϵ small, we have $g_\star^2 \ll 1$ and the analysis above can be trusted. We learn that the low-energy physics is described by a weakly coupled field theory which, as a fixed point of RG, is invariant under scale transformations. This is known as the *Banks-Zaks fixed point*. There is a general expectation (although not yet a complete proof) that relativistic theories in $d = 3+1$ which are scale invariant are also invariant under a larger conformal symmetry.

At any such fixed point, the scale invariance is enough to ensure that both external magnetic and electric probes experience a Coulomb force

$$V(r) \sim \frac{1}{r}$$

Such a phase could be described as a non-Abelian Coulomb phase, comprised of massless gluons and fermions.

What happens if we now lower N_f with fixed N_c? The formal result above says that the fixed point remains (at least until $N_f \approx 34N_c^3/(13N_c^2 - 3)$ but the value of the coupling $g_\star^2$ gets larger so that we can no longer trust the analysis. In general, we expect there to be a conformal fixed point for

$$N_\star < N_f < \frac{11N_c}{2} \tag{2.85}$$

for some critical value $N_\star$. This range of N_f is referred to as the *conformal window*. The obvious question is: what is the value of $N_\star$?

We don't currently know the answer to this question. At the lower end of the conformal window, the theory is necessarily strongly coupled which makes it difficult to get a handle on the physics. There is evidence from numerical work that when $N_c = 3$ (which is the case for QCD) then the lower end of the conformal window sits somewhere in the window $N_\star \in [8, 12]$, and probably closer to the middle than the edges. One would also expect the conformal window to scale with N_c, so one could guess that $N_\star \approx 3N_c$ to $4N_c$. There are various arguments that give values of $N_\star$ in this range, but none of them are particularly trustworthy.

We've seen that there are a set of conformal fixed point, labelled by N_c and N_f in the range (2.85). We met such fixed points before in the course on Statistical Field Theory. In that context, we came across the powerful idea of *universality*: many different ultra-violet theories all flow to the same fixed point. This is responsible for the observation that all gases, regardless of their microscopic make-up, have exactly the same divergence in the heat capacity at their critical point. We could ask: is there a form of universality in gauge theories? In other words, can we write down two gauge theories which look very different in the ultra-violet, but nonetheless flow to the same infra-red fixed point?

We don't yet know of any examples of such universality in the QCD-like gauge theories that we discuss in these lectures, although this is most likely due to our ignorance. However, such examples are known in supersymmetric theories, which consist of gauge fields, scalars and fermions interacting with specific couplings. In that context, it is known that supersymmetric $SU(N_c)$ gauge theories coupled to N_f fundamental flavours flows to the same fixed point as $SU(N_f - N_c)$ gauge theory coupled to N_f flavours. (The latter flavours should also be coupled to a bunch of gauge neutral fields.) Furthermore, the two descriptions can be identified as electric and magnetic variables for the system. This phenomenon is known as *Seiberg duality*. However, it is a topic for a different course.

Confinement and Chiral Symmetry Breaking

What happens when $N_f \leq N_\star$ and we are no longer in the conformal window? The expectation is that for $N_f < N_\star$ the coupling is once again strong enough to lead to confinement, in the sense that all finite energy excitations are gauge singlets.

Most of the degrees of freedom will become gapped, with a mass that is set parametrically by $\Lambda_{QCD} = \mu e^{1/2\beta_0 g^2(\mu)}$. However, there do remain some massless modes. These occur because of the formation of a vacuum condensate

$$\langle \bar{\psi}_i \psi_j \rangle \sim \delta_{ij} \qquad i, j = 1, \ldots, N_f$$

This spontaneously breaks the global symmetry of the model, known as the chiral symmetry. The result is once again a gapless phase, but now with the massless fields arising as Goldstone bosons. We will have a lot to say about this phase. We will say it in Section 5.

For pure Yang-Mills, we saw in Section 2.5 that a Wilson line, $W[C] = \mathrm{tr}\,\mathcal{P}\exp\left(i\oint_C A\right)$ in the fundamental representation provides an order parameter for the confining phase, with the area law, $\langle W[C]\rangle \sim e^{-\sigma A}$, the signature of confinement. However, in the presence of dynamical, charged fundamental matter – whether fermions or scalars – this criterion is no longer useful. The problem is that, for a sufficiently long flux tube, it is energetically preferable to break the string by producing a particle-anti-particle pair from the vacuum. If the flux tube has tension σ and the particles have mass m, this will occur when the length exceeds $L > 2m/\sigma$. For large loops, we therefore expect $\langle W[C]\rangle \sim e^{-\mu L}$. This is the same behaviour that we previously argued for in the Higgs phase. To see how they are related, we next turn to theories with scalars.

2.7.3 The Higgs vs Confining Phase

We now consider scalars. These can do something novel: they can condense and spontaneously break the gauge symmetry. This is the *Higgs phase*.

Consider an $SU(N_c)$ gauge theory with N_s scalar fields transforming in the fundamental representation. If the scalars are massless, then the gauge coupling runs as

$$\frac{1}{g^2(\mu)} = \frac{1}{g_0^2} - \frac{1}{(4\pi)^2}\left[\frac{11N_c}{3} - \frac{N_s}{6}\right]\log\left(\frac{\Lambda_{UV}^2}{\mu^2}\right)$$

and, correspondingly, the coefficient of the one-loop beta function is

$$\beta_0 = \frac{1}{(4\pi)^2}\left(-\frac{11N_c}{3} + \frac{N_s}{6}\right)$$

For $N_s < 22N_c$, the coupling becomes strong at an infra-red scale, $\Lambda_{QCD} = \Lambda_{UV} e^{1/2\beta_0 g_0^2}$. It is thought that the theory confines and develops a gap at this scale. We expect no massless excitations to survive.

What now happens if we give a mass m^2 to the scalars? For $m^2 > 0$, we expect these to shift the spectrum of the theory, but not qualitatively change the physics. Indeed, for $m^2 \gg \Lambda_{QCD}^2$, we can essentially ignore the scalars at low-energies and we revert to pure Yang-Mills. The real interest comes when we have $m^2 < 0$ so that the scalars condense. What happens then?

Suppose that we take $m^2 \ll -\Lambda_{QCD}^2$. This means that the scalars condense at a scale where the theory is still weakly coupled, $g^2(|m|) \ll 1$, and we can trust our semi-classical analysis. If we have enough scalars to fully Higgs the gauge symmetry ($N_s \geq N_c - 1$ will do the trick), then all the gauge bosons and scalars again become massive.

It would seem that the Higgs mechanism and confinement are two rather different ways to give a mass to the gauge bosons. In particular, the Higgs mechanism is something that we can understand in a straightforward way at weak coupling while confinement is shrouded in strongly coupled mystery. Intuitively, we may feel that the Higgs phase is not the same as the confining phase. But are they really different?

The sharp way to ask this question is: does the theory undergo a phase transition as we vary m^2 from positive to negative? We usually argue for the existence of a phase transition by exhibiting an order parameter which has different behaviour in the two phases. For pure Yang-Mills, the signature for confinement is the area law for the Wilson loop. But, as we argued above, in the presence of dynamical fundamental matter the confining string can break, and the area law goes over to a perimeter law. But this is the expected behaviour in the Higgs phase. In the absence of an order parameter to distinguish between the confining and Higgs phases, it seems plausible that they are actually the same, and one can vary smoothly from one phase to another. To illustrate this, we turn to an example.

An Example: $SU(2)$ with Fundamental Matter

Consider $SU(2)$ gauge theory with a single scalar ϕ in the fundamental representation. For good measure, we'll also throw in a single fermion ψ, also in the fundamental representation. We take the action to be

$$S = \int d^4x \; -\frac{1}{2g^2}\mathrm{tr}\, F^{\mu\nu}F_{\mu\nu} + |\mathcal{D}_\mu\phi|^2 - \frac{\lambda}{4}(\phi^\dagger\phi - v^2)^2 + i\bar{\psi}\,\not{D}\psi + m\bar{\psi}\psi$$

Note that it's not possible to build a gauge invariant Yukawa interaction with the matter content available. We will look at how the spectrum changes as we vary v^2 from positive to negative.

Higgs Phase, $v^2 > 0$: When $v^2 \gg \Lambda_{QCD}$ we can treat the action semi-classically. To read off the spectrum in the Higgs phase, it is simplest to work in unitary gauge in which the vacuum expectation value takes the form $\langle\phi\rangle = (v, 0)$. We can further use the gauge symmetry to focus on fluctuations of the form $\phi = (v + \tilde{\phi}, 0)$ with $\tilde{\phi} \in \mathbf{R}$. You can think of the other components of ϕ as being eaten by the Higgs mechanism to give mass to the gauge bosons. The upshot is that we have particles of spin 0,1/2 and 1, given by

- A single, massive, real scalar $\tilde{\phi}$.
- Two Dirac fermions $\psi_i = (\psi_1, \psi_2)$. Since the $SU(2)$ gauge symmetry is broken, these no longer should be thought of as living in a doublet. As we vary the mass $m \in \mathbf{R}$, there is a point at which the fermions become massless. (Classically, this happens at $m = 0$ of course.)
- Three massive spin 1 W-bosons A^a_μ, with $a = 1, 2, 3$ labelling the generators of $su(2)$.

Confining Phase, $v^2 < 0$: When $v^2 < 0$, the scalar has mass $m^2 > 0$ and does not condense. Now we expect to be in the confining phase, in the sense that only gauge singlets have finite energy. We can list the simplest such states: we will see that they are in one-to-one correspondence with the spectrum in the Higgs phase

- A single, real scalar $\phi^\dagger\phi$. This is expected to be a massive excitation. If we were to evaluate this in the Higgs phase then, in unitary gauge, we have $\phi^\dagger\phi = v^2 + v\tilde{\phi} + \ldots$ and so the quadratic operator corresponds to the single particle excitation $\tilde{\phi}$, plus corrections.

  There are further scalar operators that we can construct, including $\mathrm{tr}\, F_{\mu\nu}F^{\mu\nu}$ and $\bar{\psi}\psi$. These have the same quantum numbers as $\phi^\dagger\phi$ and are expected to mix with it. In the confining phase, the lightest spin 0 excitation is presumably created by some combination of these.

- Two Dirac fermions. The first is $\Psi_1 = \phi^\dagger\psi$. The second comes from using the ϵ^{ij} invariant tensor of $SU(2)$, which allows us to build $\Psi_2 = \epsilon^{ij}\phi_i\psi_j$. If we expand these operators in unitary gauge in the Higgs phase, we have $\Psi_1 = v\psi_1 + \ldots$ and $\Psi_2 = v\psi_2 + \ldots$.

  It's now less obvious that each of these fermions becomes massless for some value of $m \in \mathbf{R}$, but it remains plausible. Indeed, one can show that this does occur. (A modern perspective is that the fermionic excitation is in a different topological phase for $m \gg 0$ and $m \ll 0$, ensuring a gapless mode as we vary the mass between the two.)

- Finally, we come to the spectrum of spin 1 excitations. Since we want these to be associated to gauge fields, we might be tempted to consider gauge invariant operators such as $\mathrm{tr}\, F^{\mu\nu}F_{\mu\nu}$, but this corresponds to a scalar glueball. Instead, we can construct three gauge invariant, spin 1 operators. We have the real operator $i\phi^\dagger\mathcal{D}_\mu\phi$, and the complex operator $\epsilon^{ij}\phi_i(\mathcal{D}_\mu\phi_j)$. In unitary gauge, these become $v^2A_\mu^3$ and $v^2(A_\mu^1 + iA_\mu^2)$ respectively.

This is a strongly coupled theory, so there may well be a slew of further bound states and these presumably differ between the Higgs and confining phases. Nonetheless, the matching of the spectrum suggests that we can smoothy continue from one phase to the other without any discontinuity. We conclude that, for this example, the Higgs and confining phases are actually the same phase.

Another Example: $SU(2)$ with an Adjoint Scalar

It's worth comparing what happened above with a slightly different theory in which we can distinguish between the two phases. We'll again take $SU(2)$, but this time with an *adjoint* scalar field ϕ. We'll also throw in a fermion ψ, but we'll keep this in the fundamental representation. The action is now

$$S = \int d^4x\ \ -\frac{1}{2g^2}\mathrm{tr}\left(F^{\mu\nu}F_{\mu\nu} + (\mathcal{D}_\mu\phi)^2\right) - \frac{\lambda}{4}\left(\mathrm{tr}\,\phi^2 - \frac{v^2}{2}\right)^2 + i\bar{\psi}\,\not{D}\psi + \lambda'\bar{\psi}\phi\psi + m\bar{\psi}\psi$$

where we've now also included a Yukawa coupling between the scalar and fermion.

Once again, we can look at whether there is a phase transition as we vary v^2. For $v^2 < 0$, the scalar field is massive and we expect the theory to be gapped and confine. Importantly, in this phase the spectrum contains only bosonic excitations. There are no fermions because it's not possible to construct a gauge invariant fermionic operator.

In contrast, when $v^2 > 0$ the scalar field will get an expectation value, breaking the gauge group $SU(2) \to U(1)$, resulting in a gapless photon. There are also now two fermionic excitations which carry charge $\pm\frac{1}{2}$. The spectrum now looks very different from the confining phase.

Clearly in this case the Higgs and confining phases are different. Yet, because we have fermions in the fundamental representation, we will still have dynamical breaking of the flux tube and so fundamental Wilson loop $W[C]$ does not provide an order parameter for confinement. Nonetheless, the existence of finite energy states which transform under the $\mathbf{Z}_2$ centre of $SU(2)$ – which here coincides with $(-1)^F$, with F the fermion number – provides a diagnostic for the phase.

2.8 't Hooft-Polyakov Monopoles

Coupling dynamical, electrically charged particles to Yang-Mills theory is straightforward, although understanding their dynamics may not be. But what about dynamical magnetically charged particles?

For Abelian gauge theories, this isn't possible: if you want to include Dirac monopoles in your theory then you have to put them in by hand. But for non-Abelian gauge theories, it is a wonderful and remarkable fact that, with the right matter content, magnetic monopoles come along for free: they are solitons in the theory.

Magnetic monopoles appear whenever we have a non-Abelian gauge theory, broken to its Cartan subalgebra by an adjoint Higgs field. The simplest example is $SU(2)$ gauge theory coupled to a single adjoint scalar ϕ. As explained previously, we use the convention in which ϕ sits in the Lie algebra, so $\phi = \phi^a T^a$. For $G = SU(2)$ the generators are $T^a = \sigma^a/2$, with σ^a the Pauli matrices. We take the action to be

$$S = \int d^4x \; -\frac{1}{2g^2}\mathrm{tr} F^{\mu\nu}F_{\mu\nu} + \frac{1}{g^2}\mathrm{tr}(\mathcal{D}_\mu\phi)^2 \; - \frac{\lambda}{4}\left(\mathrm{tr}\,\phi^2 - \frac{v^2}{2}\right)^2 \tag{2.86}$$

Note that we've rescaled the scalar ϕ so that it too has a $1/g^2$ sitting in front of it.

The potential is positive definite. The vacuum of the theory has constant expectation value $\langle\phi\rangle$. Up to a gauge transformation, we can take

$$\langle\phi\rangle = \frac{1}{2}\begin{pmatrix} v & 0 \\ 0 & -v \end{pmatrix} \tag{2.87}$$

This breaks the gauge group $SU(2) \to U(1)$. The spectrum consists of a massless photon – which, in this gauge, sits in the T^3 part of the gauge group — together with massive W-bosons and a massive scalar.

There are, however, more interesting possibilities for the expectation value. Any finite energy excitation must approach a configuration with vanishing potential at spatial infinity. Such configurations obey $\mathrm{tr}\,\phi^2 \to v^2/2$ as $|\mathbf{x}| \to \infty$. Decomposing the Higgs field into the generators of the Lie algebra, $\phi = \phi^a T^a$, $a = 1, 2, 3$, the requirement that the potential vanishes defines a sphere in field space,

$$\mathbf{S}^2 := \left\{ \phi : \phi^a \phi^a = v^2 \right\} \tag{2.88}$$

We see that for any finite energy configuration, we must specify a map which tells us the behaviour of the Higgs field asymptotically,

$$\phi : \mathbf{S}^2_\infty \mapsto \mathbf{S}^2$$

The fact that these maps fall into disjoint classes should no longer be a surprise: it's the same idea that we met in Sections 2.2 and 2.3 when discussing theta vacua and instantons, and again in Section 2.5.2 when discussing vortices. This time the relevant homotopy group is

$$\Pi_2(\mathbf{S}^2) = \mathbf{Z}$$

Given a configuration ϕ, the winding number is computed by

$$\nu = \frac{1}{8\pi v^3} \int_{\mathbf{S}^2_\infty} d^2 S_i \; \epsilon^{ijk} \epsilon_{abc} \phi^a \partial_j \phi^b \partial_k \phi^c \in \mathbf{Z} \tag{2.89}$$

In a sector with $\nu \neq 0$, the gauge symmetry breaking remains $SU(2) \to U(1)$. The difference is that now the unbroken $U(1) \subset SU(2)$ changes as we move around the asymptotic $\mathbf{S}^2_\infty$.

The next step is to notice that if the Higgs field has winding $\nu \neq 0$, then we must also turn on a compensating gauge field. The argument is the same as the one we saw for vortex strings. Suppose that we try to set $A_i = 0$. Then, the covariant derivatives are simply ordinary derivatives and, asymptotically, we have $(\mathcal{D}_i\phi)^2 = (\partial_i\phi)^2 \sim (\partial_\theta\phi)^2/r^2$, with ∂_θ denoting the (necessarily non-vanishing) variation as we move around the angular directions of the asymptotic $\mathbf{S}^2_\infty$. The energy of the configuration will then include the term

$$E = \frac{1}{g^2} \int d^3x \; \mathrm{tr}\,(\partial_i \phi)^2 \sim \frac{1}{g^2} \int_{\mathbf{S}^2_\infty} d^2\Omega \int dr \; r^2 \, \mathrm{tr} \frac{(\partial_\theta\phi)^2}{r^2}$$

This integral diverges linearly. We learn that if we genuinely want a finite energy excitation in which the Higgs field winds asymptotically then we must also turn on the gauge fields A_i to cancel the $1/r$ asymptotic fall-off of the angular gradient terms, and ensure that $\mathcal{D}_\theta\phi \to 0$ as $r \to \infty$. We want to solve

$$\mathcal{D}_i\phi = \partial_i\phi - i[A_i, \phi] \to 0 \quad \Rightarrow \quad A_i \to \frac{i}{v^2}[\phi, \partial_i\phi] + \frac{a_i}{v}\phi$$

Here the first term works to cancel the fall-off from $\partial_i\phi$. To see this, you will need to use the fact that $\mathrm{tr}\,\phi^2 \to v^2$, and so $\mathrm{tr}(\phi\partial_i\phi) \to 0$, as well as the $su(2)$ commutation relations. The second term in A_i does not contribute to the covariant derivative $\mathcal{D}_i\phi$. The function a_i is the surviving, massless $U(1)$ photon which can be written in a gauge invariant way as

$$a_\mu = \frac{1}{v}\mathrm{tr}(\phi A_\mu) \tag{2.90}$$

We can also compute the asymptotic form of the field strength. The same kinds of manipulations above show that this lies in the same direction in the Lie algebra as ϕ,

$$F_{ij} = \frac{1}{v}\mathcal{F}_{ij}\,\phi$$

with

$$\mathcal{F}_{ij} = f_{ij} + \frac{i}{v^3}\mathrm{tr}\left(\phi\,[\partial_i\phi, \partial_j\phi]\right)$$

Here $f_{ij} = \partial_i a_j - \partial_j a_i$ is the Abelian field strength that we may have naively expected. But we see that there is an extra term, and this brings a happy surprise, since it contributes to the magnetic charge m of the $U(1)$ field strength. This is given by

$$m = -\int d^2S_i\ \frac{1}{2}\epsilon^{ijk}\mathcal{F}_{jk} = \frac{1}{2v^3}\int d^2S_i\ \epsilon^{ijk}\epsilon^{abc}\phi^a\partial_j\phi^b\partial_k\phi^c = 4\pi\nu \tag{2.91}$$

with ν the winding number defined in (2.89). We learn that any finite energy configuration in which the Higgs field winds asymptotically necessarily carries a magnetic charge under the unbroken $U(1) \subset SU(2)$. This object is a soliton and goes by the name of the *'t Hooft-Polyakov monopole.*

The topological considerations above have led us to a quantised magnetic charge. However, at first glance, the single 't Hooft-Polyakov monopole with $\nu = 1$ seems to have twice the charge required by Dirac quantisation (1.3), since the W-bosons have electric charge $q = 1$. But there is nothing to stop us including matter in the fundamental representation of $SU(2)$ with $q = \pm\frac{1}{2}$, with respect to which the 't Hooft-Polyakov monopole has the minimum allowed charge.

2.8.1 Monopole Solutions

We have not yet solved the Yang-Mills-Higgs equations of motion with a given magnetic charge. In general, no static solutions are expected to exist with winding $\nu > 1$, because magnetically charged objects typically repel each other. For this reason, we restrict attention to the configurations with winding $\nu = \pm 1$.

We can write an ansatz for a scalar field with winding $n = 1$,

$$\phi^a = \frac{x^a}{r^2} h(r) \quad \text{with } h(r) \to \begin{cases} 0 & r \to 0 \\ vr & r \to \infty \end{cases}$$

This is the so-called "hedgehog" ansatz, since the direction of the scalar field $\phi = \phi^a T^a$ is correlated with the direction x^a in space. Just like a hedgehog. In particular, this means that the $SU(2)$ gauge action on ϕ^a and the $SO(3)$ rotational symmetry on x^a are locked, so that only the diagonal combination are preserved by such configurations. We can make a corresponding ansatz for the gauge field which preserves the same diagonal $SO(3)$,

$$A_i^a = -\epsilon_{aij} \frac{x^j}{r^2} [1 - k(r)] \quad \text{with } k(r) \to \begin{cases} 1 & r \to 0 \\ 0 & r \to \infty \end{cases}$$

We can now insert this ansatz into the equations of motion

$$\mathcal{D}^\mu F_{\mu\nu} - i[\phi, \mathcal{D}_\nu \phi] = 0 \quad \text{and} \quad \mathcal{D}^2 \phi = 2g^2 \lambda (\operatorname{tr} \phi^2 - v^2) \phi \tag{2.92}$$

This results in coupled, ordinary differential equations for $h(r)$ and $k(r)$. In general, they cannot be solved analytically, but it is not difficult to find numerical solutions for the minimal 't Hooft-Polyakov monopole.

BPS Monopoles

Something special happens when we set $\lambda = 0$ in (2.86). Here the scalar potential vanishes which means that, at least classically, we can pick any expectation value v for the scalar. The choice of v should be thought of as extra information needed to define the vacuum of the theory. (In the quantum theory, one typically expects to generate a potential for ϕ. The exception to this is in supersymmetric theories, where cancellations ensure that the quantum potential also vanishes. Indeed, the monopoles that we describe below have a nice interplay with supersymmetry, although this is beyond the scope of these lectures.)

When the potential vanishes, it is possible to use the Bogomolnyi trick to rewrite the energy functional. In terms of the non-Abelian magnetic field $B_i = -\frac{1}{2}\epsilon_{ijk}F_{jk}$, the energy of a static configuration with vanishing electric field is

$$\begin{aligned} E &= \frac{1}{g^2}\int d^3x \; \mathrm{tr}\left(B_i^2 + (\mathcal{D}_i\phi)^2\right) \\ &= \frac{1}{g^2}\int d^3x \; \mathrm{tr}\left(B_i \mp \mathcal{D}_i\phi\right)^2 \pm 2\,\mathrm{tr}\, B_i\mathcal{D}_i\phi \\ &\geq \pm\frac{2}{g^2}\int d^3x \; \partial_i\, \mathrm{tr}\, B_i\phi \end{aligned}$$

where, to get to the last line, we have discarded the positive definite term and integrated by parts, invoking the Bianchi identity $\mathcal{D}_iB_i = 0$. We recognise the final expression as the magnetic charge. We find that the energy of a configuration is bounded by the magnetic charge

$$E \geq \frac{2v|m|}{g^2} \tag{2.93}$$

A configuration which saturates this bound is guaranteed to solve the full equations of motion. This is achieved if we solve the first order Bogomolnyi equations

$$B_i = \pm\mathcal{D}_i\phi \tag{2.94}$$

with the $\pm$ sign corresponding to monopoles (with $m > 0$) and anti-monopoles (with $m < 0$) respectively. It can be checked that solutions to (2.94) do indeed solve the full equations of motion (2.92) when $\lambda = 0$.

Solutions to (2.94) have a number of interesting properties. First, it turns out that the equations of motion for a single monopole have a simple analytic solution,

$$h(r) = vr\coth(vr) - 1 \quad \text{and} \quad k(r) = \frac{vr}{\sinh vr}$$

This was first discovered by Prasad and Sommerfield. In general, solutions to (2.94) are referred to as BPS monopoles, with Bogomolnyi's name added as well.

A warning on terminology: these BPS monopoles have rather special properties in the context of supersymmetric theories where they live in short multiplets of the supersymmetry algebra. The term "BPS" has since been co-opted and these days is much more likely to refer to some kind of protected object in supersymmetry, often one that has nothing to do with the monopole.

The Bogomolnyi equations (2.94) also have solutions corresponding to monopoles with higher magnetic charges. These solutions include configurations that look like far separated single charge monopoles. This is mildly surprising. Our earlier intuition told us that such solutions should not exist because the repulsive force between magnetically charged particles would ensure that the energy could be lowered by moving them further apart. That intuition breaks down in the Bogomolnyi limit because we have a new massless particle – the scalar ϕ – and this gives rise to a compensating attractive force between monopoles, one which precisely cancels the magnetic repulsion. You can learn much more about the properties of these solutions, and the role they play in supersymmetric theories, in the lectures on Solitons.

Monopoles in Other Gauge Groups

It is fairly straightforward to extend the discussion above to other gauge groups G. We again couple a scalar field ϕ in the adjoint representation and give it an expectation value that breaks $G \to H$ where $H = U(1)^r$, with r the rank of the gauge group.

Given an expectation value for ϕ, we can always rotate it by acting with G. However, by definition, H leaves the scalar untouched which means that configurations are now classified by maps from $\mathbf{S}^2_\infty$ into the space G/H. (In our previous discussion we had $G/H = SU(2)/U(1) = \mathbf{S}^2$ which coincides with what we found in (2.88).) A result in homotopy theory tells us that, for simply connected G,

$$\Pi_2(G/H) = \Pi_1(H) = \mathbf{Z}^r$$

We learn that the 't Hooft-Polyakov monopoles are labelled by an r-dimensional magnetic charge vector $\mathbf{m}$. This agrees with our analysis of 't Hooft lines in Section 2.6. A closer look reveals that the 't Hooft-Polyakov monopoles have magnetic charge $\mathbf{m} \in 2\pi\, \Lambda_{\rm co-root}(\mathfrak{g})$, as required by the Goddard-Nuyts-Olive quantisation (2.80).

2.8.2 The Witten Effect Again

We saw in Section 1.2.3 that, in the presence of a θ term, a Dirac monopole picks up an electric charge. As we now show this phenomenon, known as the Witten effect, also occurs for the 't Hooft-Polyakov monopole.

To see this, we simply need to be careful in identifying the electric charge operator in the presence of a monopole. We saw in (2.90) that the unbroken $U(1) \subset SU(2)$ is determined by the ϕ. The corresponding global gauge transformation is

$$\delta A_\mu = \frac{1}{v}\mathcal{D}_\mu \phi$$

But we already did the hard work and computed the Noether charge Q associated to such a gauge transformation in (2.30), where we saw that it picks up a contribution from the θ term (2.22); we have

$$Q = \frac{1}{g^2}\int d^3x \ \mathrm{tr}\left(E_i + \frac{\theta g^2}{8\pi^2}B_i\right)\frac{1}{v}\mathcal{D}_i\phi$$

In our earlier discussion, around equation (2.30), we were working in the vacuum and could discard the contribution from θ. However, in the presence of a monopole both terms contribute. The total electric charge Q is now

$$Q = q + \frac{\theta g^2 m}{8\pi^2} \tag{2.95}$$

with the naive electric charge q defined as

$$q = \frac{1}{v}\int d^3x \ \mathrm{tr}\,\mathcal{D}_i\phi\, E_i$$

and the magnetic charge m defined, as in (2.91), by

$$m = \frac{1}{v}\int d^3x \ \mathrm{tr}\,\mathcal{D}_i\phi\, B_i$$

We see that the theta term does indeed turn the monopole into a dyon. This agrees with our previous discussion of the Witten effect (1.19), with the seemingly different factor of 2 arising because, as explained above, q is quantised in units of $1/2$ in the non-Abelian gauge theory.

2.9 Further Reading

Trinity College, Cambridge boasts many great scientific achievements. The discovery of Yang-Mills theory is not among the most celebrated. Nonetheless, in January 1954 a graduate student at Trinity named Ronald Shaw wrote down what we now refer to as the Yang-Mills equations. Aware that the theory describes massless particles, which appear to have no place in Nature, Shaw was convinced by his supervisor, Abdus Salam, that the result was not worth publishing. It appears only as a chapter of his thesis [181].

Across the Atlantic, in Brookhaven National Laboratory, two office mates did not make the same mistake. C. N. Yang and Robert Mills constructed the equations which now bear their name [232]. It seems likely that they got the result slightly before Shaw, although the paper only appeared afterwards. Their original motivation now seems somewhat misguided: their paper suggests that global symmetries of quantum field theory – specifically $SU(2)$ isospin – are not consistent with locality. They write

> "It seems that this [global symmetry] is not consistent with the localized field concept that underlies the usual physical theories"

From this slightly shaky start, one of the great discoveries of 20th century physics emerged.

In those early days, the role played by Yang-Mills theory was, to say the least, confusing. Yang gave a famous seminar in Princeton in which Pauli complained so vociferously about the existence of massless particles that Yang refused to go on with the talk and had to be coaxed back to the blackboard by Oppenheimer. (Pauli had a headstart here: in 1953 he did a Kaluza-Klein reduction on $\mathbf{S}^2$, realising an $SU(2)$ gauge theory but discarding it because of the massless particle [151]. A similar result had been obtained earlier by Klein [122].)

It took a decade to realise that the gauge bosons could get a mass from the Higgs mechanism, and a further decade to realise that the massless particles were never really there anyway: they are an artefact of the classical theory and gain a mass automatically when $\hbar \neq 0$. Below is a broad brush description of this history. A collection of reminiscences, "50 Years of Yang-Mills" [108], contains articles by a number of the major characters in this story.

Asymptotic Freedom

As the 1970s began, quantum field theory was not in fashion. Fundamental laws of physics, written in the language of field theory, languished in the literature, unloved and uncited [77, 205]. The cool kids were playing with bootstraps.

The discovery of asymptotic freedom was one of the first results that brought field theory firmly into the mainstream. The discovery has its origins in the deep inelastic scattering experiments performed in SLAC in the late 1960s. Bjorken [19] and subsequently Feynman [56] realised that the experiments could be interpreted in terms of the momentum distribution of constituents of the proton. But this interpretation held only if the interactions between these constituents became increasingly weak at high energies. Feynman referred to the constituents as "partons" rather than "quarks" [57]. It is unclear whether this was because he wanted to allow for the possibility of other constituents, say gluons, or simply because he wanted to antagonise Gell-Mann.

In Princeton, David Gross set out to show that no field theory could exhibit asymptotic freedom [86]. Having ruled out field theories based on scalars and fermions, all that was left was Yang-Mills. He attacked this problem with his new graduate student

Frank Wilczek. The minus signs took some getting right, but by April 1973 they realised that they had an asymptotically free theory on their hands [83] and were keenly aware of its importance.

Meanwhile, in Harvard, Sidney Coleman was interested in the same problem. He asked his graduate student Erick Weinberg to do the calculation but, content that he had enough for his thesis, Erick passed it on to another graduate student, David Politzer. Politzer finished his calculation at the same time as the Princeton team [156]. In 2004, Gross, Politzer and Wilczek were awarded the Nobel prize. Politzer's Nobel lecture contains an interesting, and very human, account of the discovery [157].

In fact, both American teams had been scooped. In June 1972, at a conference in Marseilles, a Dutch graduate student named Gerard 't Hooft sat in a talk by Symanzik on the SLAC experiments and their relation to asymptotic freedom. After the talk, 't Hooft announced that Yang-Mills theory is asymptotically free. Symanzik encouraged him to publish this immediately but, like Shaw 20 years earlier, 't Hooft decided against it. His concern was that Yang-Mills theory could not be relevant for the strong force because it had no mechanism for the confinement of quarks [107].

The failure to publish did not hurt 't Hooft's career. By that stage he had already shown that Yang-Mills was renormalisable, a fact which played a large role in bringing the theory out of obscurity [93, 94, 95]. This was enough for him to be awarded his PhD [96]. It was also enough for him to be awarded the 1999 Nobel prize, together with his advisor Veltman. We will be seeing much more of the work of 't Hooft later in these lectures.

The analogy between asymptotic freedom and paramagnetism was made by N. K. Nielsen [148], although the author gives private credit to 't Hooft. In these lectures, we computed the one-loop beta function using the background field method. This method was apparently introduced by (of course) 't Hooft in lectures which I haven't managed to get hold of. It first appears in published form in a paper by Larry Abbott [1] (now a prominent theoretical neuroscientist) and is covered in the textbook by Peskin and Schroeder [154].

Confinement and the Mass Gap

Asymptotic freedom gave a dynamical reason to believe that Yang-Mills was likely responsible for the strong force. Earlier arguments that quarks should have three colour degrees of freedom meant that attention quickly focussed on the gauge group $SU(3)$ [84, 65]. But the infra-red puzzles still remained. Why are the massless particles predicted by Yang-Mills not seen? Why are individual quarks not seen?

Here things were murky. Was the $SU(3)$ gauge group broken by a scalar field? Or was it broken by some internal dynamics? Or perhaps the gauge group was actually unbroken but the flow to strong coupling does something strange. This latter possibility was mooted in a number of papers [84, 207, 208, 65]. This from Gross and Wilczek in 1973,

> "Another possibility is that the gauge symmetry is exact. At first sight this would appear ridiculous since it would imply the existence of massless, strongly coupled vector mesons. However, in asymptotically free theories these naive expectations might be wrong. There may be little connection between the "free" Lagrangian and the spectrum of states."

This idea was slowly adopted over the subsequent year. The idea of dimensional transmutation, in which dimensionless constants combine with the cut-off to give a physical scale, was known from the 1973 work of Coleman and E. Weinberg [27]. Although they didn't work with Yang-Mills, their general mechanism removed the most obvious hurdle for a scale-invariant theory to develop a gap. A number of dynamical explanations were mooted for confinement, but the clearest came only in 1974 with Wilson's development of lattice gauge theory [214]. This paper also introduced what we now call the Wilson line. We will discuss the lattice approach to confinement in some detail in Section 4.

The flurry of excitement surrounding these developments also serves to highlight the underlying confusion, as some of the great scientists of the 20th century clamoured to disown their best work. For example, in an immediate response to the discovery of asymptotic freedom, and six years after his construction of the electroweak theory [205], Steven Weinberg writes [208]

> "Of course, these very general results will become really interesting only when we have some specific gauge model of the weak and electromagnetic interactions which can be taken seriously as a possible description of the real world. This we do not yet have."

Not to be outdone, in the same year Gell-Mann offers [65]

> "We do not accept theories in which quarks are real, observable particles."

It's not easy doing physics.

Semi-Classical Yang-Mills

In these lectures, we first described the classical and semi-classical structure of Yang-Mills theory, and only then turned to the quantum behaviour. This is the logical way through the subject. It is not the historical way.

Our understanding of the classical vacuum structure of Yang-Mills theory started in 1975, when Belavin, Polyakov, Schwartz and Tyupkin discovered the Yang-Mills instanton [14]. Back then, Physical Review refused to entertain the name "instanton", so they were referred to in print as "pseudoparticles".

't Hooft was the first to perform detailed instanton calculations [101, 102], including the measure $K(\rho)$ that we swept under the carpet in Section 2.3.3. Among other things, his work clearly showed that physical observables depend on the theta angle. Motivated by this result, Jackiw and Rebbi [113], and independently Callan, Dashen and Gross [23], understood the semi-classical vacuum structure of Yang-Mills that we saw in Section 2.2.

Jackiw's lectures [115] give a very clear discussion of the theta angle and were the basis for the discussion here. Reviews covering a number of different properties of instantons can be found in [182, 191, 197].

Magnetic Yang-Mills

The magnetic sector of Yang-Mills theory was part of the story almost from the beginning. Monopoles in $SU(2)$ gauge theories were independently discovered by 't Hooft [99] and Polyakov [158] in 1974. The extension to general gauge groups was given in 1977 by Goddard, Nuyts and Olive [80]. This paper includes the GNO quantisation condition that we met in our discussion of 't Hooft lines, and offers some prescient suggestions on the role of duality in exchanging gauge groups. (These same ideas rear their heads in mathematics in the Langlands program.)

Bogomolnyi's Bogomolnyi trick was introduced in [20]. Prasad and Sommerfeld then solved the resulting equations of motion for the monopole [162], and the initials BPS are now engraved on all manner of supersymmetric objects which have nothing to do with monopoles. (A more appropriate name for BPS states would be Witten-Olive states [217].) Finally, Witten's Witten effect was introduced in [216]. Excellent reviews of 't Hooft-Polyakov monopoles, both with focus on the richer BPS sector, can be found in Harvey's lecture notes [89] and in Manton and Sutcliffe's book [133]. There are also some TASI lectures [191].

The Nielsen-Olesen vortex was introduced in 1973 [145]. Their motivation came from string theory, rather than field theory. The fact that such strings would confine magnetic monopoles was pointed out by Nambu [142] and the idea that this is a useful analogy for quark confinement, viewed in dual variables, was made some years later by Mandelstam [130] and 't Hooft [100].

The 't Hooft line as a magnetic probe of gauge theories was introduced in [103]. This paper also emphasises the importance of the global structure of the gauge group. A more modern perspective on line operators was given by Kapustin [120]. A very clear discussion of the electric and magnetic line operators allowed in different gauge groups, and the way this ties in with the theta angle, can be found in [4].

Towards the end of the 1970s, attention began to focus on more general questions of the phases of non-Abelian gauge theories [103, 104]. The distinction, or lack thereof, between Higgs and confining phases when matter transforms in the fundamental representation of the gauge group was discussed by Fradkin and Shenker [63] and by Banks and Rabinovici [9]; both rely heavily on the lattice. The Banks-Zaks fixed point, and its implications for the conformal window, was pointed out somewhat later in 1982 [10].

3. Anomalies

We learn as undergraduates that particles come in two types: bosons and fermions. Of these, the bosons are the more straightforward since they come back to themselves upon a 2π rotation. Fermions, however, return with a minus sign, a fact which has always endowed them with something of an air of mystery. In this section and the next, we will begin to learn a little more about the structure of fermions, and we will see the interesting and subtle phenomena that arise when fermions are coupled to gauge fields.

Our interest in this chapter lies with a phenomenon known as a *quantum anomaly*. In fact, there are a number of related phenomena that carry this name. For example, later, in Section 3.5, we will describe the so-called *'t Hooft anomaly* which can be viewed as an obstruction to gauging a global symmetry and, in many ways, this is the key idea that underlies this chapter. However, rather than jump straight in with this, we will instead build up more slowly. In doing so, our first introduction to an anomaly will be slightly different: we will start by describing an *anomaly* as a symmetry of the classical theory which does not survive to the quantum theory.

Stated in this way, we have already seen an example of an anomaly: classical Yang-Mills theory is scale invariant, but this is ruined in the quantum theory by the running of the coupling constant and the emergence of the scale Λ_{QCD}. In this section we will primarily be interested in anomalies associated to fermions. We will learn that these are intimately connected to various topological aspects of gauge theories and give rise to some surprising and beautiful phenomena.

3.1 The Chiral Anomaly: Building Some Intuition

Later in this chapter we will describe both the physical intuition and the detailed technical calculations that underly the anomaly. But we start here by describing, without proof, the key formula.

A particularly simple example of an anomaly arises when we have a massless Dirac fermion in $d = 3+1$ dimensions, coupled to an electromagnetic gauge field. The action for the fermion is

$$S = \int d^4x \; i\bar{\psi}\not{D}\psi \tag{3.1}$$

If the gauge field is dynamical, we would add to this the Maxwell action. Alternatively, we could think of the gauge field as a non-fluctuating background field, something fixed and under our control.

As we know from our first course on Quantum Field Theory, the action (3.1) has two global symmetries, corresponding to vector and axial rotations of the fermion. The first of these simply rotates the phase of ψ by a constant, $\psi \to e^{i\alpha}\psi$, with the corresponding current

$$j^\mu = \bar{\psi}\gamma^\mu\psi$$

The action (3.1) includes the coupling $A_\mu j^\mu$ of this current to the background gauge field. If we want the action to be invariant under gauge transformations $A_\mu \to A_\mu + \partial_\mu \alpha$ (and we do!) then its imperative that the current is conserved, so $\partial_\mu j^\mu = 0$. We'll see more about the interplay between anomalies and gauge symmetries in Section 3.4.

The other symmetry of (3.1) is the axial rotation, $\psi \to e^{i\alpha\gamma^5}\psi$, with associated current

$$j^\mu_A = \bar{\psi}\gamma^\mu\gamma^5\psi$$

In the classical theory, the standard arguments of Noether tells us that $\partial_\mu j^\mu_A = 0$. While this is true in the classical theory, it is not true in the quantum theory. Instead, it turns out that the divergence of the current is given by

$$\partial_\mu j^\mu_A = \frac{e^2}{16\pi^2}\epsilon^{\mu\nu\rho\sigma}F_{\mu\nu}F_{\rho\sigma} \tag{3.2}$$

where $F_{\mu\nu}$ is the electromagnetic field strength. This is known as the *chiral anomaly*. (It is sometimes called the *ABJ anomaly*, after Adler, Bell and Jackiw who first discovered it.) The anomaly tells us that in the presence of parallel electric and magnetic fields, the axial charge density can change.

Later in this section, we will derive (3.2). In fact, because it's important, we will derive it twice, using different methods. However, it's easy to get bogged down by complicated mathematics in this subject, so we will first try to build some intuition for why axial charge is not conserved.

3.1.1 Massless Fermions in Two Dimensions

Although our ultimate interest lies in four dimensional fermions (3.1), there is a slightly simpler example of the anomaly that arises for a Dirac fermion in $d = 1+1$ dimensions. (We'll see a lot more about physics in $d = 1+1$ dimensions in Section 7.) The Clifford algebra,

$$\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu} \qquad \mu,\nu = 0,1$$

with $\eta^{\mu\nu} = \mathrm{diag}(+1, -1)$ is satisfied by the two-dimensional Pauli matrices

$$\gamma^0 = \sigma^1 \quad \text{and} \quad \gamma^1 = i\sigma^2$$

The Dirac spinors are then two-component objects, ψ. The action for a massless spinor is

$$S = \int d^2x \; i\bar{\psi}\,\partial\!\!\!/\psi \tag{3.3}$$

Quantisation of this action will give rise to a particle and an anti-particle. Note that, in contrast to fermions in $d = 3+1$ dimensions, these particles have no internal spin. This is for the simple reason that there is no spatial rotation group in $d = 1+1$ dimensions.

We can write the action as

$$S = \int d^2x \; i\psi^\dagger\gamma^0(\gamma^0\partial_t + \gamma^1\partial_x)\psi = \int d^2x \; i\psi^\dagger(\partial_t - \gamma^5\partial_x)\psi \tag{3.4}$$

where

$$\gamma^5 = -\gamma^0\gamma^1 = -i\sigma^1\sigma^2 = \sigma^3$$

The name "γ^5" is slightly odd in this $d = 1+1$ dimensional context, but it is there to remind us that this matrix is analogous to the γ^5 that arises for four dimensional fermions. Just like in four-dimensions, we can decompose a massless Dirac fermion into chiral constituents, determined by its eigenvalue under γ^5. We write

$$\psi_\pm = \frac{1}{2}\left(1 \pm \gamma^5\right)\psi$$

With our choice of basis, the components are

$$\psi_+ = \begin{pmatrix} \chi_+ \\ 0 \end{pmatrix} \quad \text{and} \quad \psi_- = \begin{pmatrix} 0 \\ \chi_- \end{pmatrix}$$

Written in terms of chiral fermions, the action (3.4) then becomes

$$S = \int d^2x \; i\chi_+^\dagger\partial_-\chi_+ + i\chi_-^\dagger\partial_+\chi_- \tag{3.5}$$

with $\partial_\pm = \partial_t \pm \partial_x$. This tells us how to interpret chiral fermions in $d = 1+1$ dimensions. The equation of motion for χ_+ is $\partial_-\chi_+ = 0$ which has the solution $\chi_+ = \chi_+(t+x)$. In other words, χ_+ is a left-moving fermion. In contrast, χ_- obeys $\partial_+\chi_- = 0$ and is a right-moving fermion: $\chi_- = \chi_-(t-x)$.

Only massless Dirac fermions can be decomposed into independent chiral constituents. This is clear in $d = 1 + 1$ dimensions since massless particles must travel at the speed of light, so naturally fall into left-moving and right-moving sectors. If we want the particle to sit still, we need to add a mass term which couples the left-moving and right-moving fermions: $m\bar{\psi}\psi = m(\chi_+^\dagger \chi_- + \chi_-^\dagger \chi_+)$

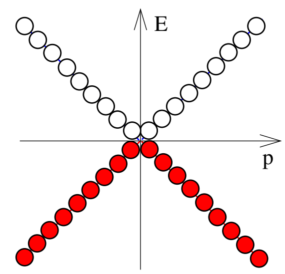


Figure 23:

We won't run through the full machinery of canonical quantisation, but the results are straightforward. One finds that there are both particles and anti-particles. Right-movers have momentum $p > 0$ and left-movers have $p < 0$. All excitations have the dispersion relation $E = |p|$.

For once, it's useful to think of this in the Dirac sea language. Here we view the states as having energy $E = \pm|p|$. The vacuum configuration consists of filling all negative energy states; these are the red states shown in the figure. Those with $E > 0$ are unfilled. In the picture we've implicitly put the system on a spatial circle, so that the momentum states are discrete, but this isn't necessary for the discussion below.

The action (3.5) has two global symmetries which rotate the individual phases of χ_+ and χ_-. Alternatively, in the language of the Dirac fermion these symmetries are $\psi \to e^{i\alpha}\psi$ and $\psi \to e^{i\alpha\gamma^5}\psi$. This means that the number of n_- of left-moving fermions and the number n_+ of right-moving fermions is separately conserved. This is referred to as a *chiral symmetry*.

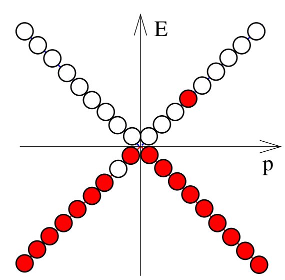


Figure 24:

Naively, we would expect that both n_+ and n_- continue to be conserved if we deform the theory, provided that both symmetries are preserved. This means that we could perturb the theory in some way which results in a right-moving particle-anti-particle pair being excited as in the picture. (Note that in this picture, the hole left in the Dirac sea has momentum $p < 0$ which, when viewed as a particle, means that it has momentum $p > 0$ as befits a right-moving excitation.) However, as long as the symmetries remain, we would not expect to be able to change a left-moving fermion into a right-moving fermion.

We will see that this expectation is wrong. One can deform the theory in such a way that both symmetries are naively preserved, and yet right-moving fermions can change into left-moving fermions.

Turning on a Background Electric Field

To see the anomaly, we need to deform our theory in some way. We do this by turning on a background electric field. This means that we replace the action (3.3) with

$$S = \int d^2x \; i\bar{\psi}\not{\mathcal{D}}\psi \tag{3.6}$$

where $\mathcal{D}_\mu = \partial_\mu - ieA_\mu$. Here A_μ is not a fluctuating, dynamical field: instead it is a fixed background field. Notice that the classical action (3.6) remains invariant under the two global symmetries and a standard application of Noether's theorem would suggest that n_+ and n_- are separately conserved. This, it turns out, is not correct.

To see the problem, we turn on an electric field $\mathcal{E}$ for some time t. We choose $\mathcal{E} > 0$ which means that it points towards the right. Because the particles are charged, the electric field will increase the momentum p, and hence the energy E, of all the filled states in the Dirac sea: they all get shifted by

$$\Delta p = e\mathcal{E}t \tag{3.7}$$

Both left and right-movers get shifted by the same amount. The net result is the Fermi surface shown in the figure to the right. But this is precisely what we thought shouldn't happen: despite the presence of the symmetry, we have created left-moving anti-particles and right-moving particles!

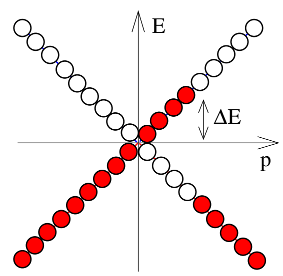


Figure 25:

We can be a little more precise about the violation of the conserved quantity. We denote by ρ_+ the density of right-moving fermions and by ρ_- the density of left-moving fermions. The shift in momentum (3.7) then becomes a shift in charge density,

$$\rho_+ = \frac{e\mathcal{E}}{2\pi}t \quad \text{and} \quad \rho_- = -\frac{e\mathcal{E}}{2\pi}t$$

where the extra factor of $1/2\pi$ comes from the density of states. The total number of fermions is conserved (counting, as usual, particles minus anti-particles). This is the conservation law that comes from the vector symmetry $\psi \to e^{i\alpha}\psi$:

$$\dot{\rho} = 0 \quad \text{where } \rho = \rho_+ + \rho_-$$

In contrast, the difference between fermion numbers is not conserved. This is the quantity that was supposed to be preserved by the axial symmetry $\psi \to e^{i\alpha\gamma^5}\psi$,

$$\dot{\rho}_A = \frac{e\mathcal{E}}{\pi} \quad \text{where } \rho_A = \rho_+ - \rho_- \tag{3.8}$$

This is known as the *axial anomaly* or the *chiral anomaly*.

We seem to have violated Noether's theorem: the axial symmetry does not give rise to a conserved quantity. How could this happen? Looking at the picture of the Dirac sea, it's clear where these extra fermions came from. They came from infinity! It was only possible to change left-movers to right-movers because the Dirac sea is infinitely deep. If we were to truncate the Dirac sea somewhere, then the excess right-movers would be compensated by a depletion of right-moving states at large, negative energy and there would be no violation of axial charge. But there is no truncation of the Dirac sea and, rather like Hilbert's hotel, the whole chain of right-moving states can be shifted up, leaving no empty spaces at the bottom.

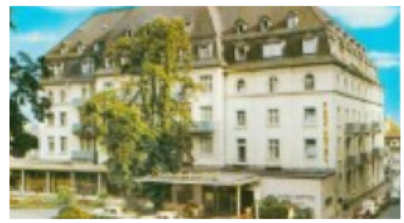


Figure 26:

This is interesting! The anomaly arises because of the infinite Dirac sea which, in turn, arises because we are dealing with continuum quantum field theory with an infinite number of states rather than a finite quantum mechanical system. Ultimately, it is this difference that allows for anomalies.

As a useless aside, here is a picture of an actual "Hilbert hotel", originally in Germany, now sadly closed. This hotel appears to be best known as a place that Elvis Presley once stayed. To my knowledge there exists no photograph that shows the full height of this hotel: you should use your imagination.

3.1.2 Massless Fermions in Four Dimensions

The discussion above seems very specific to $d = 1+1$ dimensions, where massless fermions split into left-movers and right-movers. However, there is an analogous piece of physics in $d = 3+1$ dimensions. For this, we must look at massless fermions in background electric and magnetic fields.

First some notation. We take the representation of gamma matrices to be

$$\gamma^0 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad , \quad \gamma^i = \begin{pmatrix} 0 & \sigma^i \\ -\sigma^i & 0 \end{pmatrix} \tag{3.9}$$

which obey the Clifford algebra $\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu}$ in signature $(+---)$. We also introduce

$$\gamma^5 = -i\gamma^0\gamma^1\gamma^2\gamma^3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

The Dirac fermion is a four-component spinor ψ. This can be split into two, two-component Weyl spinors $\psi_\pm$ which are eigenvectors of γ^5. In components we write

$$\psi = \begin{pmatrix} \psi_+ \\ \psi_- \end{pmatrix}$$

We now couple the spinor to a background electromagnetic field A_μ. The action is

$$S = \int d^4x \; i\bar{\psi}\not{\mathcal{D}}\psi = \int d^4x \; i\psi_+^\dagger \bar{\sigma}^\mu \mathcal{D}_\mu \psi_+ + i\psi_-^\dagger \sigma^\mu \mathcal{D}_\mu \psi_- \tag{3.10}$$

where $\mathcal{D}_\mu = \partial_\mu - ieA_\mu$ and $\sigma^\mu = (1, \sigma^i)$ and $\bar{\sigma}^\mu = (1, -\sigma^i)$. (Note that we've resorted to the convention where the electric charge sits inside the covariant derivative.)

We'll proceed in steps. We'll first see how these fermions respond to a background magnetic field $\mathbf{B}$. Setting $A_0 = 0$, the equation of motion for the chiral spinor $\psi_+^\dagger$ is

$$i\partial_t \psi_+ = i\sigma^i \mathcal{D}_i \psi_+ \tag{3.11}$$

Once again, we don't want to run through the whole process of canonical quantisation. Instead we'll cheat and think of this equation in the way that Dirac originally thought of the Dirac equation: as a one-particle Schrödinger equation for a particle with spin. In this framework, the Hamiltonian is

$$H = -i\sigma^i \mathcal{D}_i = (\mathbf{p} - e\mathbf{A}) \cdot \boldsymbol{\sigma}$$

The spin of the particle is determined by the operator $\mathbf{S} = \frac{1}{2}\boldsymbol{\sigma}$. (For massless particles, it's better to refer to this as *helicity*; we'll see its interpretation below.) Squaring the Hamiltonian, and using the fact that $\sigma^i \sigma^j = \delta^{ij} + i\epsilon^{ijk}\sigma^k$, we find

$$H^2 = (\mathbf{p} - e\mathbf{A})^2 - 2e\mathbf{B} \cdot \mathbf{S}$$

The first term is the Hamiltonian for non-relativistic particles in a magnetic field. (See, for example, the lectures on Applications of Quantum Mechanics.) The second term leads to a Zeeman splitting between spin states. Let's choose the magnetic field to lie in the z-direction, $\mathbf{B} = (0, 0, B)$, and work in Landau gauge so $\mathbf{A} = (0, Bx, 0)$. Then we have

$$H^2 = p_x^2 + (p_y - eBx)^2 + p_z^2 - 2eBS_z$$

Quantisation of motion in the (x, y)-plane leads to the familiar Landau levels. Each of these has a large degeneracy: in a region of area A there are $eBA/2\pi$ states which, in Landau gauge, are distinguished by the quantum number p_y. The resulting energy spectrum is

$$E^2 = eB(2n + 1) + p_z^2 - 2eBS_z \quad \text{with } n = 0, 1, 2, \ldots$$

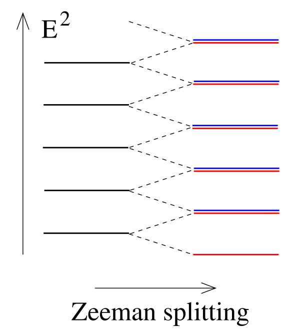


Figure 27:

At this point, there's a rather nice interplay between the energies of the Landau levels and the Zeeman splitting. This occurs because the eigenvalues of the spin operator S_z are $\pm\frac{1}{2}$. This means that the states with $S_z = +\frac{1}{2}$ in the $n = 0$ Landau level have precisely zero energy $E = 0$. Such states are, quite reasonably, referred to as *zero modes*. Meanwhile, the $n = 0$ states with $S_z = -\frac{1}{2}$ have the same energy as the $n = 1$ states with $S_z = +\frac{1}{2}$, and so on. Ignoring p_z, the resulting energy spectrum is shown in the figure. Note, in particular, that the $n = 0$ Landau level has exactly half the states of the other levels.

In very high magnetic fields, it is sensible to restrict to the zero modes in the $n = 0$ Landau level. As we've seen, these have spin $+\frac{1}{2}$. This means that they take the form

$$\psi_+(x, y; z, t) = \begin{pmatrix} \chi_+(x, y; z, t) \\ 0 \end{pmatrix}$$

where the notation is there to highlight that these states have a very specific dependence on (x, y) as they are zero-energy solutions of the Weyl equation (3.11). Meanwhile, their dependence on z and t is not yet fixed. We can determine this by plugging the ansatz back into the original action (3.10) to find

$$S = A \int dzdt \; i\bar{\chi}_+(\partial_t - \partial_z)\chi_+$$

We see that the zero modes arising from χ_+ are all right-movers in the z-direction.

States in higher Landau levels also have an effective description in terms of two-dimensional fermions. Because they have particles of both spins, the states include both left- and right-movers. Moreover, the non-zero energy of the Landau level results in an effective mass for the 2d fermion, coupling the left-movers to the right-movers.

We can repeat this story for the chiral fermions ψ_-. We once again find zero modes, but the change in minus sign in the kinetic term (3.10) ensures that they are now left-movers. Putting both together, the low-energy physics of the lowest Landau level is governed by the effective action

$$S = A\int dtdz\; i\chi_+^\dagger \mathcal{D}_-\chi_+ + i\chi_-^\dagger \mathcal{D}_+\chi_-$$

where we've re-introduced background gauge fields A_0 and A_z which can still couple to these zero modes. However, we've seen this action before: it is the action for a two-dimensional massless fermion coupled to an electromagnetic field. And, as we've seen, despite appearances it does not have a conservation law associated to chiral symmetry.

We computed the violation of axial charge in two dimensions in (3.8). This immediately translates into the violation of four-dimensional axial charge. We need only remember that the lowest Landau level has a degeneracy per area of $eB/2\pi$, and each of these states contributes to the anomaly. The upshot is that, in four dimensions, the axial charge changes if we turn on both a magnetic field B and electric field $\mathcal{E}$ lying in the same (or opposite) direction.

$$\dot{\rho}_A = \frac{eB}{2\pi}\frac{e\mathcal{E}}{\pi} = \frac{e^2}{2\pi^2}\,\mathbf{E}\cdot\mathbf{B} \tag{3.12}$$

This is the *chiral anomaly* for four-dimensional massless fermions. It is equivalent to our earlier, advertised result (3.2).

3.2 Deriving the Chiral Anomaly

In the previous section, we've seen that the axial charge of a massless fermion is not conserved in the presence of background electric and magnetic fields. This lack of conservation seems to be in direct contradiction to Noether's theorem, which states that the axial symmetry should result in a conserved charge. What did we miss?

3.2.1 Noether's Theorem and Ward Identities

Let's first remind ourselves how we prove Noether's theorem, and how it manifests itself in the quantum theory. We start by considering a general theory of a scalar field ϕ with a symmetry; we will later generalise this to a fermion and the axial symmetry of interest.

Noether's Theorem in Classical Field Theory

Consider the transformation of a scalar field ϕ

$$\delta\phi = \epsilon X(\phi) \tag{3.13}$$

Here ϵ is a constant, infinitesimally small parameter. This transformation is a *symmetry* if the change in the Lagrangian is

$$\delta L = 0$$

(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 ϵ to depend on spacetime: $\epsilon = \epsilon(x)$. Now the Lagrangian is no longer invariant, but changes as

$$\begin{aligned}\delta\mathcal{L} &= \frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi)}\,\partial_\mu(\epsilon X(\phi)) + \frac{\partial\mathcal{L}}{\partial\phi}\,\epsilon X(\phi) \\ &= (\partial_\mu\epsilon)\,\frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi)}\,X(\phi) + \left[\frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi)}\partial_\mu X(\phi) + \frac{\partial\mathcal{L}}{\partial\phi}\,X(\phi)\right]\epsilon\end{aligned}$$

But we know that $\delta\mathcal{L} = 0$ when ϵ is constant, which means that the term in square brackets must vanish. We're left with the expression

$$\delta\mathcal{L} = (\partial_\mu\epsilon)J^\mu \quad \text{with} \quad J^\mu = \frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi)}\,X(\phi)$$

The action $S = \int d^dx\ \mathcal{L}$ then changes as

$$\delta S = \int d^dx\ \delta\mathcal{L} = \int d^dx\ (\partial_\mu\epsilon)J^\mu = -\int d^dx\ \epsilon\,\partial_\mu J^\mu \tag{3.14}$$

where we pick $\epsilon(x)$ to decay asymptotically so that we can safely discard the surface term.

The expression (3.14) 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 (3.13) with $\epsilon(x)$ a function of spacetime. This means that when the equations of motion are satisfied we have the conservation law

$$\partial_\mu J^\mu = 0$$

This is Noether's theorem.

Ward Identities in Quantum Field Theory

Let's now see how this argument plays out in the framework of quantum field theory. Our tool of choice is the Euclidean path integral,

$$Z[K] = \int \mathcal{D}\phi\ \exp\left(-S[\phi] + \int d^dx\ K\phi\right) \tag{3.15}$$

where $K(x)$ is a background source for ϕ. (This is usually called $J(x)$ but I didn't want to confuse it with the current.) We again consider the symmetry (3.13), this time writing it as the transformation

$$\phi \longrightarrow \phi' = \phi + \epsilon(x) X(\phi) \tag{3.16}$$

We view this as a change of variables in the partition function, which now reads

$$Z[K] \longrightarrow \int \mathcal{D}\phi' \ \exp\left(-S[\phi'] + \int d^d x \ K\phi'\right)$$

The field in the partition function is nothing more than a dummy variable. This means that the new partition function is exactly the same as the original partition function (3.15). Nonetheless, we can manipulate this into a useful form. Using the transformation (3.16), together with (3.14), and expanding to leading order in ϵ, we have

$$\begin{aligned} Z[K] &= \int \mathcal{D}\phi' \ \exp\left(-S[\phi] + \int d^d x \ K\phi\right) \exp\left(-\int d^d x \ \epsilon \left(\partial_\mu J^\mu - KX\right)\right) \\ &\approx \int \mathcal{D}\phi' \ \exp\left(-S[\phi] + \int d^d x \ K\phi\right) \left[1 - \int d^d x \ \epsilon \left(\partial_\mu J^\mu - KX\right)\right] \end{aligned} \tag{3.17}$$

At this point we need to make a further assumption about the transformation that was not needed to derive Noether's theorem in the classical theory: not only should (3.16) be a symmetry of the action, but also a symmetry of the measure. This means that we require

$$\mathcal{D}\phi = \mathcal{D}\phi' \tag{3.18}$$

Ultimately, this will be the assumption that breaks down for axial transformations. But, for now, let's assume that (3.18) holds and derive the consequences. The first term in (3.17) (meaning the "1" in the square brackets) is simply our original partition function (3.15). This means that we have

$$\int \mathcal{D}\phi \ \exp\left(-S[\phi] + \int d^d x \ K\phi\right) \left[\int d^d x \ \epsilon(x) \left(\partial_\mu J^\mu - KX\right)\right] = 0$$

But this is true for all $\epsilon(x)$. This means that we can lose the integral to leave ourselves an expression for each spacetime point,

$$\int \mathcal{D}\phi \ \exp\left(-S[\phi] + \int d^d x \ K\phi\right) \left(\partial_\mu J^\mu - K(x) X(\phi)\right) = 0$$

We can now play with the source K to derive various expressions that involve correlation functions of $\partial_\mu J^\mu$ and ϕ. For example, setting $K = 0$ gives us

$$\langle \partial_\mu J^\mu \rangle = 0$$

Alternatively, we can differentiate with respect to $K(x')$ before setting $K = 0$ to find

$$\partial_\mu \langle J^\mu(x) \phi(x') \rangle = \delta(x - x') \langle X(\phi) \rangle \tag{3.19}$$

Differentiating more times gives us the expression

$$\partial_\mu \langle J^\mu(x) \phi(x^1) \ldots \phi(x^n) \rangle = 0 \quad \text{ for } x \neq x^i$$

while, if x does coincide with one of the insertion points x^i we pick up a term proportional to $\delta\phi$ on the right-hand side as in (3.19). These expressions are collectively known as *Ward identities*. They are sometimes expressed as the operator-valued continuity equation

$$\partial_\mu J^\mu = 0$$

which is to be viewed as saying that $\partial_\mu J^\mu$ vanishes inside any correlation function, as long as its position does not coincide with the insertion point of other fields.

The Axial Symmetry

We can apply all of the above ideas to the theory that we're really interested in – a massless Dirac fermion in $d = 3 + 1$ dimensions with action (3.1). For now, we will take A_μ to be a background gauge field, without its own dynamics. As we reviewed in the beginning of this section, this theory has both vector and axial symmetry. The infinitesimal action of the vector rotation $\psi \to e^{i\alpha}\psi$ is

$$\delta\psi = i\epsilon\psi \quad , \quad \delta\bar{\psi} = -i\epsilon\psi \tag{3.20}$$

with the corresponding current

$$j^\mu = \bar{\psi}\gamma^\mu\psi$$

The infinitesimal version of the axial rotation $\psi \to e^{i\alpha\gamma^5}\psi$ is

$$\delta\psi = i\epsilon\gamma^5\psi \quad , \quad \delta\bar{\psi} = i\epsilon\bar{\psi}\gamma^5 \tag{3.21}$$

Note that now both ψ and $\bar{\psi}$ transform in the same way. In Minkowski space, this follows from the definition $\bar{\psi} = \psi^\dagger\gamma^0$; in Euclidean space ψ and $\bar{\psi}$ are viewed as independent variables and this is simply the transformation necessary to be a symmetry of the action (3.1). An application of Noether's theorem as described above gives the current

$$j_A^\mu = i\bar{\psi}\gamma^\mu\gamma^5\psi$$

Repeating the rest of the path integral manipulations seems to tell us that the Ward identities hold with $\partial_\mu j_A^\mu = 0$. But, as we've seen in the previous section, this can't be the case: despite the presence of the axial symmetry (3.21), there are situations where the axial charge is not conserved.

3.2.2 The Anomaly lies in the Measure

As we mentioned above, in deriving the Ward identities it's not enough for the action to be invariant under a symmetry; the path integral measure must also be invariant. This approach to the anomaly is usually called the *Fujikawa method*.

For fermions this measure is schematically

$$\int \mathcal{D}\psi\mathcal{D}\bar{\psi} \tag{3.22}$$

When we change to new variables

$$\psi' = \psi + i\epsilon\gamma^5\psi \quad , \quad \bar{\psi}' = \bar{\psi} + i\epsilon\bar{\psi}\gamma^5 \tag{3.23}$$

this measure will pick up a Jacobian factor. As we now show, it is this Jacobian that gives rise to the anomaly.

Our first task is to explain what we mean by the field theoretic measure (3.22). To do this, let's consider the Dirac operator $\not{D}$ for a spinor in the background of a fixed electromagnetic field A_μ. This operator will have eigenspinors; these are c-number (i.e. not Grassmann-valued) four-component spinors ϕ_n satisfying

$$i\not{D}\phi_n = \lambda_n\phi_n \tag{3.24}$$

We expand a general spinor ψ in terms of these eigenspinors,

$$\psi(x) = \sum_n a_n\phi_n(x) \tag{3.25}$$

where a_n are Grassmann-valued numbers. Similarly, we can expand the $\bar{\psi}$ in terms of eigenspinors

$$\bar{\psi}(x) = \sum_n \bar{b}_n\bar{\phi}_n(x)$$

As usual, eigenspinors with distinct eigenvalues are orthogonal, and those with the same eigenvalues can be chosen to be orthogonal. In the present context, this means

$$\int d^4x\ \bar{\phi}_n\phi_m = \delta_{nm} \tag{3.26}$$

In terms of the eigenspinor expansion, the action reads

$$S = \int d^4x\ i\bar{\psi}\not{D}\psi = \sum_n \lambda_n \bar{b}_n a_n$$

In this language, the fermion measure (3.22) is defined to be

$$\prod_n \int d\bar{b}_n da_n$$

Of course, Grassmann integrations are easy. We have $\int da = 0$ and $\int da\ a = 1$, with similar expressions for b. If we wished to evaluate the Euclidean partition function in this language, we would have

$$\int \mathcal{D}\bar{\psi}\mathcal{D}\psi\ e^{-S} = \prod_n \int d\bar{b}_n da_n e^{-\sum_m \lambda_m \bar{b}_m a_m} = \prod_n \lambda_n \equiv \det i\not{D}$$

This approach hasn't rescued us from the usual infinities that arise in continuum quantum field theory: we're left with an infinite product which will, in general, diverge. To make sense of this expression we will have to play the usual regularisation games. We'll see a particular example of this below.

The Jacobian

Now that we've got a slightly better definition of the fermion measure, we can see how it fares under the position-dependent chiral rotation

$$\delta\psi = i\epsilon(x)\gamma^5\psi$$

Such a transformation changes the Grassmann parameters a_n in our expansion (3.25),

$$\sum_n \delta a_n\, \phi_n = i\epsilon(x)\sum_m a_m\gamma^5\phi_m$$

Using the orthogonality relation (3.26), we have

$$\delta a_n = X_{nm}a_m \quad \text{with} \quad X_{nm} = i\int d^4x\ \epsilon(x)\bar{\phi}_n\gamma^5\phi_m$$

We want to compute the Jacobian for the transformation from a_n to $a'_n = a_n + X_{nm}a_m$. Fortunately, the transformation is linear in a_n which means that the Jacobian will not depend on the value of a_n. If we were dealing with commuting, c-number objects this would be $\det(1+X)$. But integration for Grassmann variables is closer to differentiation and, for this reason, the Jacobian is actually the inverse determinant. We therefore have

$$J = \det{}^{-1}(\delta_{nm} + X_{nm})$$

Because the axial symmetry (3.21) acts on both ψ and $\bar{\psi}$ in the same way, we get the same Jacobian for the transformation of b_n. This means that we have

$$\prod_n \int d\bar{b}_n da_n = \prod_n \int d\bar{b}'_n da'_n \, J^2$$

Before we proceed, it's worth pausing to point out why the vector and axial transformations differ. For the vector transformation (3.20), we have $\delta\psi = i\epsilon\psi$ and $\delta\bar{\psi} = -i\epsilon\bar{\psi}$. This extra minus sign means that the Jacobian factors for ψ and $\bar{\psi}$ have the form $\det{}^{-1}(1+Y)$ and $\det{}^{-1}(1-Y)$ respectively, with Y similar to X but without the γ^5 matrix. This extra minus sign means that the Jacobian vanishes to leading order in ϵ; as we will see below, this is sufficient to ensure that it does not contribute to the Ward identities.

Returning to the axial symmetry, we need only evaluate the Jacobian to leading order in ϵ; the group structure of the symmetry will do the rest of the work for us. At this level, we can write

$$J = \det{}^{-1}(1+X) \approx \det(1-X) \approx \det e^{-X} = e^{-\mathrm{Tr}\, X}$$

where Tr here means the trace over spinor indices, as well as integration over space. Written in full, we have

$$J = \exp\left(-i\int d^4x \; \epsilon(x) \sum_n \bar{\phi}_n(x)\gamma^5\phi_n(x)\right) \tag{3.27}$$

Our task is to calculate this.

Calculating the Jacobian

We have to be a little careful in evaluating J. To illustrate this, here are two naive, non-careful arguments for the value of J:

- The first argument says that $J = 0$. This is because it involves a trace over spinor indices and $\mathrm{tr}\,\gamma^5 = 0$.

- The second argument says that $J = \infty$. This is because, at each point x, we're summing over an infinite number of modes ϕ_n and there is no reason to think that this sum converges.

The truth, of course, is that neither of these arguments is quite right. Instead, they play off against each other: when we understand how to regulate the sum, we will see why we're not left with $\mathrm{tr}\,\gamma^5$. And when we take the resulting trace, we'll see why the sum is not infinite.

Let's first worry about the divergence. We want to regulate the sum over modes in a manner consistent with gauge invariance. The one useful, gauge invariant, piece of information that we have about each mode is its eigenvalue λ_n. This motivates us to write

$$\begin{aligned}\int d^4x\ \epsilon(x)\sum_n \bar{\phi}_n\gamma^5\phi_n &= \lim_{\Lambda\to\infty}\int d^4x\ \epsilon(x)\sum_n \bar{\phi}_n\gamma^5\phi_n\, e^{-\lambda_n^2/\Lambda^2} \\ &= \lim_{\Lambda\to\infty}\int d^4x\ \epsilon(x)\sum_n \bar{\phi}_n\gamma^5\, e^{-(i\not{D})^2/\Lambda^2}\phi_n \qquad (3.28)\end{aligned}$$

where Λ is a regularisation scale. It has dimension of energy and, as shown above, we will ultimately send $\Lambda \to \infty$.

Notice that, already, we can see how we evade our first naive argument. The regulator has introduced extra gamma matrix structure into our expression, which means that we no longer get to argue that J is proportional to $\mathrm{tr}\,\gamma^5$ and so necessarily vanishes. Instead, the trace over gamma matrices will greatly restrict the form of J.

In the expression above, we're taking a sum over states $\phi_n(x)$. Such a sum can be viewed as a trace of whatever operator $\mathcal{O}$ is inserted between these states. But we equally well write the trace in any basis. The most familiar is the basis of plane waves $e^{ik\cdot x}$, together with a trace over spinor indices. Implementing this change of basis means that we can write

$$\sum_n \bar{\phi}_n(x)\,\gamma^5 e^{+\not{D}^2/\Lambda^2}\phi_n(x) = \int \frac{d^4k}{(2\pi)^4}\,\mathrm{tr}\left(\gamma^5 e^{-ik\cdot x}e^{+\not{D}^2/\Lambda^2}e^{ik\cdot x}\right) \qquad (3.29)$$

where now tr denotes only the trace over spinor indices.

(If the step (3.29) seems confusing, it might make you more comfortable to mention that it's the kind of manipulation that we do all the time in quantum mechanics. In that context, we have a basis of states $|\phi_n\rangle$ with wavefunction $\phi_n(x)$. We would write $\sum_n \phi_n^\dagger(x)\mathcal{O}\phi_n(x) = \sum_n \langle\phi_n|x\rangle\langle x|\mathcal{O}|\phi_n\rangle = \langle x|\mathcal{O}|x\rangle = \int \frac{dk}{2\pi}\,\langle k|x\rangle\langle x|\mathcal{O}|k\rangle = \int \frac{dk}{2\pi}\, e^{-ikx}\mathcal{O}e^{ikx}$. Note, however, that in the present context, the eigenspinors $\phi_n(x)$ are a basis of fields rather than states in a Hilbert space.)

The expression (3.29) still looks like it's difficult to evaluate. But we've got two things going for us, both descendants of the naive arguments we tried to use previously:

- The trace tr over spinor indices vanishes when taken over most products of gamma matrices. In particular, we have

  $$\operatorname{tr}\gamma^5 = \operatorname{tr}\gamma^5\gamma^\mu\gamma^\nu = 0$$

  However, if we multiply all five (Euclidean) gamma matrices together we get the identity matrix. This is captured by the expression

  $$\operatorname{tr}\gamma^5\gamma^\mu\gamma^\nu\gamma^\rho\gamma^\sigma = 4\epsilon^{\mu\nu\rho\sigma}$$

  We'll need this expression shortly.

- We still want to send $\Lambda \to \infty$ to compute the Jacobian (3.27). Our strategy will be to Taylor expand the exponential $e^{\not{\mathcal{D}}^2/\Lambda^2}$. But higher powers come with higher powers of Λ in the denominator which, as we will see, will eventually ensure that they vanish.

Let's now see how this works. First, we need a couple of identities involving the covariant derivative. The first is

$$\begin{aligned}\not{\mathcal{D}}^2 = \gamma^\mu\gamma^\nu\mathcal{D}_\mu\mathcal{D}_\nu &= \frac{1}{2}\{\gamma^\mu,\gamma^\nu\}\mathcal{D}_\mu\mathcal{D}_\nu + \frac{1}{2}[\gamma^\mu,\gamma^\nu]\mathcal{D}_\mu\mathcal{D}_\nu \\ &= \mathcal{D}^2 + \frac{1}{4}[\gamma^\mu,\gamma^\nu][\mathcal{D}_\mu,\mathcal{D}_\nu] \\ &= \mathcal{D}^2 - \frac{ie}{2}\gamma^\mu\gamma^\nu F_{\mu\nu}\end{aligned}$$

The second is

$$e^{-ik\cdot x}\mathcal{D}_\mu e^{+ik\cdot x} = \mathcal{D}_\mu + ik_\mu$$

Combining these, we have

$$\begin{aligned}e^{-ik\cdot x}e^{\not{\mathcal{D}}^2/\Lambda^2}e^{ik\cdot x} &= e^{-ik\cdot x}e^{\mathcal{D}^2/\Lambda^2 - \frac{ie}{2}\gamma^\mu\gamma^\nu F_{\mu\nu}/\Lambda^2}e^{ik\cdot x} \\ &= e^{(\mathcal{D}_\mu+ik_\mu)^2/\Lambda^2 - \frac{ie}{2}\gamma^\mu\gamma^\nu F_{\mu\nu}/\Lambda^2} \\ &= e^{(\mathcal{D}_\mu+ik_\mu)^2/\Lambda^2}\, e^{-\frac{ie}{2}\gamma^\mu\gamma^\nu F_{\mu\nu}/\Lambda^2}\, e^{\cdots}\ldots\end{aligned} \tag{3.30}$$

Here the extra terms in the ... follow from the BCH formula. They do not vanish but, as we will see, we will not need them.

We want to Taylor expand this exponent. In particular, we have

$$\gamma^5 \, e^{-\frac{ie}{2}\gamma^\mu\gamma^\nu F_{\mu\nu}/\Lambda^2} = \gamma^5 \left(1 - \frac{ie}{2}\gamma^\mu\gamma^\nu F_{\mu\nu}\frac{1}{\Lambda^2} - \frac{e^2}{8}\gamma^\mu\gamma^\nu\gamma^\rho\gamma^\sigma F_{\mu\nu}F_{\rho\sigma}\frac{1}{\Lambda^4} + \ldots\right) \tag{3.31}$$

From our arguments above about the spinor traces, we see that only the last of these terms contributes. This term scales as $1/\Lambda^4$ and we clearly need to compensate for this before we take the $\Lambda \to \infty$ in (3.28). Fortunately, this compensation comes courtesy of the $\int d^4k$ which will give the Λ^4 term that we need. (You may want to first shift $k_\mu \to k_\mu + A_\mu(x)$ to absorb the potential in the covariant derivative.)

There will also be other terms in the expansion (3.31) which are non-zero after the trace. There will also be further terms from the BCH contributions in (3.30). However, all of these will scale with some power $1/\Lambda^n$ with $n > 4$ and so will vanish when we take the $\Lambda \to \infty$ limit. A similar argument holds for the e^{∂^2/Λ^2} terms in the first exponent in (3.30). We end up with

$$\begin{aligned}
\sum_n \bar{\phi}_n \gamma^5 \phi_n &= \lim_{\Lambda\to\infty} \int \frac{d^4k}{(2\pi)^4} \, \mathrm{tr}\left(\gamma^5 e^{-ik\cdot x} e^{+\not{D}^2/\Lambda^2} e^{ik\cdot x}\right) \\
&= \lim_{\Lambda\to\infty} \int \frac{d^4k}{(2\pi)^4} \, e^{-k^2/\Lambda^2} \, \left(\frac{e^2}{2}\epsilon^{\mu\nu\rho\sigma} F_{\mu\nu}F_{\rho\sigma}\frac{1}{\Lambda^4} + \ldots\right) \\
&= \frac{e^2}{32\pi^2}\epsilon^{\mu\nu\rho\sigma} F_{\mu\nu}F_{\rho\sigma}
\end{aligned} \tag{3.32}$$

This is what we need.

The Anomalous Ward Identity

Let's put these pieces together. We've learned that under a chiral transformation $\delta\psi = i\epsilon(x)\gamma^5\psi$, the fermion measure picks up a Jacobian factor (3.27) which is calculated in (3.32). The transformation $\delta\bar{\psi} = i\epsilon(x)\bar{\psi}\gamma^5$ gives us another factor of this Jacobian so, in total, the measure transforms as

$$\int \mathcal{D}\psi\mathcal{D}\bar{\psi} \;\longrightarrow \int \mathcal{D}\psi\mathcal{D}\bar{\psi} \; \exp\left(-\frac{ie^2}{16\pi^2}\int d^4x \; \epsilon(x)\epsilon^{\mu\nu\rho\sigma}F_{\mu\nu}F_{\rho\sigma}\right) \tag{3.33}$$

It is a simple matter to follow the fate of this term when deriving the Ward identities described in Section 3.2.1. We find that the current $j^\mu_A = i\bar{\psi}\gamma^\mu\gamma^5\psi$ associated to axial transformations is no longer conserved: instead it obeys

$$\partial_\mu j^\mu_A = \frac{e^2}{16\pi^2}\epsilon^{\mu\nu\rho\sigma}F_{\mu\nu}F_{\rho\sigma} \tag{3.34}$$

This is our promised result (3.2) for the chiral anomaly.

We saw in Section 1.2 that the right-hand side of (3.34) is itself a total derivative,

$$\epsilon^{\mu\nu\rho\sigma} F_{\mu\nu} F_{\rho\sigma} = 4\partial_\mu(\epsilon^{\mu\nu\rho\sigma} A_\nu\, \partial_\rho A_\sigma)$$

It's tempting to attempt to define a new conserved current that is, roughly, $j_A - {}^\star AdA$. But this is illegal because it's not gauge invariant. Hopefully our discussion in Sections 3.1.1 and 3.1.2 has already convinced you that there's no escaping the anomaly: it is a real physical effect.

There are a number of straightforward generalisations of this result. First, if we have N_f massless Dirac fermions, then the anomaly becomes

$$\partial_\mu j^\mu_A = \frac{e^2 N_f}{16\pi^2} \epsilon^{\mu\nu\rho\sigma} F_{\mu\nu} F_{\rho\sigma}$$

Alternatively, we could return to a single Dirac fermion, but give it a mass m. This explicitly breaks the axial symmetry. Nonetheless, the anomaly remains and the divergence of the axial current is now given by

$$\partial_\mu j^\mu_A = -2im\bar{\psi}\gamma^5\psi + \frac{e^2}{16\pi^2} \epsilon^{\mu\nu\rho\sigma} F_{\mu\nu} F_{\rho\sigma}$$

For the purpose of our discussion above, we took the fermions to be dynamical (in the sense that we integrated over them in the path integral), while the gauge field A_μ took some fixed, background value. However, nothing stops us promoting the gauge field to also be dynamical, in which case we are discussing QED. The calculation above goes through without a hitch, and the result (3.34) still holds.

With dynamical gauge fields, one might wonder if there are extra corrections to the chiral anomaly. In fact, this is not the case. For deep reasons, the result (3.34) is exact; it receives neither perturbative nor non-perturbative corrections. We will start to get a sense of why this is in Section 3.3.1.

The Anomaly in Non-Abelian Gauge Theories

It is a simple matter to adapt the above arguments to non-Abelian gauge theories. For example, we may have a Dirac fermion transforming in some representation R of a non-Abelian gauge group, with field strength $F_{\mu\nu}$. The Lagrangian for the fermion is

$$\mathcal{L} = i\bar{\psi}\gamma^\mu(\partial_\mu - iA_\mu)\psi$$

The calculation that we did above goes through essentially unchanged; we need only include a trace over the colour indices. We now have

$$\partial_\mu j^\mu_A = \frac{1}{16\pi^2} \epsilon^{\mu\nu\rho\sigma}\, \mathrm{tr}_R\, F_{\mu\nu} F_{\rho\sigma} \tag{3.35}$$

Note that the overall factor of e^2 has disappeared because we are here working in the conventions described in Section 2.1.1 in which the coupling constant sits as an overall factor in the action.

The Anomaly in Two Dimensions

It is also a simple matter to adapt the above arguments for fermions in $d = 1 + 1$ dimensions (or, indeed, for fermions in any even number of spacetime dimensions). Now the gamma matrices are 2×2 and, in Euclidean space, we have

$$\mathrm{tr}\,\gamma^5\gamma^\mu\gamma^\nu = 2i\epsilon^{\mu\nu}$$

which means that the term linear in $F_{\mu\nu}$ in (3.31) is now non-vanishing. The factor $1/\Lambda^2$ is compensated by the divergent factor coming from the $\int d^2k$ integral. Repeating the derivation above, we this time find

$$\partial_\mu j^\mu_A = \frac{e}{\pi} F_{01} \tag{3.36}$$

This agrees with our earlier, heuristic derivation (3.8). Note that, in $d = 1 + 1$, one only gets an anomaly for Abelian gauge groups. Attempting to repeat the calculation for, say, $SU(N)$ would give $\mathrm{tr}\, F_{01} = 0$ on the right-hand side.

3.2.3 Triangle Diagrams

There are many different approaches to computing the anomaly. The path integral approach that we saw above is arguably the most useful for our purposes. But it is worthwhile to see how the anomaly arises in other contexts. In this section, we see how the anomaly appears in perturbation theory. Indeed, this is how the anomaly was first discovered.

We will start by considering a free, massless Dirac fermion,

$$S = \int d^4x\; i\bar{\psi}\,\partial\!\!\!/\psi$$

The essence of the argument is as follows. We will look at a certain class of one-loop Feynman diagrams known as "triangle diagrams". These are special because they involve both $U(1)_V$ current $j^\mu = \bar{\psi}\gamma^\mu\psi$ and the $U(1)_A$ current $j^\mu_A = \bar{\psi}\gamma^\mu\gamma^5\psi$. Even in our free theory, these triangle diagrams are UV divergent and need regulating. The crux of the argument is that any regulation necessarily violates either the $U(1)_V$ symmetry or the $U(1)_A$ symmetry; these is no way to make sense of the triangle diagram preserving both symmetries. As we remove the regulator, its memory lingers through the loss of one of these symmetries. This is the anomaly.

Let's now see this in detail. We focus on the three-point correlator containing two vector currents and a single axial current:

$$\Gamma^{\mu\nu\rho}(x_1, x_2, x_3) = \langle 0|T(j^\mu(x_1)\, j^\nu(x_2)\, j^\rho_A(x_3))|0\rangle$$

where, as usual, T denotes time-ordering, for Minkowski space correlators. In Euclidean space, no such ordering is necessary.

With hindsight, it is possible to see why we should look at this particular correlator because the anomaly equation (3.34) includes a single axial current j_A and two gauge fields, each of which couples to the vector current j.

It is simplest to work in momentum space. The Fourier transform is

$$\int d^3x_1 d^3x_2 d^3x_3\ \Gamma^{\mu\nu\rho}(x_1, x_2, x_3) e^{ip_1\cdot x_1 + ip_2\cdot x_2 + iq\cdot x_3} = \Gamma^{\mu\nu\rho}(p_1, p_2, q)\, \delta^3(p_1 + p_2 + q)$$

where we're using the notation that the function and its Fourier transform are distinguished only by the arguments. The delta-function on the right-hand side arises because our theory is translational invariant. Tracing their origin, we note that the momenta p_1 and p_2 refer to the vector current, while q refers to the axial current.

Before we explore the anomaly, let's first see what we would naively expect the conservation of currents to imply for $\Gamma^{\mu\nu\rho}(p_1, p_2, q)$. Consider

$$\begin{aligned} p_{1\mu}\Gamma^{\mu\nu\rho}(p_1, p_2, q) &= -i\int d^3x_1 d^3x_2 d^3x_3\ \Gamma^{\mu\nu\rho}(x_1, x_2, x_3)\frac{\partial}{\partial x_1^\mu} e^{ip_1\cdot x_1 + ip_2\cdot x_2 + iq\cdot x_3} \\ &= +i\int d^3x_1 d^3x_2 d^3x_3\ \frac{\partial \Gamma^{\mu\nu\rho}(x_1, x_2, x_3)}{\partial x_1^\mu} e^{ip_1\cdot x_1 + ip_2\cdot x_2 + iq\cdot x_3} \end{aligned}$$

But this is the kind of expression that we computed in Section 3.2.1. The Ward identity tells us that $\partial_\mu j^\mu = 0$ holds as an operator equation. There is a delta-function, contact term that arises when $x_1 = x_2$ or $x_1 = x_3$ — this can be seen on the right-hand side of (3.19) — but it vanishes in this case because neither of the currents j^μ nor j^μ_A transforms under the symmetry. (The fact that j^μ does not transform is the statement that the symmetry is Abelian.) The result is that the Ward identity for the conserved vector current takes a particularly simple form in momentum space,

$$p_{1\mu}\Gamma^{\mu\nu\rho}(p_1, p_2, q) = 0 \qquad (3.37)$$

and, equivalently,

$$p_{2\nu}\Gamma^{\mu\nu\rho}(p_1, p_2, q) = 0$$

Meanwhile, we can run exactly the same argument for the conservation of the axial symmetry to find

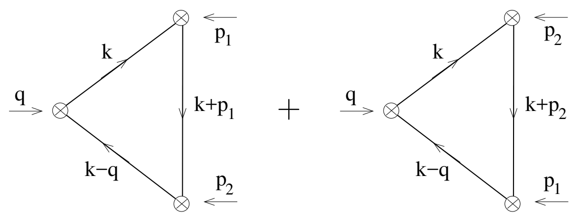
$$q_\rho \Gamma^{\mu\nu\rho}(p_1, p_2, q) = 0 \quad \Leftrightarrow \quad -(p_{1\rho} + p_{2\rho})\Gamma^{\mu\nu\rho}(p_1, p_2, q) = 0 \tag{3.38}$$

where the equivalence of these expressions comes from 4-momentum conservation: $p_1 + p_2 + q = 0$. (Note that a different index is contracted so this final expression does not follow from the previous two.) As we will now see, the anomaly means that things aren't quite this simple.

Triangle Diagrams

The leading order contribution to our three-point function comes from one-loop triangle diagrams,

$$-i\Gamma^{\mu\nu\rho}(p_1, p_2, q) = \tag{3.39}$$



In terms of equations, these diagrams read

$$-i\Gamma^{\mu\nu\rho}(p_1, p_2, q) = -\int \frac{d^4k}{(2\pi)^4}\ \mathrm{tr}\left[\frac{i}{\not{k}}\,\gamma^\rho\gamma^5\,\frac{i}{\not{k} - \not{q}}\,\gamma^\nu\,\frac{i}{\not{k} + \not{p}_1}\,\gamma^\mu\right] + \begin{pmatrix} p_1 \leftrightarrow p_2 \\ \mu \leftrightarrow \nu \end{pmatrix}$$

where the overall minus sign comes from Wick contracting the fermions and the trace is over the gamma matrix structure.

We will check all three of the Ward identities above. We start with the one we are most nervous about: (3.38). This now reads

$$-iq_\rho\Gamma^{\mu\nu\rho}(p_1, p_2, q) = i\int \frac{d^4k}{(2\pi)^4}\ \mathrm{tr}\left[\frac{1}{\not{k}}\,\not{q}\,\gamma^5\,\frac{1}{\not{k} - \not{q}}\,\gamma^\nu\,\frac{1}{\not{k} + \not{p}_1}\,\gamma^\mu\right] + \begin{pmatrix} p_1 \leftrightarrow p_2 \\ \mu \leftrightarrow \nu \end{pmatrix}$$

To proceed, we use the identity

$$\not{q}\gamma^5 = -\gamma^5\not{q} = \gamma^5(\not{k} - \not{q}) + \not{k}\gamma^5$$

to find

$$-iq_\rho\Gamma^{\mu\nu\rho}(p_1, p_2, q) = i\int \frac{d^4k}{(2\pi)^4}\ \mathrm{tr}\left[\frac{1}{\not{k}}\left(\gamma^5(\not{k} - \not{q}) + \not{k}\gamma^5\right)\frac{1}{\not{k} - \not{q}}\,\gamma^\nu\,\frac{1}{\not{k} + \not{p}_1}\,\gamma^\mu\right] + \begin{pmatrix} p_1 \leftrightarrow p_2 \\ \mu \leftrightarrow \nu \end{pmatrix}$$

$$
\begin{aligned}
= i\int \frac{d^4k}{(2\pi)^4}\,\mathrm{tr}\Bigg[&\left(\frac{1}{\not k}\,\gamma^5+\gamma^5\frac{1}{\not k-\not q}\right)\gamma^\nu\,\frac{1}{\not k+\not p_1}\,\gamma^\mu \\
&+\ \left(\frac{1}{\not k}\,\gamma^5+\gamma^5\frac{1}{\not k-\not q}\right)\gamma^\mu\,\frac{1}{\not k+\not p_2}\,\gamma^\nu\Bigg]
\end{aligned}
$$

We're left with four terms. We gather them like this:

$$
-iq_\rho\Gamma^{\mu\nu\rho}(p_1,p_2,q)=\Delta_1^{\mu\nu}+\Delta_2^{\mu\nu}
$$

where

$$
\begin{aligned}
\Delta_1^{\mu\nu} &= i\int \frac{d^4k}{(2\pi)^4}\,\mathrm{tr}\left[\frac{1}{\not k}\,\gamma^5\,\gamma^\nu\,\frac{1}{\not k+\not p_1}\,\gamma^\mu+\gamma^5\frac{1}{\not k-\not q}\,\gamma^\mu\,\frac{1}{\not k+\not p_2}\,\gamma^\nu\right] \\
&= i\int \frac{d^4k}{(2\pi)^4}\,\mathrm{tr}\left[\frac{1}{\not k}\,\gamma^5\,\gamma^\nu\,\frac{1}{\not k+\not p_1}\,\gamma^\mu-\frac{1}{\not k+\not p_2}\,\gamma^5\gamma^\nu\frac{1}{\not k-\not q}\,\gamma^\mu\right]
\end{aligned}
$$

and

$$
\begin{aligned}
\Delta_2^{\mu\nu} &= i\int \frac{d^4k}{(2\pi)^4}\,\mathrm{tr}\left[\gamma^5\frac{1}{\not k-\not q}\gamma^\nu\,\frac{1}{\not k+\not p_1}\,\gamma^\mu+\frac{1}{\not k}\,\gamma^5\gamma^\mu\,\frac{1}{\not k+\not p_2}\,\gamma^\nu\right] \\
&= i\int \frac{d^4k}{(2\pi)^4}\,\mathrm{tr}\left[-\frac{1}{\not k+\not p_1}\,\gamma^5\gamma^\mu\,\frac{1}{\not k-\not q}\gamma^\nu+\frac{1}{\not k}\,\gamma^5\gamma^\mu\,\frac{1}{\not k+\not p_2}\,\gamma^\nu\right]
\end{aligned}
$$

where in each case we go to the second line by using the cyclicity of the trace and the fact that $\{\gamma^\mu,\gamma^5\}=0$. The advantage of collecting the terms in this way is that it naively looks as if both $\Delta_1^{\mu\nu}$ and $\Delta_2^{\mu\nu}$ cancel. For example, in $\Delta_1^{\mu\nu}$, all we need to do is shift the integration variable in the first term from k to $k+p_2$. Using momentum conservation $p_1+p_2=-q$, we see that the two terms then cancel. Something similar happens for $\Delta_2^{\mu\nu}$. Taken at face value, it looks like we've succeeded in showing the Ward identity (3.38). Right? Well, no.

The problem with this argument is that all the integrals above are divergent. Indeed, all the terms in Δ_1 and Δ_2 have two powers of k in the numerator, yet we integrate over d^4k, suggesting that they diverge quadratically. In fact, as we'll see below, the gamma-matrix structure means that the divergence is actually linear. When dealing with such objects we need to be more careful.

There are a number of ways to deal with these differences of divergent integrals. Here we'll pick a particular path. Consider the general integral of the form

$$
\tilde\Delta = i\int \frac{d^4k}{(2\pi)^4}\,\Big[f(k)-f(k+a)\Big] \tag{3.40}
$$

where $f(k)$ is such that each individual integral $\int d^4k\ f(k)$ is linearly divergent. If we Taylor expand for small a, we have

$$\tilde{\Delta} = -i \int \frac{d^4k}{(2\pi)^4}\ \left[a^\mu \partial_{k^\mu} f + \frac{1}{2} a^\mu a^\nu \partial_{k^\mu}\partial_{k^\nu} f + \dots\right]$$

Each term above is a boundary term. Moreover, each term in the expansion is less and less divergent. If the original integral is only linearly divergent we need keep only the first of these terms. We have

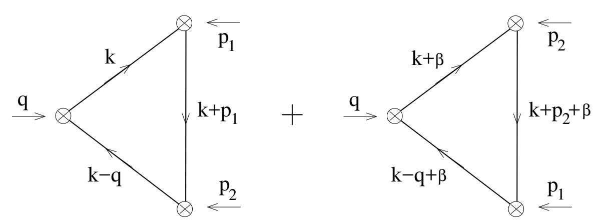
$$\tilde{\Delta} = -i \int_{\mathbf{S}^3_\infty} \frac{d\hat{k}_\mu}{(2\pi)^4}\ a^\mu\, |k|^3 f(k) \tag{3.41}$$

where the integral is taken over the boundary $\mathbf{S}^3$ at $|k| \to \infty$. We'll now look at what this surface integral gives us for our triangle diagram.

An Ambiguity in the Integrals

To proceed, let's first go back to the beginning and allow a general offset, β^μ, between the momenta that run in the two loops. We then replace (3.39) with

$$-i\Gamma^{\mu\nu\rho}(p_1, p_2, q) =$$



We will first find that the final answer depends on this arbitrary parameter β. We will then see how to resolve the ambiguity.

Following our manipulations above, we write this as

$$-iq_\rho \Gamma^{\mu\nu\rho}(p_1, p_2, q) = \tilde{\Delta}_1^{\mu\nu} + \tilde{\Delta}_2^{\mu\nu} \tag{3.42}$$

where

$$\tilde{\Delta}_1^{\mu\nu} = i \int \frac{d^4k}{(2\pi)^4}\ \mathrm{tr}\left[\frac{1}{\not{k}}\,\gamma^5\,\gamma^\nu\,\frac{1}{\not{k}+\not{p}_1}\,\gamma^\mu - \frac{1}{\not{k}+\not{\beta}+\not{p}_2}\,\gamma^5\gamma^\nu\,\frac{1}{\not{k}+\not{\beta}-\not{q}}\,\gamma^\mu\right] \tag{3.43}$$

and

$$\tilde{\Delta}_2^{\mu\nu} = i \int \frac{d^4k}{(2\pi)^4}\ \mathrm{tr}\left[-\frac{1}{\not{k}+\not{p}_1}\,\gamma^5\gamma^\mu\,\frac{1}{\not{k}-\not{q}}\gamma^\nu + \frac{1}{\not{k}+\not{\beta}}\,\gamma^5\gamma^\mu\,\frac{1}{\not{k}+\not{\beta}+\not{p}_2}\,\gamma^\nu\right] \tag{3.44}$$

Each of these is of the form (3.40). For the $\tilde{\Delta}_1^{\mu\nu}$, we have a difference of two divergent integrals, with integrand

$$f^{\mu\nu}(k) = \text{tr}\,\left[\frac{1}{\not{k}}\,\gamma^5\,\gamma^\nu\,\frac{1}{\not{k}+\not{p}_1}\,\gamma^\mu\right] = \frac{1}{k^2(k+p_1)^2}\,\text{tr}\,\left[\not{k}\gamma^5\gamma^\nu(\not{k}+\not{p}_1)\gamma^\mu\right]$$

We now use the gamma matrix identity

$$\text{tr}\,(\gamma^\nu\gamma^\rho\gamma^\mu\gamma^\sigma\gamma^5) = -4i\epsilon^{\nu\rho\mu\sigma}$$

to write

$$f^{\mu\nu}(k) = -4i\epsilon^{\nu\rho\mu\sigma}\,\frac{(k+p_1)^\rho k^\sigma}{k^2(k+p_1)^2} = -4i\epsilon^{\nu\rho\mu\sigma}\,\frac{p_1^\rho k^\sigma}{k^2(k+p_1)^2}$$

In the second equality, we've used the anti-symmetry of the epsilon tensor to remove the $k^\rho k^\sigma$ term. This is why – as advertised above – our integrals are actually linearly divergent rather than quadratically divergent. We can now simply apply the result (3.41) to the cases of interest. For the integral $\tilde{\Delta}_1^{\mu\nu}$, the off-set is given by $a = \beta + p_2$, and we have

$$\tilde{\Delta}_1^{\mu\nu} = -4\int_{\mathbf{S}^3_\infty}\frac{d\hat{k}^\lambda}{(2\pi)^4}\;\epsilon^{\nu\rho\mu\sigma}(\beta+p_2)_\lambda p_{1\rho}k_\sigma\frac{|k|^3}{k^2(k+p_1)^2}$$

To perform the integration over $\mathbf{S}^3$, we use

$$\int_{\mathbf{S}^3} d\hat{k}^\lambda k^\sigma = \frac{1}{4}\,\delta^{\lambda\sigma}\,\text{Vol}(\mathbf{S}^3)$$

with $\text{Vol}(\mathbf{S}^3) = 2\pi^2$. We find

$$\tilde{\Delta}_1^{\mu\nu} = -\frac{1}{8\pi^2}\epsilon^{\mu\nu\rho\sigma}p_{1\rho}(\beta+p_2)_\sigma$$

We can go through the same steps to evaluate $\tilde{\Delta}_2^{\mu\nu}$ in (3.44). This time we have the off-set $a = p_1 - \beta$ and find

$$\tilde{\Delta}_2^{\mu\nu} = +\frac{1}{8\pi^2}\epsilon^{\mu\nu\rho\sigma}p_{2\rho}(p_1-\beta)_\sigma$$

The Ward identity for the axial symmetry (3.42) then becomes

$$-iq_\rho\Gamma^{\mu\nu\rho} = -\frac{1}{8\pi^2}\epsilon^{\mu\nu\rho\sigma}\Big[2p_{1\rho}p_{2\sigma} + (p_1+p_2)_\rho\beta_\sigma\Big] \tag{3.45}$$

As we suspected, this depends on our arbitrary 4-momentum β. The question is: how do we fix β?

Resolving the Ambiguity

The answer comes by looking at the Ward identity (3.37) for the vector symmetry. It turns out that this too depends on β. Indeed, we have

$$-ip_{1\mu}\Gamma^{\mu\nu\rho} = i\int \frac{d^4k}{(2\pi)^4}\,\mathrm{tr}\left[\frac{1}{\not{k}}\gamma^\rho\gamma^5\frac{1}{\not{k}-\not{q}}\gamma^\nu\frac{1}{\not{k}+\not{p}_1}\not{p}_1 + \frac{1}{\not{k}}\gamma^\rho\gamma^5\frac{1}{\not{k}-\not{q}}\not{p}_1\frac{1}{\not{k}+\not{p}_2}\gamma^\nu\right]$$

Playing the same kind of games that we saw above, we have an anomalous Ward identity for the vector current

$$-ip_{1\mu}\Gamma^{\mu\nu\rho} = \frac{1}{8\pi^2}\epsilon^{\rho\nu\mu\sigma}\,p_{1\mu}(\beta - p_2)_\sigma$$

Similarly, the other vector Ward identity reads

$$-ip_{2\nu}\Gamma^{\mu\nu\rho} = \frac{1}{8\pi^2}\epsilon^{\rho\mu\nu\sigma}p_{2\nu}(\beta + p_1)_\sigma$$

We learn that all three Ward identities depend on the arbitrary 4-momentum β. This provides the clue that we need in order to determine β. Suppose that we wish to insist that the vector current survives quantisation. Indeed, this must be the case if we wish to couple this to a background gauge field. In this case, we must choose a β such that the two vector Ward identities are non-anomalous. For this, we must have

$$\beta - p_2 \sim p_1 \quad \text{and} \quad \beta + p_1 \sim p_2 \quad \Rightarrow \quad \beta = p_2 - p_1$$

With this choice

$$-ip_{1\mu}\Gamma^{\mu\nu\rho} = -ip_{2\nu}\Gamma^{\mu\nu\rho} = 0$$

while the axial Ward identity (3.45) becomes

$$-iq_\rho\Gamma^{\mu\nu\rho} = -\frac{1}{2\pi^2}\epsilon^{\mu\nu\rho\sigma}\,p_{1\rho}p_{2\sigma} \tag{3.46}$$

This is the anomaly for the free fermion.

Our discussion above looks rather different from the path integral approach of Section 3.2.2. We see that we have an arbitrary parameter β which allows us to shift the anomaly between the axial and vector currents. Why did we miss this before? The reason is that we chose a specific regulator – first introduced in (3.28) – which was gauge invariant. By construction, this ensures that the vector symmetry is preserved at the expense of the axial symmetry.

More generally, different regulators will violate some linear combination of the symmetry. Usually, it is the axial symmetry which suffers. For example, if we use Pauli-Villars, we should need to introduce a massive fermion and the mass term explicitly breaks the axial symmetry.

Including Gauge Fields

So far, the anomaly in momentum space (3.46) looks rather different from our original version (3.34)

$$\partial_\mu j^\mu_A = \frac{e^2}{16\pi^2}\epsilon^{\mu\nu\rho\sigma}F_{\mu\nu}F_{\rho\sigma} \tag{3.47}$$

However, they are actually the same formula in disguise. To see this, we couple the vector current $j^\mu = \bar{\psi}\gamma^\mu\psi$ to a $U(1)$ gauge field A_μ, so the fermions are now described by the action

$$S = \int d^4x \; i\bar{\psi}\gamma^\mu(\partial_\mu - ieA_\mu)\psi \tag{3.48}$$

For the purposes of our discussion, A_μ could be either a fixed, background field or, alternatively, a dynamical gauge field. From our previous definitions we have

$$-iq_\rho\Gamma^{\mu\nu\rho} = \int d^3x_1 d^3x_2 d^3x_3 \; \langle 0|T(j^\mu\, j^\nu\, \partial_\rho j^\rho_A)|0\rangle e^{ip_1\cdot x_1 + ip_2\cdot x_2 + iq\cdot x_3}$$

where we've omitted the delta-function $\delta^3(p_1+p_2+q)$ from the left-hand-side, as well as various arguments. Using the chiral anomaly in the form (3.47), we can write

$$\begin{aligned}\langle 0|T(j^\mu\, j^\nu\, \partial_\rho j^\rho_A)|0\rangle &= \frac{e^2}{4\pi^2}\epsilon^{\rho\sigma\lambda\tau}\langle 0|T(j^\mu\, j^\nu\, \partial_\rho A_\sigma\, \partial_\lambda A_\tau|0\rangle \\ &= \frac{e^2}{4\pi^2}\epsilon^{\rho\sigma\lambda\tau}\langle 0|j^\mu\, \partial_\rho A_\sigma|0\rangle\langle 0|j^\nu\, \partial_\lambda A_\tau|0\rangle + \text{permutation}\end{aligned}$$

But the two-point function of the current and gauge field can be read off from the Feynman rules for the action (3.48)

$$e\langle 0|j^\mu(x_1)A_\sigma(x_3)|0\rangle = -i\delta^\mu_{\ \sigma}\delta^4(x_1-x_3)$$

A little algebra then allows us to reproduce the anomaly in momentum space,

$$-iq_\rho\Gamma^{\mu\nu\rho} = -\frac{1}{2\pi^2}\epsilon^{\mu\nu\rho\sigma}\, p_{1\rho}p_{2\sigma}$$

As we mentioned in Section 3.2.2, when the gauge fields are dynamical one might worry about higher order corrections to the anomaly. It turns out that these don't arise. This was first proven by Adler and Bardeen by explicit analysis of the higher-loop Feynman diagrams. We will give a more modern, topological viewpoint on this in Section 3.3.1.

3.2.4 Chiral Anomalies and Gravity

There is a second, related contribution to the axial anomaly. This doesn't arise when the theory is coupled to background electric fields, but instead when the theory is coupled to curved spacetime. As before, this effect arises either for quantum field theory in a fixed, background spacetime, or for quantum field theory coupled to gravity which, of course, means dynamical spacetime.

Let's first review how to couple spinors to a curved spacetime. The starting point is to decompose the metric in terms of vierbeins,

$$g_{\mu\nu}(x) = e^a_\mu(x)\, e^b_\nu(x)$$

There is an arbitrariness in our choice of vierbein, and this arbitrariness introduces an $SO(3,1)$ gauge symmetry into the game. The associated gauge field ω^{ab}_μ is called the *spin connection*. It is determined by the requirement that the vierbeins are covariantly constant

$$\mathcal{D}_\mu e^a_\nu \equiv \partial_\mu e^a_\nu - \Gamma^\rho_{\mu\nu} e^a_\lambda + \omega^a_{\mu\,b} e^b_\nu = 0$$

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

$$(R_{\mu\nu})^a{}_b = \partial_\mu \omega^a_{\nu\,b} - \partial_\nu \omega^a_{\mu\,b} + [\omega_\mu, \omega_\nu]^a{}_b$$

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

$$\mathcal{D}_\mu \psi_\alpha = \partial_\mu \psi_\alpha + \frac{1}{2}\omega^{ab}_\mu (S_{ab})^\beta{}_\alpha \psi_\beta$$

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 spinors to a curved spacetime looks very similar to the coupling to electromagnetic fields. It is not surprising, therefore, that there is a gravitational contribution to the anomaly. The kind of manipulations we performed previously now give

$$\mathcal{D}_\mu j^\mu_A = -\frac{1}{384\pi^2}\epsilon^{\mu\nu\rho\sigma} R_{\mu\nu\lambda\tau} R_{\rho\sigma}{}^{\lambda\tau} \tag{3.49}$$

3.3 Fermi Zero Modes

The anomaly was first discovered in the early 1970s in an attempt to make sense of the observed decay rate of the neutral pion to a pair of photons. We will tell this story in Section 5.4.3 where we describe some aspects of the spectrum of QCD.

Here, instead, we focus on ways in which the anomaly fits into our general understanding of fermions coupled to gauge fields.

3.3.1 The Atiyah-Singer Index Theorem

The anomaly has a rather nice mathematical interpretation: it is a manifestation of the famous Atiyah-Singer index theorem.

Consider again the Dirac operator in Euclidean space in the background of a general gauge field A_μ. The operator $i\not{D}$ is Hermitian and so has real eigenvalues.

$$i\not{D}\phi_n = \lambda_n \phi_n$$

with $\lambda_n \in \mathbf{R}$. Whenever we have an eigenfunction ϕ_n with $\lambda_n \neq 0$ then $\gamma^5\phi_n$ is also an eigenfunction. This follows because $\gamma^\mu\gamma^5 = -\gamma^5\gamma^\mu$ for $\mu = 1, 2, 3, 4$ so

$$i\not{D}(\gamma^5\phi_n) = -i\gamma^5\,\not{D}\phi_n = -\lambda_n\gamma^5\phi_n \tag{3.50}$$

We see that all non-zero eigenvalues come in $\pm\lambda_n$ pairs. Moreover, ϕ_n and $\gamma^5\phi_n$ must be orthogonal functions. Evidently, the eigenfunctions with $\lambda_n \neq 0$ cannot also be eigenfunctions of γ^5.

However, the zero eigenvalues are special because the argument above no longer works. The corresponding eigenfunctions are called *zero modes*. Now, it may well be that ϕ_n and $\gamma^5\phi_n$ are actually the same functions. More generally, for the zero modes we can simultaneously diagonalise $i\not{D}$ and γ^5 (because both ϕ_n and $\gamma^5\phi_n$ have the same $i\not{D}$ eigenvalue, namely zero). Since $(\gamma^5)^2 = 1$, the possible eigenvalues of γ^5 are ± 1. We then define n_+ and n_- to be the number of zero modes of $i\not{D}$ with γ^5 eigenvalue $+1$ and -1 respectively. The total number of zero modes is obviously $n_+ + n_-$. The index of the Dirac operator is defined to be

$$\text{Index}(i\not{D}) = n_+ - n_-$$

But we have actually computed this index as part of our derivation of the anomaly above! To see this, consider again the result (3.32)

$$\sum_n \bar{\phi}_n\gamma^5\phi_n = \frac{e^2}{32\pi^2}\epsilon^{\mu\nu\rho\sigma}F_{\mu\nu}F_{\rho\sigma}$$

This is rather formal since, in $\mathbf{R}^4$ there will be a continuum of eigenvalues labelled by the index n. However, we can always compactify the theory on your favourite four-manifold and the spectrum will become discrete. If we then integrate this equation

$$\int d^4x \ \sum_n \bar{\phi}_n \gamma^5 \phi_n = \frac{e^2}{32\pi^2} \int d^4x \ \epsilon^{\mu\nu\rho\sigma} F_{\mu\nu} F_{\rho\sigma}$$

Then we note that only the zero modes contribute to the left-hand side. This is because, as we saw above, whenever $\lambda_n \neq 0$ then ϕ_n and $\gamma^5\phi_n$ are orthogonal functions. This means that the left-hand-side is the index that we want to compute

$$\int d^4x \ \sum_n \bar{\phi}_n \gamma^5 \phi_n = \int d^4x \sum_{\text{zero modes}} \bar{\phi}_n \gamma^5 \phi_n = n_+ - n_-$$

We get our final result

$$\text{Index}(i\not{D}) = \frac{e^2}{32\pi^2} \int d^4x \ \epsilon^{\mu\nu\rho\sigma} F_{\mu\nu} F_{\rho\sigma}$$

This is the *Atiyah-Singer index theorem*. Mathematicians usually state this in units where $e = 1$. Note that the right-hand side is exactly the quantity that we showed to be an integer in Section 1.2.4 when considering the theta angle in Maxwell theory.

The connection to the index theorem is our first hint that there is something deep about the anomaly. To illustrate this in physical terms, consider our theory on the space $\mathbf{R} \times X$, where X is a closed spatial 3-manifold. We define the axial charge $Q_A = \int_X j^0_A$. We also parameterise $\mathbf{R}$ by t (think "time" even though we're in Euclidean space). Then the integrated anomaly equation tells us the change in the charge,

$$\Delta Q_A = Q_A\Big|_{t=+\infty} - Q_A\Big|_{t=-\infty} = \int d^4x \ \frac{e^2}{16\pi^2} \epsilon^{\mu\nu\rho\sigma} F_{\mu\nu} F_{\rho\sigma} \tag{3.51}$$

The left-hand side is an integer because of quantum mechanics. Meanwhile, the right-hand side is an integer because of topology. The anomaly equation relates these two ideas.

This connection to topology also explains why the anomaly equation (3.34) (or, for non-Abelian gauge theories, (3.35)) is exact, and does not get corrected at higher order in perturbation theory. It is simply because the right-hand side of (3.51) is an integer and any corrections — say, at order e^4 — would change this.

3.3.2 Instantons Revisited

The anomaly tells us that, in spite of classical appearances, $U(1)_A$ is not really a symmetry of our theory. This, in turn, means that the axial charge is not conserved. The result (3.51) tells us that we expect to see violation of this charge when $\int d^4x\ F^\star F$ is non-zero. This tallies with the picture we built up in Section 3.1.2, where we needed to turn on constant background electric and magnetic fields to see that the axial charge is not conserved.

At this point, there is an important difference between Abelian and non-Abelian theories. This arises because non-Abelian theories have finite action configurations with $\int d^4x\ F^\star F \neq 0$. Among these are the classical instanton solutions that we described in Section 2.3. This means that the path integral about the vacuum state will include configurations which give rise to the violation of axial charge.

In contrast, Abelian theories have no finite action configurations which change the axial charge; such a process will not happen dynamically about the vacuum, but must be induced by turning on background fields as in Section 3.1.2. (This is true at least on $\mathbf{R}^4$; the situation changes on compact manifolds and the Abelian theories are closer in spirit to their non-Abelian counterparts.)

It's worth understanding in more detail how instantons can give rise to violation of axial charge. Let's start by revisiting the calculation of Section 2.3, where we showed that instantons provide a semi-classical mechanism to tunnel between the $|n\rangle$ vacua of Yang-Mills. The end result of that calculation was that the true physical ground states of Yang-Mills are given by the theta vacua (2.43)

$$|\theta\rangle = \sum_n e^{i\theta n}|n\rangle$$

Now what happens if we have a massless fermion in the game? As we've seen above, in the background of an instanton a massless quark will have a zero mode. Performing the path integral over the fermion fields then gives the amplitude for tunnelling between two $|n\rangle$ ground states. Schematically, we have

$$\begin{aligned}\langle n|n+\nu\rangle &\sim \int \mathcal{D}A\mathcal{D}\psi\mathcal{D}\bar{\psi}\ \exp\left(-\int d^4x\ \frac{1}{2g^2}\mathrm{tr}\, F^{\mu\nu}F_{\mu\nu} + i\bar{\psi}\,\not{D}\psi\right)\\ &\sim \int \mathcal{D}A\ \det(i\,\not{D})\exp\left(-\int d^4x\ \frac{1}{2g^2}\mathrm{tr}\, F^{\mu\nu}F_{\mu\nu}\right)\end{aligned}$$

Previously, this amplitude received a non-vanishing contribution from instantons with winding number ν. Now, however, the fermion has a zero mode in any such configuration. This means that $\det(i\not{D}) = 0$. We see that the presence of a massless fermion suppresses the vacuum tunnelling of Section 2.3.

While instantons no longer give rise to vacuum tunnelling, they do still have a role to play for, as we anticipated above, they now violate axial charge. To see how this happens, let's tease apart the calculation above. Following (3.25), we expand our fermion fields in terms of eigenspinors ϕ_n and $\bar{\phi}_n$,

$$\psi(x) = \sum_n a_n \phi_n(x) \quad \text{and} \quad \bar{\psi}(x) = \sum_n \bar{b}_n \bar{\phi}_n(x)$$

where a_n and b_n are Grassmann-valued numbers and the eigenspinors obey

$$i\not{D}\phi_n = \lambda_n \phi_n$$

The action for the fermions is

$$S = \int d^4x \; i\bar{\psi}\not{D}\psi = \sum_n \lambda_n \bar{b}_n a_n$$

A fermion zero mode is an eigenspinor – which we will denote as ϕ_0 – with $\lambda_0 = 0$. This means that the corresponding Grassmann parameters a_0 and b_0 do not appear in the action. When we compute the fermionic path integral, we have

$$\begin{aligned}\int \mathcal{D}\psi\mathcal{D}\bar{\psi} \; \exp\left(-\int d^4x \; i\bar{\psi}\,\not{D}\psi\right) &= \prod_n \int da_n d\bar{b}_n \; \exp\left(\sum_m \lambda_m \bar{b}_m a_m\right) \\ &= \prod_n \int da_n d\bar{b}_n \; \prod_m \left(1 + \lambda_m \bar{b}_m a_m\right)\end{aligned}$$

But Grasmmann integrals are particularly easy: they're either zero or one, with $\int da = 0$ and $\int da \; a = 1$. The integration above vanishes whenever there is a fermi zero mode because there's nothing to soak up the integration over the associated Grassmann variables a_0 and b_0. This is why massless fermions cause the instanton tunnelling amplitude to vanish.

We learn that we're only going to get a non-vanishing answer from instantons if we compute a correlation function that includes the fermion zero mode. This leads to a rather pretty superselection rule. Consider the correlation function

$$\langle \bar{\psi}_- \psi_+ \rangle$$

This is known as a *chiral condensate*. This has axial charge +2. If $U(1)_A$ is a good, unbroken symmetry of our theory then we would expect this to vanish in the vacuum. However, we know that $U(1)_A$ is, instead, anomalous. We will now see that this is reflected in a non-vanishing expectation value for the chiral condensate.

Written in terms of our eigenbasis, the chiral condensate becomes

$$\bar{\psi}_- \psi_+ = \frac{1}{2} \sum_{l,l'} \bar{b}_l a_{l'} \bar{\phi}_l (1 + \gamma^5) \phi_{l'}$$

where we're using the fact that $\gamma^5 \psi_+ = \psi_+$ to write ψ_+ as a projection of ψ onto the $+1$ eigenvalue of γ^5. We can then write the correlation function as

$$\begin{aligned} \langle \bar{\psi}_- \psi_+ \rangle &= \prod_n \int da_n d\bar{b}_n \prod_m \left(1 + \lambda_m \bar{b}_m a_m\right) \frac{1}{2} \sum_{l,l'} \bar{b}_l a_{l'} \bar{\phi}_l (1 + \gamma^5) \phi_{l'} \\ &= \left(\prod_n \lambda_n \right) \frac{1}{2} \left(\sum_l \frac{1}{\lambda_l} \bar{\phi}_l (1 + \gamma^5) \phi_l \right) \end{aligned} \tag{3.52}$$

We can look at the contributions to this from each instanton sector, ν. When we're in the trivial, $\nu = 0$, sector there are generically no zero modes so the product $\prod_n \lambda_n \neq 0$. (One might wonder whether perhaps $n_+ = n_- \neq 0$. This is possible, but generically will not be the case.) However, as we saw in (3.50), the eigenvalues λ_n come in $\pm$ pairs, a fact which follows from the existence of γ^5. This means that the sum over λ_l^{-1} will contain equal and opposite contributions, and the contribution from the trivial instanton sector is $\langle \bar{\psi}\psi \rangle_{\nu=0} = 0$.

In contrast, interesting things happen when we have winding $\nu = 1$. Now there is a single zero mode which obeys $\gamma^5 \phi_0 = +\phi_0$. But the multiplication by λ_0 in the product is precisely cancelled by the $\bar{\phi}_0 \phi_0$ term in the sum. We see that, in this semi-classical approximation,

$$\langle \bar{\psi}_- \psi_+ \rangle_{\nu=1} = \det{}'(i\not{D})\, \bar{\phi}_0 \phi_0$$

where $\det{}'$ means that you multiply over all eigenvalues, but omit the zero modes.

In fact, this is the only topological sector that contributes to $\langle \bar{\psi}_- \psi_+ \rangle$. When $\nu = -1$, we also have a zero mode but it has opposite chirality, $\gamma^5 \phi_0 = -\phi_0$, and so does not contribute. Instead, this sector will contribute to $\langle \bar{\psi}_+ \psi_- \rangle$.

Meanwhile, when $|\nu| \geq 2$, we have more than one zero mode and the integral (3.52) again vanishes. Instead, these sectors will contribute to correlators of the form $\langle\, (\bar{\psi}_- \psi_+)^\nu \rangle$.

3.3.3 The Theta Term Revisited

We saw above that the existence of massless fermions – and, in particular, their fermi zero modes – quashes the tunnelling between $|n\rangle$ vacua. This leaves us with a question: what becomes of the theta angle?

The answer to this is hiding within our path integral derivation of the anomaly. Consider a single Dirac fermion coupled to a gauge field (either Abelian or non-Abelian, it doesn't matter) and make a chiral rotation (3.21). On left- and right-handed spinors, this acts as

$$\psi_+ \to e^{i\alpha}\psi_+ \quad \text{and} \quad \psi_- \to e^{-i\alpha}\psi_- \tag{3.53}$$

The upshot of our long calculation in Section 3.2.2 is that the measure transforms as (3.33),

$$\int \mathcal{D}\psi\mathcal{D}\bar{\psi} \;\longrightarrow \int \mathcal{D}\psi\mathcal{D}\bar{\psi}\; \exp\left(-\frac{ie^2\alpha}{16\pi^2}\int d^4x\; \epsilon^{\mu\nu\rho\sigma}F_{\mu\nu}F_{\rho\sigma}\right)$$

But this is something that we've seen before: it is the theta-term that we introduced for Maxwell theory in Section 1.2 and for Yang-Mills in Section 2.2! We see that a chiral rotation (3.53) effectively shifts the theta-angle by

$$\theta \to \theta - 2\alpha \tag{3.54}$$

This means that the theta angle isn't really physical: it can be absorbed by changing the phase of the fermion.

(There is a caveat here: the mass for a single fermion might undergo additive renormalisation that shifts it away from zero. So it's not quite right to say that the theta angle ceases to exist when $m = 0$. Rather, we should say that for $m \in \mathbf{R}$, there is a single value where the theta-angle becomes unphysical. Note that this issue doesn't arise if multiple fermions become massless because then we get an enhanced chiral symmetry which prohibits an additive mass renormalisation.)

This ties in with our discussion of instantons in the previous section. We saw that the chiral condensate $\langle\bar{\psi}_-\psi_+\rangle$ receives a contribution only from topological sectors with winding $\nu = 1$. If we added a theta term in the action, we would find $\langle\bar{\psi}_-\psi_+\rangle \sim e^{i\theta}$, since $e^{i\theta}$ is the sign of a single instanton. This agrees with our result (3.54).

The discussion above shows that the parameter θ can be absorbed into a dynamical field, which is the phase of the fermion. But we can also turn this idea on its head. Suppose that we hadn't realised that $U(1)_A$ was anomalous, but we knew that $\langle\bar{\psi}_-\psi_+\rangle \neq 0$. We might be tempted to conclude that this condensate has broken a global symmetry and would be entitled to expect the existence of an associated Goldstone boson, which is the phase of the condensate. Yet no such Goldstone boson exists. One can view the would-be Goldstone boson as θ, but it is a parameter of the theory, rather than a dynamical field!

With more than one massless fermion, there are also fermionic condensates that break the non-anomalous part of the chiral flavour symmetry. These are not due to instantons and, this time, we do get Goldstone modes. Their story is interesting enough that it gets its own chapter: it will be told in Section 5.

So far we have focussed on massless fermions. What happens for a massive fermion? Does the θ angle suddenly become active again? Well, sort of. For a Dirac fermion, we have two choices of mass term: either $\bar{\psi}\psi$ or $i\bar{\psi}\gamma^5\psi$. Only the former is invariant under parity, but both are allowed. Written in terms of the Weyl fermions, these two mass parameters naturally split into a modulus and complex phase,

$$\mathcal{L}_{\rm mass} = m\left(e^{i\phi}\psi_+^\dagger\psi_- + e^{-i\phi}\psi_-^\dagger\psi_+\right)$$

However, the anomaly means that we can trade the phase ϕ for a theta angle, or vice-versa. Only the linear combination $\theta + \phi$ has physical meaning. 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.

The Witten Effect Revisited

We spent quite a lot of time in earlier sections understanding how the theta angle is physical. Now we have to return to these arguments to understand why they fail in the presence of massless fermions. For example, in Section 1.2.3 we discussed the Witten effect, in which a magnetic monopole picks up an electric charge proportional to θ. What happens in the presence of a massless fermion?

The answer to this question is a little more subtle. For fermions of mass m, one finds that the fermions form a condensate around the monopole of size $\sim 1/m$ and, in the presence of a theta angle, this condensate carries an electric charge that is proportional to θ as expected by the Witten effect. As the mass $m \to 0$, this electric charge spreads out into an increasingly diffuse cloud until, in the massless limit, it is no longer possible to attribute it to the monopole.

3.3.4 Topological Insulators Revisited

The ideas above also give us a different perspective on the topological insulator that we met in Section 1.2.1. Consider a Dirac fermion in $d = 3+1$ dimensions, whose mass varies as a function of one direction, say $x^3 = z$. We couple this fermion to a $U(1)$ gauge field, so the action is

$$S = \int d^4x\; i\bar{\psi}\not{D}\psi - m(z)\bar{\psi}\psi$$

We take the mass to have the profile shown in the figure. In particular, we have

$$m(z) \to \begin{cases} +m & \text{as } z \to \infty \\ -m & \text{as } z \to -\infty \end{cases}$$

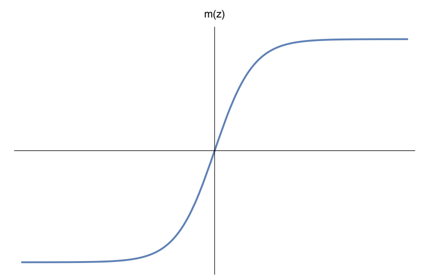


Figure 28:

with $m > 0$. If we perform a chiral rotation only in the region $z < 0$, we can make the mass positive again, but only at the expense of introducing a non-trivial $\theta = \pi$. In other words, the massive fermion above provides a microscopic realisation of the topological insulator. Note that the mass term $\bar{\psi}\psi$ is compatible with time reversal invariance as expected from the topological insulator. (In contrast, a mass term $\bar{\psi}\gamma^5\psi$ breaks time reversal.)

This set-up also brings something new. Let's turn off the gauge fields and study the Dirac equation. Using the gamma matrices (3.9), the Dirac equation is

$$\begin{aligned} i\partial_0\psi_- + i\sigma^i\partial_i\psi_- &= m\psi_+ \\ i\partial_0\psi_+ - i\sigma^i\partial_i\psi_+ &= m\psi_- \end{aligned} \tag{3.55}$$

Solutions to these equations include excitations propagating in the asymptotic $|z| \to \infty$ region, but these all cost energy $E \geq m$. However, there can be solutions with energy $E < m$ that are bound to the region $z \approx 0$. In general, the number of such bound states will depend on the properties of $m(z)$. But there is one special solution that always exists, providing the profile obeys (3.55). This is given by the ansatz

$$\psi_+ = i\sigma^3\psi_- = \exp\left(-\int^z dz'\; m(z')\right)\; \chi(x^0, x^1, x^2)$$

Note that this ansatz is localised around $z \approx 0$, dropping off exponentially as $e^{-m|z|}$ as $z \to \pm\infty$. It has the property that the ∂_z variation in (3.55) cancels the $m(z)$ dependence, leaving us with the 2-component spinor $\chi(x)$ which must satisfy

$$\partial_0\chi + \sigma^1\partial_1\chi + \sigma^2\partial_2\chi = 0$$

But this is the Dirac equation for a massless spinor in $d = 2 + 1$ dimensions. This is a Fermi zero mode, similar in spirit to those that we saw above associated to instantons. In the present context, such zero modes were first discovered by Jackiw and Rebbi.

We learn that, in this realisation, the boundary of the topological insulator houses a single gapless fermion. Indeed, these surface states can be observed in ARPES experiments and have become the poster boy for topological insulators. An example is shown on the right, beautifully revealing the relativistic $E = |k|$ dispersion relation.

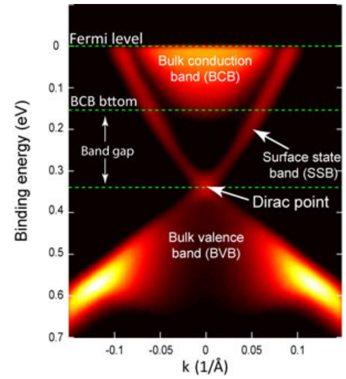


Figure 29:

Note that the surface of the topological insulator only houses a single, 3d Dirac fermion. The other putative zero mode would come from $\psi_+ = -i\sigma^3\psi_-$ but this solves the equations of motion only if $\psi_+ \sim \exp\left(+\int dz\; m(z)\right)$, and this is not normalisable.

There is an important technicality in the above story. As we have stressed, the topological insulator preserves time-reversal invariance. Yet it turns out that a single Dirac fermion in $d = 2+1$ dimensions does not preserve time-reversal. (We will discuss this in some detail in Section 8.5.) However, as the topological insulator shows, it is possible for time-reversal invariance to be preserved providing that the 3d fermion is housed as part of a larger 4d world. This is an example of a more general mechanism called *anomaly inflow* that will be described in Section 4.4.1.

3.4 Gauge Anomalies

The chiral anomaly of section 3.1 is an anomaly in a global symmetry: the naive conservation law of axial charge is violated in the quantum theory in the presence of gauge fields coupled to the vector current. Such anomalies in global symmetries are interesting: as we've seen, they are closely related to ideas of topology in gauge theory, and give rise to novel physical effects. (We will see the effect of the anomaly on pion decay in Section 5.4.3.)

In this section, we will focus on anomalies in *gauge* symmetries. While anomalies in global symmetries are physically interesting, anomalies in gauge symmetries kill all physics completely: they render the theory mathematically inconsistent! This is because "gauge symmetries" are not really symmetries at all, but redundancies in our description of the theory. Moreover, as we sketched in Section 2.1.2, these redundancies are necessary to make sense of the theory. An anomaly in gauge symmetry removes this redundancy. If we wish to build a consistent theory, then we must ensure that all gauge anomalies vanish.

There is a straightforward way to ensure that gauge symmetries are non-anomalous: only work with Dirac fermions, and with gauge fields which are coupled in the same manner to both left- and right-handed fermions. Such theories are called *vector-like*. Nothing bad happens.

Here we will be interested in a more subtle class of theories, in which left- and right-handed fermions are coupled differently to gauge fields. These are called *chiral gauge theories* and we have to work harder to ensure that they are consistent. Note that chiral gauge theories are necessarily coupled to only massless fermions. This is because a mass term requires both left- and right-handed Weyl fermions and is gauge invariant only if they transform in the same way under the gauge group. In other words, mass terms are only possible for vector-like matter.

We describe how to build chiral gauge theories with $U(1)$ gauge groups in section 3.4.1, with non-Abelian gauge groups in section 3.4.2 and with $SU(2)$ gauge groups (which turns out to be special) in section 3.4.3.

3.4.1 Abelian Chiral Gauge Theories

Here is an example of a bad theory: take a Dirac fermion and try to gauge both axial and vector symmetries. We know from our discussion in Section 3.1 that some combination of these will necessarily be anomalous.

Equivalently, we could consider a single $U(1)$ gauge theory coupled to just a single Weyl fermion, either left- or right-handed. This too will be anomalous, and therefore a sick theory.

So how can we construct a chiral gauge theory with a single $U(1)$ gauge field? We will have N_L left-handed Weyl fermions with charges $Q_i^L \in \mathbf{Z}$ and N_R right-handed Weyl fermions with charges $Q_j^R \in \mathbf{Z}$. To ensure that the triangle diagram vanishes, we require

$$\sum_{a=1}^{N_L} (Q_a^L)^3 = \sum_{j=1}^{N_R} (Q_j^R)^3 \tag{3.56}$$

There are obvious solutions to this equation with $N_L = N_R$ and $Q_a^L = Q_i^R$. These are the vector-like theories. Here we are interested in the less-obvious solutions, corresponding to chiral theories. We will assume that we have removed all vector-like matter, so that the left-handed and right-handed fermions have no charges in common.

We can simplify (3.56) a little. In $d = 3+1$ dimensions, the anti-particle of a right-handed fermion is left-handed: This means that we can always work with a set of purely left-handed fermions which have charges $Q_a = \{Q_i^L, -Q_j^R\}$. The requirement of anomaly cancellation is then

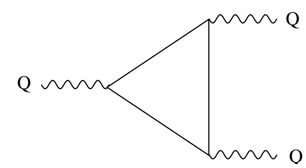


Figure 30:

$$\sum_{a=1}^{N} Q_a^3 = 0 \tag{3.57}$$

We would like to understand the possible solutions to this equation. In particular, what is the simplest set of charges that satisfies this?

Clearly for $N = 2$ fermions, the charges must come in a $\pm$ pair which is a vector-like theory. So let's look at $N = 3$. We must have two positive charges and one negative (or the other way round). Set $Q_a = (x, y, -z)$ with x, y, z positive integers. The condition for anomaly cancellation then becomes

$$x^3 + y^3 = z^3$$

Rather famously, this equation has no solutions: this is the result of Fermat's last theorem.

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

$$x^3 + y^3 + z^3 = w^3 \tag{3.58}$$

The simplest integers satisfying this are 3,4,5 and 6. Mathematicians have constructed a number of different parametric solutions to this equation, although not one that gives the most general solution. The simplest is due to Ramanujan,

$$\begin{aligned} x = 3n^2 + 5nm - 5m^2 \quad &, \quad y = 4n^2 - 4nm + 6m^2 \\ z = 5n^2 - 5nm - 3m^2 \quad &, \quad w = 6n^2 - 4nm + 4m^2 \end{aligned} \tag{3.59}$$

with n and m positive integers.

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

$$x^3 + y^3 = z^3 + w^3 \tag{3.60}$$

This equation is also closely associated to Ramanujan and the famous story of G. H. 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: $1^3 + 12^3 = 9^3 + 10^3$. The most general solution to (3.60) is known. Some of these can be generated by putting $m = n + 1$ into the Ramanujan formula (3.59) which, for $n \geq 3$, gives $x < 0$, and so yields solutions to (3.60) rather than (3.58)

Avoiding the Mixed Gravitational Anomaly

So far, we have been concerned only with cancelling the gauge anomaly. However, if we wish to place our theory on curved spacetime, then we must require that the mixed gauge-gravitational anomaly (3.49) also vanishes. For this, the diagram shown in the figure must also vanish when summed over all fermions, requiring

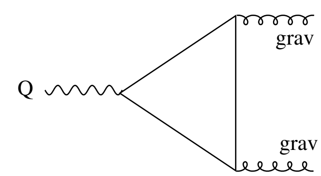


Figure 31:

$$\sum_{a=1}^{N} Q_a = 0 \tag{3.61}$$

Note that the diagram with two gauge fields and a single graviton vanishes because diffeomorphism symmetry is a non-Abelian group, and the trace of a single generator vanishes.

Our goal now is to find a set of charges which solve both (3.57) and (3.61)[7]. Let's first see that these cannot be satisfied by a set of $N = 4$ integers. To show that there can be no solutions with three positive integers and one negative, we could either plug in the explicit solution (3.59) or, alternatively use (3.61) to write $w = x + y + z$ which then implies that $w^3 > x^3 + y^3 + z^3$ in contradiction to (3.58). To see that no taxicab numbers can solve (3.61), write one pair as $x, y = a \pm b$ and the other pair as $z, w = c \pm d$ with $a, b, c, d \in \frac{1}{2}\mathbf{Z}^+$. Then (3.61) tells us that $a = c$, while (3.57) requires $b = d$.

It turns out that some questions we can ask about the solutions to (3.57) and (3.61) are hard. For example if you fix N it may be difficult to determine if there is a solution with a specified subset of charges. In contrast, it is straightforward to classify solutions if we place a bound, $|Q_a| \leq q$ on the charges. Consider the set of charges

$$\{Q_a\} = \{1^{[d_1]}, 2^{[d_2]}, \ldots, q^{[d_q]}\}$$

[7] I'm grateful to Imre Leader for explaining how to solve these equations.

where we use notation that d_p is the multiplicity of the charge p if $d_p > 0$, while $|d_p|$ is the multiplicity of $-p$ if $d_p < 0$. This notation has the advantage of removing any non-chiral matter since we can't have both charges p and $-p$. The two conditions (3.57) and (3.61) become

$$\sum_{p=1}^{q} p^3 d_p = 0 \quad \text{and} \quad \sum_{p=1}^{q} p d_p = 0 \tag{3.62}$$

This can be thought of as specifying two q-dimensional vectors which lie perpendicular to d_p. Solutions to these linear equations for $d_p \in \mathbf{Z}$ span a $(q-2)$-dimensional lattice. Each lattice point corresponds to a solution with the number of fermions given by $N = \sum_{p=1}^{q} |d_p|$.

Now we can address the question: what is the simplest chiral gauge theory? Of course, the answer depends on what you mean by "simple". For example, you may want the theory that contains the lowest charge q. In this case, the answer is the set of $N = 10$ fermions with charge

$$\{Q_a\} = \{1^{[5]}, 2^{[-4]}, 3\}$$

Alternatively, you may instead want to minimize the number of Weyl fermions N in the theory. The smallest solutions to (3.57) and (3.61) have $N = 5$ Weyl fermions. There are many such solutions, but the one with the lowest q is

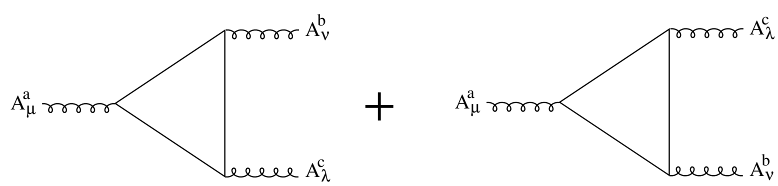
$$\{Q_a\} = \{1, 5, -7 - 8, 9\}$$

In general, the trick of changing the non-linear Diophantine equations (3.57) and (3.61) into the much simpler linear equations (3.62) means that it is simple to generate consistent chiral Abelian gauge theories.

Finally, to paraphrase Coleman, if you want your Hilbert space to contain structures capable of knowing joy, then the set of $N = 15$ fermions with charges $\{1^{[6]}, 2^{[3]}, 3^{[-2]}, 4^{[-3]}, 6\}$ is a good place to start; we'll see the importance of these charges in Section 3.4.4.

3.4.2 Non-Abelian Gauge Anomalies

We now turn to non-Abelian gauge theories with gauge group G. We have to worry about the familiar triangle diagrams, now with non-Abelian currents on each of the external legs:



The anomaly must be symmetric under $\nu \leftrightarrow \lambda$, and this symmetry then imposes itself on the group structure. The result is that a Weyl fermion in a representation $\mathbf{R}$, with generators T^a, contributes a term to the anomaly proportional to the totally symmetric group factor

$$d^{abc}(\mathbf{R}) = \mathrm{tr}\, T^a\{T^b, T^c\}$$

Furthermore, left and right-handed fermions contribute to the anomaly with opposite signs.

We will consider a bunch of left-handed Weyl fermions, transforming in representations $\mathbf{R}_{L\,i}$, with $i = 1\,\ldots, N_L$ and a bunch of right-handed Weyl fermions transforming in $\mathbf{R}_{R\,j}$ with $j = 1, \ldots, N_R$. The requirement for anomaly cancellation is then

$$\sum_{i=1}^{N_L} d^{abc}(\mathbf{R}_{L\,i}) = \sum_{j=1}^{N_R} d^{abc}(\mathbf{R}_{R\,j}) \tag{3.63}$$

As long as the gauge group is simply laced (i.e. contains no $U(1)$ factors) then there is no analog of the mixed gauge-gravitational anomaly (3.61) because $\mathrm{tr}\, T^a = 0$.

How can we satisfy (3.63)? One obvious way is to have an equal number of left- and right-handed fermions transforming in the same representations of the gauge group. A prominent example is QCD, which consists of $G = SU(3)$, coupled to $N_f = 6$ quarks, each of which is a Dirac fermion. For such vector-like theories, there is no difficulty in assigning mass terms to fermions which fits in with our theme that anomalies are associated only to massless fermions.

There are other, straightforward ways to solve (3.63). The anomaly vanishes for any representation that is either real (e.g. the adjoint) or pseudoreal (e.g. the fundamental of $SU(2)$). Here "pseudoreal" means that the conjugate representation $\bar{T}^a$ is related to the original T^a by a unitary matrix U, acting as

$$\bar{T}^a = UT^aU^{-1}$$

If we denote a group element by $e^{i\alpha^a T^a}$ then, in the conjugate representation, the same group element is given by $e^{-i\alpha^a T^{a\star}}$. This means that the conjugate representation can be written as $\bar{T}^a = -T^{a\star} = -(T_a)^T$, where the last equality follows because we can always take T^a to be Hermitian. The upshot of these arguments is that, for a real or pseudoreal representation,

$$\mathrm{tr}\, T^a\{T^b, T^c\} = \mathrm{tr}\, \bar{T}^a\{\bar{T}^b, \bar{T}^c\} = -\mathrm{tr}\, (T^a)^T\{(T^b)^T, (T^c)^T\} = -\mathrm{tr}\, T^a\{T^b, T^c\}$$

where the final equality comes from the fact that $\operatorname{tr} A = \operatorname{tr} A^T$. We learn that for any real or pseudoreal representation $\operatorname{tr} T^a\{T^b, T^c\} = 0$. Once again, this tallies nicely with the fact that anomalies are associated to fermions that are necessarily massless, since we can always write down a Majorana mass term for fermions in real representations.

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

$$G = \begin{cases} SU(N) \text{ with } N \geq 3 \\ SO(4N+2) \\ E_6 \end{cases}$$

We should add to this list $G = U(1)$ which we discussed separately in the previous section.

The list of gauge groups which might suffer perturbative gauge anomalies is short. But it turns out that it is shorter still, since the anomaly coefficient $\operatorname{tr} T^a\{T^b, T^c\}$ vanishes for both $G = E_6$ and $G = SO(4N+2)$ with $N \geq 2$. (Note that the Lie algebra $so(6) \cong su(4)$ so this remains.) We learn that we need only care about these triangle anomalies when

$$G = SU(N) \text{ with } N \geq 3$$

Interestingly, these are the gauge groups which appear most prominently in the study of particle physics.

Let's now look at solutions to the anomaly cancellation condition (3.63). At first glance, this looks as if it is a tensor equation and if each representation $\mathbf{R}$ had a different tensor structure for d^{abc} it would be tricky to solve. Fortunately, that is not the case. One can show that

$$d^{abc}(\mathbf{R}) = A(\mathbf{R})\, d^{abc}(\mathbf{N})$$

where $\mathbf{N}$ is the fundamental representation of $SU(N)$. The coefficient $A(\mathbf{R})$ is sometimes called simply the *anomaly* of the representation. To see this, first note that we have

$$A(\mathbf{R}_1 \oplus \mathbf{R}_2) = A(\mathbf{R}_1) + A(\mathbf{R}_2) \tag{3.64}$$

But an arbitrary representation can be constructed by taking tensor products of the fundamental. The representation $\mathbf{R}_1 \otimes \mathbf{R}_2$ is generated by $\mathbf{1}_1 \otimes T_2^a + T_1^a \otimes \mathbf{1}_2$, so we have

$$A(\mathbf{R}_1 \otimes \mathbf{R}_2) = \dim(\mathbf{R}_1)\, A(\mathbf{R}_2) + \dim(\mathbf{R}_2)\, A(\mathbf{R}_1) \tag{3.65}$$

Finally, note that our calculation above tells us that $A(\bar{\mathbf{R}}) = -A(\mathbf{R})$.

The formulae (3.64) and (3.65) allow us to compute the anomaly coefficient for different representations providing that we know how to take tensor products. Consider, for example, representations of $G = SU(3)$. By definition $A(\mathbf{3}) = -A(\bar{\mathbf{3}}) = 1$. If we use the fact that $\mathbf{3} \otimes \mathbf{3} = \mathbf{6} \oplus \bar{\mathbf{3}}$ then we have

$$A(\mathbf{6}) = A(\mathbf{3} \otimes \mathbf{3}) - A(\bar{\mathbf{3}}) = 3A(\mathbf{3}) + 3A(\mathbf{3}) - A(\bar{\mathbf{3}}) = 3 + 3 - (-1) = 7$$

Similarly, $\mathbf{3} \otimes \bar{\mathbf{3}} = \mathbf{8} \oplus \mathbf{1}$, which gives

$$A(\mathbf{8}) = 3A(\mathbf{3}) + 3A(\bar{\mathbf{3}}) - A(\mathbf{1}) = 3 + (-3) - 0 = 0$$

as expected since the adjoint $\mathbf{8}$ is a real representation.

What is the Simplest Non-Abelian Chiral Gauge Theory?

A chiral gauge theory is one in which the left-handed and right-handed Weyl fermions transform in different representations of the gauge group. This prohibits a tree-level mass term for the fermions, since it is not possible to write down a fermion bilinear. Theories of this type comprise some of the most interesting quantum field theories, both for theoretical and phenomenological reasons. (We'll see a particularly interesting chiral gauge theory in Section 3.4.4.) Notably, there are obstacles to placing these theories on the lattice, which means that we have no numerical safety net when trying to understand their strong coupling dynamics.

We can use our results above to construct some simple non-Abelian chiral gauge theories. One can show that the anomaly coefficients for the symmetric $\square\!\square$ and anti-symmetric $\begin{smallmatrix}\square\\[-2pt]\square\end{smallmatrix}$ representations are:

$$A(\square\!\square) = N + 4 \quad \text{and} \quad A\left(\begin{smallmatrix}\square\\[-2pt]\square\end{smallmatrix}\right) = N - 4$$

From this, we learn that we can construct a number of chiral gauge theories by taking, for $N \geq 5$,

$$G = SU(N) \text{ with a } \begin{smallmatrix}\square\\[-2pt]\square\end{smallmatrix} \text{ and } N - 4\ \bar{\square} \text{ Weyl fermions}$$

where $\overline{\Box}$ is shorthand for the anti-fundamental. Alternatively, we could have, for $N \geq 3$,

$$G = SU(N) \text{ with a } \Box\!\Box \text{ and } N+4\ \overline{\Box}$$

or

$$G = SU(N) \text{ with a } \Box\!\Box,\text{ a } \begin{smallmatrix}\Box\\\Box\end{smallmatrix} \text{ and } 2N\ \overline{\Box}$$

The simplest of these theories is:

$$SU(5) \text{ with a } \bar{\mathbf{5}} \text{ and } \mathbf{10} \tag{3.66}$$

This is a prominent candidate for a grand unified theory, incorporating the Standard Model gauge group and one generation of matter fields. We'll return to these chiral gauge theories in Section 5.6.4 where we describe their likely dynamics.

Alternatively, we can build a chiral gauge theory by taking either E_6 or $SO(4N+2)$ with complex representations, where the anomaly coefficients all vanish. The simplest such example is $SO(10)$ with a single Weyl fermion in the **16** spinor representation. This too is a prominent candidate for a grand unified theory.

The chiral gauge theories described above are the simplest to write down. But it turns out that there is one chiral gauge theory which has fewer fields. This will be described in section 3.4.4.

Some Curious Anomaly-Free Representations

Most complex irreps of $SU(N)$ have an anomaly. But not all. For $N \geq 5$, there are some rather special complex representations that do not have an anomaly. Moreover, these are related to the non-anomalous $U(1)$ gauge theories that we met above. I don't know what these representations are useful for, if anything, but their mere existence is interesting.

First, we need to introduce some notation. A general representation of $SU(N)$ can be described by $(N-1)$ non-negative integers $\mathbf{m} = (m_1, \ldots, m_{N-1})$ which can be thought of as the labels attached to the nodes of a Dynkin diagram. (Alternatively, you can think of m_i as the number of columns in the Young tableaux with i boxes.) So, for example, $\mathbf{m} = (1, 0, \ldots, 0)$ is the fundamental representation while $\mathbf{m} = (0, \ldots, 0, 1)$ is the anti-fundamental representation.

Now we make two changes of variables. First we introduce the $(N-1)$-dimensional vector

$$\mathbf{q} = (m_1+1, m_2+1, \ldots, m_{N-1}+1) \ . \tag{3.67}$$

From these, we construct the N-dimensional *Okubo vector* with entries

$$\sigma_i = -\sum_{k=1}^{i-1} k q_k + \sum_{k=1}^{n-1} (n-k) q_k \qquad i = 1, \ldots, N. \tag{3.68}$$

This is designed so that the complex conjugate of a representation $\boldsymbol{\sigma}$ is $-\boldsymbol{\sigma}$. We have constructed N variables σ_i from the $(N-1)$ variables m_i so there must be a redundancy. This manifests itself in the simple relation

$$\sum_{i=1}^{N} \sigma_i = 0 \ . \tag{3.69}$$

Meanwhile, the requirement that the anomaly vanishes for this representation, $A(\mathbf{m}) = 0$, is given by

$$\sum_{i=1}^{N} \sigma_i^3 = 0 \ . \tag{3.70}$$

But we recognise (3.69) and (3.70) as the requirements for anomaly cancellation in a $U(1)$ theory (3.57) and (3.61). This means that we can import our results from Section 3.4.1 and, given any anomaly free collection of N fermions charged under a single $U(1)$ gauge group, we can always construct a non-anomalous irrep of $SU(N)$.

These anomaly-free complex irreps don't appear to have much use! For example, the simplest such irrep arises for $SU(5)$ and corresponds to the Abelian charges $\sigma_i = \{1, 5, -7-8, 9\}$. The corresponding Dynkin indices are $\mathbf{m} = (0, 7, 3, 3)$. This irrep has dimension over a million!

3.4.3 The $SU(2)$ Anomaly

The list of gauge groups that suffer a perturbative anomaly does not include $G = SU(2)$. This is because all representations are either real or pseudoreal. For example, the fundamental **2** representation, with the generators given by the Pauli matrices σ^a, is pseudoreal. In agreement with our general result above, it is simple to check that

$$d^{abc} = \mathrm{tr}\, \sigma^a \{\sigma^b, \sigma^c\} = 0$$

This would naively suggest that we don't have to worry about anomalies in such theories. But this is premature. There is one further, rather subtle anomaly that we need to take into account. This was first discovered by Witten and, unlike our previous anomalies, cannot be seen in perturbation theory. It is a non-perturbative anomaly.

Here is the punchline. An $SU(2)$ gauge theory with a single Weyl fermion in the fundamental representation is mathematically inconsistent. Furthermore, an $SU(2)$ gauge theory with any odd number of Weyl fermions is inconsistent. To make sense of the theory, Weyl fermions must come in pairs. In other words, they must be Dirac fermions.

To see why, let's start with a theory which makes sense. We will take a Dirac fermion Ψ in the fundamental representation of $SU(2)$. The partition function in Euclidean space is, schematically,

$$\begin{aligned} Z &= \int \mathcal{D}\Psi \mathcal{D}\bar{\Psi}\mathcal{D}A \ \exp\left(-\int d^4x \ \frac{1}{2g^2}\,\mathrm{tr}\, F^{\mu\nu}F_{\mu\nu} + i\bar{\Psi}\, \not{D}\Psi\right) \\ &= \int \mathcal{D}A \ \det(i\not{D}) \ \exp\left(-\int d^4x \ \frac{1}{2g^2}\,\mathrm{tr}\, F^{\mu\nu}F_{\mu\nu}\right) \end{aligned}$$

This determinant is an infinite product over eigenvalues of $i\not{D}$ and, as such, we have to regulate this product in a gauge invariant way. We met one such regularisation in 3.2.2 where we discussed the measure in the path integral. Another simple possibility for a Dirac fermion is Pauli-Villars regularisation.

Let's now repeat this for a Weyl fermion. For concreteness, let's take a left-handed fermion ψ. Following (3.10), we have the path integral,

$$Z = \int \mathcal{D}\psi \mathcal{D}\bar{\psi}\mathcal{D}A \ \exp\left(-\int d^4x \ \frac{1}{2g^2}\,\mathrm{tr}\, F^{\mu\nu}F_{\mu\nu} + i\bar{\psi}\sigma^\mu\mathcal{D}_\mu\psi\right)$$

Integrating out the fermions, it looks like we're left with the object $\det(i\sigma^\mu\mathcal{D}_\mu)$. But this is rather subtle, because $i\sigma^\mu\mathcal{D}_\mu$ doesn't map a vector space back to itself; instead it maps left-handed fermions onto right-handed fermions. To proceed, it's best to think of the Weyl fermion as a projection $\psi = \frac{1}{2}(1+\gamma^5)\Psi$. We then have

$$Z = \int \mathcal{D}A \ \det\left(i\not{D}\,\frac{1+\gamma^5}{2}\right) \ \exp\left(-\int d^4x \ \frac{1}{2g^2}\,\mathrm{tr}\, F^{\mu\nu}F_{\mu\nu}\right) \qquad (3.71)$$

As we discussed in Section 3.3.1, $i\not{D}$ is a Hermitian operator and therefore has real eigenvalues. The existence of the γ^5 matrix ensures that these eigenvalues come in $\pm$ pairs,

$$i\not{D}\phi_n = \lambda_n\phi_n \quad \Rightarrow \quad i\not{D}(\gamma^5\phi_n) = -\lambda_n(\gamma^5\phi_n)$$

Let us assume that we have a gauge potential with no zero eigenvalues. Then the spectrum of eigenvalues of $i\not{D}$ looks something like that shown on the left-hand axis of the figure below. Formally, $\det(i\not{D}) = \prod_n \lambda_n$. To define the determinant $\det(i\not{D}(1+\gamma^5)/2)$, we should just take the product over half of these eigenvalues. In other words,

$$\det\left(i\not{D}\frac{1+\gamma^5}{2}\right) = \det{}^{1/2}(i\not{D})$$

This formula is intuitive because a Dirac fermion consists of two Weyl fermions. Our job is to make sense of it. The difficulty is that there is a $\pm$ ambiguity when we take the square-root $\det^{1/2}(i\not{D})$. This, as we will see, will be our downfall.

Let's try to define a consistent sign for our determinant $\det^{1/2}(i\not{D})$. To do so, we need to pick half of these eigenvalues in a consistent way. Here is how we will go about it. We start with some specific gauge configuration $A^\star_\mu$. For this particular choice, we define $\det^{1/2}(i\not{D})$ to be the product of the positive eigenvalues only, throwing away the negative eigenvalues. As we vary the A_μ away from $A^\star_\mu$, we follow this set of preferred eigenvalues and continue to take their product. It may be that as we vary A_μ, some of these chosen eigenvalues cross zero and become negative. Whenever this happens, $\det^{1/2}(i\not{D})$ changes sign. If we're lucky, this method has succeeded in assigning a particular sign to $\det^{1/2}(i\not{D})$ for each configuration A_μ.

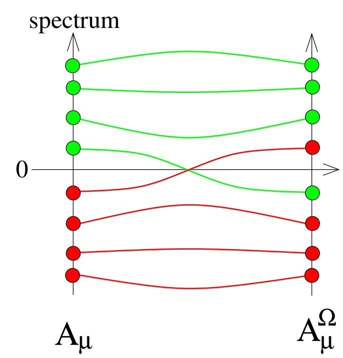


Figure 32:

Now we come to the important question: is our choice of sign gauge invariant? In particular, suppose that we start with a gauge connection A_μ and smoothly vary it until we come back to a new gauge connection which is gauge equivalent to the original,

$$A_\mu \mapsto A^\Omega_\mu = \Omega(x)A_\mu\Omega^{-1}(x) + i\Omega(x)\partial_\mu\Omega^{-1}(x)$$

For our theory to be consistent, we need that the sign of $\det^{1/2}(i\not{D})$ is the same for these two gauge equivalent configurations. If this fails to be true, then the integral over A_μ in the partition function (3.71) will give us $Z=0$ and our theory is empty.

How could this fail to work? We know that the total spectrum of $\not{D}$ is the same for gauge equivalent configurations. The concern is that as we vary smoothly from A_μ to A^Ω_μ, an odd number of eigenvalues may cross the origin, as shown in the figure. This would result in a change to the sign of the determinant.

To proceed, we need to classify the kinds of gauge transformations $\Omega(x)$ that we can have. We will consider gauge transformations such that $\Omega(x) \to \mathbf{1}$ as $x \to \infty$. This effectively compactifies $\mathbf{R}^4$ to $\mathbf{S}^4$ and all such gauge transformations provide a map $\Omega : \mathbf{S}^4 \mapsto SU(2)$. These maps are characterised by the homotopy group

$$\Pi_4(SU(2)) = \mathbf{Z}_2 \tag{3.72}$$

Note that in our discussion of instantons in Section 2.3 we used $\Pi_3(SU(2)) = \mathbf{Z}$. That's fairly intuitive to understand because $SU(2) \cong \mathbf{S}^3$, so the third homotopy group counts winding from a 3-sphere to a 3-sphere. The fourth homotopy group is less intuitive[8]: it tells us that there are topologically non-trivial maps from $\mathbf{S}^4$ to $\mathbf{S}^3$.

The homotopy group (3.72) means that all $SU(2)$ gauge transformations fall into two classes: trivial or non-trivial. We will see that under a non-trivial gauge transformation

$$\det{}^{1/2}(i\not{D}) \;\mapsto\; -\det{}^{1/2}(i\not{D}) \tag{3.73}$$

This is the non-perturbative $SU(2)$ anomaly that renders the theory inconsistent. (Rather annoyingly, because the anomaly is related to the global structure of the gauge group, it is sometimes referred to as a "global anomaly", even though it is an anomaly in a gauge symmetry instead of a global symmetry.)

Follow the Eigenvalue

It remains to show that $\det^{1/2}(i\not{D})$ indeed flips sign under a non-trivial gauge transformation as in (3.73). To do so, we consider a gauge connection $\mathcal{A}$ on the 5d space $\mathcal{M}_5 = \mathbf{R} \times \mathbf{S}^4$. We parameterise the $\mathbf{R}$ factor by τ and work in a gauge with $\mathcal{A}_\tau = 0$. Meanwhile, for $\mu = 1, 2, 3, 4$ labelling a direction on $\mathbf{S}^4$ we choose a gauge configuration such that

$$\mathcal{A}_\mu(x, \tau) \to A_\mu(x) \quad \text{as } \tau \to -\infty \tag{3.74}$$

and

$$\mathcal{A}_\mu(x, \tau) \to A^\Omega_\mu(x) \quad \text{as } \tau \to +\infty \tag{3.75}$$

Our 5d gauge field $\mathcal{A}(x, t)$ smoothly interpolates between a 4d gauge configuration at $\tau \to -\infty$ and a gauge equivalent configuration at $\tau \to +\infty$, related by a non-trivial gauge transformation.

[8]Higher homotopy groups only get more counter-intuitive! See, for example, the Wikipedia article on the homotopy groups of spheres.

We now consider the five-dimensional Dirac operator

$$\not{\mathcal{D}}_5\Psi = \gamma^\tau\frac{\partial\Psi}{d\tau} + \not{D}\Psi$$

The operator $\not{\mathcal{D}}_5$ is both real and anti-symmetric. (Both the spinor representation of $SO(5)$ and the fundamental representation of the gauge group $SU(2)$ are pseudo-real, but their tensor product is real.) There are two possibilities for the eigenvalues of such an operator: either they are zero, or they are purely imaginary and come in conjugate pairs. This means that as we vary the gauge connection $\mathcal{A}_\mu$, and the eigenvalues smoothly change, the number of zero eigenvalues can only change in pairs. The number of zero eigenvalues, mod 2, is therefore a topological invariant.

This $\mathbf{Z}_2$ topological invariant can be computed by a variant of the Atiyah-Singer index theorem. For any gauge configuration with boundary conditions (3.74) and (3.75), the index theorem tells us that the number of zero modes is necessarily odd.

Let's now see why this $\mathbf{Z}_2$ index of the five-dimensional Dirac operator $\not{\mathcal{D}}_5$ tells us that the determinant necessarily flips sign as in (3.73). Any zero mode of the $\not{\mathcal{D}}$ obeys

$$\frac{\partial\Psi}{\partial\tau} = -\gamma^\tau\not{D}\Psi \tag{3.76}$$

We will assume that the gauge configuration $\mathcal{A}_\mu(x,\tau)$ varies slowly enough in τ that we can use the adiabatic approximation for the eigenfunctions. This means that the eigenfunction $\Psi(x,\tau)$ can be written as

$$\Psi(x,\tau) = f(\tau)\,\phi(x;\tau)$$

where, for each fixed τ, $\phi(x;\tau)$ is an eigenfunction of the 4d Dirac operator

$$\gamma^\tau\not{D}\phi(x;\tau) = \lambda_n(\tau)\phi(x;\tau)$$

In this adiabatic approximation, the zero mode equation (3.76) becomes

$$\frac{df}{d\tau} = -\lambda(\tau)f(\tau) \quad\Rightarrow\quad f(\tau) = f_0\,\exp\left(-\int^\tau d\tau'\,\lambda(\tau')\right)$$

But $f(\tau)$ must be normalisable. This requires that $\lambda(\tau) > 0$ as $\tau\to+\infty$, but $\lambda(\tau) < 0$ as $\tau\to-\infty$.

We learn that for every normalisable zero mode of $\not{D}_5$, there must be an eigenvalue of the four-dimensional Dirac operator $\not{D}$ which crosses from positive to negative as we vary τ. Since the index theorem tells us that there are an odd number of zero modes, there must be an odd number of eigenvalues that cross the origin. And this, in turn, means that the determinant flips sign under a non-trivial gauge transformation as in (3.73). This is why $SU(2)$ gauge theory with a single Weyl fermion — and, indeed, with any odd number of Weyl fermions — is inconsistent.

Other Gauge Groups

Although advertised here as an anomaly of $SU(2)$ gauge groups, the same argument holds for any gauge group with non-trivial Π_4. This is not relevant for other unitary or orthogonal groups: $\Pi_4(SU(N)) = 0$ for $N \geq 3$ and $\Pi_4(SO(N)) = 0$ for all $N \geq 5$. However, $SU(2)$ is also the start of the symplectic series: $SU(2) = Sp(1)$. More generally,

$$\Pi_4(Sp(N)) = \mathbf{Z}_2 \quad \text{for all } N$$

The same arguments as above tell us that $Sp(N)$ with a single Weyl fermion in the fundamental representation has a non-perturbative anomaly and is therefore mathematically inconsistent.

3.4.4 Anomaly Cancellation in the Standard Model

We saw earlier how to build chiral, non-Abelian gauge theories with gauge group $SU(N)$. The simplest of these is the $SU(5)$ grand unified candidate (3.66). However, it turns out that there is a chiral gauge theory which is simpler than this, in the sense that it has fewer fields. This theory has gauge group

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

We denote the chiral matter as $(\mathbf{R}_1, \mathbf{R}_2)_Y$ where $\mathbf{R}_1$ and $\mathbf{R}_2$ are the representations under $SU(2)$ and $SU(3)$ respectively, and the subscript Y denotes the $U(1)$ gauge charge. The left- and right-handed fermions transform as

$$\begin{aligned} &\text{Left-Handed:} \quad l_L : \ (\mathbf{2},\mathbf{1})_{-3} \quad , \quad q_L : \ (\mathbf{2},\mathbf{3})_{+1} \\ &\text{Right-Handed:} \quad e_R : \ (\mathbf{1},\mathbf{1})_{-6} \quad , \quad u_R : \ (\mathbf{1},\mathbf{3})_{+4} \quad , \quad d_R : \ (\mathbf{1},\mathbf{3})_{-2} \end{aligned} \tag{3.77}$$

This is perhaps the most famous of all quantum field theories, for it describes the world we live in. It is, of course, the Standard Model with a single generation of fermions. (It is missing the Higgs field and associated Yukawa couplings which do not affect the anomalies. Note also that we have chosen a normalisation so that the $U(1)$ hypercharges are integers; this differs by a factor of 6 from the conventional normalisation.) Here l_L are the left-handed leptons (electron and neutrino) and e_R is the right-handed electron. Meanwhile, q_L is the left-handed doublet of up and down quarks while u_R and d_R are the right-handed up and down quarks. We may add to this a right-handed neutrino ν_R which is a singlet under all factors of G.

Let's see how anomaly cancellation plays out in the Standard Model. First the non-Abelian anomalies. The $[SU(3)]^3$ diagram is anomaly free because there are two left-handed and two right-handed quarks. Similarly, there is no problem with the non-perturbative $SU(2)$ anomaly because there are 4 fermions transforming in the $\mathbf{2}$.

This leaves us only with anomalies that involve the Abelian factor. Here things are more interesting. The $U(1)^3$ anomaly requires that the sum of charges $\sum_{\text{left}} Y^3 - \sum_{\text{right}} Y^3 = 0$. (In all of these calculations, we must remember to multiply by the dimension of the representation of the non-Abelian factors.) We have

$$[U(1)]^3: \quad \Big[2 \times (-3)^3 + 6 \times (+1)^3\Big] - \Big[(-6)^3 + 3 \times (4)^3 + 3 \times (-2)^3\Big] = 0$$

where we have arranged left- and right-handed fermions into separate square brackets. We see already that the cancellation happens in a non-trivial way. Similarly, the mixed $U(1)$-gravitational anomaly tells us that the sum of the charges $\sum_{\text{left}} Y - \sum_{\text{right}} Y = 0$ must vanish

$$U(1) \times \text{gravity}^2: \quad \Big[2 \times (-3) + 6\Big] - \Big[-6 + 3 \times 4 + 3 \times (-2)\Big] = 0$$

Finally, we have the mixed anomalies between two factors of the gauge group. The non-Abelian factors must appear in pairs, otherwise the contribution vanishes after taking the trace over group indices. But we're left with two further anomalies which must cancel:

$$\begin{aligned} &[SU(2)]^2 \times U(1): \quad -3 + 3 \times (+1) = 0 \\ &[SU(3)]^2 \times U(1): \quad 2 \times (+1) - \Big[4 - 2\Big] = 0 \end{aligned}$$

We see that all gauge anomalies vanish. Happily, our Universe is mathematically consistent!

The Standard Model is arguably the simplest chiral gauge theory that one can write down (at least with a suitable definition of the word "simple"). It is rather striking that this theory is the one that describes our Universe at energy scales $\lesssim 1$ TeV or so.

Could it have been otherwise?

There are alternative games that we can play here. For example, we could take the matter fields of the Standard Model, but assign then arbitrary hypercharges.

$$l_L : (\mathbf{2},\mathbf{1})_l\,, \quad q_L : (\mathbf{2},\mathbf{3})_q\,, \quad e_R : (\mathbf{1},\mathbf{1})_x\,, \quad u_R : (\mathbf{1},\mathbf{3})_u\,, \quad d_R : (\mathbf{1},\mathbf{3})_d$$

We then ask what values of the hypercharges $\{l,q,x,u,d\}$ give rise to a consistent theory? We have constraints from the non-Abelian anomalies:

$$\begin{aligned} [SU(2)]^2 \times U(1):& \quad 3q + l = 0 \\ [SU(3)]^2 \times U(1):& \quad 2q - u - d = 0 \end{aligned} \tag{3.78}$$

and the purely Abelian anomaly

$$[U(1)]^3: \quad 6q^3 + 2l^3 - 3u^3 - 3d^3 - x^3 = 0 \tag{3.79}$$

On their own, these are not particularly restrictive. However, if we also add the mixed gauge-gravitational anomaly

$$U(1) \times \text{gravity}^2: \quad 6q + 2l - 3(u+d) - x = 0 \tag{3.80}$$

then it is straightforward to show that there are only two possible solutions. The first of these is a trivial, non-chiral assignment of the hypercharges,

$$q = l = x = 0 \ \text{ and } \ u = -d \tag{3.81}$$

The second is, up to an overall rescaling, the charge assignment (3.77) seen in Nature,

$$x = 2l = -3(u+d) = -6q \ \text{ and } \ u - d = \pm 6q$$

This is interesting. Notice that we didn't insist on quantisation of the hypercharges above, yet the restrictions imposed by anomalies ensure that the resulting hypercharges are, nonetheless, quantised in the sense that the ratios of all charges are rational.

We could also turn this argument around. Suppose that we instead insist from the outset that the hypercharges $\{l,q,x,u,d\}$ take integer values. This is the statement that the $U(1)$ gauge group of the Standard Model is actually $U(1)$, rather than $\mathbf{R}$. We can use the first equation in (3.78) to eliminate $l = -3q$. The first equation in (3.78) tells us that the sum $u+d$ is even which means that the difference is also even: we write $u - d = 2y$. The cubic $U(1)^3$ anomaly equation (3.79) then becomes

$$x^3 + 18qy^2 + 54q^3 = 0 \tag{3.82}$$

We now want to find integer solutions to this equation. There is the trivial solution with $x = q = 0$; this gives us (3.81). Any further solution necessarily has $q \neq 0$. Because (3.82) is a homogeneous polynomial we may rescale to set $q = 1$ and look for rational solutions to the curve

$$x^3 + 18y^2 + 54 = 0 \qquad x, y \in \mathbb{Q} \tag{3.83}$$

This is a rather special elliptic curve. To see this, we introduce two new coordinates $v, w \in \mathbb{Q}$, defined by

$$x = -\frac{6}{v+w} \quad , \quad y = \frac{3(v-w)}{v+w}$$

This reveals the elliptic curve (3.83) to be the Fermat curve

$$v^3 + w^3 = 1$$

Any non-trivial rational solution to this equation would imply a non-trivial integer solution to the equation $v^3 + w^3 = z^3$. Famously, there are none. The trivial solutions are $v = 1, w = 0$ and $v = 0, w = 1$. These reproduce the hypercharge assignments (3.77) of the Standard Model.

Notice that at no point in the above argument did we make use of the mixed gauge-gravitational anomaly. We learn that if we insist on quantised hypercharge then consistent solutions of the gauge anomalies are sufficient to guarantee that the mixed gauge-gravitational anomaly is also satisfied. This is a rather unusual property for a quantum field theory.

It is well known that the Standard Model gauge group and matter content fits nicely into a grand unified framework — either $SU(5)$ with a $\bar{\mathbf{5}}$ and $\mathbf{10}$; or $SO(10)$ with a $\mathbf{16}$ — and it is sometimes said that this is evidence for grand unification. This, however, is somewhat misleading: the matter content of the Standard Model is determined by mathematical consistency alone. To find evidence for grand unification, we must look at more dynamical issues, such as the running of the three coupling constants.

Global Symmetries in the Standard Model

The Standard Model consists of more than just the matter content described above. There is also the Higgs field, a scalar transforming as $(\mathbf{2}, \mathbf{1})_3$, and the associated Yukawa couplings. After the dust has settled, the classical Lagrangian enjoys two global symmetries: baryon number B and lepton number L. The charges are:

	l_L	q_L	e_R	u_R	d_R	ν_R
B	0	$\frac{1}{3}$	0	$\frac{1}{3}$	$\frac{1}{3}$	0
L	1	0	1	0	0	1

Both B and L are anomalous. There is a contribution from both the $SU(2)$ gauge fields, and also from the $U(1)$ hypercharge. For the latter, the anomaly is given by

$$\sum_{\text{left}} BY^2 - \sum_{\text{right}} BY^2 = \frac{1}{3}\left(6 - 3\times 4^2 - 3\times(-2)^2\right) = -18$$

and

$$\sum_{\text{left}} LY^2 - \sum_{\text{right}} LY^2 = 2\times(-3)^2 - 6^2 = -18$$

Note, however, the anomalies for B and L are the same. This is true both for the mixed anomaly with $U(1)_Y$ – as shown above – and also for the mixed anomaly with $SU(2)$. This means that the combination $B - L$ is non-anomalous. It is the one global symmetry of the Standard Model.

We still have to check if there is a gravitational contribution to the $B - L$ anomaly. This vanishes only if there is a right-handed neutrino.

A More General Chiral Gauge Theory

The Standard Model is the start of a 2-parameter family of chiral gauge theories, with gauge group

$$G = U(1)\times Sp(r)\times SU(N)$$

with N odd. The matter content is a generalisation of (3.77), except there are now r copies of each of the right-handed fermions, including the right-handed neutrino. The chiral fermions transform in the representations

$$\begin{array}{rl} \text{Left-Handed:} & l_L:\ (\mathbf{2r},\mathbf{1})_{-N}\,, \quad q_L:\ (\mathbf{2r},\mathbf{N})_{+1} \\ \text{Right-Handed:} & (e_\alpha)_R:\ (\mathbf{1},\mathbf{1})_{-2\alpha N}\,, \quad (\nu_\alpha)_R:\ (\mathbf{1},\mathbf{1})_{(2\alpha-2)N} \\ & (u_\alpha)_R:\ (\mathbf{1},\mathbf{N})_{1+(2\alpha-1)N}\,, \quad (d_\alpha)_R:\ (\mathbf{1},\mathbf{N})_{1-(2\alpha-1)N} \end{array}$$

For $r = 1$ and $N = 3$, the matter content coincides with that of the Standard Model. One can check that all mixed gauge and gravitational anomalies vanish for arbitrary integer r and odd integer N.

3.5 't Hooft Anomalies

So far we have classified our anomalies into two different types: anomalies in global symmetries (which are interesting) and anomalies in gauge symmetries (which are fatal).

However, a closer look at the triangle diagrams suggests a better classification of these anomalies. Global anomalies (like the chiral anomaly) have a single global current and two gauge currents on the vertices of the triangle. They are better thought of as mixed global-gauge anomalies. What we have called gauge anomalies have gauge currents on all three vertices. But here too we have seen examples with mixed anomalies between different gauge symmetries.

This begs the question: do we gain anything by thinking about triangle diagrams with global symmetries on all three vertices? If the sum over triangle diagrams does not vanish, then the global symmetry is said to have a *'t Hooft anomaly*.

A global symmetry with a 't Hooft anomaly remains a symmetry in the quantum theory. The charges that you think are naively conserved are, indeed, conserved. You only run into trouble if you couple the symmetry to a background gauge field, in which case the charge is no longer conserved. You run into real trouble if you try to couple the symmetry to a dynamical gauge field because then the 't Hooft anomaly becomes a gauge anomaly and the theory ceases to make sense. In other words, the 't Hooft anomaly is an obstruction to gauging a global symmetry.

We've already met examples of global symmetries with a 't Hooft anomaly above. For example, a free Dirac fermion has two global symmetries $U(1)_V$ and $U(1)_A$, and there is a mixed 't Hooft anomaly between the two.

So far, it doesn't sound like a 't Hooft anomaly buys us very much. However, a very simple and elegant argument, due to 't Hooft, means that these symmetries can be a rather powerful tool to help us understand the dynamics of strongly coupled gauge theories. Suppose that we have some theory which, at high-energies, has a continuous global symmetry group G_F (here F stands for "flavour"). We are interested in the low-energy dynamics and, in particular, the spectrum of massless particles. For strongly coupled gauge theories, this is typically a very hard problem. As we've seen in Section 2, the physical spectrum need not look anything like the fields that appear in the Lagrangian. In particular, the quarks that appear at high-energies are often confined into bound states at low-energies. In this way, seemingly massless fields may get a mass through quantum effects. Conversely, it may be that some of these confined bound states themselves turn out to be massless. In short, the spectrum rearranges itself, often in a dramatic fashion, and we would like to figure out what's left at very low energies.

The 't Hooft anomaly doesn't solve this question completely, but it does provide a little bit of an insight. Here is the key idea: we gauge the global symmetry G_F. This means that we introduce new gauge fields coupled to the G_F-currents. Now, as we explained above, the 't Hooft anomaly means that such a gauging is not possible since the theory will no longer be consistent. To proceed, we must therefore also introduce some new massless Weyl fermions which do not interact directly with the original fields, but are coupled only to the G_F gauge fields. Their role is to cancel the G_F anomaly, rendering the theory consistent. We will call these new fields *spectator fermions.*

What is the dynamics of this new theory? We choose the new gauge coupling to be very small so that these gauge fields do not affect the massless spectrum of the original theory. In particular, if the new G_F gauge field itself becomes strongly coupled at some scale $\Lambda_{\rm new}$, we will pick the gauge coupling so that $\Lambda_{\rm new}$ is much smaller than any other scale in the game. The upshot is that at low energies — either in the strict infra-red, or at energies $E \gtrsim \Lambda_{\rm new}$ — there are two choices:

- The symmetry group G_F is spontaneously broken by the original gauge dynamics. In this case, the original theory, in which G_F is a global symmetry, must have massless Goldstone modes.

- The symmetry group G_F is not spontaneously broken. In this case, we are left with a G_F gauge theory which must be free from anomalies. By construction, the spectator fermions contribute towards the G_F anomaly which means that the low energy spectrum of the original theory must contain extra massless fermions which conspire to cancel the anomaly. This gives us a handle on the spectrum of massless fermions and is known as 't Hooft anomaly matching.

The essence of anomaly matching is that one can follow the anomaly from the ultra-violet to the infra-red. If the 't Hooft anomaly in the ultra-violet is A_{UV} then the spectator fermions must provide an anomaly $A_{\rm spectator}$ such that

$$A_{UV} + A_{\rm spectator} = 0$$

But if the symmetry survives in the infra-red, the anomaly persists. Now the massless fermions may look very different from those in the UV – for example, if the theory confines then they will typically be bound states — but they must contribute A_{IR} to the anomaly with

$$A_{IR} + A_{\text{spectator}} = 0 \quad \Rightarrow \quad A_{UV} = A_{IR}$$

The anomaly is special because it is an exact result, yet can be seen at one-loop in perturbation theory.

Anomaly matching has many uses. The standard application is to a $SU(N)$ gauge theory coupled to N_f massless Dirac fermions, each in the fundamental representation. This is a vector-like theory, so doesn't suffer any gauge anomaly. The global symmetry of the classical Lagrangian is

$$G_F = U(N_f)_L \times U(N_f)_R$$

where each factor acts on the left-handed or right-handed Weyl fermions. However, we've seen in Section 3.1 that the chiral anomaly means that the axial $U(1)_A$ does not survive in the quantum theory. The non-anomalous global symmetry of the theory is

$$G_F = U(1)_V \times SU(N_f)_L \times SU(N_f)_R$$

We can see immediately that G_F is likely to enjoy a 't Hooft anomaly since the $SU(N_f)$ factors act independently on left- and right-handed fermions. The question is: what does this tell us about the low-energy dynamics of our theory? The answer to this question will be the topic of Section 5, so we will delay giving the full analysis until Section 5.6 where we will show that often there is no confined bound state spectrum which can reproduce the 't Hooft anomaly in G_F. This means that G_F must be spontaneously broken, and there are massless Goldstone bosons in the theory.

An Aside: Symmetry Protected Topological Phases

In condensed matter physics, there is the notion of a *symmetry protected topological* (SPT) phase. We won't describe this in detail, but provide a few words to explain how this is related to 't Hooft anomalies.

An SPT phase is a gapped phase which, if we disregard the global symmetry, can be continuously connected to the trivial phase. However, if we insist that we preserve the global symmetry structure then it is not possible to deform an SPT phase into a trivial theory without passing through a quantum phase transition.

SPT phases can be rephrased in the language of 't Hooft anomalies. An SPT phase in spatial dimension d has a global symmetry G such that, when placed on a manifold with boundary, the $(d-1)$-dimensional theory on the boundary has a 't Hooft anomaly for G.

3.6 Anomalies in Discrete Symmetries

In this section, we turn to a slightly different topic: anomalies in discrete symmetries. Unlike our previous examples, these will have nothing to do with chiral fermions, or ultra-violet divergences in quantum field theory. Instead, our main example is an anomaly in pure Yang-Mills theory.

I should mention up front that this material is somewhat more specialised than the rest of this chapter. We will need to invoke a whole bunch of new machinery which, while fun and interesting in its own right, will not be needed for the rest of these lectures. And, at the end of the day, we will only apply this machinery to learn something new about $SU(N)$ Yang-Mills at $\theta = \pi$.

For those who are nervous that the effort is worth it, here is the gist of the story. Recall from Section 2.6 that there are (at least) two different versions of $SU(N)$ Yang-Mills theory that differ in the global structure of the gauge group. These are $G = SU(N)$ and $G = SU(N)/\mathbf{Z}_N$. Moreover, as we explained previously, the theta angles take different ranges in these two cases:

$$\begin{aligned} G = SU(N) \quad &\Rightarrow \quad \theta \in [0, 2\pi) \\ G = SU(N)/\mathbf{Z}_N \quad &\Rightarrow \quad \theta \in [0, 2\pi N) \end{aligned}$$

The discrete symmetry that we're going to focus on is time reversal. As explained in Section 1.2.5, under time reversal $\theta \to -\theta$. This means that the theory with $\theta = 0$ is invariant under time reversal. But so too is the theory when θ takes half its range, i.e. the time-reversal invariant values are

$$\begin{aligned} \theta = \pi \quad &\text{when} \quad G = SU(N) \\ \theta = \pi N \quad &\text{when} \quad G = SU(N)/\mathbf{Z}_N \end{aligned}$$

Clearly these differ. This means that if we start with $G = SU(N)$ and $\theta = \pi$ then we have time reversal invariance. If we subsequently "divide the gauge group by $\mathbf{Z}_N$" (whatever that means) keeping θ unchanged, we lose time reversal invariance. This smells very much like a mixed 't Hooft anomaly: we do something to one symmetry and lose the other. Roughly speaking, we want to say that there is a mixed 't Hooft anomaly between time reversal and the $\mathbf{Z}_N$ centre symmetry of the gauge group.

It turns out that the language above is not quite correct. There is a mixed 't Hooft anomaly, but it is between rather different symmetries, known as *generalised symmetries*. We will describe these in Section 3.6.2 below. But first it will be useful to highlight how a very similar 't Hooft anomaly arises in a much simpler example: bosonic quantum mechanics.

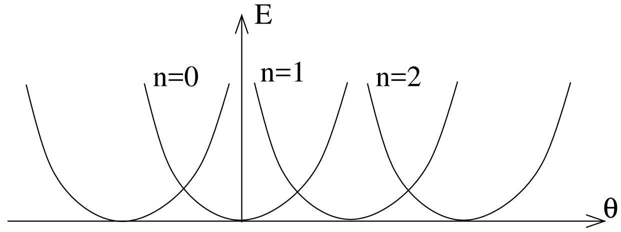


Figure 33: The energy spectrum for a particle moving around a solenoid.

3.6.1 An Anomaly in Quantum Mechanics

Many of the key features of discrete anomalies appear already in the quantum mechanics of a particle moving on a ring, around a flux tube. This is an example that we first met in the lectures on Applications of Quantum Mechanics when introducing the Aharonov-Bohm effect. We also briefly introduced this system in Section 2.2.3 of these lectures when discussing the theta angle.

We start with the Lagrangian

$$L = \frac{m}{2}\dot{x}^2 + \frac{\theta}{2\pi}\dot{x} \tag{3.84}$$

where we take the coordinate x to be periodic $x \in [0, 2\pi)$. This describes a particle of mass m moving around a solenoid with flux θ. (We'll also see this same quantum mechanical system arising later in Section 7.1 when we consider electromagnetism in $d = 1+1$ dimensions compactified on a spatial circle.)

The theta term is a total derivative. This ensures that it does not affect the equations of motion and so plays no role in the classical system. However, famously, it does change the quantum theory. To see this, we introduce the momentum

$$p = \frac{\partial L}{\partial \dot{x}} = m\dot{x} + \frac{\theta}{2\pi}$$

in terms of which, the Hamiltonian reads

$$H_\theta = \frac{1}{2m}\left(p - \frac{\theta}{2\pi}\right)^2 = \frac{1}{2m}\left(-i\frac{\partial}{\partial x} - \frac{\theta}{2\pi}\right)^2$$

where, in the second equality, we've used the canonical commutation relations $[x, p] = i$.

It is simple to solve for the spectrum of this Hamiltonian. We will ask that the wavefunctions are single-valued in x. In this case, they are given by

$$\psi_n(x) = \frac{1}{\sqrt{2\pi}}\, e^{inx}$$

where the requirement that ψ is single valued around the circle means that we must take $n \in \mathbf{Z}$. Plugging this into the time independent Schrödinger equation $H\psi = E\psi$, we find the spectrum

$$E_n(\theta) = \frac{1}{2m}\left(n - \frac{\theta}{2\pi}\right)^2 \qquad n \in \mathbf{Z}$$

The spectrum is shown in the figure as a function of θ. The key point is that the spectrum remains invariant under $\theta \to \theta + 2\pi$. However, it does so by shifting all the states $|n\rangle \to |n+1\rangle$. This is an example of *spectral flow.*

The fact that our system is periodic in θ will be important. Because of this, here are two further explanations. First, the path integral. Consider the Euclidean path integral with temporal $\mathbf{S}^1$ parameterised by $\tau \in [0, \beta)$. Then the field configurations include instantons, labelled by the winding number of the map $x : \mathbf{S}^1 \to \mathbf{S}^1$,

$$\int_{\mathbf{S}^1} d\tau\, \partial_\tau x = 2\pi k \qquad k \in \mathbf{Z}$$

Because the θ-term has a single time derivative, it comes with a factor of i in the Euclidean path integral, which is weighted by $e^{i\theta k}$ with $k \in \mathbf{Z}$. We see that the partition function is invariant under $\theta \to \theta + 2\pi$.

Next, Hamiltonian quantisation. Here, the fact that H_θ and $H_{\theta+2\pi}$ are equivalent quantum systems can be stated formally by the conjugation

$$e^{ix} H_\theta e^{-ix} = H_{\theta+2\pi}$$

Note that the operator e^{ix} is particularly natural. Indeed, the classical periodicity of x really means that x is not a good quantum operator; instead, we should only work with e^{ix}.

Symmetries

It will prove useful to describe the symmetries of the model. First, for all values of θ, there is an $SO(2) \cong U(1)$ symmetry which, classically, acts as translations: $x \to x + \alpha$.

In the quantum theory, we implement this by the operator T_α, with $\alpha \in [0, 2\pi)$. It acts on operators as

$$T_\alpha \, e^{ix} \, T_{-\alpha} = e^{i\alpha} e^{ix}$$

and on states as

$$T_\alpha |n\rangle = e^{i\alpha n} |n\rangle$$

For the two special values $\theta = 0$ and $\theta = \pi$, the system also enjoys a parity symmetry which acts classically as $P : x \to -x$. In the quantum theory, this acts on the operator as

$$P \, e^{ix} \, P = e^{-ix} \quad \text{with} \quad P^2 = 1$$

One could also view this as charge conjugation since it flips the charge of the particle moving around the solenoid; in addition, the theory has an anti-unitary time-reversal invariance at $\theta = 0$ and π but this does not seem to buy us anything new.

The action of parity on the states depends on whether $\theta = 0$ or $\theta = \pi$. Let's look at each in turn.

$\underline{\theta = 0}$: Here we have $P : |n\rangle \to |-n\rangle$. There is a unique ground state, $|0\rangle$, so parity is unbroken. However, all higher states come in pairs $|\pm n\rangle$, related by parity. We can now look at the interplay of parity and translations. It is simple to see that

$$P \, T_\alpha \, P = T_{-\alpha}$$

Mathematically, the $SO(2)$ symmetry and $\mathbf{Z}_2$ combine into $O(2) \cong \mathbf{Z}_2 \rtimes SO(2)$ where the semi-direct product $\rtimes$ is there because, as we see above, P and T_α do not commute.

$\underline{\theta = \pi}$: Now there are two ground states: $|0\rangle$ and $|1\rangle$. They have different charges under translations, with

$$T_\alpha |0\rangle = |0\rangle \quad \text{and} \quad T_\alpha |1\rangle = e^{i\alpha} |1\rangle$$

Clearly the action of parity can no longer be the same as when $\theta = 0$, because the states $|n\rangle$ and $|-n\rangle$ are not degenerate. Instead, parity now acts as

$$P : |n\rangle \to |-n+1\rangle$$

In particular, $P|0\rangle = |1\rangle$ and $P|1\rangle = |0\rangle$. This shift also shows up when we see how parity mixes with translations. We now have

$$P \, T_\alpha \, P = e^{i\alpha} T_{-\alpha}$$

This is no longer the group $O(2)$; it is sometimes referred to as the central extension of $O(2)$. Said slightly differently, we have a projective representation of $O(2)$ on the Hilbert space $\mathcal{H}$ of the theory. We can define a representation of $O(2)$ on the rays $\mathcal{H}/\mathbf{C}^\star$, but this does not lift to a representation on the Hilbert space itself.

$\underline{\theta \neq 0, \pi}$: When θ does not take a special value, there is no $\mathbf{Z}_2$ symmetry and a unique ground state. For $\theta < \pi$, the ground state is $|0\rangle$; for $\theta > \pi$ is is $|1\rangle$.

Coupling to Background Gauge Fields

For the chiral anomaly, the breakdown of the symmetry showed up most clearly when we coupled to background gauge fields (3.34). Our quantum mechanical example is no different. We turn on a background gauge field for the $U(1)$ symmetry $x \to x + \alpha$. This means that we return to our original Lagrangian (3.84) and replace it with the action

$$S_{\theta,k} = \int dt\ \frac{m}{2}\left(\dot{x} + A_0\right)^2 + \frac{\theta}{2\pi}\left(\dot{x} + A_0\right) + pA_0$$

This Lagrangian is invariant under the symmetry $x \to x + \alpha(t)$ and $A_0 \to A_0 - \dot{\alpha}(t)$. We've also included an extra term, pA_0. This is an example of a quantum mechanical Chern-Simons term. (We'll spend some time discussing the $d = 2 + 1$ version of this term in Section 8.4.) We've already encountered terms like this before in Section 2.1.3, where we argued that it was compatible with gauge invariance provided

$$p \in \mathbf{Z}$$

Our new action is not quite invariant under $\theta \to \theta + 2\pi$. We now have

$$S_{\theta+2\pi,p} = S_{\theta,p+1}$$

Equivalently, we should identify $(\theta, p) \sim (\theta + 2\pi, p - 1)$.

Now let's look at the action of parity. We still have $x \to -x$, but this must now be augmented by $P : A_0 \to -A_0$. At $\theta = 0$, this is still a good symmetry of the theory provided that $p = 0$. However, at $\theta = \pi$, we have a problem. The action of parity maps $\theta = \pi$ to $\theta = -\pi$ and $p \to -p$. We then need to shift θ back to π which, in turn, shifts $p \to p - 1$. In other words,

$$P : (\theta, p) = (\pi, p) \to (-\pi, -p) \sim (\pi, -p - 1)$$

But there is no $p \in \mathbf{Z}$ for which $-p - 1 = p$. This fact that the Chern-Simons levels necessarily differ after parity means that the theory is not parity invariant at $\theta = \pi$: it suffers a mixed 't Hooft anomaly between parity and translations.

The Partition Function

Here is yet another way to say the same thing. Let's consider the Euclidean partition function, with Euclidean time $\mathbf{S}^1$ of radius β. We introduce the chemical potential $\int d\tau A_0 = \mu$. Large gauge transformations mean that $\mu \sim \mu + 2\pi$.

We can compute the partition function

$$Z = \mathrm{Tr}\, e^{-\beta E + i\mu Q}$$

where Q is the $U(1)$ charge of the state. We will compute the partition function at $\theta = \pi$. For our purposes it will suffice to focus on the ground states $|0\rangle$ and $|1\rangle$ which we take to have $E = 0$. These have charges $Q = 0$ and $Q = 1$ respectively. We have

$$Z_{\text{ground}} = 1 + e^{i\mu}$$

Under parity, we have $P : \mu \to -\mu$. We see again that the partition function is not invariant under parity, $\mu \to -\mu$. This is not surprising: the two states have different charges under the $U(1)$ symmetry.

There is, however, once again a loophole. The two states $|0\rangle$ and $|1\rangle$ have charge that differs by 1. We can make the theory parity invariant if we assign these two states with charges $\pm\frac{1}{2}$. The partition function is then

$$Z_{\text{new}} = e^{-i\mu/2} + e^{i\mu/2} = e^{-i\mu/2} Z_{\text{ground}}$$

Now we have a partition function that is invariant under parity. But there's a price we've paid: it is no longer invariant under $\mu \to \mu + 2\pi$. This is reminiscent of the story of chiral fermions, where we could shift the anomaly between the $U(1)_V$ and $U(1)_A$ symmetries.

Adding a Potential

So far we've argued that there is a subtle interplay between parity and translations when $\theta = \pi$, which we can think of as a 't Hooft anomaly. But what is it good for? As we now explain, anomalies of this kind can be used to restrict the dynamics of the theory.

So see this, we remove the background gauge field but, in its place, turn on a potential for x. Clearly any potential must be invariant under $x \to x + 2\pi$. However, we will request something more: we will ask that the potential is invariant under $x \to x + \pi$. For example, we consider the potential

$$L = \frac{m}{2}\dot{x}^2 + \frac{\theta}{2\pi}\dot{x} + \lambda\cos(2x)$$

This has two classical ground states at $x = 0$ and $x = \pi$. Moreover, the $U(1)$ translation symmetry is broken to

$$U(1) \rightarrow \mathbf{Z}_2$$

This means that at $\theta = 0$ and $\theta = \pi$ we have two discrete symmetries: $T_\pi : x \rightarrow x + \pi$ and $P : x \rightarrow -x$.

At $\theta = 0$, the operators obey the algebra $T_\pi P = P T_\pi$. This is the algebra $\mathbf{Z}_2 \times \mathbf{Z}_2$.

But at $\theta = \pi$ there is a subtlety. The central extension means that these generators obey

$$P\, T_\pi\, P = -T_\pi \tag{3.85}$$

We can define the two elements $a = P$ and $b = T_\pi P$. These obey $a^2 = 1$ and $b^2 = T_\pi P T_\pi P = -1$ so that $b^4 = 1$. Also, we have $aba = b^{-1}$. This is the D_8 algebra; it is the symmetries of rotations of a square.

The D_8 algebra can't act on a single ground state. In particular, if both T_π and P act as phases on a state, then we can't satisfy the algebra (3.85). That means that the quantum mechanics must have two ground states for all values of λ. We can reach the same conclusion for any potential that retains T_π as a symmetry.

This argument is slick, but it is powerful. Usually we learn that double-well quantum mechanics has just a single ground state, with the two classical ground states split by instantons. The argument above says that this doesn't happen in the present situation when $\theta = \pi$. This is perhaps rather surprising. At a more prosaic level, it arises because there are two instantons which tunnel between the two vacua, one which goes one way around the circle and one which goes the other. At $\theta = \pi$, these two contributions should cancel.

3.6.2 Generalised Symmetries

We want to build up to understanding discrete anomalies in Yang-Mills theory. However, the anomalies turn out to lie in a class of symmetries that are a little unfamiliar. These go by the name of *generalised symmetries*.

We will first discuss generalised global symmetries (as opposed to gauge symmetries). We're very used to dealing with such symmetries as acting on fields or, more generally, local operators of the theory. We have both continuous and discrete symmetries. Continuous symmetries have an associated current J which is a 1-form obeying $d^\star J = 0$.

(In contrast to the rest of the lectures, throughout this section we use the notation of forms.) The charge is constructed from J, together with a co-dimension 1 submanifold $M \subset X$ of spacetime X,

$$Q = \int_M {}^*J \tag{3.86}$$

This charge then acts on local operators, defined at a point x, by

$$e^{iQ}\phi^i(x) = R^i{}_j\phi^j(x)$$

where $R^i{}_j$ is the generator of the group element and $x \in M$.

If we have a discrete symmetry, there is no current but, nonetheless, the generator is still associated to a co-dimension one manifold. We will refer to both continuous and discrete symmetries of this type as 0-form symmetries. These are the usual, familiar symmetries of quantum field theories that we have happily worked with our whole lives.

The idea of a *generalised symmetry* is to extend the ideas above to higher-form symmetries. We define a *q-form symmetry* to be one such that the generator is associated to a co-dimension $q+1$ manifold $M \subset X$. If the symmetry is continuous, then there is a $q+1$-form current J and the generator can again be written as (3.86).

For $q > 0$, these generalised symmetries are always Abelian. This follows from the group multiplication, $Q_{g_1}(M)Q_{g_2}(M)$. When $q = 0$, the manifolds M are co-dimension one and we can make sense of this product by time ordering the manifolds M. For $q > 0$, there is no such ordering. This means that the operators must all commute with each other.

A q-form symmetry acts on an operator associated to a q-dimensional manifold C. Here our interest lies in 1-form symmetries. These act on line operators such as the Wilson and 't Hooft lines. Take, for example, a Wilson line W. The action of a 1-form symmetry takes the form $QW = rW$ where r is a phase and the manifolds M and C have linking number 1.

Generalised Symmetries in Maxwell Theory

Our ultimate interest is in generalised symmetries in Yang-Mills theory. But it will prove useful to first discuss generalised symmetries in the context of pure Maxwell theory.

There are two 2-forms which are conserved. Each can be thought of as the current for a global 1-form symmetry

$$
\begin{aligned}
\text{Electric 1-form symmetry:} \quad & J^e = \frac{2}{g^2} F \\
\text{Magnetic 1-form symmetry:} \quad & J^m = \frac{1}{2\pi} {}^\star F
\end{aligned} \tag{3.87}
$$

Each of these currents is conserved, in the sense that they obey $d{}^\star J = 0$. The electric 1-form symmetry shifts the gauge field by a flat connection: $A \to A + d\alpha$. In contrast, the action of the magnetic 1-form symmetry is difficult to see in the electric description; instead, it shifts the magnetic gauge field $\tilde{A}$ by a flat connection. Relatedly, the electric 1-form symmetry acts on Wilson lines W; the magnetic 1-form symmetry acts on 't Hooft lines T.

The fate of these symmetries depends on the phase of the theory which, as explained in Sections 2.5 and 2.6, is governed by the Wilson and 't Hooft line expectation values. These typically give either area law, or perimeter law. We will say:

$$
\begin{aligned}
\text{Area law:} \quad & \langle W \rangle \sim e^{-A} \quad \Rightarrow \quad \langle W \rangle = 0 \\
\text{Perimeter law:} \quad & \langle W \rangle \sim e^{-L} \quad \Rightarrow \quad \langle W \rangle \neq 0
\end{aligned}
$$

This may look a little arbitrary, but it is a natural generalisation of what we already know. A traditional, 0-form symmetry, is said to be spontaneously broken if a charged operator $\mathcal{O}$ has expectation value $\lim_{|x-y|\to\infty} \langle \mathcal{O}(x)\mathcal{O}(y)\rangle = \langle \mathcal{O}(x)\rangle\langle \mathcal{O}(y)\rangle \neq 0$. In other words, the expectation value depends only on the edge points x and y. The analogy for a 1-form symmetry is that the expectation value depends only on the perimeter.

With this convention, in the Coulomb phase we have $\langle W \rangle \neq 0$ and $\langle T \rangle \neq 0$, so that both symmetries are spontaneously broken. But a broken global symmetry should give rise to an associated massless Goldstone boson. This is nothing but the photon itself,

$$
\langle 0 | F_{\mu\nu} | \epsilon, p \rangle \sim (\epsilon_\mu p_\nu - \epsilon_\nu p_\mu) e^{ipx}
$$

This gives a rather surprising new perspective on an old question. Whenever we have massless degrees of freedom, there is usually some underlying reason. For massless scalar fields, Goldstone's theorem typically provides the reason. But we see that we can also invoke Goldstone's theorem to explain why the photon is gapless: we just need to extend its validity to higher form symmetries.

We can also think about the fate of these symmetries when we add matter to the theory. Suppose, first, that we introduce charged electric degrees of freedom. This explicitly breaks the electric one-form symmetry since $d^\star J \sim d^\star F$ which no longer vanishes because the Maxwell equations now have a source. However, the magnetic symmetry, which follows from the Bianchi identity, survives. It is spontaneously broken in the Coulomb phase, but unbroken in the Higgs phase. Moreover, here we have magnetic vortex strings described in Section 2.5.2, that carry charge under the 1-form symmetry.

In contrast, if we introduce magnetic degrees of freedom then only the electric 1-form symmetry survives. This is broken in the Coulomb phase, but unbroken in the Higgs phase where the confining electric strings carry charge.

There is a variant of this. Suppose that we add electrically charged matter but with charge N. Then there is a $\mathbf{Z}_N$ electric 1-form symmetry which shifts the gauge field by a flat connection with $\mathbf{Z}_N$ holonomy which leaves the matter invariant. In the Coulomb phase, both this symmetry and the magnetic 1-form symmetry are broken, as before. But something novel happens in the Higgs phase where the gauge symmetry breaks $U(1) \to \mathbf{Z}_N$. Now $\langle W \rangle \neq 0$ reflecting the fact that the $\mathbf{Z}_N$ electric 1-form symmetry is spontaneously broken, while the magnetic 1-form symmetry survives. Alternatively, we could also add charge 1 monopoles which condense, so that the gauge theory confines. Now $\langle W \rangle = 0$ but $\langle W^N \rangle \neq 0$ since the dynamical matter can screen, causing the string to break. We see that the $\mathbf{Z}_N$ electric 1-form symmetry is unbroken in this phase.

The various dynamics on display above suggests the following relationship:

$$\text{Spontaneously broken 1-form symmetry } H \;\Rightarrow\; \text{Unbroken gauge symmetry } H$$

This is interesting. A discrete gauge symmetry in the infra-red is a form of topological order. This is because, when compactified on non-trivial manifolds, we can have flat connections for this discrete gauge symmetry — which is another way of saying holonomy around cycles. These flat connections can then give rise to multiple ground states.

Generalised Symmetries in Yang-Mills

Finally, we turn to our main topic of interest. We will study the generalised symmetries in Yang-Mills theory with two different gauge groups, $G = SU(N)$ and $SU(N)/\mathbf{Z}_N$. The latter group is sometimes referred to as $PSU(N) \equiv SU(N)/\mathbf{Z}_N$. Much of what we have to say will be a recapitulation of the ideas we saw in Section 2.6.2 regarding 't Hooft and Wilson lines, now viewed in the language of generalised symmetries.

$\underline{G = SU(N)}$

The Abelian story above has a close analog in non-Abelian gauge dynamics. We start by considering the case of simply connected gauge group, $G = SU(N)$. We can have Wilson lines in all representations of G, with charges lying anywhere in the electric weight lattice. If we denote the Wilson line in the fundamental representation by W, this means that we have W^l for all $l = 1, 2, \ldots$. In contrast, the 't Hooft lines must carry charges in the magnetic root lattice. If we denote the "fundamental" 't Hooft line as T, this means that we only have T^N and multiples thereof.

As long as there is no matter transforming under the $\mathbf{Z}_N$ centre of $SU(N)$, then the theory also has an electric $\mathbf{Z}_N$ one-form symmetry. This acts by shifting the gauge field by a flat $\mathbf{Z}_N$ gauge connection or, equivalently, inducing a holonomy in the $\mathbf{Z}_N$ centre of $SU(N)$. Another way of saying this is that the Wilson line W picks up a phase ω with $\omega^N = 1$ under this 1-form symmetry.

When the theory lies in the confining phase, the $\mathbf{Z}_N$ 1-form symmetry is unbroken. Here we have $\langle W \rangle \sim e^{-A}$, with A the area of the loop, and the theory has electric flux tubes which, due to the absence of fundamental matter, cannot break. These electric flux tubes are $\mathbf{Z}_N$ strings which carry charge under the $\mathbf{Z}_N$ one-form symmetry.

This theory also has a different phase. We can access this if we introduce scalar fields ϕ transforming in the adjoint of the gauge group, so that the $\mathbf{Z}_N$ one-form symmetry remains. Then by going to a Higgs phase with $\langle \phi \rangle \neq 0$, we have $\langle W \rangle \sim e^{-L}$, with L the perimeter of the loop. Now the $\mathbf{Z}_N$ symmetry is broken. Correspondingly, there are no electric flux tubes in this phase. However, we now have a topological field theory at low energies because $G = SU(N) \rightarrow \mathbf{Z}_N$, so a discrete $\mathbf{Z}_N$ gauge symmetry remains.

Summarising, we can view the Wilson line as an order parameter for the electric one-form symmetry

$$\text{Electric } \mathbf{Z}_N \text{ one-form symmetry: } \begin{cases} \text{unbroken if } \langle W \rangle \sim e^{-A} \\ \text{broken if } \langle W \rangle \sim e^{-L} \end{cases}$$

A broken $\mathbf{Z}_N$ one-form symmetry gives rise to a $\mathbf{Z}_N$ gauge symmetry.

$\underline{G = SU(N)/\mathbf{Z}_N}$

Now let's consider how this story changes when $G = SU(N)/\mathbf{Z}_N$. The Wilson lines are now restricted to lie in the electric root lattice, so only multiples of W^N survive. In contrast, the whole range of 't Hooft lines T^l with $l = 1, 2, \ldots$ are allowed. (Strictly speaking, this is true at $\theta = 0$; we'll look at the role of the θ angle below.)

The theory now has a magnetic $\mathbf{Z}_N$ one-form symmetry, whose order parameter is the 't Hooft line T. We have

$$\text{Magnetic } \mathbf{Z}_N \text{ one-form symmetry: } \begin{cases} \text{unbroken if } \langle T \rangle \sim e^{-A} \\ \text{broken if } \langle T \rangle \sim e^{-L} \end{cases}$$

So this magnetic $\mathbf{Z}_N$ one-form symmetry is broken in the confining phase, resulting in an emergent $\mathbf{Z}_N$ magnetic gauge symmetry.

3.6.3 Discrete Gauge Symmetries

We're going to need one final piece of technology for our story. This is the idea of a gauge symmetry based on a discrete, rather than continuous, group.

It's tempting to think of a gauge symmetry as something in which the transformation can take different values at different points in space. But this approach clearly runs into problems for a discrete group since the transformation parameter cannot vary continuously. Instead, we should remember the by-now familiar mantra: gauge symmetries are redundancies. A discrete gauge symmetry simply means that we identify configurations related by this symmetry.

There is a simple, down-to-earth method to arrive at a discrete gauge theory: we start with a continuous gauge theory, and subsequently break it down to $\mathbf{Z}_N$. Indeed, we already saw two examples of this above. In the first, we start with $U(1)$ gauge theory, with a scalar of charge N. Upon condensation, we have $U(1) \to \mathbf{Z}_N$. Alternatively, we could take $SU(N)$ gauge theory with adjoint Higgs fields, giving rise to $SU(N) \to \mathbf{Z}_N$.

Here we take the $U(1)$ gauge theory as our starting point. We can focus on the phase, $\phi \in [0, 2\pi)$ of the scalar field. We have a gauge symmetry

$$\phi \to \phi + N\alpha$$

where $\alpha \sim \alpha + 2\pi$ is also periodic. In the Higgs phase, the scalar kinetic term is

$$\mathcal{L}_1 = t^2 (d\phi - NA) \wedge {}^\star(d\phi - NA)$$

for some $t \in \mathbf{R}$ which is set by the expectation value of the scalar. In the low-energy limit, $t^2 \to \infty$ and we have $A = \frac{1}{N} d\phi$ which tells us that the connection must be flat. However, something remains because the holonomy around any non-contractible loop can be $\frac{1}{2\pi} \oint A \in \frac{1}{N}\mathbf{Z}$.

It is useful to dualise ϕ. We do this by first introducing a 3-form H and writing

$$\mathcal{L}_{1.5} = \frac{1}{(4\pi)^2 t^2} H \wedge {}^\star H + \frac{i}{2\pi} H \wedge (d\phi - NA)$$

Integrating out H through the equation of motion ${}^\star H = 4\pi i t^2 (d\phi - NA)$ takes us back to the original Lagrangian $\mathcal{L}_1$. Meanwhile, if we send $t^2 \to \infty$ at this stage, we get the Lagrangian

$$\mathcal{L}_{1.5} \to \frac{i}{2\pi} H \wedge (d\phi - NA)$$

where H now plays the role of a Lagrange multiplier, imposing $A = \frac{1}{N} d\phi$. Alternatively, we can instead integrate out ϕ in $\mathcal{L}_{1.5}$. The equation of motion requires that $dH = 0$. This means that we can write $H = dB$ locally. We're then left with the Lagrangian

$$\mathcal{L}_2 = \frac{1}{(4\pi)^2 t^2} H \wedge {}^\star H + \frac{iN}{2\pi} B \wedge dA$$

In the limit $t^2 \to \infty$, this becomes

$$\mathcal{L}_{BF} = \frac{iN}{2\pi} B \wedge dA$$

This Lagrangian is known as *BF theory*. It is deceptively simple and, as we have seen above, is ultimately equivalent to a $\mathbf{Z}_N$ discrete gauge symmetry. Our task now is to elucidate how this works. The subtleties arise from the fact that the two gauge fields have quantised periods, so when integrated over appropriate cycles yield

$$\int_{\Sigma^2} F \in 2\pi \mathbf{Z} \quad \text{and} \quad \int_{\Sigma^3} H \in 2\pi \mathbf{Z}$$

The BF theory has two gauge symmetries: $A \to A + d\alpha$ and $B \to B + d\lambda$. However, as we've seen, the $U(1)$ gauge theory for A is actually Higgsed down to $\mathbf{Z}_N$, a fact which is clear in our initial formulation in $\mathcal{L}_1$, but less obvious in the BF theory formulation. Similarly, the 1-form gauge symmetry for B is also Higgsed down to a $\mathbf{Z}_N$ 1-form gauge symmetry. To see this, we dualise A. We first add a Maxwell term for $F = dA$ and consider the Lagrangian

$$\mathcal{L}_{2.5} = \frac{1}{2e^2} F \wedge {}^\star F - \frac{i}{2\pi} F \wedge (d\hat{A} - NB)$$

If we integrate out the 1-form $\hat{A}$, we recover the fact that $F = dA$ locally. Note that if we send $e^2 \to \infty$, to remove the Maxwell term, we're left with

$$\mathcal{L}_{2.5} \to -\frac{i}{2\pi} F \wedge (d\hat{A} - NB) \tag{3.88}$$

where F now plays the role of a Lagrange-multiplier 2-form. Alternatively, we can instead integrate out F using its equations of motion ${}^\star F = -\frac{ie^2}{2\pi}(d\hat{A} - NB)$ to get

$$\mathcal{L}_3 = \frac{e^2}{8\pi^2}(d\hat{A} - NB) \wedge {}^\star(d\hat{A} - NB)$$

This now takes a similar form as the action $\mathcal{L}_1$ that we started with. We should view the dual gauge field $\hat{A}$ as a matter field which is charged under B. Correspondingly, the $U(1)$ 1-form gauge symmetry is Higgsed down to $\mathbf{Z}_N$.

What we learn from this is that a $\mathbf{Z}_N$ discrete gauge theory also comes with a $\mathbf{Z}_N$ 1-form gauge symmetry.

The Operators

Our theory has two gauge symmetries, under which

$$\begin{aligned} \phi \rightarrow \phi + N\alpha \quad &\text{and} \quad A \rightarrow A + d\alpha \\ \hat{A} \rightarrow \hat{A} + N\lambda \quad &\text{and} \quad B \rightarrow B + d\lambda \end{aligned}$$

As we've seen, both are Higgsed down to $\mathbf{Z}_N$. Nonetheless, all operators that we write down must be invariant under these symmetries. Examples of such operators include

$$d\phi - NA \sim {}^\star H \quad \text{and} \quad d\hat{A} - NB \sim {}^\star F \tag{3.89}$$

where the equations of motion show that these are actually related to the dual fields H and F respectively. However, these are all trivial in the theory. To find something more interesting, we must turn to line and surface operators.

There are two electric operators, a Wilson line $W_A[C]$ and a "Wilson surface", $W_B[S]$,

$$W_A[C] = \exp\left(i\int_C A\right) \quad \text{and} \quad W_B[S] = \exp\left(i\int_S B\right)$$

As usual, the Wilson line describes the insertion of a probe particle of charge 1 with worldline C. Meanwhile, the Wilson surface describes the insertion of a vortex string with worldsheet S. The scalar ϕ has winding $\int d\phi = 2\pi$ around the vortex which, using $A = \frac{1}{N}d\phi$, means that the vortex string carries magnetic flux $1/N$. A particle of charge 1 picks up a holonomy $2\pi/N$ through the Aharonov-Bohm effect. This is captured in the correlation function

$$\Big\langle W_A[C]\, W_B[S] \Big\rangle = \exp\left(\frac{2\pi i}{N} n(C,S)\right)$$

where $n(C,S)$ is the linking number of C and S. This correlation function is the non-trivial content of the $\mathbf{Z}_N$ gauge theory. We see, in particular, that the operators W_A^N and W_B^N are both trivial in the sense that they commute with all other operators. This can also be understood by a $\mathbf{Z}_N$ gauge transformation which takes a general operator

$$W_A^q[C] = \exp\left(iq\int_C A\right)$$

and shifts $q \to q + N$. Note that we can also think of this as a $\mathbf{Z}_N$ global 1-form symmetry. Because $\langle W_A[C]\rangle \sim e^{-L}$, this 1-form symmetry is spontaneously broken, in agreement with our previous discussion that this should accompany a $\mathbf{Z}_N$ gauge symmetry.

One might think that there are also 't Hooft operators in the theory, constructed by exponentiating the gauge invariant operators (3.89). The magnetic gauge field dual to A is $\hat{A}$, and we can write

$$T_A[C,S] = \exp\left(i\int_C \hat{A} - iN\int_S B\right) \tag{3.90}$$

where, now, S is a surface which ends on the line C. The insertion of a 't Hooft line is equivalent to cutting out a tube $\mathbf{S}^2\times\mathbf{R}$ around C and imposing $\int_{\mathbf{S}^2} F = 2\pi$. However, the operator $T_A[C,S]$ is trivial in the theory. First, note that the attached surface operator has charge N and so is invisible. Moreover, by a gauge transformation we can always set $\hat{A} = 0$ locally. The real meaning of the 't Hooft operator $T_A[C,S]$ is simply that N Wilson surface operators can end on a line.

We can view this in a slightly different way. Suppose that there are magnetic monopoles of charge 1 under the $U(1)$ gauge symmetry. This gauge symmetry is Higgsed which means that these monopoles are attached to strings. But the minimum string has charge $1/N$, so the monopole is attached to N strings.

An analogous operator can be constructed using the magnetic dual to B. We have

$$T_B[P,C] = \exp\left(i\phi(P) - iN\int_C A\right)$$

where now C is a line which ends at the point P. The same arguments as above mean that this operator is also trivial. It is telling us only that N Wilson line operators can end at a point.

3.6.4 Gauging a $\mathbf{Z}_N$ One-Form Symmetry

Finally we can start to put the pieces together. Recall that $G = SU(N)$ Yang-Mills has a $\mathbf{Z}_N$ global electric one-form symmetry that acts on Wilson lines. We will show that if we promote this one-form symmetry to a gauge symmetry then we end up with $G = SU(N)/\mathbf{Z}_N$ Yang-Mills.

We can also play this game in reverse. Starting with $G = SU(N)/\mathbf{Z}_N$ Yang-Mills, we can gauge the global magnetic one-form symmetry to return to $G = SU(N)$ Yang-Mills.

To this end, let's start with $SU(N)$ Yang-Mills. We have a proliferation of gauge fields of various kinds, and we're running out of letters. So, for this section only, we will refer to the $SU(N)$ gauge connection as a. We will couple this to a BF theory which we write in the form (3.88),

$$\mathcal{L}_{BF} = \frac{i}{2\pi}\int Z\wedge(d\hat{V} - NB)$$

The trick is to combine the $SU(N)$ gauge connection a with the $U(1)$ gauge connection $\hat{V}$ to form a $U(N) \cong (U(1)\times SU(N))/\mathbf{Z}_N$ connection

$$\mathcal{A} = a + \frac{1}{N}\hat{V}\,\mathbf{1}_N$$

Here's what's going on. We could try to construct a flat connection a from a $SU(N)/\mathbf{Z}_N$ bundle which is not an $SU(N)$ bundle. This is not allowed in the $SU(N)$ theory. However, we can compensate this with a gauge connection $\hat{V}$ which would not be allowed in a pure $U(1)$ theory. The obstructions cancel between the two, so we're left with a good $U(N)$ gauge connection. We then define the $U(N)$ field strength

$$\mathcal{G} = d\mathcal{A} + \mathcal{A}\wedge\mathcal{A}$$

This field strength is not invariant under the 1-form gauge symmetry of the BF theory, namely $\hat{V}\to\hat{V}+N\lambda$ and $B\to B+d\lambda$; it transforms as

$$\mathcal{G}\to\mathcal{G}+d\lambda$$

This means that we can't simply write down the usual Yang-Mills term for $\mathcal{G}$. Instead, we need to form the gauge invariant combination $\mathcal{G}-B$ and write the action

$$S_{SU(N)/\mathbf{Z}_N} = \frac{1}{2g^2}\int \mathrm{Tr}\,(\mathcal{G}-B)\wedge{}^\star(\mathcal{G}-B) + \frac{i}{2\pi}\int Z\wedge(d\hat{V}-NB) \qquad (3.91)$$

Note that we have set the theta term to zero here because it comes with its own story which we will tell later. To see what's happening, we can look at the line operators. We started with an $SU(N)$ gauge theory with Wilson line

$$W[C] = \mathrm{Tr}\,\mathcal{P}\exp\left(i\int_C a\right) \tag{3.92}$$

However, this is not invariant under the $U(N)$ gauge transformations that lie in $SU(N)/\mathbf{Z}_N$ rather than $SU(N)$. So we need to augment it to get a gauge invariant operator. The obvious thing to do is to replace a with the $U(N)$ connection $\mathcal{A}$, but now this fails to be gauge invariant under the 1-form symmetry. To resolve this, we need to work with

$$\mathcal{W}[C,\Sigma] = W[C]\exp\left(\frac{i}{N}\int_C \hat{V} - i\int_\Sigma B\right)$$

where $\partial\Sigma = C$. This is now gauge invariant, but it comes with its own woes because it's not a line operator but a surface operator, depending on the choice of Σ. To get an honest line operator, we need to take

$$\mathcal{W}^N[C,\Sigma] = W^N[C]\exp\left(i\int_C \hat{V} - iN\int_\Sigma B\right)$$

As before, the constraint from integrating out Z tells us that $N\int_\Sigma B = \int_\Sigma dA$. But on any closed manifold, $\int dA \in 2\pi\mathbf{Z}$. This means that the line operator $\mathcal{W}^N[C,\Sigma]$ doesn't really depend on the choice of Σ. But this is exactly the class of Wilson lines which are allowed in $SU(N)/\mathbf{Z}_N$.

From our discussion in the previous section (and in Section 2.6.2), we know that the $SU(N)/\mathbf{Z}_N$ theory has more 't Hooft lines that the $SU(N)$ theory that we started from. These are easy to write down in our new formulation: they are

$$T[C] = \exp\left(i\int_\Sigma Z\right) \tag{3.93}$$

The Theta Term

Now let's add a theta term into the game. One of the key distinctions between $SU(N)$ and $SU(N)/\mathbf{Z}_N$ Yang-Mills is that $\theta \in [0, 2\pi)$ in the former, while $\theta \in [0, 2\pi N)$ in the latter. How does this distinction arise when transforming from one theory to another?

We start by writing the obvious, gauge invariant theta term

$$S_\theta = \frac{i\theta}{8\pi^2}\int \mathrm{Tr}\,(\mathcal{G} - B)\wedge(\mathcal{G} - B)$$

where $\theta \in [0, 2\pi)$. Under the shift $\theta \to \theta + 2\pi$, we apparently have

$$\Delta S_\theta = \frac{i}{4\pi}\int \mathrm{Tr}\,\mathcal{G}\wedge\mathcal{G} - \frac{i}{2\pi}\int \mathrm{Tr}\,\mathcal{G}\wedge B + \frac{iN}{4\pi}\int B\wedge B$$

The equation of motion for Z tells us that $\mathrm{Tr}\,\mathcal{G} = d\hat{V} = NB$. Using this relation, we have

$$\Delta S_\theta = \frac{i}{4\pi}\int \mathrm{Tr}\,\mathcal{G}\wedge\mathcal{G} - \frac{iN}{4\pi}\int B\wedge B$$

The first term above is an integer multiple of 2π, so we have

$$\Delta S_\theta = -\frac{iN}{4\pi}\int B\wedge B + 2\pi i\mathbf{Z}$$

We see that the action isn't invariant under the shift $\theta \to \theta + 2\pi$ but, as we've seen in other contexts, what we really care about is e^{iS_θ}. And this too is not quite invariant, but shifts by a contact term for B. For this reason, we augment our theta angle action to become

$$S_\theta = \frac{i\theta}{8\pi^2}\int \mathrm{Tr}\,(\mathcal{G}-B)\wedge(\mathcal{G}-B) - \frac{ipN}{4\pi}\int B\wedge B \tag{3.94}$$

We will ultimately see that p plays the role of a discrete theta angle. First, we note again that the effect of sending $\theta \to \theta + 2\pi$ is

$$p \to p - 1$$

At first glance, the $B\wedge B$ term doesn't look gauge invariant under shifts $B \to B + d\lambda$. But this is misleading: the term is gauge invariant provided that $p \in \mathbf{Z}$. Indeed, our original θ term is manifestly gauge invariant, so this contact term must also be. To see this explicitly, note that under a gauge transformation, we have

$$\frac{ipN}{4\pi}\int B\wedge B \;\to\; \frac{ipN}{4\pi}\int B\wedge B + \frac{ipN}{2\pi}\int d\lambda\wedge B + \frac{ipN}{4\pi}\int d\lambda\wedge d\lambda$$

Here the 1-form has $\int d\lambda \in 2\pi\mathbf{Z}$ which means that the last term is an integer multiple of 2π. (Actually, for N even this is true, while for N odd it is true only on spin manifolds.) Meanwhile, using the constraint $NB = d\hat{V}$, we also have $\int B \in (2\pi/N)\mathbf{Z}$, so the second term is also an integer multiple of 2π and the partition function is gauge invariant.

Finally, note that this same integrality constraint means that $\frac{1}{4\pi}\int B\wedge B \in (2\pi/N^2)\mathbf{Z}$. This means that the discrete theta angle p in (3.94) can take values

$$p = 0, 1, \ldots, N-1$$

As we would expect. The theta angle of the $SU(N)/\mathbf{Z}_N$ theory will be

$$\theta_{SU(N)/\mathbf{Z}_N} = 2\pi p + \theta \in [0, 2\pi N) \tag{3.95}$$

in agreement with our earlier discussion in Section 2.6.2.

We would next like to see how the discrete theta angle p shifts the electric charge of 't Hooft lines. First there is a fairly straightforward, albeit slightly handwaving argument. If we rewrite

$$\frac{ipN}{4\pi}\int B\wedge B = \frac{ip}{4\pi N}\int d\hat{V}\wedge d\hat{V}$$

then we see that this looks like a standard theta term $\hat{\theta} = 2\pi p/N$ for $\hat{V}$. This will give electric charge to 't Hooft lines of $\hat{V}$ which are, equivalently, the Wilson lines of the dual gauge field V. These are precisely the operators (3.93) which we identified as the new emergent 't Hooft lines of the $SU(N)/\mathbf{Z}_N$ theory.

There is a more direct way to see this. We can also directly require that Z transforms under the 1-form gauge symmetry as

$$Z \to Z + p\,d\lambda \tag{3.96}$$

The integrality condition $\int Z \in 2\pi\mathbf{Z}$ and $\int d\lambda \in 2\pi\mathbf{Z}$ is retained if $p \in \mathbf{Z}$. This renders the theory gauge invariant without imposing the constraint $d\hat{V} = NB$. With the gauge transformation on Z, we see immediately that the 't Hooft lines (3.93) are no longer gauge invariant, transforming as $T[C] \to e^{ip\int_C \lambda}T[C]$. To compensate, we're forced to use the line operators

$$\tilde{T}[C] = T[C]\,\mathrm{Tr}\,\mathcal{P}\exp\left(-ip\int_C \mathcal{A}\right)$$

This is the dyonic line operator, in which the magnetic 't Hooft line picks up an electric charge. This is precisely the expected effect of the discrete theta angle.

3.6.5 A 't Hooft Anomaly in Time Reversal

It's been rather a long road to put together all the machinery that we need. But, finally, we can put these ideas together to tell us something new.

We sketched the main idea at the beginning of this section. We start with $G = SU(N)$ Yang-Mills which, as we now know, enjoys a $\mathbf{Z}_N$ global, electric one-form symmetry. At two special values $\theta = 0$ and $\theta = \pi$ it also enjoys time reversal invariance, as we reviewed in Section 1.2.5.

Suppose that we work in the theory with $\theta = 0$. If we gauge the $\mathbf{Z}_N$ one-form symmetry, then we find ourselves left with the $G = SU(N)/\mathbf{Z}_N$ Yang-Mills theory, now with $\theta_{SU(N)/\mathbf{Z}_N} = 2\pi p$ with p the discrete theta angle that appeared in (3.95). We are always free to pick $p = 0$ and we end up with a theory which preserves time reversal invariance.

However, life is different if we sit at $\theta = \pi$. Now if we gauge the $\mathbf{Z}_N$ one-form symmetry, we're left with the $G = SU(N)/\mathbf{Z}_N$ Yang-Mills, but now with

$$\theta_{SU(N)/\mathbf{Z}_N} = (2p + 1)\pi$$

For some $p \in \mathbf{Z}$. This theory is time reversal invariant only when $\theta_{SU(N)/\mathbf{Z}_N} = 0$ and $\theta_{SU(N)/\mathbf{Z}_N} = \pi N$.

Let's first consider N even. In this case, there is no choice of $p \in \mathbf{Z}$ for which our final theory is time reversal invariant. We learn that if we start with $\theta = \pi$, then we can gauge the $\mathbf{Z}_N$ one-form symmetry at the cost of losing time reversal invariance. In other words, we have a mixed 't Hooft anomaly between the $\mathbf{Z}_N$ one-form symmetry and time reversal.

So what are the consequences? Importantly, this anomaly must be reproduced in the low-energy physics. At $\theta = 0$, we expect Yang-Mills theory to be in a gapped, boring phase, with nothing interesting going on beyond the strong coupling scale Λ_{QCD}. But this cannot also be the case at $\theta = \pi$: whatever physics occurs there has to account for the anomaly. There are three options: the first two options are entirely analogous to our discussion of 't Hooft chiral anomalies in Section 3.5, but the third is novel:

- Time reversal invariance is spontaneously broken at $\theta = \pi$. This means that the theory is gapped, but with two degenerate ground states. There can be domain walls between these two states.

  Note that there is a theorem, due to Vafa and Witten, which says that parity cannot be spontaneously broken in vector-like gauge theories, but this theorem explicitly applies only at $\theta = 0$.

- The theory is gapless at $\theta = \pi$, and the resulting theory reproduces the discrete 't Hooft anomaly.

- The theory is topological at $\theta = \pi$. This means that it is gapped, with no low-energy propagating degrees of freedom, but still has interesting things going on. One way to probe the subtle behaviour of the theory is to place it on a non-trivial background manifold. For example, the number of ground states depends on the topology of the manifold

What about when N is odd? Here it looks as if we are in better shape, because we can always pick $p = (N-1)/2$ to end up with $\theta_{SU(N)/\mathbf{Z}_N} = N\pi$. This means that, strictly speaking, there is no 't Hooft anomaly in this case. However, there is a global inconsistency, because there is no choice of p which preserves time reversal for both $\theta = 0$ and $\theta = \pi$. If we assume that the theory is confining, gapped and boring when $\theta = 0$ then there is always the possibility that the theory undergoes a first order phase transition as we vary θ from 0 to π. However, if there is no such phase transition, then the theory at $\theta = \pi$ must again be non-trivial, in the sense that it falls into one of the three categories listed above. Thus, in the absence of a first order phase transition, there is no difference between N even and N odd.

So which of these possibilities occurs? We don't know for sure, but we can take some hints from large N. In Section 6.2.5, we will show that when $N \gg 1$, the first option above occurs, and time reversal is spontaneously broken at $\theta = \pi$. There is a general expectation that this behaviour persists for most, if not all, N, simply on the grounds that it appears to be the simplest option.

There is, however, one tantalising possibility for $G = SU(2)$ Yang-Mills. It has been suggested that the theory at $\theta = \pi$ is actually gapless, and its dynamics is described by a single $U(1)$ gauge field. We currently have no way to determine whether this phase is realised, or if time reversal is again spontaneously broken.

3.7 Further Reading

The anomaly is one of the more subtle aspects of quantum field theory. Like much of the subject, it has its roots in a combination of experimental particle physics, and a healthy dose of utter confusion.

The story starts with an attempt to understand the decay of the neutral pion π^0 into two photons. (This story will be told in more detail in Section 5.4.3.) The neutral pion is uncharged, so does not couple directly to photons. In 1949, Steinberger suggested that the decay occurs through a loop process, with the $SU(2)$ isospin triplet of pions π^a coupling to the proton and neutron doublet N through the interaction

$$G_{\pi N}\pi^a \bar{N}^a \gamma^5 \sigma^a N \tag{3.97}$$

The resulting amplitude gets pretty close to the measured pion decay rate of 10^{-16} s. It appeared that all was good.

The trouble came some decades later with the realisation that the pion is a Goldstone boson. (We will explain this when we discuss chiral symmetry breaking in Section 5.) This means that couplings of the form (3.97) are not allowed: the pion can have only derivative couplings. Indeed, one can show that if all the symmetries of the classical Lagrangian hold, then a genuinely massless pion would be unable to decay into two photons [190, 198]. The previous success in predicting the decay of the pion suddenly appeared coincidental.

The anomaly provides the resolution to this puzzle, as first pointed out in 1969 by Bell and Jackiw [16] (yes, *that* Bell [15]) and, independently, by Adler [2]. The extension to non-Abelian gauge groups was made by Bardeen in the same year [12]. (At this point in time, his dad had only one Nobel prize.)

The gravitational contribution to the chiral anomaly was computed as early as 1972 by Delbourgo and Salam [39]. The fact that anomalies cancel in the Standard Model was first shown in [82, 21], albeit phrased as avoiding a lack of renormalisability rather than avoiding a fatal inconsistency. (In fairness, non-renormalisability was thought to be fatal at the time.)

The first hint that the anomaly was related to something deeper can first be seen in a proof, by Adler and Bardeen, that it is one-loop exact. But the full picture took some years to emerge. The relation between instantons and the anomaly was first realised by 't Hooft [101], and the connection to the Atiyah-Singer index theorem was made in [114].

The path integral approach that we described in these lectures is due to Fujikawa and was developed ten years after the anomaly was first discovered [68, 69]. This was, perhaps, the first time that properties of the path integral measure were shown to play an important role in quantum field theory; this has been a major theme since, not least with Witten's discovery in 1982 of the $SU(2)$ anomaly [225].

Excellent reviews of anomalies can be found in lectures by Bilal [18] and Harvey [90].

The idea of a 't Hooft anomaly as an important constraint on low energy physics was introduced by 't Hooft in the lectures [105]; its application to chiral symmetry breaking will be described in Section 5.6.

Section 3.6 on anomalies in discrete symmetries contains somewhat newer material. Discrete gauge symmetries have a long history on the lattice and, in the continuum, were discussed in a number papers studying geometry through the lens of QFT. The presentation of BF given here was largely taken from [11] and generalised higher form symmetries from [70]. The fact that these higher form symmetries can have mixed anomalies with discrete symmetries, such as time reversal, was described in [71]. (The theorem which says that time reversal or parity cannot be spontaneously broken at $\theta = 0$ can be found in [196].) The quantum mechanics analogy of a particle on a circle is taken from the appendix of [71].

4. Lattice Gauge Theory

Quantum field theory is hard. Part of the reason for our difficulties can be traced to the fact that quantum field theory has an infinite number of degrees of freedom. You may wonder whether things get simpler if we can replace quantum field theory with a different theory which has a finite, albeit very large, number of degrees of freedom. We will achieve this by discretizing space (and, as we will see, also time). The result goes by the name of *lattice gauge theory*.

There is one, very practical reason for studying lattice gauge theory: with a discrete version of the theory at hand, we can put it on a computer and study it numerically. This has been a very successful programme, especially in studying the mass spectrum of Yang-Mills and QCD, but it is not our main concern here. Instead, we will use lattice gauge theory to build better intuition for some of the phenomena that we have met in these lectures, including confinement and some subtle issues regarding anomalies.

There are different ways that we could envisage trying to write down a discrete theory:

- Discretize space, but not time. We could, for example, replace space with a cubic, three dimensional lattice. This is known as *Hamiltonian lattice gauge theory*.

  This has the advantage that it preserves the structure of quantum mechanics, so we can discuss states in a Hilbert space and the way they evolve in (continuous) time. The resulting quantum lattice models are conceptually similar to the kinds of things we meet in condensed matter physics. The flip side is that we have butchered Lorentz invariance and must hope that it emerges at low energies.

  This is the approach that we will use when we first introduce fermions in Section 4.3. But, for other fields, we will be even more discrete...

- Discretize spacetime. We might hope to do this in such a way that preserves some remnant of Lorentz invariance, and so provide a natural discrete approximation to the path integral.

  There are two ways we could go about doing this. First, we could try to construct a lattice version of Minkowski space. This, it turns out, is a bad idea. Any lattice clearly breaks the Lorentz group. However, while a regular lattice will preserve some discrete remnant of the rotation group $SO(3)$, it preserves no such remnant of the Lorentz boosts. The difference arises because $SO(3)$ is compact, while $SO(3,1)$ is non-compact. This means that if you act on a lattice with $SO(3)$, you will come back to your starting point after, say, a π rotation. In contrast, acting with a Lorentz boost in $SO(3,1)$ will take you further and further away from your starting point. The upshot is that lattices in Minkowski space are not a good idea.

The other option is to work with Euclidean spacetime. Here there is no problem in writing down a four dimensional lattice that preserves some discrete subgroup of $SO(4)$. The flip side is that we have lost the essence of quantum mechanics; there is no Hilbert space, and no concept of entanglement. Instead, we have what is essentially a statistical mechanics system, with the Euclidean action playing the role of the free energy. Nonetheless, we can still compute correlation functions and, from this, extract the spectrum of the theory and we may hope that this is sufficient for our purposes.

Throughout this section, we will work with a cubic, four-dimensional Euclidean lattice, with lattice spacing a. We introduce four basis vectors, each of unit length. It is useful, albeit initially slightly unfamiliar, to denote these as $\hat{\mu}$, with $\mu = 1,2,3,4$. A point x in our discrete Euclidean spacetime is then restricted to lie on the lattice Γ, defined by

$$\Gamma = \left\{ x : x = \sum_{\mu=1}^{4} a n_\mu \hat{\mu} \ , \ n_\mu \in \mathbf{Z} \right\} \tag{4.1}$$

The lattice spacing plays the role of the ultra-violet cut-off in our theory

$$a = \frac{1}{\Lambda_{UV}}$$

For the lattice to be a good approximation, we will need a to be much smaller than any other physical length scale in our system.

Because our system no longer has continuous translational symmetry, we can't invoke Noether's theorem to guarantee conservation of energy and momentum. Instead we must resort to Bloch's theorem which guarantees the conservation of "crystal momentum", lying in the Brillouin zone, $|k| \leq \pi/a$. (See, for example, the lectures on Applications of Quantum Mechanics.) Umklapp processes are allowed in which the lattice absorbs momentum, but only in units of $2\pi/a$. This means that provided we focus on low-momentum processes, $k \ll \pi/a$, we effectively have conservation of momentum and energy.

(An aside: the discussion above was a little quick. Bloch's theorem is really a statement in quantum mechanics in which we have continuous time. It applies directly only in the framework of Hamiltonian lattice gauge theory. In the present context, we really mean that the implications of momentum conservation on correlation functions will continue to hold in our discrete spacetime lattice, provided that we look at suitably small momentum.)

4.1 Scalar Fields on the Lattice

To ease our way into the discrete world, we start by considering a real scalar field $\phi(x)$. A typical continuum action in Euclidean space takes the form

$$S = \int d^4x \; \frac{1}{2}(\partial_\mu \phi)^2 + V(\phi) \tag{4.2}$$

Our first task is to construct a discrete version of this, in which the degrees of freedom are

$$\phi(x) \text{ with } x \in \Gamma$$

This is straightforward. The kinetic terms are replaced by the finite difference

$$\partial_\mu \phi(x) \;\longrightarrow\; \frac{\phi(x + a\hat{\mu}) - \phi(x)}{a} \tag{4.3}$$

while the integral over spacetime is replaced by the sum

$$\int d^4x \;\longrightarrow\; a^4 \sum_{x\in\Gamma}$$

Our action (4.2) then becomes

$$S = a^4 \sum_{x\in\Gamma} \; \frac{1}{2} \sum_\mu \left(\frac{\phi(x + a\hat{\mu}) - \phi(x)}{a} \right)^2 + V(\phi(x))$$

As always, this action sits in the path integral, whose measure is now simply a whole bunch of ordinary integrals, one for each lattice point:

$$Z = \int \prod_{x\in\Gamma} d\phi(x) \; e^{-S}$$

With this machinery, computing correlation functions of any operators reduces to performing a large but (at least for a lattice of finite size) finite number of integrals.

It's useful to think about the renormalisation group (RG) in this framework. Suppose that we start with a potential that takes the form

$$V(\phi) = \frac{m_0^2}{2}\phi^2 + \frac{\lambda_0}{4}\phi^4 \tag{4.4}$$

As usual in quantum field theory, m_0^2 and λ_0 are the "bare" parameters, appropriate for physics at the lattice scale. We can follow their fate under RG by performing the kind of blocking transformation that was introduced in statistical mechanics by Kadanoff. This is a real space RG procedure in which one integrates out the degrees of freedom on alternate lattice sites, say all the sites in (4.1) in which one or more n_μ is odd. This then leaves us with a new theory defined on a lattice with spacing $2a$. This will renormalise the parameters in the action. In particular, the mass term will typically shift to

$$m^2 \sim m_0^2 + \frac{\lambda_0}{a^2}$$

This is the naturalness issue for scalar fields. If we want to end up with a scalar field with physical mass $m^2_{\rm phys} \ll 1/a^2$, then the bare mass must be delicately tuned to be of order the cut-of, $m_0^2 \sim -\lambda_0/a^2$, so that it cancels the contribution that arises when performing RG. This makes it rather difficult in practice to put scalar fields on the lattice. As we will see below, life is somewhat easier for gauge fields and, after jumping through some hoops, for fermions.

As usual, RG does not leave the potential in the simple, comfortable form (4.4). Instead it will generate all possible terms consistent with the symmetries of the theory. These include higher terms such as ϕ^6 and ϕ^8 in the potential, as well as higher derivative terms such as $(\partial_\mu\phi\partial^\mu\phi)^2$. (Here, and below, we use the derivative notation as shorthand for the lattice finite difference (4.3).) This doesn't bother us because all of these terms are irrelevant (in the technical sense) and so don't affect the low-energy physics.

However, this raises a concern. The discrete rotational symmetry of the lattice is less restrictive than the continuous rotational symmetry of $\mathbf{R}^4$. This means that RG on the lattice will generate some terms involving derivatives $\partial\phi$ that would not arise in the continuum theory. If these terms are irrelevant then they will not affect the infra-red physics and we can sleep soundly, safe in the knowledge that the discrete theory will indeed give a good approximation to the continuum theory at low energies. However, if any of these new terms are relevant then we're in trouble: now the low-energy physics will not coincide with the continuum theory.

So what are the extra terms that arise from RG on the lattice? They must respect the $\mathbf{Z}_2$ symmetry $\phi \to -\phi$ of the original action, which means that they have an even number of ϕ fields. They must also respect the discrete rotation group that includes, for example, $x_1 \to x_2$. This rules out lone terms like $(\partial_1 \phi)^2$. The lowest dimension term involving derivatives that respects these symmetries is

$$\sum_{\mu=1}^{4} (\partial_\mu \phi)^2$$

But this is, of course, the usual derivative term in the action. The first operator that is allowed on the lattice but prohibited in the continuum is

$$\sum_{\mu=1}^{4} \phi\, \partial_\mu^4\, \phi \tag{4.5}$$

This has dimension 6, and so is irrelevant. Happily, we learn that the lattice scalar field theory differs from the continuum only by irrelevant operators. Provided that we fine tune the mass, we expect the long wavelength physics to well approximate a continuum theory of a light scalar field.

4.2 Gauge Fields on the Lattice

We now come to Yang-Mills. Our task is to write down a discrete theory on the lattice that reproduces the Yang-Mills action. For concreteness, we will restrict ourselves to $SU(N)$ gauge theory, with matter in the fundamental representation.

As a first guess, it's tempting to follow the prescription for the scalar field described above and introduce four, Lie algebra valued gauge fields $A_\mu(x)$, with $\mu = 1, 2, 3, 4$ at each point $x \in \Gamma$. This, it turns out, is not the right way to proceed. At an operational level, it is difficult to implement gauge invariance in such a formalism. But, more importantly, this approach completely ignores the essence of the gauge field. It misses the idea of holonomy.

4.2.1 The Wilson Action

Mathematicians refer to the gauge field as a connection. This hints at the fact that the gauge field is a guide, telling the internal, colour degrees of freedom of a particle or field how to evolve through parallel transport. The gauge field "connects" these internal degrees of freedom at one point in space to those in another.

We saw this idea earlier in Section 2 after introducing the Yang-Mills field. (See Section 2.1.3.) Consider a test particle which carries an internal vector degree of freedom w_i, with $i = 1, \ldots, N$. As the particle moves along a path C, from x_i to x_f, this vector will evolve through parallel transport

$$w(\tau_f) = U[x_i, x_f]w(\tau_i)$$

where the holonomy, or Wilson line, is given by the path ordered exponential

$$U[x_i, x_f] = \mathcal{P} \exp\left(i \int_{x_i}^{x_f} A\right) \tag{4.6}$$

Note that $U[x_i, x_f]$ depends both on the end points, and on the choice of path C.

This is the key idea that we will implement on the lattice. We will not treat the Lie-algebra valued gauge fields A_μ as the fundamental objects. Instead we will work with the group-valued Wilson lines U. These Wilson lines are as much about the journey as the destination: their role is to tell other fields how to evolve. The matter fields live on the sites of the lattice. In contrast, the Wilson lines live on the links.

Specifically, on the link from lattice site x to $x + \hat{\mu}$, we will introduce a dynamical variable

$$\text{link } x \to x + \hat{\mu}: \qquad U_\mu(x) \in G$$

The fact that the fundamental degrees of freedom are group valued, rather than Lie algebra valued, plays an important role in lattice gauge theory. It means, for example that there is an immediate difference between, say, $SU(N)$ and $SU(N)/\mathbf{Z}_N$, a distinction that was rather harder to see in the continuum. We will see other benefits of this below.

At times we will wish to compare our lattice gauge theory with the more familiar continuum action. To do this, we need to re-introduce the A_μ gauge fields. These are related to the lattice degrees of freedom by

$$U_\mu(x) = e^{iaA_\mu(x)} \tag{4.7}$$

The placing of the μ subscripts on the left and right hand side of this equation should make you feel queasy. It looks bad because if one side transforms covariantly under $SO(4)$ rotations, then the other does not. But we don't want these variables to transform under continuous symmetries; only discrete ones. This is the source of your discomfort.

We will wish to identify configurations related by gauge transformations. In the continuum, under a gauge transformation $\Omega(x)$, the Wilson line (4.6) transforms as

$$U[x_i, x_f; C] \to \Omega(x_i)\, U[x_i, x_f; C]\, \Omega^\dagger(x_f)$$

We can directly translate this into our lattice. The link variable transforms as

$$U_\mu(x) \to \Omega(x)\, U_\mu(x)\, \Omega^\dagger(x + \hat{\mu}) \tag{4.8}$$

The next step is to write down an action that is invariant under gauge transformations. We can achieve this by multiplying together a string of neighbouring Wilson lines, and then taking the trace. With no dangling ends, this is guaranteed to be gauge invariant. This is the lattice version of the Wilson loop (2.15) that we met in Section 2.

We can construct a Wilson loop for any closed path C in the lattice. When the path goes from the site x to $x + \hat{\mu}$, we include a factor of $U_\mu(x)$; when the path goes from site x to site $x - \hat{\mu}$, we include a factor of $U_\mu^\dagger(x - \hat{\mu})$. The simplest such path is a square which traverses a single plaquette of the lattice as shown in the figure. The corresponding Wilson loop is

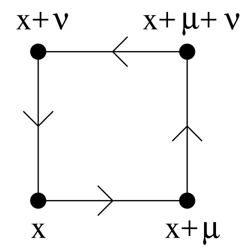


Figure 34:

$$W_\Box = \mathrm{tr}\, U_\mu(x)\, U_\nu(x + \hat{\mu})\, U_\mu^\dagger(x + \hat{\nu})\, U_\nu^\dagger(x)$$

To get some intuition for this object, we can write it in terms of the gauge field (4.7). We will assume that we can Taylor expand the gauge field so that, for example, $A_\nu(x + \hat{\mu}) \approx A_\nu(x) + a\partial_\mu A_\nu(x) + \ldots$. Then we have

$$\begin{aligned} W_\Box &\approx \mathrm{tr}\, e^{iaA_\mu(x)}\, e^{ia(A_\nu(x) + a\partial_\mu A_\nu(x))}\, e^{-ia(A_\mu(x) + a\partial_\nu A_\mu(x))}\, e^{-iaA_\nu(x)} \\ &\approx \mathrm{tr}\, e^{ia(A_\mu(x) + A_\nu(x) + a\partial_\mu A_\nu(x) + \frac{ia}{2}[A_\mu(x), A_\nu(x)])}\, e^{-ia(A_\nu(x) + A_\mu(x) + a\partial_\nu A_\mu(x)) - \frac{ia}{2}[A_\mu(x), A_\nu(x)])} \end{aligned}$$

where, to go to the second line, we've used the BCH formula $e^A e^B = e^{A + B + \frac{1}{2}[A,B] + \ldots}$. On both lines we've thrown away terms of order a^3 in the exponent. Using BCH just once more, we have

$$\begin{aligned} W_\Box &= \mathrm{tr}\, e^{ia^2 F_{\mu\nu}(x) + \ldots} = \mathrm{tr}\, \left(1 + ia^2 F_{\mu\nu} - \frac{a^4}{2} F_{\mu\nu} F_{\mu\nu} + \ldots\right) \\ &= -\frac{a^4}{2} \mathrm{tr}\, F_{\mu\nu} F_{\mu\nu} + \ldots \end{aligned}$$

where, as usual, $F_{\mu\nu}(x) = \partial_\mu A_\nu(x) - \partial_\nu A_\mu(x) - i[A_\mu(x), A_\nu(x)]$ and the $\ldots$ include both a constant term and terms higher in order in a^2. Note that there is no sum over μ, ν in this expression; instead these μ, ν indices tell us of the orientation of the plaquette.

By summing over all possible plaquettes, we get something that reproduces the Yang-Mills action at leading order. The Wilson loop $W_\Box$ itself is not real so we need to add the conjugate $W_\Box^\dagger$, which is the loop with the opposite orientation. This then gives us the *Wilson action*

$$S_{\text{Wilson}} = -\frac{\beta}{2N}\sum_{\Box}\left(W_\Box + W_\Box^\dagger\right) = \frac{a^4\beta}{4N}\int d^4x \ \mathrm{tr}\, F_{\mu\nu}F_{\mu\nu} + \ldots \tag{4.9}$$

where now we are again using summation convention for μ, ν indices. An extra factor of $1/2$ has appeared because the sum over plaquettes differs by a factor of 2 from the sum over μ, ν. It is convention to put a factor of $1/N$ in front of the action. The coupling β is related to the continuum Yang-Mills coupling (2.8) by

$$\frac{\beta}{2N} = \frac{1}{g^2}$$

The Wilson action only coincides with the Yang-Mills action at leading order. Expanding to higher orders in a will give corrections. The next lowest dimension operator to appear is $F_{\mu\nu}\mathcal{D}_\mu^2 F_{\mu\nu}$. It has dimension 6 and does not correspond to an operator that respects continuous $O(4)$ rotational symmetry. In this way, it is analogous to the operator (4.5) that we saw for the scalar field. Happily, it is irrelevant.

The Wilson action is far from unique. For example, we could have chosen to sum over double plaquettes $\Box\!\Box$ as opposed to single plaquettes. Expanding these, or any such Wilson loop, will result in a $F_{\mu\nu}F_{\mu\nu}$ term simply because this is the lowest dimension, gauge invariant operator. These Wilson loops differ in the relative coefficients of the expansion.

For numerical purposes, this lack of uniqueness can be exploited. We could augment the Wilson action with additional terms corresponding to double, or larger, plaquettes. This can be done in such a way that the Yang-Mills action survives, but the higher dimension operators, such as $F_{\mu\nu}\mathcal{D}_\mu^2 F_{\mu\nu}$ cancel. This means that the leading higher derivative terms are even more irrelevant, and helps with numerical convergence. We won't pursue this (or, indeed, any numerics) here.

Adding Dynamical Matter

As we mentioned before, matter fields live on the sites of the lattice. Consider a scalar field $\phi(x)$ transforming in the fundamental representation of the gauge group. (Fermions will come with their own issues, which we discuss in Section 4.3.) Under a gauge transformation we have

$$\phi(x) \rightarrow \Omega(x)\phi(x)$$

We can now construct gauge invariant objects by topping and tailing the Wilson line with particle and anti-particle matter insertions. The simplest example has the particle and anti-particle separated by just one lattice spacing, $\phi^\dagger(x)U_\mu(x)\phi(x+\hat{\mu})$. More generally, we can separate the two as much as we like, as the long as the Wilson line forges a continuous path between them.

To write down a kinetic term for this scalar, we need the covariant version of the finite difference (4.3). This is given by

$$\int d^4x \; |\mathcal{D}_\mu\phi(x)|^2 \;\longrightarrow\; a^2 \sum_{(x,\mu)} \left[2\phi^\dagger(x)\phi(x) - \phi^\dagger(x)U_\mu(x)\phi(x+\hat{\mu}) - \phi^\dagger(x+\hat{\mu})U_\mu^\dagger(x)\phi(x)\right]$$

In this way, it is straightforward to couple scalar matter to gauge fields. We won't have anything more to say about dynamical matter here, but we'll return to the question in Section 4.3 when we discuss fermions on the lattice.

4.2.2 The Haar Measure

To define a quantum field theory, it's not enough to give the action. We also need to specify the measure of the path integral.

Of course, usually in quantum field theory we're fairly lax about this, and the measure certainly isn't defined at the level of rigour that would satisfy a mathematician. The lattice provides us an opportunity to do better, since we have reduced the path integral to a large number of ordinary integrals. For lattice gauge theory, the appropriate measure is something like

$$\prod_{(x,\hat{\mu})} dU_\mu(x) \tag{4.10}$$

so that we integrate over the $U \in G$ degree of freedom on each link. The question is: what does this mean?

Thankfully this is a question that is well understood. We want to define an integration measure over the group manifold G. We will ask that the measure obeys the following requirements:

- Left and right invariance. This means that for any function $f(U)$, with $U \in G$, and for any $\Omega \in G$,

$$\int dU \;\, f(U) = \int dU \;\, f(\Omega U) = \int dU \;\, f(U\Omega) \tag{4.11}$$

This will ensure that our path integral respects the gauge symmetry (4.8). By a change of variables, this is equivalent to the requirement that $d(U\Omega) = d(\Omega U) = dU$ for all $\Omega \in G$.

- Linearity:

$$\int dU\ \left(\alpha f(U) + \beta g(U)\right) = \alpha \int dU\ f(U) + \beta \int dU\ g(U)$$

This is something that we take for granted in integration, and we would very much like to retain it here.

- Normalisation condition:

$$\int dU\ 1 = 1 \tag{4.12}$$

A difference between gauge theory on the lattice and in the continuum is that the dynamical degrees of freedom live in the group G, rather than its Lie algebra. The group manifold is compact, so that $\int dU\ 1$ just gives the volume of G. There's no real meaning to this volume, so we choose to normalise it it to unity.

It turns out that there is a unique measure with these properties. It is known as the *Haar measure*.

We won't need to explicitly construct the Haar measure in what follows, because the properties above are sufficient to calculate what we'll need. Nonetheless, it may be useful to give a sense of where it comes from. We start in a neighbourhood of the identity. Here we can write any $SU(N)$ group element as

$$U = e^{i\alpha^a T^a}$$

with T^a the generators of the $su(N)$ algebra. In this neighbourhood, the Haar measure becomes (up to normalisation)

$$\int dU = \int d^{N^2-1}\alpha\, \sqrt{\det \gamma} \tag{4.13}$$

where γ is the canonical metric on the group manifold,

$$\gamma_{ab} = \operatorname{tr}\left(U^{-1}\frac{\partial U}{\partial \alpha^a} U^{-1}\frac{\partial U}{\partial \alpha^b}\right)$$

This measure is both left and right invariant in the sense of (4.11), since the group action corresponds to shifting $\alpha^a \to \alpha^a + \text{constant}$.

Now suppose that we want to construct the measure in the neighbourhood of any other point, say U_0. We can do this by using the group multiplication to transport the neighbourhood around the identity to a corresponding neighbourhood around U_0. In this way, we can construct the measure over various patches of the group manifold.

One way to transport the measure from one neighbourhood to another is by right multiplication. We write

$$U = e^{i\alpha^a T^a}\, U_0 \tag{4.14}$$

We then again use the definition (4.13) to define the measure. This measure is left invariant, satisfying $dU = d(\Omega U)$ since multiplying U on the left by Ω corresponds to shifting $\alpha^a \to \alpha^a + \text{constant}$. In fact, this is the unique left invariant measure.

But is the measure right invariant? If we multiply U on the right then the group element Ω must make its way past U_0 before we can conclude that it shifts α^a by a constant. But Ω and U_0 do not necessarily commute. Nonetheless, the measure is right invariant. This follows from the fact that we have constructed the unique left invariant measure which means that, if we consider the measure $d(U\Omega)$, which is also left invariant then, by uniqueness, it must be the same as the original. So $d(U\Omega) = dU$.

Integrating over the Group

In what follows, we will need results for some of the simpler integrations.

We start by computing the integral $\int dU\; U$. Because the measure is both left and right invariant, we must have

$$\int dU\; U = \int dU\; \Omega_1 U \Omega_2$$

for any Ω_1 and $\Omega_2 \in G$. But there's only one way to achieve this, which is

$$\int dU\; U = 0 \tag{4.15}$$

More generally, we will only get a non-vanishing answer if we integrate objects which are invariant under G. This will prove to be a powerful constraint, and we'll discuss it further below.

The simplest, non-trivial integral is therefore $\int dU\; U^\dagger_{ij} U_{kl}$, where we've included the gauge group indices $i, j = 1, \ldots, N$. This must be proportional to an invariant tensor, and the only option is

$$\int dU\; U^\dagger_{ij} U_{kl} = \frac{1}{N}\delta_{jk}\delta_{il} \tag{4.16}$$

To see that the $1/N$ factor is correct, we can contract the jk indices and reproduce the normalisation condition (4.12). One further useful integral comes from the baryon vertex, which gives

$$\int dU\; U_{i_1 j_1} \ldots U_{i_N j_N} = \frac{1}{N!}\,\epsilon_{i_1 \ldots i_N}\,\epsilon_{j_1 \ldots j_N}$$

Elitzur's Theorem

Let's now return to our lattice gauge theory. We wish to compute expectation values of operators $\mathcal{O}$ by computing

$$\langle \mathcal{O} \rangle = \frac{1}{Z} \int \prod_{(x,\mu)} dU_\mu(x)\; \mathcal{O}\; e^{-S_{\rm Wilson}}$$

This is simply lots of copies of the group integration defined above. The fact that any object which transforms under G necessarily vanishes when integrated over the group manifold has an important consequence for our gauge theory: it ensures that we have

$$\langle \mathcal{O} \rangle = 0$$

for any operator $\mathcal{O}$ that is not gauge invariant. This is known as *Elitzur's theorem*. Note that this statement has nothing to do with confinement. It is just as valid for electromagnetism as for Yang-Mills, and is a statement about the operators we should be considering in a gauge theory.

Elitzur's theorem follows in a straightforward manner from (4.15). To illustrate the basic idea, we will show how it works for a link variable, $\mathcal{O} = U_\nu(y)$. We want to compute

$$\langle U_\nu(y) \rangle = \frac{1}{Z} \int \prod_{(x,\mu)} dU_\mu(x)\; U_\nu(y)\; e^{-S_{\rm Wilson}}$$

The specific link variable $U_\nu(y)$ will appear in a bunch of different plaquettes that arise in the Wilson action. For example, we could focus on the plaquette Wilson loop

$$W_\Box = {\rm tr}\, U_\nu(y)\, U_\rho(y+\hat{\nu})\, U^\dagger_\nu(y+\hat{\rho})\, U^\dagger_\rho(y) \tag{4.17}$$

But we know that the measure is invariant under group multiplication of any link variable. We can therefore make the change of variable

$$U_\rho(y) \to U_\nu^\dagger(y) U_\rho(y) \tag{4.18}$$

in which case the particular plaquette Wilson loop (4.17) becomes

$$W_\Box \;\to\; \operatorname{tr} U_\rho(y+\hat{\nu})\, U_\nu^\dagger(y+\hat{\rho})\, U_\rho^\dagger(y)$$

and is independent of $U_\nu(y)$. You might think that this will screw up some other plaquette action, where $U_\nu(y)$ will reappear. There are 8 links emanating from the site y, as shown in the disappointingly 3d figure on the right. You can convince yourself that if you make the same change of variables (4.18) for each of them then S_{Wilson} no longer depends on the specific link variable $U_\nu(y)$. We can then isolate the integral over the link variable $U_\nu(y)$, to get

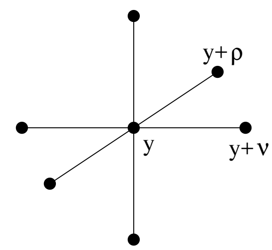


Figure 35:

$$\langle U_\nu(y)\rangle = \text{other stuff} \times \int dU_\nu(y)\; U_\nu(y) = 0$$

which, as shown, vanishes courtesy of (4.15). This tells us that a single link variable cannot play the role of an order parameter in lattice gauge theory. But this is something we expected from our discussion in the continuum.

We see that the Wilson action is rather clever. It's constructed from link variables $U_\nu(y)$, but doesn't actually depend on them individually. Instead, it depends only on gauge invariant quantities that we can construct from the link variables. These are the Wilson loops.

A Comment on Gauge Fixing

The integration measure (4.10) will greatly overcount physical degrees of freedom: it will integrate over many configurations all of which are identified by gauge transformations. What do we do about this? The rather wonderful answer is: nothing at all.

In the continuum, we bend over backwards worrying about gauge fixing. This is because we are integrating over the Lie algebra and will get a divergence unless we fix a gauge. But there is no such divergence in the lattice formulation because we are integrating over the compact group G. Instead, the result of failing to fix the gauge will simply be a harmless normalisation constant.

4.2.3 The Strong Coupling Expansion

We now have all the machinery to define the partition function of lattice gauge theory,

$$Z = \int \prod_{(x,\mu)} dU_\mu(x)\; e^{-S_{\rm Wilson}} \tag{4.19}$$

with $S_{\rm Wilson}$ the sum over plaquette Wilson loops,

$$S_{\rm Wilson} = -\frac{\beta}{2N} \sum_{\Box} \left(W_\Box + W_\Box^\dagger \right) \tag{4.20}$$

Because we're in Euclidean spacetime, the parameter β plays the same role as the inverse temperature in statistical mechanics. It is related to the bare Yang-Mills coupling as $\beta = 2N/g^2$.

We expect this theory to give a good approximation to continuum Yang-Mills when the lattice spacing a is suitably small. Here "small" is relative to the dynamically generated scale Λ_{QCD}. Thinking of $1/a$ as the UV cut-off of the theory, the physical scale is defined by

$$\Lambda_{QCD} = \frac{1}{a} e^{1/2\beta_0 g^2} \tag{4.21}$$

where β_0 is the one-loop beta-function which, despite the unfortunate similarity in their names, has nothing to do with the lattice coupling β that we introduced in the Wilson action. We calculated the one-loop beta function in Section 2.4 and, importantly, $\beta_0 < 0$.

In the expression (4.21), g^2 is the bare gauge coupling. We see that we have a separation of scales between Λ_{QCD} and the cut-off provided our theory is weakly coupled in the UV,

$$g^2 \ll 1 \quad \Leftrightarrow \quad \beta \gg 1$$

In this case, we expect the lattice gauge theory to closely match the continuum. We only have to do some integrals. Lots of integrals. I can't do them. You probably can't either. But a computer can.

We could also ask: what happens in the opposite regime, namely

$$g^2 \gg 1 \quad \Leftrightarrow \quad \beta \ll 1$$

It's not obvious that this regime is of interest. From (4.21), we see that there is no separation between the physical scale, Λ_{QCD}, and the cut-off scale $1/a$, so this is unlikely to give us quantitative insight into continuum Yang-Mills. Nonetheless, it does have one thing going for it: we can actually calculate in this regime! We do this by expanding the partition function (4.19) in powers of β. This is usually referred to as the *strong coupling expansion*; it is analogous to the high temperature expansion in statistical lattice models. (See the lectures on Statistical Physics for more details of how this works in the Ising model.)

Confinement

We'll use the strong coupling expansion to compute the expectation value of a large rectangular Wilson loop, $W[C]$,

$$W[C] = \frac{1}{N}\,\mathrm{tr}\left(\mathcal{P}\prod_{(x,\mu)\in C} U_\mu(x)\right) \tag{4.22}$$

Here the factor of $1/N$ is chosen so that if all the links are $U = 1$ then $W[C] = 1$. We'll place this loop in a plane of the lattice as shown in the figure, and give the sides length L and T. (Each of these must be an integer multiple of a.)

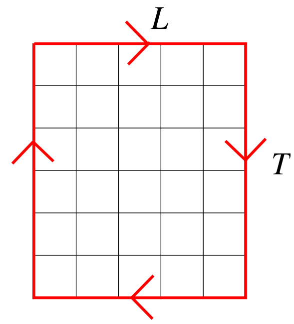


Figure 36:

We would like to calculate

$$\langle W[C]\rangle = \frac{1}{Z}\int \prod_{(x,\mu)} dU_\mu(x)\; W[C]\, e^{-S_{\text{Wilson}}}$$

In the strong coupling expansion, we achieve this by expanding $e^{-S_{\text{Wilson}}}$ in powers of $\beta \ll 1$. What is the first power of β that will give a non-zero answer? If a given link variable U appears in the integrand just once then, as we've seen in (4.15), it will integrate to zero. This means, for example, that the β^0 term in the expansion of $e^{-S_{\text{Wilson}}}$ will not contribute, since it leaves each of the links in $W[C]$ unaccompanied.

The first term in the expansion of $e^{-S_{\text{Wilson}}}$ that will give a non-vanishing answer must contribute a $U^\dagger$ for each link in C. But any $U^\dagger$ that appears in the expansion of S_{Wilson} must be part of a plaquette of links. The further links in these plaquettes must also have companions, and these come from further plaquettes. It is best to think graphically. The links U of the Wilson loop are shown in red. They must be compensated by a corresponding $U^\dagger$ from S_{Wilson} plaquettes; these are shown in blue in the next figure. The simplest way to make sure that no link is left behind is to tile a surface bounded by C by plaquettes. We have shown some of these tiles in the figure. Note that each of the plaquettes $W_\square$ must have a particular orientation to cancel the Wilson loop on the boundary; this orientation then dictates the way further tiles are laid.

There are many different surfaces S that we could use to tile the interior of C. The simplest is the one that lies in the same plane as C and covers each lattice plaquette exactly once. However, there are other surfaces, including those that do not lie in the plane. We can compute the contribution to $\langle W[C]\rangle$ from any given surface S. Only the plaquettes of a specific orientation in the Wilson action (4.20) will contribute (e.g. $W_\Box$, but not $W_\Box^\dagger$). The beta dependence is therefore

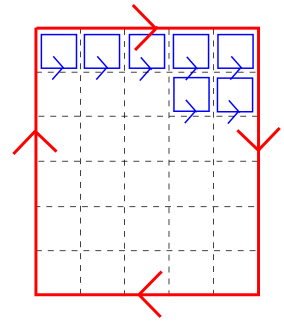


Figure 37:

$$\left(\frac{\beta}{2N}\right)^{\#\text{ of plaquettes}}$$

Each link in the surface (including those in the original C) will give rise to an integral of the form (4.16). This then gives a term of the form

$$\left(\frac{1}{N}\right)^{\#\text{ of links}}$$

Finally, for every site on the surface (including those on the original C), we'll be left with a summation $\delta_{ij}\delta_{ji} = N$. This gives a factor of

$$N^{\#\text{ of sites}}$$

Including the overall factor of $1/N$ in the normalisation of the Wilson loop (4.22), we have the contribution to the Wilson loop

$$\langle W[C]\rangle = \frac{1}{N}\left(\frac{\beta}{2N}\right)^{\#\text{ of plaquettes}}\left(\frac{1}{N}\right)^{\#\text{ of links}} N^{\#\text{ of sites}}$$

where we've used the fact that $Z = 1$ at leading order in β. This is the answer for a general surface. The leading order contribution comes from the minimal, flat surface which bounds C which has

$$\#\text{ of plaquettes} = \frac{LT}{a^2}$$

and

$$\#\text{ vertical links} = \frac{(L+a)T}{a^2} \quad \text{and} \quad \#\text{ horizontal links} = \frac{L(T+a)}{a^2}$$

and

$$\#\text{ sites} = \frac{(L+a)(T+a)}{a^2}$$

The upshot is that the leading order contribution to the Wilson loop is

$$\langle W[C]\rangle = \left(\frac{\beta}{2N^2}\right)^{LT/a^2}$$

But this is exactly what we expect from a confining theory: it is the long sought area law (2.75) for the Wilson loop,

$$\langle W[C]\rangle = e^{-\sigma A}$$

where $A = LT$ is the area of the minimal surface bounded by C and the string tension σ is given by

$$\sigma = -\frac{1}{a^2}\log\left(\frac{\beta}{2N^2}\right)$$

At the next order, this will get corrections of $\mathcal{O}(\beta)$. Note that the string tension is of order the UV cut-off $1/a$, which reminds us that we are not working in a physically interesting regime. Nonetheless we have demonstrated, for the first time, the promised area law of Yang-Mills, the diagnostic for confinement.

A particularly jarring way to illustrate that we're not computing in the continuum limit is to note that the computation above makes no use of the non-Abelian nature of the gauge group. We could repeat everything for Maxwell theory, in which the link variables are $U \in U(1)$. Nothing changes. We again find an area law in the strong coupling regime, indicating the existence of a confining phase.

What are we to make of this? For $U(1)$ gauge theory, there clearly must be a phase transition as we vary the coupling from $\beta \ll 1$ to $\beta \gg 1$ where we have the free, continuum Maxwell theory that we know and love. But what about Yang-Mills? We may hope that there is no phase transition for non-Abelian gauge groups G, so that the confining phase persists for all values of β. It seems that this hope is likely to be dashed. At least as far as the string tension is concerned, it appears that there is a finite radius of convergence around $\beta = 0$, and the string tension exhibits an essential singularity at a finite value of β. It is not known if there is a different path – say by choosing a different lattice action – which avoids this phase transition.

The Mass Gap

We can also look for the existence of a mass gap in the strong coupling expansion. Since we're in Euclidean space, we have neither Hilbert space nor Hamiltonian so we can't talk directly about the spectrum. However, we can look at correlation functions between two far separated objects.

The objects that we have to hand are the Wilson loops. We take two, parallel plaquette Wilson loops $W_\square$ and $W_{\square'}$, separated along a lattice axis by distance R. We expect the correlation function of these Wilson loops to scale as

$$\langle W_\square W_{\square'} \rangle \sim e^{-mR} \tag{4.23}$$

with m the mass of the lightest excitation. If the theory turns out to be gapless, we will instead find power-law decay.

We can compute this correlation function in the strong coupling expansion. The argument is the same as that above: to get a non-zero answer, we must form a tube of plaquettes. The minimum such tube is depicted in the figure, with the source Wilson loops shown in red, and the tiling from the action shown in blue. (This time we have not shown the orientation of the Wilson loops to keep the figure uncluttered.) It has

$$\#\text{ of plaquettes} = \frac{4R}{a}$$

$$\#\text{ links} = \frac{4(2R+a)}{a}$$

$$\#\text{ sites} = \frac{4(R+a)}{a}$$

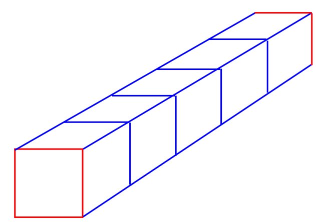


Figure 38:

The leading order contribution to the connected part of the correlation function is therefore

$$\langle W_\square W_{\square'} \rangle = \frac{1}{N^2}\left(\frac{\beta}{2N^2}\right)^{4R/a}$$

Comparing to the expected form (4.23), we see that we have a mass gap

$$m = -\frac{4}{a}\log\left(\frac{\beta}{2N^2}\right)$$

Once again, it's comforting to see the expected behaviour of Yang-Mills. Once again, we see the lack of physical realism highlighted in the fact that the mass scale is the same order of magnitude as the UV cut-off $1/a$.

4.3 Fermions on the Lattice

Finally we turn to fermions. Here things are not so straightforward. The reason is simple: anomalies.

Even before we attempt any calculations, we can anticipate that things might be tricky. Lattice gauge theory is a regulated version of quantum field theory. If we work on a finite, but arbitrarily large lattice, we have a finite number of degrees of freedom. This means that we are back in the realm of quantum mechanics. There is no room for the subtleties associated to the chiral anomaly. There is no infinite availability at the Hilbert hotel.

This means that we're likely to run into trouble if we try to implement chiral symmetry on the lattice or, at the very least, if we attempt to couple gapless fermions to gauge fields. We might expect even more trouble if we attempt to put chiral gauge theories on the lattice. In this section, we will see the form that this trouble takes.

4.3.1 Fermions in Two Dimensions

We can build some intuition for the problems ahead by looking at fermions in $d = 1+1$ dimensions. Here, Dirac spinors are two-component objects. We work with the gamma matrices

$$\gamma^0 = \sigma^1 \quad , \quad \gamma^1 = i\sigma^2 \quad , \quad \gamma^3 = -\gamma^0\gamma^1 = \sigma^3$$

The Dirac fermion then decomposes into chiral fermions $\chi_\pm$ as

$$\psi = \begin{pmatrix} \psi_+ \\ \psi_- \end{pmatrix}$$

In the continuum, the action for a massless fermion is

$$S = \int d^2x \; i\bar{\psi}\,\partial\!\!\!/\psi = \int d^2x \; i\psi_+^\dagger \partial_- \psi_+ + i\psi_-^\dagger \partial_+ \psi_- \tag{4.24}$$

with $\partial_\pm = \partial_t \pm \partial_x$. The equations of motion tell us $\partial_-\psi_+ = \partial_+\psi_- = 0$. This means that ψ_+ is a left-moving fermion, while ψ_- is a right-moving fermion.

As in Section 3.1, it is useful to think in the language of the Dirac sea. The dispersion relation $E(k)$ for fermions in the continuum is drawn in the left hand figure. All states with $E < 0$ are to be thought of as filled; all states with $E > 0$ are empty.

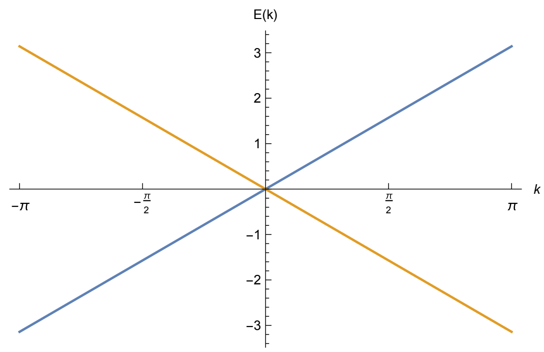


Figure 39: The dispersion relation for a Dirac fermion in the continuum

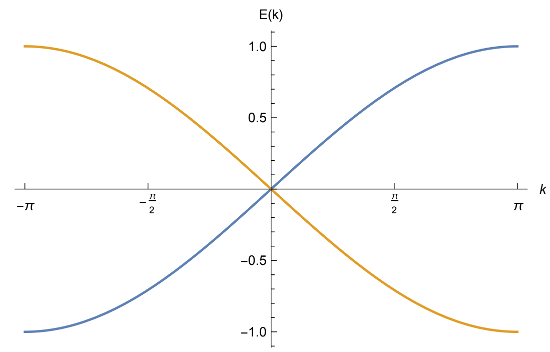


Figure 40: A possible deformation to keep the dispersion periodic in the Brillouin zone (with $a = 1$).

The (blue) line with positive gradient describes the excitations of the right-moving fermion ψ_-: the particles have momentum $k > 0$ while the filled states have momentum $k < 0$ which means that the anti-particles (a.k.a holes) again have momentum $k > 0$. Similarly, the (orange) line with negative gradient describes the excitations of the left-moving fermion ψ_+.

The chiral symmetry of the action (4.24) means that the left- and right-handed fermions are individually conserved. As we have seen Section 3.1, this is no longer the case in the presence of gauge fields. But, for now, we will consider only free fermions so the chiral symmetry remains a good symmetry, albeit one that has a 't Hooft anomaly.

So much for the continuum. What happens if we introduce a lattice? We will start by keeping time continuous, but making space discrete with lattice spacing a. This is familiar from condensed matter physics, and we know what happens: the momentum takes values in the Brillouin zone

$$k \in \left[-\frac{\pi}{a}, \frac{\pi}{a}\right)$$

Importantly, the Brillouin zone is periodic. The momentum $k = +\pi/a$ is identified with the momentum $k = -\pi/a$.

What does this mean for the dispersion relation? We'll look at some concrete models shortly, but first let's entertain a few possibilities. We require that the dispersion relation $E(k)$ remains a continuous, smooth function, but now with $k \in \mathbf{S}^1$ rather than $k \in \mathbf{R}$. This means that the dispersion relation must be deformed in some way.

One obvious possibility is shown in the right hand figure above: we deform the shape of the dispersion relation so that it is horizontal at the boundary of the Brillouin zone

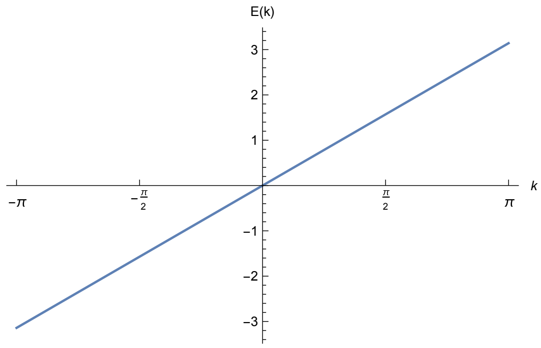


Figure 41: The dispersion relation for a right-handed fermion in the continuum

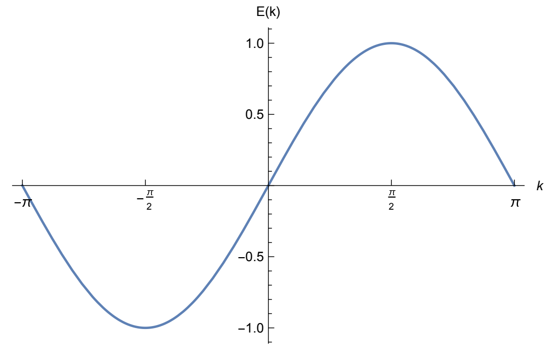


Figure 42: A possible deformation to keep the dispersion periodic in the Brillouin zone (with $a = 1$).

$|k| = \pi/a$. We then identify the states at $k = \pm\pi/a$. Although this seems rather mild, it's done something drastic to the chiral symmetry. If we take, say, a right-moving excitation with $k > 0$ and accelerate it, it will eventually circle the Brillouin zone and come back as a left-moving excitation. This is shown graphically by the fact that the blue line connects to the orange line at the edge of the Brillouin zone. (This is similar to the phenomenon of Bloch oscillations observed in cold atom systems; see the lectures on Applications of Quantum Mechanics.) Said another way, to get such a dispersion relation we must include an interaction term between ψ_+ and ψ_-. This means that, even without introducing gauge fields, there is no separate conservation of left and right-moving particles: we have destroyed the chiral symmetry. Note, however, that we have to excite particles to the maximum energy to see violation of chiral symmetry, so it presumably survives at low energies.

Suppose that we insist that we wish to preserve chiral symmetry. In fact, suppose that we try to be bolder and put just a single right-moving fermion ψ_+ on a lattice. We know that the dispersion relation $E(k)$ crosses the $E = 0$ axis at $k = 0$, with $dE/dk > 0$. But now there's no other line that it can join. The only option is that the dispersion relation also crosses the $E = 0$ at some other point $k \neq 0$, now with $dE/dk < 0$. An example is shown in right hand figure above. Now the lattice has an even more dramatic effect: it generates another low energy excitation, this time a left-mover. We learn that we don't have a theory of a chiral fermion at all: instead we have a theory of two Weyl fermions of opposite chirality. Moreover, once again a right-moving excitation can evolve continuously into a left-moving excitation. This phenomenon is known as *fermion doubling*.

You might think that you can simply ignore the high momentum fermion. And, of course, in a free theory you essentially can. But as soon as we turn on interactions — for example, by adding gauge fields — these new fermions can be pair produced just as easily as the original fermions. This is how the lattice avoids the gauge anomaly: it creates new fermion species!

More generally, it is clear that the Brillouin zone must house as many gapless left-moving fermions as right-moving fermions. This is for a simple reason: what goes up, must come down. This is a precursor to the Nielsen-Ninomiya theorem that we will discuss in Section 4.3.3

Quantising a Chiral Fermion

Let's now see how things play out if we proceed in the obvious fashion. The Hamiltonian for a chiral fermion on a line is

$$H = \pm \int dx \; i\psi_\pm^\dagger \partial_x \psi_\pm$$

The form of the Hamiltonian is the same for both chiralities; only the $\pm$ sign out front determines whether the particle is left- or right-moving. As we will see below, the requirement that the Hamiltonian is positive definite will ultimately translate this sign into a choice of vacuum state above which all excitations move in a particular direction.

For concreteness, we'll work with right-moving fermions ψ_-. We discretise this system in the obvious way: we consider a one-dimensional lattice with sites at $x = na$, where $n \in \mathbf{Z}$, and take the Hamiltonian to be

$$H = -a \sum_{x \in a\mathbf{Z}} i\psi_-^\dagger(x) \left[\frac{\psi_-(x+a) - \psi_-(x-a)}{2a} \right]$$

The Hamiltonian is Hermitian as required. We introduce the usual momentum expansion

$$\psi_-(x) = \int_{-\pi/a}^{+\pi/a} \frac{dk}{2\pi} \; e^{ikx} \, c_k$$

Note that we have momentum modes for both $k > 0$ and $k < 0$, even though this is a purely right-moving fermion. Inserting the mode expansion into the Hamiltonian gives

$$H = \frac{1}{2a} \int_{-\pi/a}^{+\pi/a} \frac{dk}{2\pi} \; 2\sin(ka) \, c_k^\dagger c_k$$

From this we can extract the one-particle dispersion relation by constructing the state $|k\rangle = c_k^\dagger|0\rangle$, to find the energy $H|k\rangle = E(k)|k\rangle$, with

$$E(k) = \frac{1}{a}\sin(ka)$$

This gives a dispersion relation of the kind we anticipated above: it has zeros at both $k = 0$ and at the edge of the Brillouin zone $k = \pi/a$. As promised, we started with a right-moving fermion but the lattice has birthed a left-moving partner.

Finally, a quick comment on the existence of states with $k < 0$. The true vacuum is not $|0\rangle$, but rather $|\Omega\rangle$ which has all states with $E < 0$ filled. This is the Dirac sea or, Fermi sea since the number of such states are finite. This vacuum obeys $c_k|\Omega\rangle = 0$ for $k > 0$ and $c_k^\dagger|\Omega\rangle = 0$ for $k < 0$. In this way, $c_k^\dagger$ creates a right-moving particle when $k > 0$, and c_k creates a right-moving anti-particle with momentum $|k|$ when $k < 0$.

4.3.2 Fermions in Four Dimensions

A very similar story plays out in $d = 3 + 1$ dimensions. A Weyl fermion $\psi_\pm$ is a 2-component complex spinor and obeys the equation of motion

$$\partial_0\psi_\pm = \pm\sigma^i\partial_i\psi_\pm$$

The Hamiltonian for a single Weyl fermion takes the form

$$H = \pm\int d^3x\ i\psi_\pm^\dagger\sigma^i\partial_i\psi_\pm$$

Once again, we wish to write down a discrete version of this Hamiltonian on a cubic spatial lattice Γ. For concreteness, we'll work with ψ_-. We take the Hamiltonian to be

$$H = -a^3\sum_{\mathbf{x}\in\Gamma} i\psi_-(\mathbf{x})\sum_{i=1,2,3}\sigma^i\left[\frac{\psi_-(\mathbf{x}+a\hat{\mathbf{i}}) - \psi_-(\mathbf{x}-a\hat{\mathbf{i}})}{2a}\right]$$

where $i = 1, 2, 3$ labels the spatial directions. In momentum space, the spinor is

$$\psi_-(\mathbf{x}) = \int_{BZ}\frac{d^3k}{(2\pi)^3}\ e^{i\mathbf{k}\cdot\mathbf{x}}\,c_{\mathbf{k}}$$

where $c_{\mathbf{k}}$ is again a two-component spinor. Here the momentum is integrated over the Brillouin zone

$$k_i \in \left[-\frac{\pi}{a}, \frac{\pi}{a}\right) \qquad i = 1, 2, 3$$

The Hamiltonian now takes the form

$$H = \frac{1}{2a}\int_{BZ}\frac{d^3k}{(2\pi)^3}\sum_{i=1,2,3} 2\sin(k_i a)\, c^\dagger_{\mathbf{k}}\,\sigma^i\, c_{\mathbf{k}} \tag{4.25}$$

If we focus on single particle excitations, the spectrum now has two bands, corresponding to a particle and anti-particle, and is given by

$$E(\mathbf{k}) = \frac{1}{a}\sum_{i=1,2,3}\sin(k_i a)\,\sigma^i$$

Close to the origin, $k \ll 1/a$, the Hamiltonian looks like that of the continuum fermion, with dispersion

$$E(\mathbf{k}) \approx \mathbf{k}\cdot\boldsymbol{\sigma} \tag{4.26}$$

This is referred to as the *Dirac cone*; it is sketched in the figure. Note that the bands cross precisely at $E = 0$ which, in a relativistic theory, plays the role of the Fermi energy. If the dispersion relation were to cross anywhere else, we would have a Fermi surface.

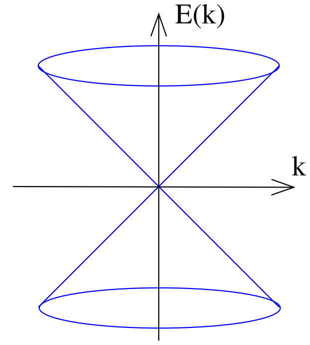


Figure 43:

The fact that the Dirac cone corresponds to a right-handed fermion ψ_- shows up only in the overall + sign of the Hamiltonian. A left-handed fermion would have a minus sign in front. In fact, our full lattice Hamiltonian (4.25) has both right- and left-handed fermions since, like the $d = 1+1$ example above, it exhibits fermion doubling. There are gapless modes at momentum

$$k_i = 0 \text{ or } \frac{\pi}{a}$$

This gives $2^3 = 8$ gapless fermions in total. If we expand the dispersion relation around, say $\mathbf{k}_1 = (\pi/a, 0, 0)$, it looks like

$$E(\mathbf{k}') \approx -\mathbf{k}'\cdot\boldsymbol{\sigma} \quad \text{ where } \mathbf{k}' = \mathbf{k}-\mathbf{k}_1$$

which is left-handed. Of the 8 gapless modes, you can check that 4 are right-handed and 4 are left-handed. We see that, once again, the lattice has generated new gapless modes. Anything to avoid that anomaly.

4.3.3 The Nielsen-Ninomiya Theorem

We saw above that a naive attempt to quantise a $d = 3+1$ chiral fermion gives equal numbers of left and right-handed fermions in the Brillouin zone. The Nielsen-Ninomiya theorem is the statement that, given certain assumptions, this is always going to be the case. It is the higher dimensional version of "what goes up must come down".

The Nielsen-Ninomiya theorem applies to free fermions. We will work in terms of the one-particle dispersion relation, rather than the many-body Hamiltonian. To begin with, we consider a dispersion relation for a single Weyl fermion (we will generalise shortly). In momentum space, the most general Hamiltonian is given by

$$H = v_i(\mathbf{k})\sigma^i + \epsilon(\mathbf{k})\,\mathbf{1}_2 \tag{4.27}$$

where $\mathbf{k}$ takes values in the Brillouin zone.

In the language of condensed matter physics, this Hamiltonian has two bands, corresponding to the fact that each term is a 2×2 matrix. The first question that we will ask is: when do the two bands touch? This occurs when each $v_i(\mathbf{k}) = 0$ for $i = 1,2,3$. This is three conditions, and so we expect to generically find solutions at points, rather than lines, in the Brillouin zone $\text{BZ} \subset \mathbf{R}^3$. Let us suppose that there are D such points, which we call $\mathbf{k}_\alpha$,

$$v_i(\mathbf{k}_\alpha) = 0 \quad , \quad \alpha = 1,\ldots,D$$

Expanding about any such point, the dispersion relation becomes

$$H \approx v_{ij}(\mathbf{k}_\alpha)\,(\mathbf{k}-\mathbf{k}_\alpha)^j\sigma^i \qquad \text{with} \quad v_{ij} = \frac{\partial v_i}{\partial k^j}$$

This now takes a similar form to (4.26), but with an anisotropic dispersion relation. The chirality of the fermion is dictated by

$$\text{chirality} = \text{sign}\det v_{ij}(\mathbf{k}_\alpha) \tag{4.28}$$

The assumption that the band crossing occurs only at points means that $\det v_{ij}(\mathbf{k}_\alpha) \neq 0$. The Nielsen-Ninomiya theorem is the statement that, for any dispersion (4.27) in a Brillouin zone, there are equal numbers of left- and right-handed fermions.

We offer two proofs of this statement. The first follows from some simple topological considerations. For $\mathbf{k} \neq \mathbf{k}_\alpha$, we can define a unit vector

$$\hat{\mathbf{v}}(\mathbf{k}) = \frac{\mathbf{v}}{|\mathbf{v}|}$$

The key idea is that this unit vector can wind around each of the degenerate points $\mathbf{k}_\alpha$. To see this, surround each such point with a sphere $\mathbf{S}^2_\alpha$. Evaluated on these spheres, $\hat{\mathbf{v}}$ provides a map

$$\hat{\mathbf{v}} : \mathbf{S}^2_\alpha \mapsto \mathbf{S}^2$$

But we know that such maps are characterised by $\Pi_2(\mathbf{S}^2) = \mathbf{Z}$. Generically, this winding will take values ± 1 only. In non-generic cases, where we have, say, winding $+2$, we can perturb the $\mathbf{v}$ slightly and the offending degenerate point will split into two points each with winding $+1$. This is the situation we will deal with.

This winding $\{+1, -1\} \subset \Pi_2(\mathbf{S}^2)$ is precisely the chirality (4.28). One quick argument for this is that the a spatial inversion will flip both the winding and the sign of the determinant.

To finish the argument, we need to show that the total winding must vanish. This follows from the compactness of the Brillouin zone. Here are some words. We could consider a sphere $\mathbf{S}^2_{\text{bigger}}$ which encompasses more and more degenerate points. The winding around this sphere is equal to the sum of the windings of the $\mathbf{S}^2_\alpha$ which sit inside it. By the time we get to a sphere $\mathbf{S}^2_{\text{biggest}}$ which encompasses all the points, we can use the compactness of the Brillouin zone to contract the sphere back onto itself on the other side. The winding around this sphere must, therefore, vanish.

Here are some corresponding equations. The winding number ν_α is given by

$$\nu_\alpha = \frac{1}{8\pi} \int_{\mathbf{S}^2_\alpha} d^2 S_i \; \epsilon^{ijk} \epsilon^{abc} \hat{v}^a \frac{\partial \hat{v}^b}{\partial k^j} \frac{\partial \hat{v}^c}{\partial k^k} = \pm 1$$

We saw this expression previously in (2.89) when discussing 't Hooft-Polyakov monopoles. Let us define BZ$'$ as the Brillouin zone with the balls inside $\mathbf{S}^2_\alpha$ excised. This means that the boundary of BZ$'$ is

$$\partial(\text{BZ}') = \sum_{\alpha=1}^{D} \mathbf{S}^2_\alpha$$

Note that this is where we've used the compactness of the Brillouin zone: there is no contribution to the boundary from infinity. We can then use Stokes' theorem to write

$$\sum_{\alpha=1}^{D} \nu_\alpha = \frac{1}{8\pi} \int_{\text{BZ}'} d^3 k \; \frac{\partial}{\partial k^i} \left(\epsilon^{ijk} \epsilon^{abc} \hat{v}^a \frac{\partial \hat{v}^b}{\partial k^j} \frac{\partial \hat{v}^c}{\partial k^k} \right)$$

But the bulk integrand is strictly zero,

$$\frac{\partial}{\partial k^i}\left(\epsilon^{ijk}\epsilon^{abc}\hat{v}^a\frac{\partial \hat{v}^b}{\partial k^j}\frac{\partial \hat{v}^c}{\partial k^k}\right) = \epsilon^{ijk}\epsilon^{abc}\frac{\partial \hat{v}^a}{\partial k^i}\frac{\partial \hat{v}^b}{\partial k^j}\frac{\partial \hat{v}^c}{\partial k^k} = 0$$

because each of the three vectors $\partial\hat{v}^a/\partial k^i$, $i = 1, 2, 3$ is orthogonal to $\hat{v}^a$ and so all three must lie in the same plane. This tells us that

$$\sum_{\alpha=1}^{D}\nu_\alpha = 0$$

as promised.

Note that the Nielsen-Ninomiya theorem counts only the points of degeneracy in the dispersion relation (4.27): it makes no comment about the energy $\epsilon(\mathbf{k}_\alpha)$ of these points. To get relativistic physics in the continuum, we require that $\epsilon(\mathbf{k}_\alpha) = 0$. This ensures that the bands cross precisely at the top of the Dirac sea, and there is no Fermi surface. This isn't as finely tuned as it appears and arises naturally if there is one electron per unit cell; we saw an example of this phenomenon in the lectures on Topics in Quantum Mechanics when we discussed graphene.

Another Proof of Nielsen-Ninomiya: Berry Phase

There is another viewpoint on the Nielsen-Ninomiya theorem that is useful. This places the focus on the Hilbert space of states, rather than the dispersion relation itself[9].

For each $\mathbf{k} \in$ BZ, there are two states. As long as $\mathbf{k} \neq \mathbf{k}_\alpha$, these have different energies. In the language of the Dirac sea, the one with lower energy is filled and the one with higher energy is empty. We focus on the lower energy, filled states which we refer to as $|\psi(\mathbf{k})\rangle$, $\mathbf{k} \neq \mathbf{k}_\alpha$. The Berry connection is a natural $U(1)$ connection on these filled states, which tells us how to relate their phases for different values of $\mathbf{k}$,

$$\mathcal{A}_i(\mathbf{k}) = -i\langle\psi(\mathbf{k})|\,\frac{\partial}{\partial k^i}\,|\psi(\mathbf{k})\rangle$$

You can find a detailed discussion of Berry phase in both the lectures on Topics in Quantum Mechanics and the lectures on Quantum Hall Effect. From the Berry phase, we can define the Berry curvature

$$\mathcal{F}_{ij} = \frac{\partial \mathcal{A}_j}{\partial k^i} - \frac{\partial \mathcal{A}_i}{\partial k^j}$$

[9]This is closely related to the Nobel winning TKKN formula that we discussed the lectures on the Quantum Hall Effect.

The Berry curvature for the dispersion relation (4.27) is the simplest example that we met when we first came across the Berry phase and is discussed in detail in both previous lectures. The chirality of the gapless fermion can now be expressed in terms of the curvature $\mathcal{F}$, which has the property that, when integrated around any degenerate point $\mathbf{k}_\alpha$,

$$\nu_\alpha = \frac{1}{2\pi}\int_{\mathbf{S}^2_\alpha} \mathcal{F} = \pm 1$$

Now we complete the argument in the same way as before. We have

$$\frac{1}{2\pi}\int_{\mathrm{BZ}'} d\mathcal{F} = \frac{1}{2\pi}\sum_{\alpha=1}^{D}\int_{\mathbf{S}^2_\alpha}\mathcal{F} = \sum_{\alpha=1}^{D}\nu_\alpha = 0$$

Again, we learn that there are equal numbers of left- and right-handed fermions.

We can extend this proof to systems with multiple bands. Suppose that we have a system with q bands, of which p are filled. This state of affairs persists apart from at points $\mathbf{k}_\alpha$ where the p^{th} band intersects the $(p+1)^{\mathrm{th}}$. Away from these points, we denote the filled states as $|\psi_a(\mathbf{k})\rangle$ with $a = 1,\ldots p$. These states then define a $U(p)$ Berry connection

$$(\mathcal{A}_i)_{ba} = -i\langle\psi_a|\,\frac{\partial}{\partial k^i}\,|\psi_b\rangle$$

and the associated $U(p)$ field strength

$$(\mathcal{F}_{ij})_{ab} = \frac{\partial(\mathcal{A}_j)_{ab}}{\partial k^i} - \frac{\partial(\mathcal{A}_i)_{ab}}{\partial k^j} - i[\mathcal{A}_i,\mathcal{A}_j]_{ab}$$

This time the winding is

$$\nu_\alpha = \frac{1}{2\pi}\int_{\mathbf{S}^2_\alpha}\mathrm{tr}\,\mathcal{F}$$

The same argument as above tells us that, again, $\sum_\alpha \nu_\alpha = 0$.

4.3.4 Approaches to Lattice QCD

So far our discussion of fermions has been in the Hamiltonian formulation, where time remains continuous. The issues that we met above do not disappear when we consider discrete, Euclidean spacetime. For example, the action for a single massless Dirac fermion is

$$S = \int d^4x\; i\bar{\psi}\gamma^\mu\partial_\mu\psi$$

The obvious discrete generalisation is

$$S = a^4 \sum_{x \in \Gamma} i\bar{\psi}(x) \sum_\mu \gamma^\mu \left[\frac{\psi(x+a\hat{\mu}) - \psi(x-a\hat{\mu})}{2a}\right] \tag{4.29}$$

Working in momentum space, this becomes

$$S = \int_{BZ} \frac{d^4k}{(2\pi)^4}\ \bar{\psi}_{-k}\, D(k)\, \psi_k \tag{4.30}$$

with the inverse propagator

$$D(k) = \frac{1}{a}\sum_\mu \gamma^\mu \sin(k_\mu a) \tag{4.31}$$

We again see the fermion doubling problem, now in the guise of poles in the propagator $D^{-1}(k)$ at $k_\mu = 0$ and $k_\mu = \pi/a$. Since we have also discretised time, the problem has become twice as bad: there are now $2^4 = 16$ poles.

The Nielsen-Ninomiya theorem that we met earlier has a direct translation in this context. It states that it is not possible to write down a $D(k)$ in (4.30) that obeys the following four conditions,

- $D(k)$ is continuous within the Brillouin zone. This means, in particular, that it is periodic in k.
- $D(k) \approx \gamma^\mu k_\mu$ when $k \ll 1/a$, so that the theory looks like a massless Dirac fermion when the momentum is small.
- $D(k)$ has poles only at $k = 0$. This is the requirement that there are no fermion doublers. As we've seen, this requirement doesn't hold if we follow the naive discretization (4.31).
- $\{\gamma^5, D(k)\} = 0$. This is the statement that the theory preserves chiral symmetry. It is true for our naive approach (4.31), but this suffered from fermionic doublers. As we will see below, if we try to remove these we necessarily screw with chiral symmetry. Indeed, we saw a very similar story in Section 4.3.1 when we discussed fermions in $d = 1 + 1$ dimensions.

What to make of this? Clearly, we're not going to be able to simulate chiral gauge theories using these methods. But what about QCD? This is a non-chiral theory that involves only Dirac fermions. Even here, we have some difficulty because if we try to remove the doublers to get the right number of degrees of freedom, then we are going to break chiral symmetry explicitly. Of course, ultimately chiral symmetry will be broken by the anomaly anyway, but there's interesting physics in that anomaly and that's going to be hard to see if we've killed chiral symmetry from the outset.

What to do? Here are some possible approaches. We will discuss a more innovative approach in the following section.

SLAC Fermions

We're going to have to violate one of the requirements of the Nielsen-Ninomiya theorem. One possibility is to give up on periodicity in the Brillouin zone. Now what goes up need not necessarily come down. We make the dispersion relation discontinuous at some high momentum. For example, you could just set $D(\mathbf{k}) = \gamma^\mu k_\mu$ everywhere, and suffer the discontinuity at the edge of the Brillouin zone. This, it turns out, is bad. A discontinuity in momentum space corresponds to a breakdown of locality in real space. The resulting theories are not local quantum field theories. They do not behave in a nice manner.

Wilson Fermions

As we mentioned above, another possibility is to kill the doublers, at the expense of breaking chiral symmetry. One way to implement this, first suggested by Wilson, is to add to the original action (4.30) the term

$$S = ar \int d^4x \; \bar{\psi}\, \partial^2 \psi = ar \sum_{x\in\Gamma} a^4 \bar{\psi}(x) \sum_\mu \left[\frac{\psi(x + a\hat{\mu}) - 2\psi(x) + \psi(x - a\hat{\mu})}{a^2} \right]$$

In momentum space, this becomes

$$S = -\frac{4r}{a} \int_{BZ} \frac{d^4k}{(2\pi)^4} \; \bar{\psi}_{-k} \sin^2\left(\frac{k_\mu a}{2}\right) \psi_k$$

and we're left with the inverse propagator

$$D(k) = \frac{1}{a}\gamma^\mu \sin(k_\mu a) - \frac{4r}{a} \sin^2\left(\frac{k_\mu a}{2}\right) \tag{4.32}$$

where there is also an implicit sum over μ in the second term. This now satisfies the first three of the four requirements above, with all the spurious fermions at $k_\mu = \pi/a$ lifted. The resulting dispersion relation is analogous to what we saw in $d = 1 + 1$ dimensions. The down side is that we have explicitly broken chiral symmetry, which can be seen by the lack of gamma matrices in the second term above. This becomes problematic when we consider interacting fermions, in particular when we introduce gauge fields. Under RG, we no longer enjoy the protection of chiral symmetry and expect to generate any terms which were previously prohibited, such as mass terms $\bar{\psi}\psi$ and dimension 5 operators $\bar{\psi}\gamma^\mu\gamma^\nu F_{\mu\nu}\psi$. Each of these must be fine tuned away, just like the mass of the scalar in Section 4.1.

Staggered Fermions

The final approach is to embrace the fermion doublers. In fact, as we will see, we don't need to embrace all 16 of them; only 4.

To see this, we need to return to the real space formalism. At each lattice site, we have a 4-component Dirac spinor $\psi(x)$. We denote the position of the lattice site as $x = a(n_1, n_2, n_3, n_4)$, with $n_\mu \in \mathbf{Z}$. We then introduce a new Dirac spinor $\chi(x)$, defined by

$$\psi(x) = \gamma_1^{n_1}\gamma_2^{n_2}\gamma_3^{n_3}\gamma_4^{n_4}\,\chi(x) \tag{4.33}$$

In the action (4.29), we have $\bar{\psi}(x)\,\gamma^\mu\,\psi(x \pm a\hat{\mu})$. Written in the χ variable, the term $\gamma^\mu\,\psi(x \pm a\hat{\mu})$ will have two extra powers of γ^μ compared to $\bar{\psi}(x)$; one from the explicit γ^μ out front, and the other coming from the definition (4.33). Since we have $(\gamma^\mu)^2 = +1$ in Euclidean space, we will find

$$\gamma^\mu\,\psi(x \pm a\hat{\mu}) = (-1)^{\text{some integer}}\gamma_1^{n_1}\gamma_2^{n_2}\gamma_3^{n_3}\gamma_4^{n_4}\chi(x \pm a\hat{\mu})$$

where the integer is determined by commuting various gamma matrices past each other. But this means that the integrand of the action has terms of the form

$$\bar{\psi}(x)\,\gamma^\mu\psi(x + a\hat{\mu}) = \eta_{x,\mu}\bar{\chi}(x)\chi(x + a\hat{\mu})$$

where there's been some more commuting and annihilating of gamma matrices going on, resulting in the signs

$$\eta_{x,1} = 1 \quad , \quad \eta_{x,2} = (-1)^{n_1} \quad , \quad \eta_{x,3} = (-1)^{n_1+n_2} \quad , \quad \eta_{x,4} = (-1)^{n_1+n_2+n_3}$$

The upshot is that the transformation (4.33) has diagonalised the action in spinor space. One can check that this same transformation goes through unscathed if we couple the fermion to gauge fields. This means that, on the lattice, we have

$$\det(i\not{D}) = \det{}^4(\hat{D})$$

for some operator $\hat{D}$. The operator $\hat{D}$ still includes contributions from the 16 fermions dotted around the Brillouin zone, but only one spinor index contribution from each. We may then take the fourth root and consider $\det(\hat{D})$ by itself. Perhaps surprisingly, one still finds a relativistic theory in the infra-red, with 4 of the 16 doublers providing the necessary spinor degrees of freedom.

Roughly speaking, you can think of the staggered fermions as arising from placing just a single degree of freedom on each lattice site. After doubling, we have 16 degrees of freedom living at the origin of momentum space and the corners of the Brillouin zone. The staggering trick is to recombine these 16 degrees of freedom back into 4 Dirac spinors. The idea that some subset of the fermion doublers may play the role of spin sounds strange at first glance, but is realised in $d = 2+1$ dimensions in graphene.

This staggered approach still leaves us with $16/4 = 4$ Dirac fermions. At high energy, these are coupled in a way which is distinct from four flavours in QCD. Nonetheless, it is thought that, when coupled to gauge fields, the continuum limit coincides with QCD with four flavours which, in this context, are referred to as *tastes*. The lattice theory has a $U(1) \times U(1)$ chiral symmetry, less than the $U(4) \times U(4)$ chiral symmetry of the (classical) continuum but still sufficient to prevent the generation of masses. This is a practical advantage of staggered fermions.

In fact, there are further reasons to be nervous about staggered fermions. As we've seen, the continuum limit results in 4 Dirac fermions. Let's call them $\psi_{\alpha i}$, where $\alpha = 1,2,3,4$ is the spinor index and $i = 1,2,3,4$ is the taste (flavour) index. However, these spinor and tase indices appear on the same footing in the lattice: both come from doubling. This suggests that they will sit on the same footing in the continuum limit. But that's rather odd. It means that, upon a Lorentz transformation Λ, the resulting Dirac spinors will transform as

$$\psi_{\alpha i} \rightarrow S[\Lambda]_\alpha^{\ \beta} S[\Lambda]_i^{\ j} \psi_{\beta j}$$

with $S[\Lambda]$ the spinor representation of the Lorentz transformation. (Since we're in Euclidean space, it is strictly speaking just the rotation group $SO(4)$.) The first term $S[\Lambda]_\alpha^{\ \beta}$ is the transformation property that we would expect of a spinor, but the second term $S[\Lambda]_i^{\ j}$ is very odd, since these are flavour indices. In particular, it means that if we rotate by 2π, we never see the famous minus sign acting on the staggered fermions. Instead we get two minus signs, one acting on the two indices, and the resulting object actually has integer spin!

What's going on here is that the object $\psi_{\alpha i}$ is really a bi-spinor, in the sense that both α and i are spinor indices. In representation theory language, a Dirac spinor transforms as $(\frac{1}{2}, 0) \oplus (0, \frac{1}{2})$. The staggered fermions then transform in

$$[(\tfrac{1}{2}, 0) \oplus (0, \tfrac{1}{2})] \otimes [(\tfrac{1}{2}, 0) \oplus (0, \tfrac{1}{2})] = 2(0,0) \oplus 2(\tfrac{1}{2}, \tfrac{1}{2}) \oplus (1,0) \oplus (0,1)$$

Here $(0,0)$ are scalars, $(\frac{1}{2}, \frac{1}{2})$ is the vector representation, while $(1,0)$ and $(0,1)$ are the self-dual and anti-self-dual representations of 2-forms. In fact, formally, the collection of objects on the right can be written as a sum of forms of different degrees,

$$\Phi = \phi^{(0)} + \phi^{(1)}_\mu dx^\mu + \phi^{(2)}_{\mu\nu} dx^\mu \wedge dx^\nu + \phi^{(3)}_{\mu\nu\rho} dx^\mu \wedge dx^\nu \wedge dx^\rho + \phi^{(4)}_{\mu\nu\rho\sigma} dx^\mu \wedge dx^\nu \wedge dx^\rho \wedge dx^\sigma$$

where Poincaré duality means that the 4-form has the same degrees of freedom as a scalar and the 3-form the same as a vector. The 16 degrees of freedom that sit in the staggered fermions $\psi_{\alpha i}$ can then be rearranged to sit in Φ. Moreover, the Dirac equation on ψ has a nice description in terms of these forms; it becomes

$$(d - *d * + m)\Phi = 0$$

This is sometimes called a Dirac-Kähler field.

The upshot is that staggered fermions don't quite give rise to Dirac fermions, but a slightly more exotic object constructed in terms of forms. Nonetheless, this doesn't stop people using them in an attempt to simulate QCD, largely because of the numerical advantage that they bring. Given the discussion above, one might be concerned that this is not quite a legal thing to do and it is, in fact, simulating a different theory.

This is not the only difficulty with staggered fermions. The four tastes necessarily have the same mass meaning that, the problems above notwithstanding, staggered fermions do not allow us to get close to a realistic QCD theory, where the masses of the four lightest quarks are very different. To evade this issue, one sometimes attempts to simulate a single quark by taking yet another fourth-root, $\det{}^{1/4}(\hat{D})$. It seems clear that this does not result in a local quantum field theory. Arguments have raged about how evil this procedure really is.

4.4 Towards Chiral Fermions on the Lattice

A wise man once said that, when deciding what to work on, you should first evaluate the importance of the problem and then divide by the number of people who are already working on it. By this criterion, the problem of putting chiral fermions on the lattice ranks highly. There is currently no fully satisfactory way of evading the Nielsen-Ninomiya theorem. This means that there is no way to put the Standard Model on a lattice.

On a practical level, this is not a particularly pressing problem. It is the weak sector of the Standard Model which is chiral, and here perturbative methods work perfectly well. In contrast, the strong coupling sector of QCD is a vector-like theory and this is where most effort on the lattice has gone. However, on a philosophical level, the lack of lattice regularisation is rather disturbing. People will bang on endlessly about whether or not we live in "the matrix'", seemingly unaware that there are serious obstacles to writing down a discrete version of the known laws of physics, obstacles which, to date, no one has overcome.

In this section, I will sketch some of the most promising ideas for how to put chiral fermions on a lattice. None of them quite works out in full – yet – but may well do in the future.

4.4.1 Domain Wall Fermions

Our first approach has its roots in the continuum, which allows us to explain much of the basic idea without invoking the lattice. We start by working in $d = 4+1$ dimensions. The fifth dimension will be singled out in what follows, and we refer to it as $x^5 = y$.

In $d = 4+1$, the Dirac fermion has four components. The novelty is that we endow the fermion with a spatially dependent mass, $m(y)$

$$i\partial\!\!\!/\psi + i\gamma^5\partial_y\psi - m(y)\psi = 0 \tag{4.34}$$

where we pick the boundary conditions

$$m(y) \rightarrow \pm M \quad \text{as } y \rightarrow \pm\infty$$

with $M > 0$. We will take the profile $m(y)$ to be monotonic, with $m(y) = 0$ only at $y = 0$. A typical form of the mass profile is shown in the figure. Profiles of this kind often arise when we solve equations which interpolate between two degenerate vacua. In that context, they are referred to as *domain walls* and we'll keep the same terminology, even though we have chosen $m(y)$ by hand.

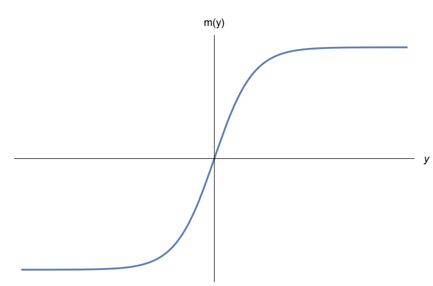


Figure 44:

The fermion excitation spectrum includes a continuum of scattering states with energies $E \geq M$ which can exist asymptotically in the y direction. At these energies, physics is very much five dimensional. But there are also states with $E < M$ which are bound to the wall. If we restrict to these energies then physics is essentially four dimensional. In this sense, the mass M can be thought of as an unconventional cut-off for the four dimensional theory on the wall.

In the chiral basis of gamma matrices,

$$\gamma^0 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad , \quad \gamma^i = \begin{pmatrix} 0 & \sigma^i \\ -\sigma^i & 0 \end{pmatrix}, \ i = 1,2,3 \quad , \quad \gamma^5 = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}$$

where the factors of i in γ^5 reflect the fact that we're working in signature $(+,-,-,-,-)$. The Dirac equation becomes

$$\begin{aligned} i\partial_0\psi_- + i\sigma^i\partial_i\psi_- - \partial_5\psi_+ &= m(y)\psi_+ \\ i\partial_0\psi_+ - i\sigma^i\partial_i\psi_+ + \partial_5\psi_- &= m(y)\psi_- \end{aligned}$$

where $\psi = (\psi_+, \psi_-)^T$. There is one rather special solution to these equations,

$$\psi_+(x,y) = \exp\left(-\int^y dy'\ m(y')\right)\chi_+(x) \quad \text{and} \quad \psi_-(x,y) = 0$$

The profile is supported only in the vicinity of the domain wall; it dies off exponentially $\sim e^{-M|y|}$ as $y \to \pm\infty$. Importantly, there is no corresponding solution for ψ_-, since the profile must be of the form $\exp\left(+\int dy' m(y')\right)$ which now diverges exponentially in both directions.

The two-component spinor $\chi_+(x)$ obeys the equation for a right-handed Weyl fermion,

$$\partial_0\chi_+ - \sigma^i\partial_i\chi_+ = 0$$

We see that we can naturally localise chiral fermions on domain walls. The existence of this mode, known as a *fermion zero mode*, does not depend on any of the detailed properties of $m(y)$. We met a similar object in Section 3.3.4 when discussing the topological insulator.

This is interesting. Our original 5d theory had no hint of any chiral symmetry. But, at low-energies, we find an emergent chiral fermion and an emergent chiral symmetry.

Implications for the Lattice

So far, our discussion in this section has taken place in the continuum. How does it help us in our quest to put chiral fermions on the lattice?

The idea to apply domain wall fermions to lattice gauge theory is due to Kaplan. At first sight, this doesn't seem to buy us very much: a straightforward discretisation of the Dirac equation (4.34) shows that the domain wall does nothing to get rid of the doublers: in Euclidean space there are now 2^4 right-handed fermions χ_+, with the new modes sitting at the corners of the Brillouin zone as usual. Moreover, on the lattice one also finds a further 2^4 left-moving fermions χ_-. This brings us right back to a vector-like theory, with 2^4 Dirac fermions.

However, the outlook is brighter when we add a 5d Wilson term (4.32) to the problem. By tuning the coefficient to lie within a certain range, we can not only remove all of the 16 left-handed fermions χ_-, but we can remove 15 of the 16 right-handed fermions. This leaves us with just a single right-handed Dirac fermion localised on the domain wall.

It is surprising that the Wilson term (4.32) can remove an odd number of gapless fermions from the spectrum since everything we learned up until now suggests that gapless modes can only be removed in pairs. But we have something new here, which is the existence of the infinite fifth dimension. This gives a novel mechanism by which zero modes can disappear: they can become non-normalisable.

There is an alternative way to view this. Suppose that we make the fifth direction compact. Then the domain wall must be accompanied by an anti-domain wall that sits at some distance L. While the domain wall houses a right-handed zero mode, the anti-domain wall has a left-handed zero mode. Now Nielsen-Ninomiya is obeyed, but the two fermions are sequestered on their respective walls, with any chiral symmetry breaking interaction suppressed by $e^{-L/a}$.

I will not present that analysis that leads to the conclusions above. But we will address a number of questions that this raises. First, what happens if we couple the chiral mode on the domain wall to a gauge field? Second, how has the single chiral mode evaded the Nielsen-Ninomiya theorem?

4.4.2 Anomaly Inflow

We have seen that a domain wall in $d = 4 + 1$ dimension naturally localises a chiral $d = 3 + 1$ fermion. This may make us nervous: what happens if we now couple the system to gauge fields?

At low energies, the only degree of freedom is the zero mode on the domain wall, so we might think it makes sense to restrict our attention to this. (We'll see shortly that things are actually a little more subtle.) Let us introduce a $U(1)$ gauge field everywhere in $d = 4+1$ dimensional spacetime, under which the original Dirac fermion ψ has charge $+1$.

We haven't yet discussed gauge theories in $d = 4 + 1$ dimensions, although we'll learn a few things below. The first statement we'll need is that there are no chiral anomalies in odd spacetime dimensions. This is because there is no analog of γ^5. We might, therefore, expect that a $U(1)$ gauge theory coupled to a single Dirac fermion is consistent in $d = 4 + 1$ dimensions. We will revisit this expectation shortly.

However, from a low energy perspective we seem to be in trouble, because there is a single massless chiral fermion χ_+ on the domain wall which has charge $+1$ under the gauge field. The fact that the gauge field extends in one extra dimension does not stop the anomaly which is now restricted to the region of the domain wall. Under the assumption that the zero mode is restricted to the $y = 0$ slice, the anomaly (3.34) for the gauge current

$$\partial_\mu j^\mu = \frac{1}{32\pi^2}\epsilon^{\mu\nu\rho\sigma}F_{\mu\nu}F_{\rho\sigma}\,\delta(y) \tag{4.35}$$

It is a factor of $1/2$ smaller than the chiral anomaly for a Dirac fermion because we have just a single Weyl fermion. This is bad: if the $U(1)$ gauge field is dynamical then this is precisely the form of gauge anomaly that we cannot tolerate. Indeed, as we saw in (3.33), under a gauge transformation $A_\mu \to A_\mu + \partial_\mu\omega(x, y)$, the measure for the 4d chiral fermion will transform as

$$\int \mathcal{D}\chi\mathcal{D}\bar{\chi} \;\longrightarrow \int \mathcal{D}\chi\mathcal{D}\bar{\chi}\;\exp\left(-\frac{i}{32\pi^2}\int d^4x\;\omega(x;0)\,\epsilon^{\mu\nu\rho\sigma}F_{\mu\nu}F_{\rho\sigma}\right) \tag{4.36}$$

Fortunately, there is another phenomenon which will save us. Let's return to $d = 4+1$ dimensions. Far from the domain wall, the fermion is massive and we can happily integrate it out. You might think that as $m \to \infty$, the fermion simply decouples from the dynamics. But that doesn't happen in odd spacetime dimensions. Instead, integrating out a massive fermions generates a term that is proportional to $\mathrm{sign}(m)$,

$$S_{CS} = -\frac{k}{24\pi^2}\int d^5x\;\epsilon^{\mu\nu\rho\sigma\lambda}A_\mu\,\partial_\nu A_\rho\,\partial_\sigma A_\lambda \tag{4.37}$$

with

$$k = \frac{1}{2}\frac{m}{|m|}$$

This is a *Chern-Simons term* and k is referred to as the *level*. We will discuss the corresponding term in $d = 2+1$ dimensions in some detail in Section 8.4. We will also perform the analogous one-loop calculation in Section 8.5 and show how the Chern-Simons term, proportional to the sign of the mass, is generated when a Dirac fermion is integrated out. The calculation necessary to generate (4.37) is entirely analogous.

Under a gauge transformation $A_\mu \to A_\mu + \partial_\mu\omega$, the Chern-Simons action (4.37) transforms as

$$\delta S_{CS} = -\frac{k}{24\pi^2}\int d^5x\;\partial_\mu\left(\epsilon^{\mu\nu\rho\sigma\lambda}\omega\partial_\nu A_\rho\,\partial_\sigma A_\lambda\right)$$

This is a total derivative. Under most circumstances, we can simply throw this away. But there are some circumstances when we cannot, and the presence of a domain wall is one such an example. We take the thin wall limit, in which we approximate

$$\frac{m}{|m|} = \begin{cases} -1 & y < 0 \\ +1 & y > 0 \end{cases}$$

Since the level is now spatially dependent, we should put it inside the integral. After some integration by parts, we then find that the change of the Chern-Simons term is then

$$\begin{aligned} \delta S_{CS} &= -\frac{3}{24\pi^2}\int d^5x\ \frac{m}{2|m|}\partial_y\left(\omega\epsilon^{\mu\nu\rho\sigma}\partial_\mu A_\nu\,\partial_\rho A_\sigma\right) \\ &= +\frac{1}{8\pi^2}\int d^5x\ \delta(y)\,\omega\epsilon^{\mu\nu\rho\sigma}\partial_\mu A_\nu\,\partial_\rho A_\sigma \end{aligned}$$

We see that this precisely cancels the gauge transformation that comes from the chiral fermion (4.36). A very similar situation occurs in the integer quantum Hall effect, where a 2d chiral fermion on the boundary compensates the lack of gauge invariance of a $d = 2+1$ dimensional Chern-Simons theory in the bulk. This was described in the lectures on the Quantum Hall Effect.

We learn that the total theory is gauge invariant, but only after we combine two subtle effects. In particular, the anomalous current (4.35) on the domain wall is real. A low energy observer, living on the wall, would see that the number of fermions is not conserved in the presence of an electric and magnetic field. But, for a higher dimensional observer there is no mystery. The current is generated in the bulk (strictly speaking, at infinity) by the Chern-Simons term,

$$J^\mu = \frac{\delta S_{CS}[A]}{\delta A_\mu} = -\frac{1}{32\pi^2}\frac{m}{|m|}\epsilon^{\mu\nu\rho\sigma\lambda}F_{\nu\rho}F_{\sigma\lambda}$$

The current is conserved in the bulk, but has a non-vanishing divergence on the domain wall where it is cancelled by the anomaly. This mechanism is referred to as *anomaly inflow*.

There is one final subtlety. I mentioned above that the five-dimensional Maxwell theory coupled to a single Dirac fermion is consistent. This is not quite true. Even in the absence of a domain wall, one can show that the 5d Chern-Simons (4.37) theory is invariant under large gauge transformations only if we take $k \in \mathbf{Z}$. (We'll explain why this is for 3d Chern-Simons theories in Section 8.4.) But integrating out a massive fermion gives rise to a half-integer k rather than integer. In other words, even in the absence of a domain wall the 5d theory is not quite gauge invariant. This doesn't invalidate our discussion above; we simply need to add a bare Chern-Simons term with level $k = 1/2$ so that, after integrating out the massive fermion, the effective level is $k = 1$ when $y > 0$ and $k = 0$ when $y < 0$. (This discussion is slightly inaccurate: we'll have more to say on these issues in Section 8.5.)

4.4.3 The Ginsparg-Wilson Relation

We have not yet addressed exactly how the domain wall fermion evades the Nielsen-Ninomiya theorem. Here we explain the loophole. The idea that follows is more general than the domain wall, and goes by the name of *overlap fermions*.

Rather than jump straight to the case of a Weyl fermion, let's first go back and think about a Dirac fermion. We take the action in momentum space to be

$$S = \frac{1}{a}\int_{BZ} \frac{d^4k}{(2\pi)^4}\ \bar{\psi}_{-k}\, D(k)\, \psi_k$$

for some choice of inverse propagator $D(k)$. As explained in Section 4.3.4, the Nielsen-Ninomiya theorem can be cast as four criterion which cannot all be simultaneously satisfied by $D(k)$. One of these is the requirement that the theory has a chiral symmetry, in the guise of

$$\{\gamma^5, D(k)\} = 0$$

The key idea is to relax this constraint, but relax it in a very particular way. We will instead require

$$\{\gamma^5, D(k)\} = aD\,\gamma^5\,D \tag{4.38}$$

This is the *Ginsparg-Wilson relation.* Note the presence of the lattice spacing a on the right-hand-side. This means that in the continuum limit, which is naively $a \to 0$, we expect to restore chiral symmetry.

In fact, the Ginsparg-Wilson relation ensures that a chiral symmetry exists at all scales. However, it's rather different from the chiral symmetry that we're used to. It's simple to check that the action is invariant under

$$\delta\psi = i\gamma^5\left(1 - \frac{a}{2}D\right)\psi \quad , \quad \delta\bar{\psi} = i\bar{\psi}\left(1 - \frac{a}{2}D\right)\gamma^5 \tag{4.39}$$

These transformation rules have the strange property that the amount a fermion is rotated depends on its momentum. In real space, this means that the symmetry does not act in the same way on all points of the lattice. In the language of condensed matter physics, it is not an *onsite* symmetry. This will cause us a headache shortly.

So the Ginsparg-Wilson relation (4.38) is sufficient to guarantee a chiral symmetry, albeit an unconventional one. The next, obvious question is: what form of D obeys this relation? It's perhaps simplest to give a solution in the continuum, where $a = 1/M$ is simply interpreted as some high mass scale. You can check that, in real (Euclidean) space, the following operator obeys the Ginsparg-Wilson relation,

$$D = \frac{1}{a}\left(1 - \frac{1 - a\partial\!\!\!/}{\sqrt{1 - a^2\partial^2}}\right) \quad \Rightarrow \quad D(k) = \frac{1}{a}\left(1 - \frac{1 - iak\!\!\!/}{\sqrt{1 + a^2k^2}}\right) \tag{4.40}$$

This is the overlap operator. It obeys the Hermiticity property $D^\dagger = \gamma^5 D\gamma^5$. At low momenta, $a\partial \ll 1$, we reproduce the usual Dirac operator,

$$D = \partial\!\!\!/ + \ldots$$

At high momentum, things look stranger. In particular, the derivatives in the denominator mean that this operator is non-local. However, it's not very non-local, and can be shown to fall off exponentially at large distances.

The Ginsparg-Wilson relation relies only on the gamma matrix structure of the operator (4.40). This means that we can also write down operators on the lattice, simply by replacing $\partial\!\!\!/$ by the operator appropriate for, say, Wilson fermions (4.32). Moreover, we can couple our fermions to gauge fields simply by replacing $\partial\!\!\!/$ with $D\!\!\!\!/$, or its lattice equivalent.

Next, we can try to use this chiral symmetry to restrict the Dirac fermion to the analogue of a Weyl fermion. Usually this is achieved by using the projection operators

$$P_\pm = \frac{1}{2}\left(1 \pm \gamma^5\right)$$

For overlap fermions, we need a different projection operator. This is

$$\hat{P}_\pm = \frac{1}{2}\left(1 \pm \gamma^5(1 - aD)\right)$$

You can check that this obey $P_\pm^2 = P_\pm$ and $P_\pm P_\mp = 0$, using the Ginsparg-Wilson relation (4.38). To write down the theory in terms of chiral fermions, we actually need both projection operators: the action can be expressed as

$$\begin{aligned} S &= \frac{1}{a}\int_{BZ} \frac{d^4k}{(2\pi)^4}\ \bar{\psi}_{-\mathbf{k}}(P_+ + P_-)\, D(\mathbf{k})\, (\hat{P}_+ + \hat{P}_-)\, \psi_{\mathbf{k}} \\ &= \frac{1}{a}\int_{BZ} \frac{d^4k}{(2\pi)^4}\ \left[\bar{\psi}_{-\mathbf{k}}P_-\, D(\mathbf{k})\, \hat{P}_+\, \psi_{\mathbf{k}} + \bar{\psi}_{-\mathbf{k}}P_+\, D(\mathbf{k})\, \hat{P}_-\, \psi_{\mathbf{k}}\right] \end{aligned}$$

Throwing away one of these terms can then be thought of as a chiral fermion. It can be shown that if one writes down a strict 4d action for the domain wall fermion, it takes a form similar to that above.

It seems like we have orchestrated a way to put a chiral fermion on the lattice, albeit with a number of concessions forced upon us by the strange Ginsparg-Wilson relation. So what's the catch? The problem comes because, although the action is invariant under (4.39), the measure is not. The measure for a Dirac fermion transforms as

$$\begin{aligned}\delta\left[\mathcal{D}\bar{\psi}\mathcal{D}\psi\right] &= \mathcal{D}\bar{\psi}\mathcal{D}\psi\ \mathrm{Tr}\left[i\gamma^5\left(1-\frac{a}{2}D\right)+i\left(1-\frac{a}{2}D\right)\gamma^5\right]\\ &= \mathcal{D}\bar{\psi}\mathcal{D}\psi\ \mathrm{Tr}\left[2i\gamma^5 - ia\gamma^5 D\right]\end{aligned}$$

This now smells like the way the anomaly shows up in the continuum. Except here, the lack of invariance shows up even before we couple to gauge fields. If we also include gauge fields, and project onto chiral fermions, then we run into trouble. In general, the measure will not be gauge invariant. This, of course, is the usual story of anomalies. However, now life has become more complicated, in large part because of the non-onsite nature of the chiral transformation. What we would like to show is that the measure remains gauge invariant if and only if the matter coupling does not suffer a gauge anomaly. This was studied in some detail by Lüscher. The current state of the art is that this technique can be shown to be consistent for Abelian, chiral gauge theories, but open questions remain in the more interesting non-Abelian case.

4.4.4 Other Approaches

There is one final assumption of the Nielsen-Ninomiya theorem that we could try to leverage in an attempt to put chiral fermions on the lattice: this is the assumption that the fermions are free, so that we can talk in terms of a one-particle dispersion relation. One might wonder if it's possible to turn on some interactions to lift collections of gapless fermions in a manner consistent with 't Hooft anomalies, while preserving symmetries which you might naively have thought should be broken. There has been a large body of work on this topic, which now goes by the name of *symmetric mass generation*, starting with Eichten and Preskill. It's interesting.

4.5 Further Reading

Kenneth Wilson is one of the more important figures in the development of quantum field theory. His work in the early 1970s on the renormalisation group, largely driven by the need to understand second order phase transitions in statistical physics, had an immediate impact on particle physics. The older ideas of renormalisation, due to Schwinger, Tomonaga, Feynman and Dyson, appeared to be little better than sweeping infinites under the carpet. Viewed through Wilson's new lens, it was realised that these infinities are telling us something deep about the way Nature appears on different length scales.

Wilson's pioneering 1974 paper on lattice gauge theory showed how to discretize a gauge theory, and demonstrated the existence of confinement in the strong coupling regime [214]. He worked only with $U(1)$ gauge group, although this was quickly generalised to a large number of non-Abelian gauge theories [8]. The Hamiltonian approach to lattice gauge theory was developed soon after by Kogut and Susskind [124].

In fact, Wilson was not the first to construct a lattice gauge theory. A few years earlier, Wegner described a lattice construction of what we now appreciate as $\mathbf{Z}_2$ gauge theory [201]. (Arguably, the first lattice gauge theory paper was earlier still: in 1966 an electrical engineer called Kane Yee wrote down a discrete version of Maxwell theory, with the gauge fields on the links, with the goal of solving the classical equations numerically rather than simulating the quantum theory [233].) The lattice continued to play a prominent role in many subsequent conceptual developments of quantum field theory, not least because such a (Hamiltonian) lattice really exists in condensed matter physics. Elitzur's theorem was proven in [52].

Wilson's original lattice gauge theory paper does not mention that a discrete version of the theory lends itself to numerical simulation, but this was surely on his mind. He later used numerical renormalisation group techniques [215] to solve the Kondo problem – a sea of electrons interacting with a spin impurity — which also exhibits asymptotic freedom [125]. It wasn't until the late 1970s that people thought seriously about simulating Yang-Mills on the lattice. The first Monte Carlo simulation of four dimensional Yang-Mills was performed by Creutz in 1980 [33].

More details on the basics of lattice gauge theory can be found in the book by Creutz [34] or the review by Guy Moore [138].

Fermions on the Lattice

Wilson introduced his approach to fermions, giving mass to the doublers at the corners of the Brillouin zone, in Erice lectures in 1975. To my knowledge, this has never been published. Other approaches soon followed: the discontinuous SLAC derivative in [48], and the staggered approach in [124]. The general problem of putting fermions on the lattice was later elaborated upon by Susskind [188]. The "rooting" trick, to reduce the number of staggered fermions, is prominently used in lattice simulations, but its validity remains controversial: see [180, 35] for arguments.

The idea that placing fermions on the lattice is a deep, rather than irritating, problem is brought into sharp focus by the theorem of Nielsen and Ninomiya [146, 147].

The story of domain wall fermions has its origins firmly in the continuum. Jackiw and Rebbi were the first to realise that domain walls house chiral fermions [111], a result which now underlies the classification of certain topological insulators. The interaction of these fermions with gauge fields was studied by Callan and Harvey who introduced the idea of anomaly inflow [25]. The fact that this continuum story can be realised in the lattice setting was emphasised by David Kaplan [117].

In a parallel development, the Ginsparg-Wilson relation was introduced in [73] in a paper that sat unnoticed for many years. Maybe it would have helped if the authors were more famous. The first (and, to date, only) solution to this relation was discovered by Neuberger [143, 144], and the resulting exact chiral symmetry on the lattice was shown by Lüscher [126]. The relationship between domain wall fermions and the Ginsparg-Wilson relation was shown in [121].

The idea that strong coupling effects could lift the fermion doublers, in a way consistent with ('t Hooft) anomalies, was first suggested by Eichten and Preskill [51]. This subject has had a renaissance of late, starting with the pioneering work of Fidkowski and Kitaev on interacting 1d topological insulators [58, 59]. They show that Majorana zero modes can be lifted, preserving a particular time reversal symmetry, only in groups of 8. There are more conjectural extensions to higher dimensions where, again, it is thought that only specific numbers of fermions can be gapped together. In $d = 3 + 1$, the conjecture is that Weyl fermions can become gapped in groups of 16; the fact that the Standard Model (with a right-handed neutrino) has $16n$ Weyl fermions has not escaped attention [202, 234].

The lectures by Witten on topological phases of matter include a clear discussion of the Nielsen-Ninomiya theorem [229]. Excellent reviews on the issues surrounding chiral fermions on the lattice have been written by Lüscher [127] and Kaplan [118].

5. Chiral Symmetry Breaking

In this section, we discuss the following class of theories: $SU(N_c)$ gauge theory coupled to N_f Dirac fermions, each transforming in the fundamental representation of the the gauge group. A particularly important member of this class is QCD, the theory of the strong nuclear interactions, and we will consider this specific theory in some detail in Section 5.4. Furthermore, throughout this section we will adopt various terminology of QCD. For example, we will refer to the fermions throughout as *quarks*.

It turns out that the most startling physics occurs when we take the fermions to be massless. For this reason, we will start our discussion with this case, and delay consideration of massive fermions to Section 5.2.3. The Lagrangian of the theory is

$$\mathcal{L} = -\frac{1}{2g^2}\mathrm{Tr}\, F_{\mu\nu}F^{\mu\nu} + \sum_{i=1}^{N_f} i\bar{\psi}_i \not{D}\psi_i \tag{5.1}$$

where $\not{D}\psi = \not{\partial}\psi - i\gamma^\mu A_\mu \psi$. Here $i = 1, \ldots, N_f$ labels the species of quark and is sometimes referred to as a *flavour* index. (Note that ψ also carries a colour index that runs from 1 to N_c and is suppressed in the expressions above.)

Much of what we have to say below will follow from the global symmetries of the theory (5.1). Indeed, the theory has a rather large symmetry group which is only manifest when we decompose the fermionic kinetic terms into into left-handed and right-handed parts

$$\sum_{i=1}^{N_f} i\bar{\psi}_i \not{D}\psi_i = \sum_{i=1}^{N_f} i\psi^\dagger_{+i}\bar{\sigma}^\mu \mathcal{D}_\mu \psi_{+i} + i\psi^\dagger_{-i}\sigma^\mu \mathcal{D}_\mu\psi_{-i}$$

Written in this way, we see that the classical Lagrangian has the symmetry

$$G_F = U(N_f)_L \times U(N_f)_R$$

which acts as

$$U(N_f)_L : \psi_{-i} \mapsto L_{ij}\psi_{-j} \quad \text{and} \quad U(N_f)_R : \psi_{+i} \mapsto R_{ij}\psi_{+j} \tag{5.2}$$

where L and R are both $N_f \times N_f$ unitary matrices. As we will see in some detail below, in the quantum theory different parts of this symmetry group suffer different fates.

Perhaps the least interesting is the overall $U(1)_V$, under which both ψ_- and ψ_+ transform in the same way: $\psi_{\pm,i} \to e^{i\alpha}\psi_{\pm,i}$. This symmetry survives and the associated conserved quantity counts the number of quark particles (minus the number of anti-quarks) of either handedness. In the context of QCD, this is referred to as baryon number.

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: $\psi_{\pm,i} \to e^{\pm i\beta}\psi_{\pm,i}$. We already saw the fate of this symmetry in Section 3.1 where we learned that it suffers an anomaly.

This means that the global symmetry group of the quantum theory is

$$G_F = U(1)_V \times SU(N_f)_L \times SU(N_f)_R \tag{5.3}$$

In this section, our interest lies in what becomes of the two non-Abelian symmetries. These act as (5.2), but where L and R are now each elements of $SU(N_f)$ rather than $U(N_f)$.

5.1 The Quark Condensate

As we've seen in Section 2.4, the dynamics of our theory depends on the values of N_f and N_c. For low enough N_f, we expect that the low-energy physics will be dominated by two logically independent phenomena. We have met the first of these phenomena already: confinement. In this section, we will explore the second of these phenomena: the formation of a *quark condensate*.

The quark condensate – also known as a *chiral condensate* – is a vacuum expectation value of the composite operators $\bar{\psi}_{-i}(x)\psi_{+j}(x)$. (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

$$\langle \bar{\psi}_{-i}\psi_{+j} \rangle = -\sigma\delta_{ij} \tag{5.4}$$

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 2 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 may differ by some order 1 number.

There are a couple of obvious questions that we can ask.

- Why does this condensate form?
- What are the consequences of this condensate?

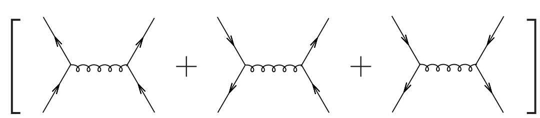
The first of these questions is, like many things in strongly coupled gauge theories, rather difficult to answer with any level of precision, and a complete understanding is still lacking. In what follows, we will give some heuristic arguments. In contrast, the second question turns out to be surprisingly straightforward to answer, because it is determined entirely by symmetry. We will explore this in Section 5.2.

Why Does the Quark Condensate Form?

The existence of a quark condensate (5.4) is telling us that the vacuum of space is populated by quark-anti-quark pairs. This is analogous to what happens in a superconductor, where pairs of electrons condense.

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}.

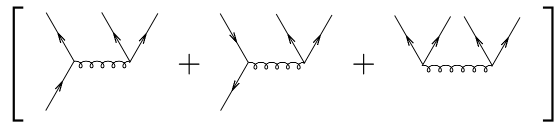
To gain some intuition for why a condensate might form, let's look at what happens at weak coupling $g^2 \ll 1$. Here we can work perturbatively and see how the gluons change the quark Hamiltonian. There are two, qualitatively different effects. The first is the kind that we already met in Section 2.5.1; a tree level exchange of gluons gives rise to a force between quarks. This takes the form

$$\Delta H_1 = g^2$$ 

As we saw in Section 2.5.1, 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

$$\Delta H_2 = g^2$$



The novelty of these terms is that they 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 (5.4) 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 5.4. 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 (5.4) must form whenever the theory confines. We will give this argument in Section 5.6.

5.1.1 Symmetry Breaking

Although the condensate (5.4) preserves the Lorentz invariance of the vacuum, it does not preserve all the global symmetries of the theory. To see this, we can act with a chiral $SU(N_f)_L \times SU(N_f)_R$ rotation, given by

$$\psi_{-i} \mapsto L_{ij}\psi_{-j} \quad \text{and} \quad \psi_{+i} \mapsto R_{ij}\psi_{+j}$$

The ground state of the our theory is not invariant. Instead, the condensate transforms as

$$\langle \bar{\psi}_{-i}\psi_{+j}\rangle \mapsto \sigma(L^\dagger R)_{ij}$$

This is an example of 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

$$G_F = U(1)_V \times SU(N_f)_L \times SU(N_f)_R \;\rightarrow\; U(1)_V \times SU(N_f)_V \tag{5.5}$$

where $SU(N_f)_V$ is the diagonal subgroup of $SU(N_f)_L \times SU(N_f)_R$. The purpose of this chapter is to explore the consequences of this symmetry breaking. As we will see, the consequences are astonishingly far-reaching.

Other Symmetry Breaking Patterns

Throughout this chapter, we will only discuss the symmetry breaking pattern (5.5), since this is what is observed in QCD. But before we move on, it's worth briefly mentioning that other gauge theories can exhibit different symmetry breaking patterns.

For example, consider a $SO(N)$ gauge theory coupled to N_f Dirac fermions in the N-dimensional vector representation. In contrast to the $SU(N)$ gauge theory described above, the vector representation of $SO(N)$ is real. This means that we can equivalently describe the system as having $2N_f$ Weyl fermions, each of which transform in the same vector representation. Correspondingly, the global symmetry group of this theory is

$$G_F = SU(2N_f)$$

A chiral condensate of the form (5.4) will spontaneously break

$$G_F = SU(2N_f) \;\rightarrow\; O(2N_f)$$

Symmetry breaking patterns of this type are typical for fermions in real representations of the gauge group.

The other representative symmetry breaking pattern occurs for $Sp(N)$ gauge groups, again coupled to N_f Dirac fermions in the fundamental ($2N$-dimensional) representation. This representation is pseudo-real; if you take the complex conjugate you can turn it back into the original representation through the use of an anti-symmetric invariant tensor J^{ab}. (A familiar example is $SU(2) \equiv Sp(1)$ where you can turn a $\mathbf{2}$ representation into a $\bar{\mathbf{2}}$ representation by multiplying by the ϵ^{ab} invariant tensor.) This meanst that, once again, the global symmetry group is $G_F = SU(2N_f)$. However, this time when the chiral condensate (5.4) forms, it spontaneously breaks

$$G_F = SU(2N_F) \;\rightarrow\; Sp(N_f)$$

Symmetry breaking patterns of this type are typical for fermions in pseudo-real representations.

5.2 The Chiral Lagrangian

The existence of a spontaneously broken symmetry (5.5) immediately implies a whole slew of interesting phenomena. First, the vacuum of our theory is not unique. Instead, there is a manifold of vacua, parameterised by the condensate

$$\langle \bar{\psi}_{-i}\psi_{+j}\rangle = -\sigma\, U_{ij}$$

where $U \in SU(N_f)$. Next, Goldstone's theorem tells us that there are massless particles in the spectrum. These are bound states of the original quarks, but are now best thought of as long-wavelength ripples of the condensate, where it's value now varies in space and time: $U = U(x)$. Note that there are $N_f^2 - 1$ such Goldstone bosons, one for each broken generator in (5.5). We parameterise these excitations by writing

$$U(x) = \exp\left(\frac{2i}{f_\pi}\,\pi(x)\right) \quad \text{with} \quad \pi(x) = \pi^a(x)\,T^a \tag{5.6}$$

Here $\pi(x)$ is valued in the Lie algebra $su(N_f)$. The matrices T^a_{ij} are the generators of the $su(N_f)$ and the component fields $\pi^a(x)$, labelled by $a = 1, \ldots, N_f^2 - 1$ are called *pions*. (As we explain in Section 5.4, these are named after certain mesons in QCD.)

We have also introduced a dimensionful constant f_π in the definition (5.6). For now, this ensures that the pions have canonical dimensions for scalar fields in four dimensions. It is sometimes called the *pion decay constant*, although this name makes very little sense in our current theory because the pions are stable, massless excitations and don't decay. We'll see where the name comes from in Section 5.4.3 when we discuss how these ideas manifest themselves in the Standard Model.

The Low-Energy Effective Action

We would now like to understand the dynamics of the massless Goldstone modes. As we will see, at low-energies, the form of this action is entirely determined by the symmetries of the theory.

To proceed, we want to construct a theory of the Goldstone modes 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

$$U(x) \to L^\dagger U(x) R$$

What kind of terms can we add to the action consistent with this symmetry? The obvious term is $\mathrm{tr}\, U^\dagger U$ but because $U \in SU(N_f)$, we have $\mathrm{tr}\, U^\dagger U = N_f$ and so cannot appear in the action. (Here the trace is over the N_f flavour indices.) Happily, this is consistent with the fact that U is a massless Goldstone field and it means that we need to look for terms which depend on the spacetime derivatives, $\partial_\mu U$. There are, of course, many such terms. However, our interest is in the low-energy dynamics which, since we have only massless particles, is the same thing as the long-wavelength physics. This means that the most important terms are those with the fewest derivatives.

The upshot of these arguments is that the low-energy effective Lagrangian can be written as a derivative expansion. The leading term has two derivatives. At first glance, it looks as if there are three different candidates:

$$(\mathrm{tr}\, U^\dagger \partial_\mu U)^2 \quad , \quad \mathrm{tr}\,(\partial^\mu U^\dagger \partial_\mu U) \quad , \quad \mathrm{tr}\,(U^\dagger \partial_\mu U)^2$$

However the first term 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, at leading order, there is a unique action that describes the dynamics of pions,

$$\mathcal{L}_2 = \frac{f_\pi^2}{4}\,\mathrm{tr}\,(\partial^\mu U^\dagger\, \partial_\mu U) \tag{5.7}$$

This is the *chiral Lagrangian*. Although the Lagrangian is very simple, this is not a free theory because U is valued in $SU(N_f)$. In fact, this is an example of an important class of scalar field theories in which the fields are coordinates on some manifold which, in the present case, is the group manifold $SU(N_f)$. Theories of this type are called *non-linear sigma models* and arise in many different areas of physics.

Historically, the chiral Lagrangian was the first example of a non-linear sigma model, first introduced by Gell-Mann and Lévy in 1960. The origin of the name "sigma-model" is rather strange: the "sigma-particle" is a particular meson in QCD which, it turns out, is the one particle that is *not* captured by the sigma-model! We will explain this a little more in Section 5.4.

For now, the fact that U is valued in $SU(N_f)$ has a rather straightforward consequence: it means that we cannot 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.

5.2.1 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 (5.6). Keeping only terms quadratic and quartic, the chiral Lagrangian $\mathcal{L}_2$ becomes

$$\mathcal{L}_2 = \mathrm{tr}\,(\partial\pi)^2 - \frac{2}{3f_\pi^2}\mathrm{tr}\,\left(\pi^2(\partial\pi)^2 - (\pi\partial\pi)^2\right) + \ldots \tag{5.8}$$

Note that if we use $\mathrm{tr}\,T^aT^b = \frac{1}{2}\delta^{ab}$ for $su(N_f)$ generators, then the kinetic term has the standard normalisation for each pion field: $\mathrm{tr}\,(\partial\pi)^2 = \frac{1}{2}\,\partial^\mu\pi^a\partial_\mu\pi^a$.

An Example: $N_f = 2$

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 interaction terms then read

$$\mathcal{L}_{\mathrm{int}} = -\frac{1}{6f_\pi^2}\left(\pi^a\pi^a\partial\pi^b\partial\pi^b - \pi^a\partial\pi^a\pi^b\partial\pi^b\right)$$

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

$$\begin{aligned} i\mathcal{A}^{abcd} = \frac{i}{6f_\pi^2}\Big[& \ \delta^{ab}\delta^{cd}\Big(4(p_a\cdot p_b + p_c\cdot p_d) + 2(p_a\cdot p_c + p_a\cdot p_d + p_b\cdot p_c + p_b\cdot p_d)\Big) \\ & + (b\leftrightarrow c) + (b\leftrightarrow d)\Big] \end{aligned}$$

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

$$\begin{aligned} s &= (p_a + p_b)^2 = 2p_a\cdot p_b = 2p_c\cdot p_d \\ t &= (p_a - p_c)^2 = -2p_a\cdot p_c = -2p_b\cdot p_d \\ u &= (p_a - p_d)^2 = -2p_a\cdot p_d = -2p_b\cdot p_c \end{aligned}$$

Using the relation $s + t + u = 0$, the amplitude takes the particularly simple form,

$$i\mathcal{A}^{abcd} = \frac{i}{f_\pi^2}\left[\delta^{ab}\delta^{cd}s + \delta^{ac}\delta^{bd}t + \delta^{ad}\delta^{bc}u\right]$$

Above we have worked at tree level, keeping only the two-derivative terms. We can try to improve our results in two ways: we can include higher derivative terms in the chiral Lagrangian, and we can try to calculate diagrams at one-loop level and higher.

At the next order in the derivative expansion, there are three independent terms. We have $\mathcal{L} = \mathcal{L}_2 + \mathcal{L}_4$ with

$$\begin{aligned}\mathcal{L}_4 = a_1 \left(\mathrm{tr}\, \partial^\mu U^\dagger\, \partial_\mu U\right)^2 + a_2 \left(\mathrm{tr}\, \partial_\mu U^\dagger\, \partial_\nu U\right) \left(\mathrm{tr}\, \partial^\mu U^\dagger\, \partial^\nu U\right) \\ + a_3 \mathrm{tr}\left(\partial_\mu U^\dagger\, \partial^\mu U \partial_\nu U^\dagger\, \partial^\nu U\right)\end{aligned} \tag{5.9}$$

Here a_i are dimensionless coupling constants. These terms will provide corrections to pion-pion scattering that are suppressed at low energy by powers of E/f_π.

Next: loops. The chiral Lagrangian (5.7) 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 fact, there is a slightly more interesting story here which I will not describe in detail. If you compute the one-loop correction to pion scattering from $\mathcal{L}_2$, you will find that it scales as $p^4 \log p^2$. The presence of the logarithm means that this term cannot be generated by a tree graph from higher order terms in the chiral Lagrangian and, indeed, at low-energies is enhanced relative to the contributions from $\mathcal{L}_4$.

Furthermore, it turn out that there is a term more important than $\mathcal{L}_4$ that we've missed. This is known as the Wess-Zumino-Witten term. It doesn't contribute to pion scattering, so we can neglect it for the purposes above. However, it plays a key role in the overall structure of the theory. We will discuss this term in detail in Section 5.5.

5.2.2 Currents

We started our discussion with the microscopic non-Abelian gauge theory (5.1) and have ended up, at low-energies, with a very different looking theory (5.7). In general, it is useful to know how operators in the UV get mapped to operators in the IR. There is one class of operators for which this map is particularly straightforward: these are the currents associated to the $SU(N_f)_L \times SU(N_f)_R$ chiral symmetry.

In the microscopic theory, the flavour currents are written most simply in terms of the vector and axial combinations: $J^a_{V\mu} = J^a_{L\mu} + J^a_{R\mu}$ and $J^a_{A\mu} = J^a_{L\mu} - J^a_{R\mu}$, with the familiar expressions

$$J^a_{V\mu} = \bar{\psi}_i T^a_{ij} \gamma_\mu \psi_j \quad \text{and} \quad J^a_{A\mu} = \bar{\psi}_i T^a_{ij} \gamma_\mu \gamma^5 \psi_j \tag{5.10}$$

where T^a_{ij} are $su(N_f)$ generators. What are the analogous expressions 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 $U \to L^\dagger U$ so, infinitesimally,

$$\delta_L U = -i\alpha^a T^a U$$

We can now compute the current using the standard trick: elevate $\alpha^a \to \alpha^a(x)$. The Lagrangian is no longer invariant, but now transforms as $\delta\mathcal{L} = \partial_\mu \alpha^a J^a_{L\,\mu}$; the function $J^a_{L\,\mu}$ is the current that we're looking for. Implementing this, we find

$$J^a_{L\,\mu} = \frac{if_\pi^2}{4}\,\mathrm{tr}\Big(U^\dagger T^a \partial_\mu U - (\partial_\mu U^\dagger) T^a U\Big) \tag{5.11}$$

We can also expand this in pion fields (5.6). To leading order we have simply

$$J^a_{L\,\mu} \approx -\frac{f_\pi}{2}\partial_\mu \pi^a$$

Similarly, under $SU(N_f)_R$, we have $\delta U = i\alpha^a U T^a$ and

$$J^a_{R\,\mu} = \frac{if_\pi^2}{4}\Big(-T^a U^\dagger \partial_\mu U + (\partial_\mu U^\dagger) U T^a\Big) \approx +\frac{f_\pi}{2}\partial_\mu \pi^a \tag{5.12}$$

Note that both currents have non-vanishing matrix elements between the vacuum $|0\rangle$ and a one-particle pion state $|\pi^a(p)\rangle$. For example

$$\langle 0|J^a_{L\,\mu}(x)|\pi^b(p)\rangle = -i\frac{f_\pi}{2}\delta^{ab}\,p_\mu e^{-ix\cdot p} \tag{5.13}$$

Historically, the approach to chiral symmetry breaking was known as *current algebra*, and this equation plays a starring role. 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.

Although the chiral symmetry is broken, the diagonal combination $SU(N_f)_V$ survives, and

$$\langle 0|J^a_{V\,\mu}|\pi^b\rangle = \langle 0|J^a_{L\,\mu} + J^a_{R\,\mu}|\pi^b\rangle = 0$$

5.2.3 Adding Masses

Our discussion so far has been for massless quarks. We now consider the effect of turning on masses. The Lagrangian is:

$$\mathcal{L} = -\frac{1}{2g^2}\mathrm{Tr}\,F_{\mu\nu}F^{\mu\nu} + \sum_{i=1}^{N_f}\left(i\bar{\psi}_i \not{D}\psi_i - m_i\bar{\psi}_i\psi_i\right)$$

If the masses are large compared to Λ_{QCD}, then the quarks play no role in the low-energy physics. Here we will be interested in the situation where the masses are small, $m_i \ll \Lambda_{QCD}$.

It is a general rule – and a deep fact about quantum field theory – that turning on a mass for fermions always breaks some global symmetry. In the present case, the masses explicitly break the chiral symmetry. If all the masses are equal, then there remains a non-Abelian $U(N_f)_V$ flavour symmetry. In contrast, if all the masses are different, we have only the Cartan subalgebra $U(1)^{N_f}$.

In the previous section, we saw that we can derive powerful statements about the low-energy physics due to the spontaneous breaking of the chiral symmetry. Now this symmetry is explicitly broken by the masses themselves, but all is not lost. For $m_i \ll \Lambda_{QCD}$, we still have an *approximate* chiral symmetry. The quark condensate is still associated to the scale Λ_{QCD}, and the masses give only a small correction. This means that we can still write

$$\langle \bar{\psi}_{-i} \psi_{+j} \rangle \approx -\sigma\, U_{ij}$$

with $U \in SU(N_f)$. We can then incorporate the masses in the chiral Lagrangian by introducing the $N_f \times N_f$ mass matrix,

$$M = \mathrm{diag}(m_1, \ldots, m_{N_f})$$

In the presence of masses, the leading order chiral Lagrangian is

$$\mathcal{L}_2 = \int d^4x\ \frac{f_\pi^2}{4}\, \mathrm{tr}\, (\partial^\mu U^\dagger\, \partial_\mu U) + \frac{\sigma}{2} \mathrm{tr}\, \left(MU + U^\dagger M^\dagger \right)$$

This lifts the vacuum manifold of the theory. It can be thought of as adding a potential to the $SU(N_f)$ vacuum moduli space, 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

$$\mathcal{L}_2 = \mathrm{tr}\, (\partial\pi)^2 - \frac{\sigma}{f_\pi^2}\, \mathrm{tr}(M + M^\dagger)\pi^2 + \ldots \tag{5.14}$$

and we see that we get a mass term for the pions as expected.

5.3 Miraculously, Baryons

The purpose of the chiral Lagrangian is to describe the low-energy dynamics of pions. These are the massless Goldstone bosons that arise after spontaneous symmetry breaking which, in terms of the original quarks take the schematic form $\bar{\psi}_i \psi_j$. These particles are all neutral under the $U(1)_V$ vector symmetry.

There are also bound states of quarks which carry quantum numbers under $U(1)_V$. These are the *baryons* that arise by contracting the $a = 1, \ldots, N_c$ colour indices. Schematically these take the form

$$\epsilon_{a_1 \ldots a_{N_c}} \psi_{i_1}^{a_1} \ldots \psi_{i_{N_c}}^{a_{N_c}} \tag{5.15}$$

where we have neglected the spinor indices. The baryons are bosons when N_c is even and fermions when N_c is odd. With our normalisation, they have charge $+N_c$ under the vector symmetry $U(1)_V$. Often one rescales the charges of the quarks to have $U(1)_V$ charge $1/N_c$ so that the baryon has charge $+1$; this re-scaled symmetry is then referred to simply as *baryon number*.

Assuming that our theory confines, the baryons are expected to have mass $\sim \Lambda_{QCD}$. Nonetheless, they are the lightest particles carrying $U(1)_V$ charge and so are stable.

There is no reason to expect that the chiral Lagrangian knows anything about the baryons. Indeed, to construct the chiral Lagrangian we intentionally threw out all but the massless excitations. It is therefore something of a wonderful surprise to learn that the baryons do arise in the chiral Lagrangian: they are solitons.

The Topological Charge

Let's first show that the chiral Lagrangian has a hidden conserved current. Static field configurations in the chiral Lagrangian are described by a map from spatial $\mathbf{R}^3$ to the group manifold $SU(N_f)$. If we insist that the field asymptotes to the same vacuum state so, for example,

$$U(\mathbf{x}) \to 1 \quad \text{as } |\mathbf{x}| \to \infty$$

then we effectively compactify $\mathbf{R}^3$ to $\mathbf{S}^3$. Now static configurations can be thought of as a map

$$U(\mathbf{x}) : \mathbf{S}^3 \mapsto SU(N_f)$$

Such configurations are characterised by their winding

$$\Pi_3(SU(N_f)) = \mathbf{Z}$$

This winding number — which we denote by $B \in \mathbf{Z}$ — is computed by the integral

$$B = \frac{1}{24\pi^2} \int d^3x \; \epsilon_{ijk} \mathrm{tr}\left(U^\dagger (\partial_i U)\, U^\dagger (\partial_j U)\, U^\dagger\, \partial_k U\right) \tag{5.16}$$

In fact, we can go further and write down a local current

$$B^\mu = \frac{1}{24\pi^2}\epsilon^{\mu\nu\rho\sigma}\mathrm{tr}\,(U^\dagger(\partial_\nu U)\,U^\dagger(\partial_\rho U)\,U^\dagger\,\partial_\sigma U)$$

which obeys $\partial_\mu B^\mu = 0$ by virtue of the anti-symmetric tensor. The winding number is then given by $B = \int d^3x\ B^0$.

It is natural to search for an interpretation of this conserved current B^μ, it terms of the microscopic theory. The only candidate is $U(1)_V$, strongly suggesting that we should identify B^μ with the baryon number current and, correspondingly, the solitons with baryons. This appears to be magic. We tried to throw away everything that wasn't massless. But if you treat the pions correctly, the baryons reappear as solitons.

A First Attempt at Solutions

What do these soliton solutions look like? Let's start with the two-derivative chiral Lagrangian. The associated energy functional for static field configurations is

$$E = \frac{f_\pi^2}{4}\int d^3x\ \mathrm{tr}\,\partial_i U^\dagger \cdot \partial_i U$$

where now $i = 1,2,3$ runs over spatial indices only. Solutions to the equations of motion are minima (or, more generally, saddle points) of this energy functional. A simple scaling argument tell us that these don't exist. To see this, consider a putative solution $U_\star(\mathbf{x})$ with energy $E_\star$. Then the new configuration $U_\lambda(\mathbf{x}) = U_\star(\lambda\mathbf{x})$ has energy

$$E_\lambda = \frac{f_\pi^2}{4}\int d^3x\ \mathrm{tr}\,\partial_i U_\star^\dagger(\lambda\mathbf{x})\cdot\partial_i U_\star(\lambda\mathbf{x}) = \frac{1}{\lambda}E_\star$$

We see that we can always lower the energy of any configurations simply by rescaling its size. This simple observation — which goes by the name of *Derrick's theorem* — means that although the chiral Lagrangian has the topology to support solitons, no static solutions exist. The reason for this is that the classical theory is scale invariant so there is nothing to set the size of the soliton. (The only dimensionful quantity, f_π, multiplies the whole action and so doesn't affect the classical equations of motion.)

5.3.1 The Skyrme Model

The situation improves when we include higher derivative terms. These will scale differently with λ, and may result in a minimum of the energy functional.

We saw previously that there are three possible terms with four derivatives (5.9),

$$\mathcal{L}_4 = a_1 \left(\mathrm{tr}\, \partial^\mu U^\dagger \, \partial_\mu U\right)^2 + a_2 \left(\mathrm{tr}\, \partial_\mu U^\dagger \, \partial_\nu U\right) \left(\mathrm{tr}\, \partial^\mu U^\dagger \, \partial^\nu U\right)$$
$$+ a_3 \mathrm{tr} \left(\partial_\mu U^\dagger \, \partial^\mu U \partial_\nu U^\dagger \, \partial^\nu U\right)$$

and we expect that the effective action contains all three terms with some choice of coefficients a_1, a_2 and a_3. However, it turns out to be much easier to discuss solitons if we take a particular linear combination of these terms. We take the effective action to be

$$\mathcal{L} = \frac{f_\pi^2}{4} \, \mathrm{tr} \, \left(\partial^\mu U^\dagger \partial_\mu U\right) + \frac{1}{32g^2} \mathrm{tr} \, \left([U^\dagger \partial^\mu U, U^\dagger \partial^\nu U][U^\dagger \partial_\mu U, U^\dagger \partial_\nu U]\right)$$

This is called the *Skyrme model*.

There is no first-principles justification for this particular 4-derivative term although it's worth mentioning that it is the unique term which contains no more than two time derivatives, making it more straightforward to interpret the classical equations of motion. Here g^2 is a dimensionless coupling constant that will ultimately determine the scale of the soliton relative to f_π.

To simplify our notation, we introduce the $su(N_f)_L$ current.

$$L_\mu = U^\dagger \partial_\mu U$$

After massaging the four-derivative terms, you can check that the static energy can be written as

$$E = \frac{f_\pi^2}{4} \int d^3x \, \mathrm{tr} \, \left(L_i L_i^\dagger - \frac{1}{4g^2 f_\pi^2} (\epsilon_{ijk} L_i L_j)(\epsilon_{lmk} L_l^\dagger L_m^\dagger) \right)$$

We now use the Bogomolnyi trick that we already employed in Section 2 for instantons, vortices and monopoles: we write the energy functional as a total square,

$$E = \frac{f_\pi^2}{4} \int d^3x \, \mathrm{tr} \left| L_i \mp \frac{1}{2g f_\pi} \epsilon_{ijk} L_j L_k \right|^2 \pm \frac{f_\pi}{4g} \int d^3x \, \epsilon_{ijk} L_i L_j L_k$$

The first term is clearly positive definite. But the second term is something that we've seen before: it is the topological winding (5.16) that we identified with the baryon number B. We learn that the energy is bounded below by the baryon number

$$E \geq \frac{6\pi^2 f_\pi}{g} |B| \tag{5.17}$$

This now looks more promising: the energy of multiple baryons grows at least linearly with B. Soliton configurations with non-trivial winding are called *Skyrmions* and are identified with baryons in the theory.

5.3.2 Skyrmions

Let's see what Skyrmion solutions look like. The usual way to proceed with bounds like (5.17) is to try to saturate them. For $B > 0$, this occurs when the fields obey the first order differential equation

$$L_i = \frac{1}{2gf_\pi}\epsilon_{ijk}L_jL_k \tag{5.18}$$

While this is usually a sensible approach, it turns out that it doesn't help in the present case. One can show that there are no solutions to (5.18). Instead, we must turn to the full, second order, equations of motion and solve

$$\partial_\mu L^\mu = \frac{1}{4f_\pi^2 g^2}\partial_\mu[L_\nu,[L^\mu,L^\nu]] \tag{5.19}$$

We will solve this for the simplest case of

$$N_f = 2$$

Here, the target space = group manifold $SU(2) = \mathbf{S}^3$. For a single Skyrmion, the field $U(\mathbf{x})$ must wrap once around the $\mathbf{S}^3$ target space as we move around the spatial $\mathbf{R}^3$. This is achieved by the so-called hedgehog ansatz,

$$U_{\text{Skyrme}}(\mathbf{x}) = \exp\left(if(r)\boldsymbol{\sigma}\cdot\hat{\mathbf{x}}\right) = \cos f(r) + i\boldsymbol{\sigma}\cdot\hat{\mathbf{x}}\sin f(r) \tag{5.20}$$

This field configuration has winding number $B = 1$ if we pick the function $f(r)$ to have boundary conditions

$$f(r) \to \begin{cases} 0 & \text{at } r = 0 \\ \pi & \text{as } r \to \infty \end{cases}$$

The equation of motion (5.19) then becomes an ordinary differential equation on $f(r)$,

$$(r^2 + 2\sin^2 f)f'' + 2rf' + \sin 2f\ f'^2 - \sin 2f - \sin^2 f\sin 2f = 0$$

which can be solved numerically; it is a monotonically increasing function whose exact form is not needed for our purposes. The energy of this solution turns out to be about 25% higher than the bound (5.17).

Our Skyrme model is built around symmetries. For $N_f = 2$, the symmetry group is $SU(2)_L \times SU(2)_R$, but if we insist (as we did above) that the field tends towards its vacuum value asymptotically, $U(\mathbf{x}) \to 1$, then it leaves us only with the diagonal

$SU(2)_V$ as a global symmetry. Including the group of spatial rotations, we have the symmetry group

$$SU(2)_{\rm rot} \times SU(2)_V \tag{5.21}$$

The single Skyrmion (5.20) is not invariant under either of these $SU(2)$ groups separately. However, it is invariant under the diagonal $SU(2)$ which acts simultaneously as a spatial and flavour rotation.

The subgroup of (5.21) which acts non-trivially on the Skyrmion solution (5.20) can be used to generate new solutions. These are trivially related to the original, and just change its embedding in the target space. Nonetheless, they have important consequences. After quantisation, they endow the Skyrmion with quantum numbers under $SU(2)_V$. For example, one can show that the simplest Skyrmion described above sits in a doublet of $SU(2)_V$. In QCD, viewed as having two light quarks, this is interpreted as the proton and neutron.

The Skyrme model has spawned a mini-industry, and there is much more to say about its quantisation, and its utility in describing both nucleons and higher nuclei. We won't say this here.

There, however, is one important aspect of Skyrmions that we have not yet understood: their quantum statistics. Since the baryon (5.15) contains N_c quarks, we would hope that the Skyrmion is a boson when N_c is even and a fermion when N_c is odd. Yet, so far, the chiral Lagrangian knows nothing about the number of colours N_c. It turns out that we have missed a rather subtle term in the effective action, known as the Wess-Zumino-Witten term. This will be introduced in section 5.5, and in section 5.5.3 we will see that it indeed makes the Skyrmion fermionic or bosonic depending on the number of colours N_c.

5.4 QCD

Until now, we have kept our discussion general. However, there is one example of the class of theories that we have been discussing whose importance dwarfs all others. This is QCD, the theory of the strong nuclear interaction.

QCD is an $SU(3)$ gauge theory coupled to $N_f = 6$ Dirac fermions that we call quarks. However, for many questions concerning the low-energy behaviour of the theory, only two — or sometimes three – of these quarks are important. To see why, we need to look at their masses. (I've included their electromagnetic charge Q for convenience.)

Quark	Charge	Mass (in MeV)
d = down	-1/3	4
u = up	+2/3	2
s = strange	-1/3	95
c = charm	+2/3	1250
b = bottom	-1/3	4200
t = top	+2/3	170,000

Note that the up quark is lighter than the down, an inversion of the hierarchy relative to the other two generations. We can compare these quark masses to the strong coupling scale,

$$\Lambda_{QCD} \approx 300 \text{ MeV}$$

We see that the masses of the two lightest quarks $m_u, m_d \ll \Lambda_{QCD}$ while the strange quark has mass $m_s < \Lambda_{QCD}$, although there is not a large separation of scales. Meanwhile, the other three quarks are clearly substantially heavier than Λ_{QCD} and play no role in the low-energy physics. This means that, for many purposes we can consider QCD to have $N_f = 3$ quarks while, for some purposes, we may want to take $N_f = 2$.

When we take $N_f = 3$, we have several different $SU(3)$ groups floating around. The gauge group is $SU(3)$ and the global symmetry group is $SU(3)_L \times SU(3)_R$, which is spontaneously broken down to $SU(3)_V$ by the chiral condensate. In this section, it is these global symmetries that are of interest.

The global flavour symmetries are not exact because they are broken explicitly by the quark masses. The fact that $m_u \approx m_d$ means that the $SU(2)_V \subset SU(3)_V$ subgroup which rotates only up and down quarks is a rather better symmetry of Nature than the full $SU(3)_V$. This approximate $SU(2)_V$ symmetry was first noticed by Heisenberg in 1932 and is called *isospin*. Heisenberg viewed this symmetry as rotating protons and neutrons, rather than quarks.

Confinement of quarks means that the particles we observe are either mesons (comprising a quark + anti-quark) or baryons (comprising three quarks). These excitations must arrange themselves in representations of the unbroken symmetries of the theory. As we noted, the global symmetries are not exact due to the different quark masses but, as we describe below, are nonetheless visible in the observed spectrum. The fact that mesons and baryons arrange themselves into approximate multiplets of $SU(3)_V$ was first noticed by Gell-Mann, who referred to this classification as the *eightfold way*.

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	$10^{-8} - 10^{-11}$

5.4.1 Mesons

Many hundreds of mesons are observed in Nature[10]. A simple model of a meson views it 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.

Let's start with the spin 0 mesons. These are all pseudoscalars, with parity -1. A number of these have masses that are lighter or comparable to the proton (which weighs in at 938 MeV). These are shown in the table above.

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 $\bar{K}^0$. The list therefore contains, in total, nine different particles + anti-particles.

All mesons are unstable, decaying via the weak force. We will describe this briefly in Section 5.4.3 but, for now, our interest lies in understanding how these mesons arise in the first place. In particular, we would like to understand why this particular pattern of masses emerges.

First, an obvious comment: the masses of the mesons are not equal to the sum of the masses of their constituent quarks! 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

[10]All the properties of all the particles in the universe can be found in the Particle Data Group website http://pdg.lbl.gov/.

understand. Thankfully, for a subset of the mesons, we have the chiral Lagrangian to help us.

Let's see what we would expect based on chiral symmetry. If we consider QCD with just two light quarks – the up and the down – then the spontaneous symmetry breaking of $SU(2)_L \times SU(2)_R$ symmetry should give us three light almost-Goldstone modes. These are the three pions, π^+, π^- and π^0.

The fact that the pions are both bound states of fundamental fermions, and yet can also be viewed as Goldstone bosons, was first suggested by Yoichiro Nambu in the early 1960s. His vision is all the more remarkable given that it came 10 years before the formulation of QCD, and several years before Gell-Mann and Zweig introduced the idea of quarks. Nambu made many further ground-breaking contributions to theoretical physics, including the realisation that quarks carry three colours (not to mention writing down one of the key equations of string theory). He had to wait until 2008 for his Nobel prize.

Suppose now that we consider $N_f = 3$ light quarks. We expect $N_f^2 - 1 = 8$ almost Goldstone-modes. These are usually referred as *pseudo-Goldstone bosons*. And, indeed, there are eight mesons which are substantially lighter than the others: these are the pions, kaons and the η. They sit inside our 3×3 matrix π like this:

$$\pi = \frac{1}{\sqrt{2}} \begin{pmatrix} \frac{\pi^0}{\sqrt{2}} + \frac{\eta}{\sqrt{6}} & \pi^+ & K^+ \\ \pi^- & -\frac{\pi^0}{\sqrt{2}} + \frac{\eta}{\sqrt{6}} & K^0 \\ K^- & \bar{K}^0 & -\frac{2\eta}{\sqrt{6}} \end{pmatrix} \tag{5.22}$$

This is not an obvious arrangement. How do we figure out which particles go where? The answer, as with everything in this game, is symmetry. Our theory has a $SU(3)_V$ symmetry, which allows us to assign two Cartan charges $U(1) \times U(1) \subset SU(3)_V$ to each element of the the matrix π. These charges are called "isospin" and "strangeness" and coincide with almost-conserved quantities of the particles that can be determined experimentally.

The eight Goldstone modes that sit in π would be exactly massless if the $SU(3)_L \times SU(3)_R$ were exact. However, chiral symmetry is broken by the quark mass matrix

$$M = \text{diag}(m_u, m_d, m_s)$$

Since 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 in the earlier table. Equation (5.14) then gives us the pion masses. Expanding this out, we find

$$\mathcal{L}_{\rm mass} = \frac{-\sigma}{f_\pi^2}\,\left[\frac{1}{2}(m_u+m_d)\left((\pi^0)^2 + 2\pi^+\pi^-\right) + (m_u+m_s)K^-K^+ \right. \tag{5.23}$$
$$\left. + (m_d+m_s)\bar{K}^0K^0 + \frac{1}{2}\left(\frac{m_u}{3}+\frac{m_d}{3}+\frac{4m_s}{3}\right)\eta^2 + \frac{1}{\sqrt{3}}(m_u-m_d)\pi^0\eta\right]$$

Note that there is mixing between π^0 and η, albeit one that disappears when $m_u = m_d$ so that isospin is restored. There is lots of interesting information in this equation. First note that we cannot directly relate the quark masses to the meson masses; they depend on the unknown ratio σ/f_π^2. Nonetheless, there are a number of simple relations between meson masses, quark masses and the chiral condensate that we can extract. For example, the mass of π^0 is given by

$$m_\pi^2 = \frac{2\sigma}{f_\pi^2}(m_u+m_d)$$

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*.

By taking ratios, we can relate meson and quark masses directly. For example, we have

$$\frac{m_{K^+}^2 - m_{K^0}^2}{m_\pi^2} = \frac{m_u - m_d}{m_u+m_d} \tag{5.24}$$

Finally, 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

$$m_K^2 \approx \frac{3}{4}m_\eta^2 + \frac{1}{4}m_\pi^2$$

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.

The η' Meson

There is one meson listed in the table that is not a Goldstone boson. This is the η' which, despite having similar quark content to the η, has almost twice the mass. Note that, in contrast to the other eight mesons, $\eta' = \frac{1}{\sqrt{3}}(u\bar{u}+d\bar{d}+s\bar{s})$ is a singlet under $SU(3)_V$. This is actually the would-be Goldstone boson associated to the $U(1)_A$ axial symmetry. However, as we have seen, this symmetry suffers from an anomaly, which means that the η' meson is not massless in the chiral limit, and is not particularly light in the real world.

The Mysterious Sigma

There is one light scalar meson listed in the particle data book that I have not yet mentioned. It 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. Moreover, it has vanishing quantum numbers (angular momentum, parity, isospin and strangeness).

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.

5.4.2 Baryons

We will briefly describe the baryon spectrum in QCD. In the non-relativistic quark model, with $G = SU(3)$ gauge group, each baryon contains three quarks. As with the mesons, this is a caricature of a baryon which, in reality, is a complicated object that contains many hundreds of gluons, quarks and anti-quarks, but with three more quarks than anti-quarks. This caricature sometimes goes by the name of the non-relativistic quark model.

If we work with $N_f = 3$ species of light quarks, each transforms in the $\mathbf{3}$ of $SU(3)_V$. We have

$$\mathbf{3} \otimes \mathbf{3} \otimes \mathbf{3} = \mathbf{1} \oplus \mathbf{8} \oplus \mathbf{8} \oplus \mathbf{10}$$

A little bit of group theory, combined with the Pauli exclusion principle, shows that those baryons which have spin 1/2 must lie in the $\mathbf{8}$ of $SU(3)_V$. Indeed, there is an octuplet of baryons whose mass differ from each other by about 30%. These are shown in the table on the next page.

Similarly, one can show that baryons with spin 3/2 lie in the $\mathbf{10}$ of $SU(3)_V$. Such a decuplet of baryons also exists: they go by the names Δ (with charges 0, ±1 and 2), $U^\star$ (with charges 0 and ±1), $\Xi^\star$ (with charges -1 and 0) and Ω^- with charge -1.

The fact that the baryons sit nicely into representations of $SU(3)_V$ was first noticed by Gell-Mann who dubbed it the *eightfold way*. At the time the Ω^- baryon — which has quark content sss — had not been discovered. Gell-Mann (and, independently, Ne'eman) used the representation properties to predict the mass, charge and decay products of this particle.

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^0	dds	1197	10^{-10}
Xi Ξ^0	uss	1315	10^{-10}
Xi Ξ^-	dss	1321	10^{-10}

For the pions, we showed how the mass splitting can be explained from the chiral Lagrangian. We will not do this for baryons, although with some work one can show that the Skyrmion spectrum indeed gives reasonable agreement.

5.4.3 Electromagnetism, the Weak Force, and Pion Decay

It's not just the quark masses that explicitly break the $SU(3)_V$ flavour symmetry of the Standard Model; the symmetry is also broken by the coupling to the other forces.

At low energies, the relevant force is electromagnetism. The $U(1)_{EM}$ of electromagnetism is a subgroup of $SU(3)_V$, generated by

$$Q = \begin{pmatrix} \frac{2}{3} & 0 & 0 \\ 0 & -\frac{1}{3} & 0 \\ 0 & 0 & -\frac{1}{3} \end{pmatrix} \tag{5.25}$$

This is enough to tell us how to couple photons to the chiral Lagrangian. We simply need to replace the derivatives in (5.7) with covariant derivatives,

$$S = \int d^4x \; \frac{f_\pi^2}{4} \mathrm{tr}\left(\mathcal{D}^\mu U^\dagger \, \mathcal{D}_\mu U\right) \tag{5.26}$$

where

$$\mathcal{D}_\mu U = \partial_\mu U - ieA_\mu [Q, U]$$

with e the electric charge of an electron.

At the classical level, this coupling preserves a $(U(1) \times SU(2))_L \times (U(1) \times SU(2))_R$ subgroup of the $SU(3)_L \times SU(3)_R$ chiral symmetry. This means that, if all quark masses vanish, the four neutral mesons π^0, η, K^0 and $\bar{K}^0$ would still be Goldstone bosons, and massless even when we include the effects of electromagnetism. In contrast, the charged pions $\pi^\pm$ and $K^\pm$ are massless only at tree level. One-loop effects give a contribution to their mass of the form $\delta m^2_{EM} \sim e^2 \mathrm{tr}(QUQU)$. The charged pion masses in (5.23) then become

$$m^2_{\pi^\pm} = \frac{2\sigma}{f_\pi^2}(m_u + m_d) + \delta m^2_{EM} \quad \text{and} \quad m^2_{K^\pm} = \frac{2\sigma}{f_\pi^2}(m_d + m_s) + \delta m^2_{EM}$$

By taking ratios of these meson masses, we can cancel the factors of σ / f_π^2 and δm^2_{EM} and learn about the quark masses. For example, taking into account electromagnetic corrections, we can generalise (5.24) to

$$\frac{(m^2_{K^\pm} - m^2_{K^0}) - (m^2_{\pi^\pm} - m^2_{\pi^0})}{m^2_{\pi^0}} = \frac{m_u - m_d}{m_u + m_d}$$

From the measured masses of the mesons, we then get that $m_d/m_u \approx 2$.

Charged Pion Decay

Although certain pions are relatively long lived – most notably the $\pi^\pm$ and the kaons – none are absolutely stable. They decay through the weak force. Happily, this too is rather straightforward to calculate using the chiral Lagrangian, because the weak gauge group coincides with $SU(2)_L$ isospin.

For example, the charged pion $\pi^+ = u\bar{d}$ has a lifetime of $\sim 10^{-8}$ seconds, decaying almost always to

$$\pi^+ \to \bar{\mu} + \nu_\mu$$

The decay is mediated by the W-boson. If we integrate out the W-boson, we can equally well describe the decay using Fermi's four-fermion interaction,

$$\mathcal{L}_{Fermi} = \frac{G_F}{\sqrt{2}} \left[\bar{u}\gamma^\mu(1-\gamma^5)d\right]\left[\bar{\mu}\gamma_\mu(1-\gamma^5)\nu_\mu\right]$$

where $G_F \approx 10^{-5}$ GeV$^{-}2$ is the Fermi constant. The computation of the decay rate now factorises into two pieces: the leptonic part $\langle \bar{\mu}\nu_\mu | \bar{\mu}\gamma_\mu(1-\gamma^5)\nu_\mu|0\rangle$ can be computed perturbatively. However, the piece involving the quarks involve strongly interacting physics, $\langle 0|\bar{u}\gamma^\mu(1-\gamma^5)d|\pi^+\rangle$. Thankfully we can compute this using the currents that we introduced in Section 5.2.2. The operator coincides with the $SU(2)_L$ current (5.10),

$$\bar{u}\gamma_\mu(1-\gamma^5)d = 2(J^1_{L\,\mu} + iJ^2_{L\,\mu})$$

We can then use our result (5.13),

$$\langle 0|J^a_{L\,\mu}(x)|\pi^b(p)\rangle = -i\frac{f_\pi}{2}\delta^{ab}\,p_\mu e^{-ix\cdot p}$$

We simply need to remember that $\pi^+ = \frac{1}{\sqrt{2}}(\pi^1 + i\pi^2)$ to find that the matrix element is determined by f_π,

$$\langle 0|\bar{u}\gamma^\mu(1-\gamma^5)d|\pi^+\rangle = -i\sqrt{2}f_\pi p^\mu e^{-ip\cdot x}$$

Recall that when we first introduced f_π in Section 5.2, we mentioned that it is called the *pion decay constant*, even though that name made little sense in the theory we were considering. Now we see why: it is the scale which directly determines the decay width of the pion.

To compute the lifetime of the pion, we must square the matrix element and integrate over the phase space of $\bar{\mu}$ and ν_μ. The end result for the rate of decay is then given by

$$\Gamma(\pi^+ \to \bar{\mu} + \nu_\mu) = \frac{G_F^2 f_\pi^2}{4\pi} m_\pi m_\mu^2 \left(1 - \frac{m_\mu^2}{m_\pi^2}\right)^2$$

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

$$\pi^0 \to \gamma\gamma$$

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.

The full history is, as with many things in this subject, rather convoluted. The pion decay was first computed in the 1940s, by assuming a coupling to the nucleons $N = (p, n)$ of the form $G_{\pi N}\pi^a \bar{N}\gamma^5\sigma^a N$. This gives a result which is pretty close to the observed value. Unfortunately, this calculation is wrong. As we've seen, the pion is really a Goldstone boson and so has only derivative couplings, at least in the limit $m_\pi \to 0$. Indeed, one can show that in a theory with an unbroken $SU(2)_L \times SU(2)_R$ chiral symmetry, the decay $\pi^0 \to \gamma\gamma$ would be forbidden. What's going on?

The answer is that we've missed something. Gauging a subgroup $U(1)_{EM} \subset SU(2)_V$ introduces an anomaly for the axial currents. We can import our calculation of the chiral anomaly from Section 3.1. For two quarks, up and down, each with $N_c = 3$ colours, we have

$$\partial^\mu J^a_{A\,\mu} = \frac{N_c}{16\pi^2}\epsilon^{\mu\nu\rho\sigma}F_{\mu\nu}F_{\rho\sigma}\,\mathrm{tr}\left(\frac{\sigma^a}{2}Q^2\right)$$

where here $F_{\mu\nu}$ denotes the electromagnetic field strength. In contrast to (5.25), we now take $U(1)_{EM} \subset SU(2)_V$ to be generated by

$$Q = \begin{pmatrix} \frac{2}{3} & 0 \\ 0 & -\frac{1}{3} \end{pmatrix}$$

Only the $a = 3$ component of the current is non-vanishing, with

$$\partial^\mu J^3_{A\,\mu} = \frac{N_c}{96\pi^2}\epsilon^{\mu\nu\rho\sigma}F_{\mu\nu}F_{\rho\sigma}$$

But this is precisely the current which, from (5.13), creates the neutral pion π^0, with $\langle 0|J^3_{A\,\mu}|\pi^0\rangle = -if_\pi\, p_\mu e^{-ix\cdot p}$. The anomaly equation then gives an amplitude for $\pi^0 \to \gamma\gamma$. This amplitude is the same as that which would arise from the coupling in the Lagrangian

$$\mathcal{L} = \frac{N_c e^2}{96\pi^2 f_\pi}\pi^0\epsilon^{\mu\nu\rho\sigma}F_{\mu\nu}F_{\rho\sigma} \tag{5.27}$$

Note that the decay amplitude is proportional to N_c, the number of colours. Comparing to the experimental data provides a way to determine $N_c = 3$. (Actually, this is a little bit quick because the $U(1)$ charge assignments above are fixed, in part, by anomaly cancellation which, as we saw in Section 3.4.4, changes if we change N_c.) Above we have used just two quarks, $N_f = 2$, but we get the same results using $N_f = 3$ if we correctly identify the current producing π^0 from within the matrix (5.22).

We have argued that the anomaly means there must be an effective coupling of the form (5.27). Yet there's something odd in this, because if we expand out the action (5.26), no such term arises. Indeed, naively this term appears to contradict the ethos of this whole section, because the Goldstone boson π^0 isn't obviously derivatively coupled, which seems very unGoldstonelike. Nonetheless, it would be nice to be able to write down a low-energy effective action that correctly captures the anomaly, rather than adding it in by hand. It turns out that there is a beautiful way to achieve this.

5.5 The Wess-Zumino-Witten Term

We have argued that, at low-energies, the dynamics of the Goldstone modes is captured by the chiral Lagrangian

$$S = \frac{f_\pi^2}{4}\int d^4x\ \mathrm{tr}(\partial_\mu U^\dagger \partial^\mu U) \tag{5.28}$$

We also briefly discussed in Section 5.2.1 the higher order terms that we could add to this action to improve its accuracy as we go to higher energies. It turns out, however, that this misses one very important term, one which, among other things, accounts for the anomaly. This is known as the Wess-Zumino-Witten term.

To motivate the need for an extra term, let's look more closely at the discrete symmetries of the chiral Lagrangian (5.28). They are:

- Charge conjugation, $C : U \mapsto U^\star$.
- "Naive parity", $P_0 : \mathbf{x} \to -\mathbf{x}$ with $t \mapsto t$ and $U \mapsto U$.
- An extra symmetry: $U \to U^\dagger$. In terms of the pion fields (5.6)

  $$U = \exp\left(\frac{2i}{f_\pi}\,\pi^a T^a\right) = 1 + \frac{2i}{f_\pi}\pi^a T^a + \ldots \tag{5.29}$$

  this symmetry acts as $\pi^a \mapsto -\pi^a$. In other words, it counts pions mod 2. For this reason, we denote the symmetry as $(-1)^{N_\pi}$ where N_π is the number of pions.

However, these are not all symmetries of the underlying QCD-like gauge theory. Indeed, the pions and other Goldstone bosons in QCD are pseudoscalars, meaning that they are odd under parity. The correct parity transformation should be

$$P = P_0(-1)^{N_\pi}$$

It is unusual – although not unheard of – to have a low-energy theory which enjoys more symmetries than its high-energy parent. It might lead us to suspect that we've missed something. Are we really sure that there are no terms that we can add to (5.28) which violate both P_0 and $(-1)^{N_\pi}$, leaving only P as a symmetry?

It is simple to look through the higher derivative terms (5.9) that we met before and convince yourself that they all preserve both P_0 and $(-1)^{N_\pi}$. Indeed, the way to get something that violates P_0 is to use the anti-symmetric tensor $\epsilon^{\mu\nu\rho\sigma}$. But if we try to form a four-derivative term in the action from this, we would have

$$\epsilon^{\mu\nu\rho\sigma}\ \mathrm{tr}\Big(U^\dagger\,(\partial_\mu U)U^\dagger\,(\partial_\nu U)U^\dagger\,(\partial_\rho U)U^\dagger\,(\partial_\sigma U)\Big) = 0 \tag{5.30}$$

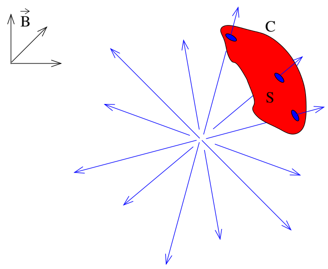


Figure 45: Integrating over S...

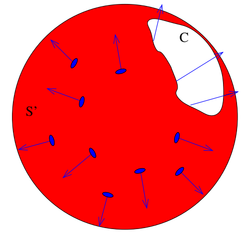


Figure 46: ...or over S'.

and, as shown, this vanishes by anti-symmetry. You can also consider higher derivative terms and see that they too preserve all these discrete symmetries. There's no way to construct terms in the action that violate P_0.

However, the story is rather different if we work with the equation of motion. The equation of motion arising from (5.28) is

$$\frac{1}{2}f_\pi^2\,\partial_\mu(U^\dagger\partial^\mu U) = 0$$

We could add to this the term

$$\frac{1}{2}f_\pi^2\,\partial_\mu(U^\dagger\partial^\mu U) = \frac{k}{48\pi^2}\epsilon^{\mu\nu\rho\sigma}\,U^\dagger\,(\partial_\mu U)U^\dagger\,(\partial_\nu U)U^\dagger\,(\partial_\rho U)U^\dagger\,(\partial_\sigma U) \tag{5.31}$$

where k is some constant which we will fix shortly and the normalisation of $48\pi^2$ is for later convenience. This is the famous Wess-Zumino-Witten term, first introduced in this context by Witten. Despite our feeble attempts above, it turns out that there is a way to write an action for this term, but not if we restrict ourselves to actions in four-dimensions!

5.5.1 An Analogy: A Magnetic Monopole

A useful analogy can be found in Dirac monopoles. This is a story that we've already met in Section 1.1. Consider a particle of mass m and unit charge moving in $\mathbf{R}^3$ in the background of a Dirac monopole. The equation of motion is

$$m\ddot{x}_i = \lambda\epsilon_{ijk}x_j\dot{x}_k$$

with λ a constant which determines the strength of the monopole. This system shares some similarities with our discussion above. First, the left-hand side is invariant under two discrete symmetries: time reversal $t \mapsto -t$ and parity $x_i \mapsto -x_i$. However, the term on the right-hand side is not separately invariant under both of these, but only if we do both at once. Furthermore, the equation of motion is invariant under $SO(3)$ rotations.

Can we construct an action for this equation of motion? If we try to do so preserving the $SO(3)$ rotational invariance, we run into trouble because the obvious term that we might try to write down to reproduce the right-hand-side is $\epsilon^{ijk} x_i x_j \dot{x}_k = 0$ by anti-symmetry. This, of course, is analogous to (5.30). However, this doesn't mean that no action exists. In fact, there are two possibilities. One is to introduce a gauge potential $A_i(x)$ and write down the action

$$S = \int_C dt \ \frac{1}{2} m\dot{x}_i^2 + \lambda \, A_i(x)\dot{x}^i$$

where C is the worldline of the particle. An example of such a gauge potential was given in (1.5). This approach has two problems: the gauge potential necessarily breaks the $SO(3)$ symmetry, which is no longer manifest in the action; and the gauge potential necessarily suffers from a Dirac string singularity.

We can circumvent both of these problems simply by using Stokes' theorem. Suppose that we take C to be a closed path. We then write

$$\int_C dt \ A_i(x)\dot{x}^i = \int_S dS^{ij} F_{ij}(x) \tag{5.32}$$

where S is a two-dimensional disc, with boundary $\partial S = C$, as shown in the figure. Now things are much nicer. The field strength $F_{ij} = \epsilon_{ijk} x^k/|x|^3$ is both $SO(3)$ invariant and, away from the origin, non-singular. However, the price that we paid is that the action is written in terms of a two-dimensional surface, rather than the one-dimensional worldline.

There is one further problem with the action (5.32) because, as we saw in Section 1.1, there is an ambiguity in the choice of surface S. There is another surface S', with the opposite orientation, that also does the job. For the path integral to be well-defined, we require that these two options give the same answer. We must have

$$\exp\left(i\lambda \int_C dt \ A_i \dot{x}^i\right) = \exp\left(i\lambda \int_S dS^{ij} F_{ij}\right) = \exp\left(-i\lambda \int_{S'} dS^{ij} F_{ij}\right)$$

Stitching together the two discs gives the closed two sphere $\mathbf{S}^2$. The condition can then be written as the requirement

$$\exp\left(i\lambda \int_{S\cup S'} dS^{ij} F_{ij}\right) = \exp\left(i\lambda \int_{\mathbf{S}^2} dS^{ij} F_{ij}\right) = 1 \tag{5.33}$$

However, the magnetic flux through any closed surface is quantised, with the minimum flux given by $\int_{\mathbf{S}^2} dS^{ij} F_{ij} = 4\pi$. We see that the path integral is consistent only if

$$\lambda \in \frac{1}{2}\mathbf{Z}$$

This is simply a restatement of the Dirac quantisation condition that we already met in Section 1.1.

5.5.2 A Five-Dimensional Action

With the discussion of the magnetic monopole fresh in our minds, let's now return to the chiral Lagrangian. We would like to ask if there is some action which respects the $SU(N_f)_L \times SU(N_f)_R$ symmetry of the chiral Lagrangian and reproduces the term on the right-hand-side of (5.31). The answer is yes, but it can only be written by invoking a fifth dimension.

We will work in the Euclidean path integral and the argument is simplest if we take our spacetime to be $\mathbf{S}^4$. We introduce a five-dimensional ball, D, such that $\partial D = \mathbf{S}^4$. We extend the fields $U(x)$ over $\mathbf{S}^4$ to $U(y)$, where y are coordinates on the ball D. We can then reproduce the equation of motion (5.31) from the action

$$S = \frac{f_\pi^2}{4} \int d^4x \ \mathrm{tr}(\partial_\mu U^\dagger \partial^\mu U) + k \int_D d^5y \ \omega \tag{5.34}$$

where

$$\omega = -\frac{i}{240\pi^2} \epsilon^{\mu\nu\rho\sigma\tau} \mathrm{tr} \left(U^\dagger \frac{\partial U}{\partial y^\mu} U^\dagger \frac{\partial U}{\partial y^\nu} U^\dagger \frac{\partial U}{\partial y^\rho} U^\dagger \frac{\partial U}{\partial y^\sigma} U^\dagger \frac{\partial U}{\partial y^\tau} \right) \tag{5.35}$$

This is the Wess-Zumino-Witten (WZW) term. There are a few things to say about this. First, it is manifestly invariant under the $SU(N_f)_L \times SU(N_f)_R$ chiral symmetry. Second, it naively appears to depend on the choice of extension of $U(x)$ to the five-dimensional space $U(y)$, but this is an illusion. The equations of motion computed from the action depend only on $U(x)$ restricted to the boundary $\mathbf{S}^4$. There are a couple of ways to see this. A somewhat involved calculation shows that the variation of ω is indeed a boundary term. Alternatively, we can expand U in the pion fields as in (5.29),

$$U^\dagger \partial_\mu U = \frac{2i}{f_\pi} \partial_\mu \pi + \mathcal{O}(\pi^2)$$

Then

$$\begin{aligned}
\int_D d^5y \ \omega &= \frac{2}{15\pi^2 f_\pi^5} \int_D d^5y \ \epsilon^{\mu\nu\rho\sigma\tau} \partial_\mu \, \mathrm{tr} \left(\pi \partial_\nu \pi \partial_\rho \pi \partial_\sigma \pi \partial_\tau \pi \right) + \mathcal{O}(\pi^6) \\
&= \frac{2}{15\pi^2 f_\pi^5} \int_{\mathbf{S}^4} d^4x \ \epsilon^{\nu\rho\sigma\tau} \, \mathrm{tr} \left(\pi \partial_\nu \pi \partial_\rho \pi \partial_\sigma \pi \partial_\tau \pi \right) + \mathcal{O}(\pi^6)
\end{aligned}$$

Written in this form, the $SU(N_f)_L \times SU(N_f)_R$ symmetry is no longer manifest. This is entirely analogous to the lack of manifest rotation symmetry in the Dirac monopole connection. Nonetheless, since it came from the term (5.34), the symmetry must be there, albeit hidden.

We see that the new term gives a five-point interaction between Goldstone modes. In the context of QCD, this mediates the decay $K^+ + K^- \to \pi^+ + \pi^- + \pi^0$, which explicitly breaks the $(-1)^{N_\pi}$ symmetry of the original chiral Lagrangian.

Quantisation of the Coefficient

Just as for the Dirac monopole, there is an ambiguity in our choice of five-dimensional ball D with $\partial D = \mathbf{S}^4$. We could just as well take a ball D', also with $\partial D' = \mathbf{S}^4$ but with the opposite orientation. We can now make the same kind of arguments that, in (5.33), gave us Dirac quantisation. We have

$$\exp\left(ik\int_D d^5y\ \omega\right) = \exp\left(-ik\int_{D'} d^5y\ \omega\right)$$

Stitching together the two five-dimensional balls now makes a five-sphere: $D \cup D' = \mathbf{S}^5$. For our path integral to make sense, we must have

$$\exp\left(ik\int_{\mathbf{S}^5} d^5y\ \omega\right) = 1 \tag{5.36}$$

By now it's probably no surprise to learn that there's some pretty topology that underlies this formula! The integrand provides a map from $\mathbf{S}^5$ to the group manifold $SU(N_f)$, parameterised by $U(y)$. Such maps are characterised by the fifth homotopy group,

$$\Pi_5(SU(N)) = \mathbf{Z} \quad \text{for } N \geq 3$$

This means that as long as we have $N_f \geq 3$ flavours, each map can be assigned a winding $n \in \mathbf{Z}$. It turns out that this winding is computed by

$$\int_{\mathbf{S}^5} d^5y\ \omega = 2\pi n$$

The quantisation condition (5.36) is then satisfied providing

$$k \in \mathbf{Z}$$

This leads us to our next question. What is k?

Rediscovering the Anomaly

The Wess-Zumino-Witten term is closely related to the chiral anomaly. This, it turns out, will give us a strategy to determine the integer k.

Here is the plan. We will gauge a $U(1)$ subgroup of $SU(N_f)_{\rm diag} \subset SU(N_f)_L \times SU(N_f)_R$. To do this, we introduce a charge matrix Q, as in (5.25), and promote the derivatives in the chiral Lagrangian to covariant derivatives

$$S = \int d^4x\, \frac{f_\pi^2}{4} {\rm tr}\, (\mathcal{D}^\mu U^\dagger\, \mathcal{D}_\mu U) + S_{WZW}$$

with $\mathcal{D}_\mu U = \partial_\mu U - ieA_\mu[Q,U]$. However, we also need to find a way to make the Wess-Zumino-Witten term gauge invariant. It's tempting to just do the same trick, and promote $\partial_\mu U$ to $\mathcal{D}_\mu U$ in (5.35). But this isn't allowed because the resulting action now depends on what's going on in five dimensions. Any gauging must take place only in four dimensions.

To proceed, we first look at how the WZW term changes under an infinitesimal transformation $\delta U = i\alpha(x)[Q,U]$, where $\alpha(x)$ depends only on the four-dimensional coordinates. We have, schematically,

$$\delta(U^\dagger \partial U) = i\alpha[Q, U^\dagger\partial U] + i\partial\alpha\, U^\dagger[Q,U]$$

The variation of the 5-form ω defined in (5.35) has terms of order $\partial\alpha^n$, with $n = 0, 1, \ldots, 5$. Of these the $n = 0$ term vanishes by cyclicity of the trace, while the $n = 2, 3, 4, 5$ terms vanish by the anti-symmetry of the $\epsilon^{\mu\nu\rho\sigma\lambda}$ symbol. After judicious use of the identity $U^\dagger\partial U = -(\partial U^\dagger)U$, we find

$$\delta\omega = (\partial_\mu\alpha)\hat{J}^\mu$$

where the current $\hat{J}^\mu$ is given by

$$\begin{aligned}
\hat{J}^\mu &= \frac{1}{48\pi^2}\, \epsilon^{\mu\nu\rho\sigma\lambda} {\rm tr}\left(\{Q, \partial_\nu U^\dagger\}\, \partial_\rho U\, U^\dagger \partial_\sigma U\, U^\dagger\partial_\lambda U\right) \\
&= \frac{1}{48\pi^2}\, \epsilon^{\mu\nu\rho\sigma\lambda} \partial_\nu {\rm tr}\left(\{Q, U^\dagger\}\, \partial_\rho U\, U^\dagger \partial_\sigma U\, U^\dagger\partial_\lambda U\right)
\end{aligned}$$

where you need to work a little bit to check that the extra terms that you get from acting with ∂_ν vanish by anti-symmetry. Because the current is a total derivative (and because $\partial\alpha$ depends only on the four-dimensional coordinates), the variation of $\int_D d^5x\ \omega$ reduces to a boundary term and, at leading order, can be cancelled by the variation of the four-dimensional gauge field $\delta A_\mu = \partial_\mu\alpha/e$. This means that we can introduce the gauged WZW term

$$S_{WZW} = k\left[\int_D d^5x\ \omega - e\int d^4x\ A_\mu(x) J^\mu\right]$$

with the four-dimensional current given by

$$J^\mu = \frac{1}{48\pi^2}\,\epsilon^{\mu\rho\sigma\tau}\mathrm{tr}\left(\{Q,U^\dagger\}\,\partial_\rho U\,U^\dagger\partial_\sigma U\,U^\dagger\partial_\lambda U\right)$$

However, it turns out that we're still not done. To get a fully gauge-invariant action, we need to work to one higher order in the gauge coupling e. Here we simply quote the result: the fully gauge invariant WZW term is given by

$$S_{WZW} = k\left[\int_D d^5x\;\omega - e\int d^4x\;A_\mu(x)J^\mu \right.$$
$$\left. + \frac{ie^2}{24\pi^2}\int d^4x\;\epsilon^{\mu\nu\rho\sigma}(\partial_\mu A_\nu)A_\rho\mathrm{tr}\left(\{Q^2,U^\dagger\}\,\partial_\sigma U + U^\dagger QUQU^\dagger\partial_\sigma U\right)\right]$$

How does this help us determine k? To see this, we need to expand out this action in terms of pion fields. For simplicity, let's do this for $N_f = 3$ quarks, with the charge matrix (5.25) appropriate for QCD. Among the order e^2 terms from above, there sits

$$\mathcal{L} = \frac{ke^2}{96\pi^2 f_\pi}\pi^0\epsilon^{\mu\nu\rho\sigma}F_{\mu\nu}F_{\rho\sigma}$$

But we've seen this before: this is the term which captures the anomaly (5.27). To agree with the anomaly, the integer k must be equal to the number of colours

$$k = N_c$$

This is a beautiful result. Until now the chiral Lagrangian has appeared to be independent of the gauge group $SU(N_c)$; all that was needed was for the gauge dynamics to initiate chiral symmetry breaking and then it seemed that it could be forgotten. We see that this isn't quite true: a memory of the underlying gauge group survives as the coefficient of the WZW term.

5.5.3 Baryons as Bosons or Fermions

We saw in section 5.3 that the chiral Lagrangian provides a lovely and surprising new perspective on baryons: they are solitons, constructed from topologically twisted pion fields. The conserved baryon current is identified with the topological current

$$B^\mu = \frac{1}{24\pi^2}\epsilon^{\mu\nu\rho\sigma}\mathrm{tr}\,(U^\dagger(\partial_\nu U)\,U^\dagger(\partial_\rho U)\,U^\dagger\,\partial_\sigma U)$$

This winding number — which we denote by $B\in\mathbf{Z}$ — is computed by the integral

$$B = \frac{1}{24\pi^2}\int d^3x\;\epsilon_{ijk}\mathrm{tr}\,(U^\dagger(\partial_i U)\,U^\dagger(\partial_j U)\,U^\dagger\,\partial_k U)$$

However, there was something lacking in our previous discussion. From the underlying quarks, we know that baryons should be bosons when N_c is even and fermions when N_c is odd. How is this basic fact reproduced in the chiral Lagrangian? Here we show that, for $N_f \geq 3$, the Wess-Zumino-Witten term is exactly what we need.

We focus on $N_f = 3$. (The story is basically unchanged for higher N_f.) Consider a static Skyrmion of the form (5.20) embedded in the $SU(3)$ matrix U as

$$U_0(\mathbf{x}) = \begin{pmatrix} U_{\rm Skyrme}(\mathbf{x}) & 0 \\ 0 & 1 \end{pmatrix}$$

We wish to compare the amplitude for two different processes to occur over some long time T. In the first process, the soliton simply sits stationary in space. In the second process, we rotate the soliton by 2π slowly about its origin. The first process has amplitude e^{iET}, where E is the energy of the soliton. We have to work a little harder to compute the amplitude for the second process. There are two contributions from the two different terms in the chiral Lagrangian (5.34). The first of these comes from the usual kinetic term. Since this involves two time derivatives, it will contribute a piece of order $\sim 1/T$ which can be ignored in the $T \to \infty$ limit. In contrast, the WZW term is linear in time derivatives and will contribute a constant piece. This is what we want.

Here we sketch the calculation. We saw in section 5.3 that the Skyrmion is invariant under a simultaneous spatial and isospin rotation. This means that we can swap our rotation in space for a flavour rotation. A suitable configuration is given by

$$U(\mathbf{x}, t) = \begin{pmatrix} e^{i\pi t/T} & & \\ & e^{-i\pi t/T} & \\ & & 1 \end{pmatrix} U_0(\mathbf{x}) \begin{pmatrix} e^{-i\pi t/T} & & \\ & e^{i\pi t/T} & \\ & & 1 \end{pmatrix}$$

We must then extend this configuration over the 5-dimensional ball D and compute the integral

$$\Gamma = -\frac{i}{240\pi^2} \int_D d^5y \; \epsilon^{\mu\nu\rho\sigma\tau} {\rm tr} \left(U^\dagger \frac{\partial U}{\partial y^\mu} \, U^\dagger \frac{\partial U}{\partial y^\nu} \, U^\dagger \frac{\partial U}{\partial y^\rho} \, U^\dagger \frac{\partial U}{\partial y^\sigma} \, U^\dagger \frac{\partial U}{\partial y^\tau} \right)$$

One finds

$$\Gamma = \pi$$

This is what we needed. It means that the amplitude for a soliton which rotates by 2π is not e^{iET} but is instead

$$e^{iET} e^{iN_c\pi} = (-1)^{N_c} e^{iET}$$

The factor of $(-1)^{N_c}$ is telling us that these solitons are bosons when N_c is even and fermions when N_c is odd.

Baryons when $N_f = 2$

When $N_f = 2$ there is no WZW term. This means that the chiral Lagrangian does not know about the underlying number of colours N_c. Nonetheless, there is a new ingredient. This arises because

$$\Pi_4(SU(2)) = \mathbf{Z}_2 \tag{5.37}$$

while $\Pi_4(SU(N)) = 0$ for $N \geq 3$. Note that this is the same homotopy group that arose in the non-perturbative anomaly described in section 3.4.3.

If we work in compactified Euclidean spacetime, then any field configuration in the chiral Lagrangian is a map from $\mathbf{S}^4$ to $SU(2)$ and so is labelled by $\nu = \pm 1$. This gives us different options for the path integral. We could either weight all configurations equally, or weight them with a factor of $(-1)^\nu$. These should be thought of as two different theories which, in analogy with section 2.2, could be said to be distinguished by a "discrete theta parameter" $\theta = 0$ or π.

Here is an example of a field configuration with $\nu = -1$: create a soliton-anti-soliton pair from the vacuum, rotate one around the other, and then annihilate them again. In the theory with $\theta = 0$ this configuration is not weighted any differently and the solitons are bosons. In the theory with $\theta = \pi$, this configuration is weighted with an extra factor of -1. Here the solitons are fermions.

We learn that in the theory with $N_f = 2$, we have a choice: we can either quantise the solitons as a boson or as a fermion. This choice arises as an extra discrete parameter which we must stipulate to fully define the path integral.

5.6 't Hooft Anomaly Matching

Until now, our strategy has been to assume that the quark condensate (5.4) forms and then explore the consequences. Our justification for the condensate itself was rather flimsy. In this section we will improve slightly on this state of affairs. While we will not give a proof that the condensate forms, we will show that it is implied by another, well-known effect of strongly coupled gauge theories: confinement. To show this, we will use the 't Hooft anomaly matching arguments of Section 3.5.

5.6.1 Confinement Implies Chiral Symmetry Breaking

By now the global symmetry group of $G = SU(N_c)$ gauge theory with N_f quarks should be very familiar: it is

$$G_F = U(1)_V \times SU(N_f)_L \times SU(N_f)_R$$

This group has a 't Hooft anomaly which, at high energies, arises from the quarks. If the theory confines, this anomaly must be reproduced by massless bound state fermions in the infra-red. The essence of the argument is that no such bound states can exist.

Let's first compute the 't Hooft anomalies in the ultra-violet, where the quarks contribute. There is no anomaly for $[U(1)_V]^3$, but there are anomalies for both $[SU(N_f)]_L^3$ and $[SU(N_f)_L]^2 \times U(1)_V$, together with the corresponding anomalies for $SU(N_f)_R$. We have

$$\begin{aligned} [SU(N_f)_L]^3 : \quad & A_1 = N_c \\ [SU(N_f)_L]^2 \times U(1)_V : \quad & A_2 = N_c \end{aligned} \tag{5.38}$$

where, in both cases, $A = N_c$ is counting the number of colours of the quarks.

What about in the infra-red? Confinement means that the quarks bind to form colour singlets. Our task is to figure out how the resulting states transform under the flavour symmetry G_F. Here the details depend on the choice of gauge group. When N_c is even, both mesons and baryons are bosons so there are no solutions to the 't Hooft anomaly conditions. This is a striking result. It tells us that there is no way to form massless bound states which match the anomaly. For the theory to be consistent, it must be that G_F is spontaneously broken in the infra-red. The simplest possibility is that the symmetry is broken down to its vector-like subgroup which is free from anomalies. This, of course, is the pattern of chiral symmetry breaking (5.5) that arises from the quark condensate.

Fermionic Baryons

When the number of colours N_c is odd the baryons are fermions. Now we have to work a little harder. Is it possible that these baryons are massless and match the anomaly? To proceed, we will restrict attention to the simplest case of

$$N_c = 3$$

The arguments that follow can be generalised to arbitrary $SU(N_c)$ gauge group.

If the 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 $U(1)_V$. The putative massless baryons therefore transform under the G_F flavour symmetry in representations given by the Young diagrams,

$$\begin{array}{|c|c|c|}\hline l & l & l \\ \hline \end{array}\ ,\quad \begin{array}{|c|}\hline l \\ \hline l \\ \hline l \\ \hline \end{array}\ ,\quad \begin{array}{|c|}\hline l \\ \hline \end{array}\otimes\begin{array}{|c|c|}\hline r & r \\ \hline \end{array}\ ,\quad \begin{array}{|c|}\hline l \\ \hline \end{array}\otimes\begin{array}{|c|}\hline r \\ \hline r \\ \hline \end{array}\ ,\quad \begin{array}{|c|c|}\hline l & l \\ \hline l & \multicolumn{1}{c}{} \\ \cline{1-1} \end{array} \tag{5.39}$$

What are the helicities of these baryons? We can take a pair of left- or right-handed fermions and form a Lorentz scalar $\epsilon^{\alpha\beta}\psi_{+\alpha}\psi_{+\beta}$ where, for once, we've explicitly written the α, β spinor indices. This means that it's possible to contract the spinor indices such that each baryon above is left-handed. Similarly, if we replace $\boxed{l}$ with $\boxed{r}$ then we have the possible set of right-handed baryons

$$\begin{array}{|c|c|c|}\hline r & r & r \\ \hline \end{array}\ ,\quad \begin{array}{|c|}\hline r \\ \hline r \\ \hline r \\ \hline \end{array}\ ,\quad \begin{array}{|c|}\hline r \\ \hline \end{array}\otimes\begin{array}{|c|c|}\hline l & l \\ \hline \end{array}\ ,\quad \begin{array}{|c|}\hline r \\ \hline \end{array}\otimes\begin{array}{|c|}\hline l \\ \hline l \\ \hline \end{array}\ ,\quad \begin{array}{|c|c|}\hline r & r \\ \hline r & \multicolumn{1}{c}{} \\ \cline{1-1} \end{array} \tag{5.40}$$

These have opposite helicity of the representations in (5.39). The $[U(1)_V]^3$ anomaly remains trivially satisfied if the spectrum of massless baryons is vector-like so we will assume that if a massless baryon of the type (5.39) arises, then its counterpart in (5.40) also arises.

Since we're dealing with a strongly coupled theory, how can we be sure that the indices are contracted so that (5.39) are left-handed and (5.40) are right-handed? First, 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 $\pm\frac{1}{2}$. But is it possible to dress these baryons with gluons which shift their helicity by ± 1?

To be on the safe side, we associate an *index*, $p_\alpha \in \mathbf{Z}$, with $\alpha = 1, \ldots, 5$ to each of the five baryons in (5.39). 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 (5.38).

Next, we need a little group theory. For a representation $\mathbf{R}$ of $SU(N_f)$, we will need to know the dimension $\dim(\mathbf{R})$, the anomaly coefficient $A(\mathbf{R})$, as well as the Dynkin index $\mu(\mathbf{R})$,

$$\operatorname{tr} T^a T^b = \frac{1}{2}\mu(\mathbf{R})\delta^{ab}$$

For each of the representations of interest, we have

$\mathbf{R}$	$\dim(\mathbf{R})$	$\mu(\mathbf{R})$	$A(\mathbf{R})$
$\Box$	N_f	1	1
$\Box\Box$	$\frac{1}{2}N_f(N_f+1)$	N_f+2	N_f+4
$\begin{smallmatrix}\Box\\\Box\end{smallmatrix}$	$\frac{1}{2}N_f(N_f-1)$	N_f-2	N_f-4
$\Box\Box\Box$	$\frac{1}{6}N_f(N_f+1)(N_f+2)$	$\frac{1}{2}(N_f+2)(N_f+3)$	$\frac{1}{2}(N_f+3)(N_f+6)$
$\begin{smallmatrix}\Box\\\Box\\\Box\end{smallmatrix}$	$\frac{1}{6}N_f(N_f-1)(N_f-2)$	$\frac{1}{2}(N_f-2)(N_f-3)$	$\frac{1}{2}(N_f-3)(N_f-6)$
$\begin{smallmatrix}\Box\Box\\\Box\end{smallmatrix}$	$\frac{1}{3}N_f(N_f^2-1)$	N_f^2-3	N_f^2-9

We can now compute the infra-red anomalies, assuming that we have p_α massless baryons of each type. For $[SU(N_f)_L]^3$ with $N_f \geq 3$, the anomaly is

$$A_1 = \frac{1}{2}(N_f+3)(N_f+6)p_1 + \frac{1}{2}(N_f-3)(N_f-6)p_2 + \left(\frac{1}{2}N_f(N_f+1) - N_f(N_f+4)\right)p_3$$
$$+ \left(\frac{1}{2}N_f(N_f-1) - N_f(N_f-4)\right)p_4 + (N_f^2-9)p_5$$

Note that the baryons with numbers p_3 and p_4 arise from tensor products and have two terms. For example, for p_3 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 $\mu(\mathbf{R})$,

$$\frac{A_2}{3} = \frac{1}{2}(N_f+2)(N_f+3)p_1 + \frac{1}{2}(N_f-2)(N_f-3)p_2 + \left(\frac{1}{2}N_f(N_f+1) - N_f(N_f+2)\right)p_3$$
$$+ \left(\frac{1}{2}N_f(N_f-1) - N_f(N_f-2)\right)p_4 + (N_f^2-3)p_5$$

To match the anomalies, we need to find p_α such that $A_1 = A_2 = 3$.

To start, let's look at $N_f = 3$. Anomaly matching gives

$$A_1 = 27p_1 - 15p_3 = 3 \quad \text{and} \quad \frac{A_2}{3} = 15p_1 - 9p_3 + 6p_5 = 1$$

We can immediately see that there can be no solutions to the second of these equations since $A_2/3$ in the infra-red theory is necessarily a multiple of 3 and cannot reproduce the ultra-violet anomaly $A_2/3 = 1$. 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:

$$\begin{aligned} A_1 &= 35p_1 - p_2 - 22p_3 + 6p_4 + 7p_5 = 3 \\ \frac{A_2}{3} &= 21p_1 + p_2 - 14p_3 - 2p_4 + 13p_5 = 1 \end{aligned}$$

Now there are solutions. For example $p_2 = 3$ and $p_4 = 1$ with $p_1 = p_3 = p_5 = 0$ does the job. This corresponds to four massless baryons in the representations

$$[3(\bar{\mathbf{4}}, \mathbf{1}) \oplus (\mathbf{4}, \mathbf{6})]_L \oplus [3(\mathbf{1}, \bar{\mathbf{4}}) \oplus (\mathbf{6}, \mathbf{4})]_R \tag{5.41}$$

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 (5.41) 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, and we will have to work harder below to justify this. But, for now, let's assume that this is true and see where it leads us.

If we give one of the quarks a mass, then the symmetry group is explicitly broken to

$$G_F = U(1)_V \times SU(4)_L \times SU(4)_R \;\longrightarrow\; G'_F = U(1)_V \times SU(3)_L \times SU(3)_R$$

What happens to our putative massless spectrum (5.41)? A little group decomposition tells us that under G'_F, the left-handed baryons transform as

$$3(\bar{\mathbf{4}}, \mathbf{1}) \to 3(\bar{\mathbf{3}}, \mathbf{1}) \oplus 3(\mathbf{1}, \mathbf{1}) \quad \text{and} \quad (\mathbf{4}, \mathbf{6}) \to (\mathbf{3}, \bar{\mathbf{3}}) \oplus (\mathbf{3}, \mathbf{3}) \oplus (\mathbf{1}, \bar{\mathbf{3}}) \oplus (\mathbf{1}, \mathbf{3})$$

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 $N_f \geq 3$.

5.6.2 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 with two vertical boxes is just the singlet of $SU(2)$, while the representation with three boxes arranged as a dogleg is the same as the fundamental. Meanwhile, $\begin{array}{|c|}\hline r \\ \hline r \\ \hline\end{array}$ is the singlet of $SU(2)$. Meanwhile, the representation with three vertical boxes does not exist. The 't Hooft matching condition for gauge group $SU(3)$ now gives

$$\frac{A_2}{3} = 10p_1 - 5p_3 + p_4 = 1$$

This has many solutions. The simplest possibility $p_1 = p_3 = 0$ and $p_4 = 1$. 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

$$(\mathbf{2}, \mathbf{1})_3 \oplus (\mathbf{1}, \mathbf{2})_3 \tag{5.42}$$

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.

But what if...

Although the lattice tells us that the chiral symmetry is broken for $N_f = 2$, it is nonetheless an interesting exercise to understand better how we could have ended up with a massless baryon. The story that we will find has a nice twist and — as we will see in Section 5.6.4 — turns out to be realised in other contexts.

To start, let's return to our calculation of the classical force between quarks. We saw in Section 2.5.1 that a quark and anti-quark attract in the singlet channel and repel in the adjoint. This played a role in our initial discussion in Section 5.1 of why a quark condensate $\langle \bar{\psi}\psi \rangle$ might form in the first place.

However, we also saw in Section 2.5.1 that the two quarks attract in the anti-symmetric channel and repel in the symmetric channel. We might wonder if it's possible to form a condensate of quark pairs, rather than quark-anti-quark pairs. Such a condensate would break the gauge group.

In more detail, for $N_c = 3$ and $N_f = 2$ the initial gauge and global group of the theory is $G = SU(3)_{\text{gauge}} \times SU(2)_L \times SU(2)_R \times U(1)_V$. The quarks transform as

$$\psi_- : \ (\mathbf{3}, \mathbf{2}, \mathbf{1})_1 \quad \text{and} \quad \psi_+ : \ (\mathbf{3}, \mathbf{1}, \mathbf{2})_1 \tag{5.43}$$

For $N_f = 2$, a condensate of quarks can take the form

$$\langle \psi^a_{+\,i} \psi^b_{+\,j} \rangle = \langle \psi^a_{-\,i} \psi^b_{-\,j} \rangle = -\epsilon^{abc} \epsilon_{ij} \sigma_c \tag{5.44}$$

Here the spinor indices are contracted so that the condensate is Lorentz invariant. The use of ϵ^{ij} means that the condensate is also invariant under the global $SU(2)_L \times SU(2)_R$ chiral symmetry. However, since the condensate σ_a transforms in the $(\mathbf{3} \otimes \mathbf{3})_{\text{anti-sym}} = \bar{\mathbf{3}}$ of $SU(3)$, it breaks the gauge symmetry

$$G = SU(3)_{\text{gauge}} \to SU(2)_{\text{gauge}}$$

where we've added the "gauge" label because the number of different $SU(2)$ groups is about to get confusing. Naively it looks like the condensate (5.44) also breaks the $U(1)_V$ symmetry, but this can be restored by combining it with a suitable $U(1) \subset SU(3)_{\rm gauge}$. For example, if we take $\sigma_c = \sigma\delta_{1c}$ then the generator

$$Q'_V = Q_V + \text{diag}(2,-1,-1)_{\rm gauge}$$

is unbroken and commutes with $SU(2)_{\rm gauge}$. This means that, at low-energies, our theory has the symmetry

$$G' = SU(2)_{\rm gauge} \times SU(2)_L \times SU(2)_R \times U(1)'_V$$

How do the quarks (5.43) transform under G'? A little bit of representation decomposition shows

$$\psi_- :\ (\mathbf{1},\mathbf{2},\mathbf{1})_3 \oplus (\mathbf{2},\mathbf{2},\mathbf{1})_0 \quad \text{and} \quad \psi_+ :\ (\mathbf{1},\mathbf{1},\mathbf{2})_3 \oplus (\mathbf{2},\mathbf{1},\mathbf{2})_0$$

The existence of the condensate can be thought of as giving mass to the fermions that sit in the $\mathbf{2}$ of $SU(2)_{\rm gauge}$. (Note that, as in the condensate (5.44), we can form a singlet from $\mathbf{2}\otimes\mathbf{2}$ so there's no problem with either gauge invariance nor chiral symmetry.) But those fermions that are singlets under $SU(2)_{\rm gauge}$ are protected from getting a mass by the surviving $U(1)'_V$ chiral symmetry. The curious fact is that these massless fermions sit in precisely the representations (5.42) which satisfy 't Hooft anomaly matching.

There's something rather odd about this. In the 't Hooft anomaly matching argument, we assumed that the theory confines and looked for massless baryons – composites of three underlying quarks. In the analysis above, however, we proposed that the quark condensate Higgses the gauge group and the massless fermion is just a single quark, albeit with $U(1)'_V$ charge $+3$.

In fact, these are two different ways of looking at the same underlying physics. In the presence of the condensate (5.44), the vacuum is filled with pairs of quarks which can mix with the lone massless quark to form the composite baryon. Moreover, as we saw in Section 2.7.3, when we have a scalar in the fundamental representation — here played by the condensate $\psi\psi$ — there is no distinction between the Higgs and confining phases. The two descriptions — in terms of massless baryons or in terms of a condensate Higgs field — use different words, but are telling us the same thing. This situation sometimes goes by the rather pretentious name of *complementarity* (a much overused word in physics, and one which is possibly better saved for other, more subtle, phenomena).

Figure 47: Two possible behaviours for the baryon mass. The Vafa-Witten theorem rules out the second option.

As we mentioned above, it appears that the scenario sketched here doesn't occur for QCD-like theories with $N_f = 2$, presumably because the condensate which breaks chiral symmetry is preferred for more subtle, dynamical reasons. Nonetheless, something similar does happen for chiral gauge theories.

5.6.3 The Vafa-Witten-Weingarten Theorems

To invoke the full power of 't Hooft anomaly matching, we needed to assume that any baryon that contains a massive quark is itself massive. This is not at all obvious in a strongly interacting theory of the kind we're dealing with. When the mass of the quark is very large, $m \gg \Lambda_{\rm QCD}$, it is certainly true that the baryon must be massive. But for small quark masses $m \ll \Lambda_{\rm QCD}$, we could well imagine a situation where the binding energy cancels the quark mass, resulting in a massless bound state that contains massive constituents.

Two possibilities are depicted in Figure 47. The first shows the mass of the baryon increasing monotonically with the constituent quark mass. This is the scenario that we assumed above. The second figure shows another plausible scenario: the baryon remains massless for some finite value of the quark mass, before the theory undergoes some kind of phase transition at $m = m_\star$. If this were to happen, it would nullify our previous conclusions.

Fortunately, the second scenario cannot happen. It is ruled out by a theorem due to Vafa and Witten. In fact, there are a number of such theorems, all of which have rather similar proofs. We prove four such theorems below, the first two due to Vafa and Witten, the second two due to Weingarten. As we will see, the second Vafa-Witten theorem can be invoked to rule out the scenario shown in the second figure.

A Positive Definite Measure

Our setting is the QCD-like theories discussed throughout this chapter. All the theorems that we will prove rely on the same property of the path integral: a positive definite measure.

When computing correlation functions of gauge invariant operators, say $\mathcal{O}(x)$, we need to do the path integral. In Euclidean space, this takes the form

$$\langle \mathcal{O}(x) \ldots \mathcal{O}(y) \rangle = \frac{1}{Z} \int \mathcal{D}A \prod_{i=1}^{N_f} \mathcal{D}\bar{\psi}_i \, \mathcal{D}\psi_i \, e^{-S_{YM} + \sum \bar{\psi}_i (\not{\partial} + m)\psi_i} \, \mathcal{O}(x) \ldots \mathcal{O}(y)$$

Here S_{YM} is the usual Yang-Mills action. For simplicity, we've given each quark a common mass, m which we take to be positive: $m > 0$. Clearly it would be simple to generalise this. For some applications below, we'll explore the chiral limit by taking $m \to 0$. In practice, we should also include gauge fixing terms in this expression, but these don't affect the discussion below so we omit them for simplicity.

It is straightforward to do the fermionic path integral, leaving us with the path integral over the gauge fields. This takes the form

$$\langle \mathcal{O}(x) \ldots \mathcal{O}(y) \rangle = \frac{1}{Z} \int \mathcal{D}A \, e^{-S_{YM}} \left[\det(\not{D} + m) \right]^{N_f} \mathcal{O}(x) \ldots \mathcal{O}(y)$$

For many applications of interest, the operators $\mathcal{O}$ will also depend on the fermions. In this case, any fermion bi-linear should be replaced by its propagator in the usual manner. We'll see examples below.

We see that the effect of the fermions is to change the measure of the path integral over the gauge field. We write the correlation functions as

$$\langle \mathcal{O}(x) \ldots \mathcal{O}(y) \rangle = \int d\mu \, \mathcal{O}(x) \ldots \mathcal{O}(y)$$

where all the trickiness has now been absorbed in the measure

$$d\mu = \frac{1}{Z} \int \mathcal{D}A \, e^{-S_{YM}} \left[\det(\not{D} + m) \right]^{N_f} \tag{5.45}$$

The key observation is that this measure is positive definite. This is clearly true for the Yang-Mills part of the action, with $S_{YM} = \frac{1}{2g^2} \int \mathrm{tr}\, F^{\mu\nu} F_{\mu\nu}$. But it's also true for the Dirac operator. This is because QCD is a vector-like theory. Suppose that, for a choice of gauge field A_μ, the Dirac operator has a non-zero eigenvalue $\lambda \in \mathbf{R}$, so there is an eigenspinor

$$i\not{D}\psi = \lambda\psi$$

Then we also have an eigenvalue $-\lambda$. This follows because $\{\gamma^5, \not{D}\} = 0$, so

$$i\not{D}(\gamma^5\psi) = -i\gamma^5\not{D}\psi = -\lambda\gamma^5\psi$$

Of course, there may also be some number, n, of zero modes of $\not{D}$. The general form of the determinant is then

$$\det(\not{D} + m) = m^n \prod_\lambda (m - i\lambda)(m + i\lambda) = m^n \prod_\lambda (m^2 + \lambda^2) \tag{5.46}$$

which is manifestly positive definite providing $m > 0$. Before we go on, it's worth pausing to make a couple of comments.

- It's important that we set the theta angle to zero, $\theta = 0$, for the following arguments. This is because the theta term comes with an $\epsilon^{\mu\nu\sigma\rho}$ symbol,

$$S_\theta = \frac{\theta}{32\pi^2} \int \mathrm{tr}\epsilon^{\mu\nu\rho\sigma} F_{\mu\nu} F_{\rho\sigma} \tag{5.47}$$

  and so, when Wick rotated to Euclidean space, appears in the path integral as e^{iS_θ}. That extra factor of i means, when $\theta \neq 0$, the measure is not positive definite.

- Relatedly, the mass should be positive $m > 0$. We can see this explicitly in the contribution from the n zero modes in (5.46). But it's simpler to note that, from the chiral anomaly discussed in Section 3.3.3, a negative mass can be viewed as a non-zero θ angle.

- Clearly we needed the fermions to sit in a vector-like representation of the gauge group to argue for a positive definite measure. This means that many of the arguments we will make below fail in chiral gauge theories.

Let's now see what a positive definite measure buys us.

Theorem 1: Parity is not Spontaneously Broken

Parity is a symmetry of QCD. One might wonder if it remains a symmetry of the ground state. The spontaneous breaking of parity would show up as a non-zero expectation value for some parity odd scalar operator, $\mathcal{O}(x)$ which plays the role of an order parameter. We will argue that, in QCD, we necessarily have

$$\langle \mathcal{O} \rangle = 0$$

for any parity odd scalar.

To see this, consider the QCD Lagrangian deformed by the addition of this parity odd scalar.

$$\mathcal{L}(\alpha) = \mathcal{L}_{QCD} + \alpha\mathcal{O}$$

To leading order in α, the energy density of the ground state is

$$E(\alpha) = E(0) + \alpha\langle\mathcal{O}\rangle$$

If parity is spontaneously broken in QCD, then there are two ground states and $\langle\mathcal{O}\rangle$ is positive in one ground state and negative in the other. This means that spontaneous breaking of parity implies that $E(\alpha) < E(0)$ for arbitrarily small α.

Let's now calculate $E(\alpha)$ in the path integral. We have

$$e^{-VE(\alpha)} = \int d\mu\; e^{i\alpha\int d^4x\,\mathcal{O}}$$

where V is the volume of (Euclidean) spacetime. The important point is the factor of i in the exponent. This arises in Euclidean space only for parity odd operators because they necessarily come with an odd number of $\epsilon_{\mu\nu\rho\sigma}$. Indeed, we already saw an example of this with the θ term (5.47). We learn that adding a parity odd operator to the action changes the path integral by a phase. Because the measure is positive definite, this phase can only decrease the value of the path integral, so

$$e^{-V(\alpha)} \leq e^{-VE(0)} \quad \Rightarrow \quad E(\alpha) \geq E(0)$$

We learn that the energy density has a minimum at $\alpha = 0$ which, in turn, tells us that parity is not spontaneously broken in vector-like theories.

As a side remark, if we apply the argument above to the theta term itself (5.47), we learn that the addition of a theta term necessarily increases the energy of the vacuum: $E(\theta) \geq E(0)$. This observation sits at the heart of axion attempts to explain why the QCD theta angle is so small in our world.

Theorem 2: A Bound on Current-Current Correlation Functions

We now turn to the promised result: a relation between the masses of bound states and the bare masses of the underlying quarks. To proceed, we're going to consider two point functions of currents. We will take

$$J^a_\mu = \bar{\psi}_i\gamma_\mu(T^a)^{ij}\psi_j$$

where T^a is some $SU(N_f)$ flavour generator. In terms of the path integral, the two point function can be written as

$$\langle J^a_\mu(x)\, J^{b\,\dagger}_\nu(y)\rangle = \int d\mu \ \mathrm{tr}\,\left(\gamma_\mu T^a S(x,y)\gamma_\nu T^b S(y,x)\right)$$

where the trace is over spinor and flavour indices, and the propagator takes the form

$$S(x,y) = \langle x|\frac{1}{\not{D}+m}|y\rangle \tag{5.48}$$

Note that this is the propagator evaluated in the background of a fixed gauge field A_μ. The hard part is to then integrate over all gauge configurations, a procedure that is swept into the innocuous looking $\int d\mu$. Of course, we're not going to be able to do this integral. But we will be able to make remarkable progress simply from the knowledge that the measure is positive definite.

We will first give a slightly rough outline of the result, together with an explanation of why it shows what we want. We will then proceed with the proof and, along the way, see a number of further subtleties that we have to address.

The basic idea is to first fix A_μ. We would then like to invoke an inequality along the lines of

$$|S(x,y)| \leq C\, e^{-m|x,y|} \tag{5.49}$$

where m is the bare mass of the quark, C is some constant, and $|S(x,y)|$ refers to the matrix norm with respect to spinor and flavour indices. Crucially, this inequality must hold for any background gauge field A_μ, with the constants C and m independent of A_μ. In other words, it should be a *uniform* bound.

Such a uniform bound survives when averaged over all gauge fields with a positive definite measure. This then gives us a bound on the correlation function that we're interested in,

$$\langle J^a_\mu(x)\, J^{b\,\dagger}_\nu(y)\rangle < C' e^{-2m|x-y|} \tag{5.50}$$

What is the interpretation of such a bound? Suppose that the lightest particle carrying the flavour quantum numbers of the current has mass M. Then, at large distances, we would expect the current-current correlation function to be dominated by exchange of this particle, meaning that

$$\langle J^a_\mu(x)\, J^{b\,\dagger}_\nu(y)\rangle \sim e^{-M|x-y|}$$

The bound above tells us that the physical mass of the particle is bounded from below

$$M \geq 2m \tag{5.51}$$

where m is the bare mass of the quarks.

There are two immediate consequences of this result:

- First, it rules out the possibility of massless bosons. This is important because an equal bare mass for all the quarks breaks the axial symmetry, but leaves behind the vector $SU(N_f)_V$ flavour symmetry. The result (5.51) tells us that this vector flavour symmetry cannot be spontaneously broken, for it it were we would have massless Goldstone bosons with $M = 0$.

  We learn that the vector flavour symmetry is not spontaneously broken when the quarks have a bare mass. But if it's not spontaneously broken for any $m > 0$, then it can only become spontaneously broken in the limit $m \to 0$ if there is some miraculous accidental degeneracy, where a Lorentz invariant excited state decreases its energy, becoming exactly degenerate with the ground state at $m = 0$. This seems implausible. Under the assumption that no accidental degeneracy of this kind occurs, the Vafa-Witten theorem shows that vector-like symmetries are not spontaneously broken.

- Secondly, the Vafa-Witten theorem rules out the existence of massless fermions when the bare mass of the quarks are non-vanishing. This, of course, was what we wanted to prove.

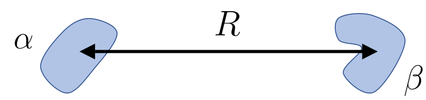
  Here, however, things are a little less straightforward and there is a subtlety that should be stressed. The calculation above holds in the presence of a finite UV cut-off, Λ. This was left implicit in the derivation, but the presence of the bare masses m in the inequality (5.51) is the hint that there is an underlying cut-off in the game. The Vafa-Witten theorem tells us that, for $m \neq 0$, there can be no massless composite fermion carrying flavour quantum numbers for any finite Λ. However, it does not rule out the possibility that a massless fermion emerges as $\Lambda \to \infty$ and the cut-off is removed. Indeed, we know that it is only in this limit that anomalies kick in so, strictly speaking, 't Hooft anomaly matching only requires the existence of massless fermions in the $\Lambda \to \infty$ limit. For QCD, it is not believed that such behaviour happens. But the Vafa-Witten theorem isn't as watertight as we might hope in showing this.

A Proof of Theorem 2

Let's now prove the Vafa-Witten theorem (5.50). The trick is not to work with the propagator (5.48) between position eigenstates $|x\rangle$ and $|y\rangle$, but instead to work with a smeared propagator

$$S(\alpha,\beta) = \langle\alpha|\frac{1}{\not{D}+m}|\beta\rangle$$

where $|\alpha\rangle$ and $|\beta\rangle$ are wavepackets that have support only in localised regions, separated by a distance R as shown below:



The "localised support" means that $A_\mu(x)|\alpha\rangle = 0$ for x outside of the region α, and similar for $|\beta\rangle$. We'll soon see the advantage of working with these smeared propagators.

To proceed, we use a standard trick to rewrite the propagator as

$$S(\alpha,\beta) = \int_0^\infty dt\ \langle\alpha|e^{-(\not{D}+m)t}|\beta\rangle = \int_0^\infty dt\ e^{-mt}\langle\alpha|e^{-i(-i\not{D})t}|\beta\rangle$$

Here t is just an artificial parameter that we've used to rewrite the integral. The next step is the clever one: we reinterpret t as a genuine time direction for a theory in $d = 4+1$ dimensions with Hamiltonian $H = -i\not{D}$. By causality, we know that a signal from region α takes at least time $t = R$ to reach the region β. This means that we must have

$$\langle\alpha|e^{-iHt}|\beta\rangle = 0 \quad \text{ for } 0 \leq t \leq R$$

Furthermore, at later times we can simply use the Cauchy-Schwarz inequality to bound

$$|\langle\alpha|e^{-iHt}|\beta\rangle| \leq \sqrt{\langle\alpha|\alpha\rangle}\,\sqrt{\langle\beta e^{+iHt}|e^{-iHt}\beta\rangle} = \sqrt{\langle\alpha|\alpha\rangle}\,\sqrt{\langle\beta|\beta\rangle}$$

where, in the second equality, we have used the unitarity of e^{iHt}. This then gives us the promised uniform bound on the propagator

$$|S(\alpha,\beta)| \leq\ |\alpha|\,|\beta| \int_R^\infty dt\ e^{-mt} = \frac{|\alpha|\,|\beta|}{m}e^{-mR} \tag{5.52}$$

This is more or less the result that we wanted. It's not quite the advertised bound on the propagator (5.49) because it uses wavepackets rather than position eigenstates. Nonetheless, it's just as good for the purposes of proving what we want. The derivation also makes it clear why we needed smeared wavepackets; it's because their norm $|\alpha|$ and $|\beta|$ appear explicitly in the bound. In contrast, position eigenstates aren't normalisable and so don't work for our purposes.

Theorem 3: The Pion is the Lightest Meson

There are yet more applications of the positive definite measure. These are inequalities between the masses of various physical particles, first introduced by Weingarten. The first of these says that the pion is the lightest meson.

We start by introducing the pseudoscalar meson field

$$\pi = \bar{\psi}_i \gamma^5 \psi_j$$

where we have picked some $i \neq j$. In QCD, we would most naturally pick $i =$ up quark and $j =$ down quark, so that π is identified with the genuine pion. Here, we'll refer to π as the pion for any i and j. We give all quarks the same mass, $m > 0$, so that the $SU(N_f)_V$ vector symmetry is unbroken. The propagator of the pion is then

$$\langle \pi(x)\pi^\dagger(y)\rangle = \int d\mu \ \mathrm{tr}\left[S(x,y)\,\gamma^5\,S(y,x)\,\gamma^5\right]$$

where $d\mu$ is the usual positive-definite measure (5.45), $S(x,y)$ is the fermion propagator introduced in (5.48) and where the trace is over spinor and colour indices. Now, because $\{\gamma^5,\, \not{D}\} = 0$, we have

$$\gamma^5 S(y,x)\gamma^5 = \gamma^5\,\langle y|(\not{D}+m)^{-1}|x\rangle\,\gamma^5 = \langle y|(-\not{D}+m)^{-1}|x\rangle = \langle x|(\not{D}+m)|y\rangle^\dagger \quad (5.53)$$

where the final equality follows because $\not{D}$ is anti-Hermitian. Note that the $|x\rangle$ and $|y\rangle$ labels got swapped as part of taking the Hermitian conjugate; the remaining $\dagger$ acts on colour and spinor indices. But this means that we have

$$\langle \pi(x)\pi^\dagger(y)\rangle = \int d\mu \sum_{\text{colour,spinor}} |S(x,y)|^2 \geq 0 \quad (5.54)$$

We learn that the pion propagator is positive definite. Note that the presence of the γ^5 matrix was crucial to make this claim, since it gave us the Hermitian conjugate in (5.53).

Let's now contrast this with the propagator for a scalar meson that doesn't include the γ^5. We have

$$\sigma = \bar{\psi}_i \psi_j$$

where we take $i \neq j$ to be the same indices as those carried by the pion. Repeating the arguments above, we now get

$$\langle \sigma(x)\sigma^\dagger(y)\rangle = \int d\mu \ \mathrm{tr}\left[S(x,y)\,S(y,x)\right] = \int d\mu \ \mathrm{tr}\left[S(x,y)\gamma^5 S(x,y)^\dagger \gamma^5\right]$$

This means that we're again summing over $|S(x,y)|^2$, but this time with different plus and minus signs for different spinor indices, coming from the presence of γ^5 matrices in the final expression. We learn that we necessarily have

$$\langle \sigma(x)\sigma^\dagger(y)\rangle \;\leq\; \langle \pi(x)\pi^\dagger(y)\rangle$$

But, at large distances, we expect each of these correlation functions to be dominated by the mass of the corresponding meson (or the mass of the lightest particle carrying the same quantum numbers). This means that the inequality above becomes, for large $|x-y|$,

$$e^{-m_\sigma|x-y|} \leq e^{-m_\pi|x-y|} \quad \Rightarrow \quad m_\sigma \geq m_\pi$$

This, of course, holds in our world because the pion is a Goldstone boson for broken chiral symmetry. Indeed, the mass inequality above can, like the Vafa-Witten theorem, be used to argue against the vector-like symmetry being broken, for then the σ meson would be a massless Goldstone boson. The result above says that this can't happen at finite m where the pion is massive, and so the sigma meson must also be massive.

It is straightforward to repeat the arguments above with a different gamma matrix structure. For example, we could look at vector mesons of the form $\rho = \bar{\psi}\gamma^\mu\psi$ and show that these too are heavier than the pions.

Theorem 4: Baryons are Heavier than Pions

The second Weingarten inequality bounds the mass of the baryon. For QCD, with three colours, the baryon takes the form

$$B = \epsilon^{abc}\psi^i_a\psi^j_b\psi^k_c$$

Here i,j and k are flavour indices and a,b and c are colour indices. The spinor indices are left implicit; they could be contracted to form a spin-$\frac{1}{2}$ baryon, or uncontracted for a spin-$\frac{3}{2}$ baryon. As we've seen, two-point correlation function takes the form

$$\langle B(x)B^\dagger(y)\rangle \sim e^{-m_B|x-y|}$$

where m_B is the mass of the lightest baryon sharing the quantum numbers of B. As previously, we take the bare masses of all quarks to be $m>0$. We then have the expression

$$\langle B(x)B^\dagger(y)\rangle = \epsilon^{abc}\epsilon^{a'b'c'}\int d\mu\ \mathrm{tr}\left[S(x,y)_{aa'}S(x,y)_{bb'}S(x,y)_{cc'}\right]$$

where $d\mu$ is again the positive-definite measure (5.45) and this time we've kept the colour indices explicit. First, we use the Cauchy-Schwarz inequality to bound

$$\langle B(x)B^\dagger(y)\rangle \leq \int d\mu\, \left(\sum_{a,a',\text{spinor}} |S(x,y)_{aa'}|^2\right)^{3/2}$$

Suppose that we could argue that, for any choice of background gauge field,

$$|S(x,y)| \leq C' e^{-m|x-y|} \tag{5.55}$$

with C' a constant, independent of the choice of gauge field, and m the bare mass of the quark. In this case, we would immediately have

$$\langle B(x)B^\dagger(y)\rangle \;\leq\; C'e^{-m|x-y|}\int d\mu \sum_{\text{colour,spinor}} |S(x,y)|^2 = C'e^{-m|x-y|}\,\langle \pi(x)\pi^\dagger(y)\rangle$$

where, in the second equality, we've used our previous expression for the pion propagator (5.54). Now, recall from the proof of the Vafa-Witten theorem that we don't quite have (5.55), but we have something almost as good: we need to replace the position eigenstates $|x\rangle$ and $|y\rangle$ with smeared wavepackets $|\alpha\rangle$ and $|\beta\rangle$ and we can then derive the uniform bound (5.52). This will do for our purposes; we therefore come to the conclusion that $m_B \geq m_\pi + m$. In the limit that the bare mass vanishes, so $m \to 0$, we learn that

$$m_B \geq m_\pi$$

Of course, this is hardly groundbreaking information given what we know about particle physics. But here it is derived from first principles, with no assumption of chiral symmetry breaking. Moreover, it tells us that if we wish chiral symmetry to be unbroken, with the 't Hooft anomaly saturated by massless baryons, then the pion must also be massless. But this seems very unlikely, since if the pion is massless then there is nothing to stop it condensing and breaking the chiral symmetry after all.

5.6.4 Chiral Gauge Theories Revisited

The existence of a global symmetry with a 't Hooft anomaly guarantees the existence of massless particles in the spectrum. If the symmetry is spontaneously broken, we have Goldstone bosons. If the symmetry is unbroken, we have massless fermions whose presence is needed to reproduce the anomaly.

So far, we have discussed situations in which 't Hooft anomaly matching ensures the existence of massless bosons (together with the case of $N_f = 2$ where anomaly matching is ambivalent, but bosons arise anyway). Here we describe situations where massless fermions arise. Perhaps unsurprisingly, this typically happens in chiral gauge theories where tree-level fermion masses are prohibited by the gauge symmetry.

We will focus on one of the simplest chiral gauge theories,

$$G = SU(5) \text{ with two Weyl spinors: } \psi_a \text{ in the } \bar{\mathbf{5}} \text{ and } \chi^{ab} \text{ in the } \mathbf{10}$$

Here $a, b = 1, \ldots, 5$ are the gauge group indices. The classical theory has two global symmetries: $U(1)_\psi$ and $U(1)_\chi$, each rotating the phase of a single fermion. One combination of these suffers a mixed anomaly with $SU(5)$. The surviving generator is

$$Q = 3Q_\psi - Q_\chi$$

This has a 't Hooft anomaly

$$A = \sum_{\text{fermions}} Q^3 = 5 \times 3^3 + 10 \times (-1)^3 = 125$$

Let us now suppose that the theory confines, leaving the $U(1)_Q$ unbroken. The simplest colour singlet is the 3-fermion bound state

$$\psi_a \psi_b \chi^{ab} \tag{5.56}$$

This has charge $Q = 5$, giving an infra-red contribution to the 't Hooft anomaly

$$A = 5^3 = 125$$

We see that it is plausible that this fermion bound state does indeed remain massless.

A Different Perspective

We can reach the same conclusion through a rather different argument. Suppose that a fermion bi-linear forms a condensate. Since any such bilinear is necessarily charged under the gauge group, the condensate will partially Higgs the gauge symmetry. What symmetry breaking patterns occur?

This is not completely straightforward. We can make a number of different fermion bilinears, each decomposing into some number of channels. Based on the computation of the classical force between quarks described in section 2.5.1, some of these channels will be attractive and some repulsive. It seems likely that the condensate forms in an attractive channel, but there are several of these.

At this point, we need to use a little guesswork. The most naive approach is to determine which quark pair has the most attractive force and assume that the condensate forms in this channel. This is clearly optimistic — after all, we're dealing with a strongly coupled theory and the classical force calculation is unlikely to provide quantitative guidance — but does give sensible answers in many cases. It is known as the *maximally attractive channel* criterion. More generally in these situations, one tries different possibilities and sees which outcomes seem the least baroque. Note that, in contrast to the QCD-like theories, we cannot turn to the lattice for help because there are various obstacles to discretising chiral fermions.

For the problem in hand, it is thought that the naive, most-attractive channel hypothesis does give rise to the correct physics. In fact, there are two channels which are equally attractive. These are:

$$\mathbf{5} \subset \bar{\mathbf{5}} \otimes \mathbf{10} \quad \text{and} \quad \bar{\mathbf{5}} \subset \mathbf{10} \otimes \mathbf{10}$$

We therefore postulate the existence of two quark condensates

$$\langle \psi_a \chi^{ab} \rangle = \sigma^b \quad \text{and} \quad \langle \chi^{ab} \chi^{cd} \rangle = \epsilon^{abcde} \Delta_e \tag{5.57}$$

These two condensates are not gauge invariant. Between them, they could break the $SU(5)$ gauge group to either $SU(4)$ (if they lie parallel to each other) or $SU(3)$. Again, we have to engage in a little guesswork. We will assume that they line up, with $\sigma^a = \sigma \delta^{a1}$ and $\Delta_a = \Delta \delta_{a1}$. The gauge group is then broken to

$$G = SU(5)_{\text{gauge}} \rightarrow SU(4)_{\text{gauge}}$$

Naively, each of the condensates breaks the non-anomalous $U(1)$ global symmetry, with $Q(\sigma) = 2$ and $Q(\Delta) = -2$. However, as in the previous section, we can define a new, unbroken global symmetry by mixing the $U(1)$ with a suitable generator of the $SU(5)$ gauge symmetry,

$$Q' = Q - \frac{1}{2}\text{diag}(4, -1, -1, -1, -1)$$

At low-energies, the gauge and global symmetry groups are

$$G = SU(4)_{\text{gauge}} \times U(1)'$$

Decomposing each fermion into representations of this new group, we have

$$\psi : \ \bar{\mathbf{5}}_3 \ \rightarrow \ \bar{\mathbf{4}}_{5/2} \oplus \mathbf{1}_5 \quad \text{and} \quad \chi : \ \mathbf{10}_{-1} \ \rightarrow \ \mathbf{6}_0 \oplus \mathbf{4}_{-5/2}$$

The $\langle \psi \chi \rangle$ condensate in (5.57) gives mass to $\bar{\mathbf{4}}_{5/2} \otimes \mathbf{4}_{-5/2}$, while the $\langle \chi \chi \rangle$ condensate gives mass to $\mathbf{6}_0 \otimes \mathbf{6}_0$. This leaves us with the gauge singlet $\mathbf{1}_5$. This has the same quantum numbers as the massless composite fermion (5.56) that we anticipated by 't Hooft anomaly matching.

Although we've had to engage in some guesses along the way, we end up with a plausible situation: the low energy dynamics of the chiral $SU(5)$ theory consists of a single, free Weyl fermion. This can either be viewed as a composite fermion (5.56) in a confining theory, or as a fundamental fermion in a theory with quark condensates (5.57): the end result is the same.

We could also ask if there are other possibilities which look equally plausible. For example, is it possible that the global $U(1)_Q$ is spontaneously broken, resulting in a massless boson instead of a massless fermion? For this to happen, we need to construct a bosonic, gauge invariant condensate. The simplest contains six fermions — $\psi_a \psi_b \chi^{ab} \psi_c \psi_d \chi^{cd}$ — and it seems unlikely that such a condensate would form.

More Chiral Gauge Theories

The $SU(5)$ gauge theory described above is not the only one which is thought to confine, giving massless composite fermions. Indeed, the same behaviour is thought to occur for the two classes of chiral gauge theories introduced in Section 3.4.2. The first of these is:

$$G = SU(N) \text{ with a } \begin{smallmatrix}\Box\\ \Box\end{smallmatrix} \text{ and } N-4\ \overline{\Box} \text{ Weyl fermions}$$

We denote the $N-4$ fermions in the anti-fundamental representation as ψ and the fermion in the anti-symmetric as χ. The theory has a $SU(N-4) \times U(1)$ global symmetry, where the $SU(N-4)$ factor rotates the ψ fields, while the non-anomalous $U(1)$ charges are given by

$$Q_\psi = N-2 \quad \text{and} \quad Q_\chi = 4-N$$

Assuming that this theory confines, the question is: what becomes of this global symmetry. As we have seen, if it is to survive unscathed then there must be a massless, composite fermion that reproduces the 't Hooft anomaly. A candidate is the collection of 3-fermion bound states that, schematically, take the form $\lambda = \psi\chi\psi$. Displaying all the indices, this is

$$(\lambda_\alpha)_{ij} = \psi_{\beta a i}\, \epsilon^{\beta\gamma}\, \chi_\gamma{}^{ab} \psi_{\alpha b j} \tag{5.58}$$

where $\alpha, \beta, \gamma = 1, 2$ are spinor indices, $i, j = 1, \ldots, N-4$ are $SU(N-4)$ flavour indices, and $a, b = 1, \ldots, N$ are $SU(N)$ gauge indices. If you track through all the symmetry properties, you'll find that λ_{ij} is symmetric in ij, so this spinor transforms in the symmetric $\Box\Box$ representation of the $SU(N-4)$ global symmetry group. It also has charge $Q = N$. It is not hard to check that these massless fermions λ do indeed saturate the 't Hooft anomalies, and therefore provide a good candidate for the infra-red physics of this theory.

As with the $SU(5)$ model, there is also a complementary approach to deriving the same result in which one first assumes that fermi bilinears $\langle\chi\psi\rangle$ and $\langle\chi\chi\rangle$ condense, breaking the gauge group $SU(N) \rightarrow SU(4)$, with all fermions pairing up except for a lone λ with the same quantum numbers that we saw above.

The second chiral gauge theory that we met earlier is similar, but has quarks in the symmetric rather than anti-symmetric representation

$$G = SU(N) \text{ with a } \Box\!\Box \text{ and } N+4\ \overline{\Box}$$

This time there is a global $SU(N+4)$ symmetry, together with a single non-anomalous $U(1)$ under which the anti-fundamental fermions ψ and the symmetric fermion χ have charges

$$Q_\psi = N+2 \quad \text{and} \quad Q_\chi = -(N+4)$$

Once again, it seems plausible that the theory confines without breaking the $SU(N+4) \times U(1)$ global symmetry, with the 't Hooft anomalies saturated by a fermion (5.58). Tracking through the symmetrisation, this time λ sits in the anti-symmetric $\begin{smallmatrix}\Box\\ \Box\end{smallmatrix}$ representation of the global symmetry group $SU(N+4)$, again with charge $Q = N$. A few short calculations show that the 't Hooft anomalies do indeed match.

5.7 Further Reading

Spontaneous symmetry breaking is a powerful and unifying idea, explaining disparate phenomena in both particle physics and condensed matter physics. It is responsible for the existence of phonons in a solid and, as we have seen, the existence of pions in the strong force. When implemented in gauge theories, it provides a unified explanation for superconductivity and the electroweak vacuum.

Jeffery Goldstone was the first to realise that a spontaneously broken global symmetry gives rise to a massless particle – what we now call the Goldstone boson. He made this conjecture, and provided examples, in a 1961 paper whose title – "Field theories with Superconductor Solutions" – reveals the early cross-fertilisation between condensed matter and particle physics [78]. The general proof of the theorem followed soon afterwards in a paper with Salam and Weinberg [79].

Goldstone's theorem was initially viewed with some dismay in particle physics. The existence of strictly massless bosons was ruled out by experiment, suggesting that spontaneous symmetry breaking had little role to play at the fundamental level. This, of course, was too hasty. Subsequent work by Higgs and others, exploring symmetry breaking in gauge theories, provided the underpinning for the Standard Model. Meanwhile, it was realised that an approximate global symmetry could be spontaneously broken, resulting in an approximate Goldstone boson. (The name pseudo-Goldstone boson was coined by Weinberg, apparently to Jeffrey's annoyance.)

The discovery of what we would now call chiral symmetry was actually made slightly before Goldstone's insight. In 1960, Yoichiro Nambu explained that an exact axial-vector current in beta decay would imply the existence of a massless pion field [140]. Like many papers of the time, it avoids the language of field theory and instead focusses on the "current algebra", in which one works with commutation relations between currents and their matrix elements. This somewhat masks the connection to spontaneous symmetry breaking, which is not emphasised in the paper. This was one of the (many!) contributions for which Nambu was awarded the 2008 Nobel prize.

A more modern formulation of the chiral Lagrangian came only in the mid-1960s. Gell-Mann and Levy introduced the sigma model [72]. In fact, they introduced two versions: the first is what we might call a "linear sigma model" and includes the field σ, related to the pion fields by a constraint $\sigma^2 + \vec{\pi}^2 = 1$. Embarrassed by the new field which had not been observed in experiments, they subsequently integrated out to derive the "non-linear sigma model", now named after a particle that does not exist and does not appear in anywhere in the theory. The group-theoretic formulation of the non-linear sigma model that we used here is due to Weinberg [206], and was extended to general groups in [22].

The idea that baryons could arise as solitons in the chiral Lagrangian was proposed by Tony Skyrme, in a remarkably prescient pair of papers written in 1960 and 1961 and [184, 185]. These papers were apparently written without any awareness of the work described above, and were essentially ignored for more than a decade while the story of chiral symmetry breaking unfolded. The papers came to prominence only in the 1980s when it was realised that they played an important role in the story. The term "skyrmion" was coined in a 1984 meeting in honour of Tony Skyrme. (In a cute twist, the second paper thanks "Mr A. J. Leggatt" for performing the calculations as an undergraduate student. This mis-spelled student went to win the Nobel prize.)

The WZW term was introduced by Witten in 1983 [226]. The arguments in Section 5.5 are largely taken from this paper. (Many of Witten's papers from this time are masterclasses in clarity; the best way to learn much of modern physics is simply to read Witten's papers.) As we saw, for $N_f = 2$ there is no WZW term, but the fact that topology can determine the quantum statistics of the Skyrmion was noted by Finkelstein and Rubinstein, back in 1968 [60].

More on the history of chiral symmetry breaking can be found in the article by Weinberg [209]. More details about the physics of chiral symmetry breaking can be found in the lecture notes of Scherer and Schindler [173] and Peskin [153].

The idea that anomalies place severe constraints on the spectrum of strongly interacting gauge theories was first emphasised by 't Hooft in the lectures [105], with the application to chiral symmetry breaking that we described in these lectures. This was elaborated on by Frishman, Schwimmer, Banks and Yankielowicz, [64]. The "persistent mass condition", prohibiting the formation of massless bound states using massive constituents, was framed by Preskill and Weinberg [163] and found a more rigorous grounding in the Vafa-Witten theorems [195, 196]. The mass inequalities, which also make use of the positive definite measure, were first introduced by Weingarten (very) slightly before the Vafa-Witten theorem [210]. The idea that the Higgs and confining phases provide complementary, but equivalent, viewpoints on the dynamics of chiral gauge theories was first enunciated in [189].

6. Large N

Non-Abelian gauge theories are hard. We may have mentioned this previously. Indeed, it's not a bad summary of the lectures so far. The difficulty stems from the lack of a small, dimensionless parameter which we can use as the basis for a perturbative expansion.

Soon after the advent of QCD, 't Hooft pointed out that gauge theories based on the group $G = SU(N)$ simplify in the limit $N \to \infty$. This can then be used as a starting point for an expansion in $1/N$. Viewed in the right way, Yang-Mills does have a small parameter after all.

At first glance, it seems surprising that the theory simplifies in the large N limit. Naively, you might think that the theory only gets more complicated as the number of fields increase. However, this intuition breaks down when the fields are related by a symmetry, in which case the collective behaviour of the fields becomes stiffer as their number increases. This results in a novel, classical regime of the theory. The weakly coupled degrees of freedom typically look very different from the gluons that we start with in the original Lagrangian.

Large N limits are now commonplace in statistical and quantum physics. As a general rule of thumb, the large N limit renders a theory tractable when the number of degrees of freedom grows linearly with N. (We shall meet two examples in Section 7 when we discuss the $\mathbf{CP}^{N-1}$ model and the Gross-Neveu model.) In contrast, when the number of degrees of freedom grows as N^2, or faster, then the theory simplifies but, apart from a few special cases, cannot be solved. This is the case for Yang-Mills where the large N limit will not allow us to demonstrate, say, confinement. Nonetheless, it does provide an approach which allows us to compute certain properties. Moreover, it points to a deep connection between gauge theory and string theory, one which underlies many of the recent advances in both subjects.

You might reasonably wonder whether the large N expansion is likely to be relevant for QCD which has $N = 3$. We'll see as we go along how useful it is. A common rebuttal, originally due to Witten, is that in natural units the fine structure constant is

$$\alpha = \frac{e^2}{4\pi} \approx \frac{1}{137} \quad \Rightarrow \quad e \approx 0.30$$

This comparison is a little unfair. The true expansion parameter in QED is better phrased as $\alpha/4\pi \sim 10^{-3}$. In contrast, there are no factors of 4π that ride to the rescue for Yang-Mills. The expansion parameter is $1/N$ or, in many situations, $1/N^2$. We might therefore hope that this approach will give us results that are quantitatively correct at the 10% level.

6.1 A Quantum Mechanics Warm-Up: The Hydrogen Atom

We start by providing a simple example where a large N limit offers a novel way to apply perturbation theory. The set-up is very familiar: the hydrogen atom.

In natural units, $\hbar = c = \epsilon_0 = 1$, the Hamiltonian of the hydrogen atom is

$$H = -\frac{1}{2m}\nabla^2 - \frac{\alpha}{r} \tag{6.1}$$

with α the fine structure constant. In our first course on Quantum Mechanics, we learn the exact solution for the bound states of this system. But suppose we didn't know this. Can we try to approximate the solutions using perturbation theory?

Since there's a small number, $\alpha \approx 1/137$, sitting in the potential term, you might think that you could expand in α. But this is misleading. In the context of atomic physics, the fine structure constant cannot be used as the basis for a perturbative expansion. This is because we can always reabsorb it by a change of scale. Define $r' = m\alpha r$. Then the Hamiltonian becomes,

$$H = m\alpha^2 \left[-\frac{1}{2}\nabla'^2 - \frac{1}{r'}\right]$$

We see that the fine structure constant simply sets the overall scale of the problem. This means that we expect the order of magnitude of bound state to be around

$$E_{\text{atomic}} = -m\alpha^2 \approx -27.2 \text{ eV}$$

In fact, the ground state energy is $E_{\text{atomic}}/2 \approx -13.6$ eV, the factor of $1/2$ coming from solving the Schrödinger equation.

For our purposes this means that the hydrogen atom is, like Yang-Mills, a theory with a scale but with no small, dimensionless parameter. How, then, to construct a perturbative solution? One possibility is to generalise the problem from three dimensions to N dimensions. The Hamiltonian remains (6.1), but now with ∇^2 denoting the Laplacian in $\mathbf{R}^N$ rather than $\mathbf{R}^3$. Clearly we have increased the number of degrees of freedom from 3 to N. We have also increased the symmetry group from $SO(3)$ to $SO(N)$.

We note in passing that we are not solving the higher dimensional version of the hydrogen atom, since in that case the Coulomb force would fall-off as $1/r^{N-2}$. Instead, we keep the Coulomb force fixed as $1/r$ and vary the dimension of space.

To see how this helps, we will focus on the s-wave sector. Here the Schrödinger equation becomes

$$H\psi = m\alpha^2\left(-\frac{1}{2}\frac{d^2}{dr'^2} - \frac{(N-1)}{2r'}\frac{d}{dr'} - \frac{1}{r'}\right)\psi = E\psi$$

At leading order in $1/N$, we can replace the $(N-1)$ factor by N. We'll do this because the equations are a little simpler, although if we were serious about pursuing perturbation theory in $1/N$, we would have to be more careful. We can now remove the term that is first order in derivatives by redefining the wavefunction as $\psi(r') = \chi(r')/r'^{N/2}$, leaving us with the rescaled Schrödinger equation

$$H\chi = m\alpha^2\left(-\frac{1}{2}\frac{d^2}{dr'^2} + \frac{N^2}{8r'^2} - \frac{1}{r'}\right)\chi = E\chi$$

We'll make one further rescaling, and define a new radial coordinate, $r' = N^2R$. The Schrödinger equation now becomes

$$H\chi = \frac{m\alpha^2}{N^2}\left(-\frac{1}{2N^2}\frac{d^2}{dR^2} + V_{\rm eff}(R)\right)\chi = E\chi \quad \text{with} \quad V_{\rm eff}(R) = \frac{1}{8R^2} - \frac{1}{R}$$

This rescaling has removed all N dependence from the effective potential. Instead, we see that it appears in two places: the overall scale of the problem; and the effective (dimensionless) mass of the particle, which can be read off from the kinetic term and is $m_{\rm eff} = N^2$.

We're left with a very heavy particle, moving in the one-dimensional effective potential $V_{\rm eff}(R)$. In this limit, we can expand the potential in a Taylor series around the minimum $R_{\rm min} = 1/4$. To leading order, we can then treat the problem as a harmonic oscillator, centred on $R_{\rm min}$. Higher order terms in the Taylor series will affect the energy only at subleading order in $1/N$.

To leading order, the ground state energy is given by $V_{\rm eff}(R_{\rm min})$. (The zero point energy of the harmonic oscillator is suppressed by $1/m_{\rm eff} \sim 1/N^2$.) This gives us our expression for the ground state of the harmonic oscillator,

$$E_{\rm ground} = \frac{m\alpha^2}{N^2}\left(2 + \mathcal{O}\left(\frac{1}{N}\right)\right)$$

If we now revert to the real world with $N = 3$, we get $E_{\rm ground} \approx 2m\alpha^2/9$. The true answer, as we mentioned above, is $E_{\rm ground} = m\alpha^2/2$.

Of course, it's a little perverse to apply perturbation theory to a problem for which there is an exact solution. But the key idea remains: the extra degrees of freedom, together with the restriction of $O(N)$ symmetry, combine to render the problem weakly coupled in the limit $N \to \infty$. We will now see how a similar effect occurs for Yang-Mills theory.

6.2 Large N Yang-Mills

The action for $SU(N)$ Yang-Mills theory is

$$S_{\rm YM} = -\frac{1}{2g^2}\int d^4x\ {\rm tr}\, F^{\mu\nu}F_{\mu\nu}$$

There is an immediate hurdle if we try to naively take the large N limit. As we saw in Section 2.4, confinement and the mass gap all occur at the strong coupling scale Λ_{QCD} which, at one-loop, is given by

$$\Lambda_{QCD} = \Lambda_{UV}\exp\left(-\frac{3}{22}\frac{(4\pi)^2}{g^2 N}\right)$$

If we keep both the UV cut-off Λ_{UV} and the gauge coupling g^2 fixed, and send $N \to \infty$, then there is no parametric separation between the physical scale Λ_{QCD} and the cut-off. This is bad. To rectify this, we define the *'t Hooft coupling*,

$$\lambda = g^2 N$$

We will consider the theory in the limit $N \to \infty$, with both Λ_{UV} and λ held fixed. This ensures that the physical scale Λ_{QCD} also remains fixed in this limit. Indeed, throughout this section we will discuss how masses, lifetimes and scattering amplitudes of various states scale with N. In all cases, it is Λ_{QCD} which fixes the dimensions of these properties.

With these new couplings, the Yang-Mills action is

$$S_{\rm YM} = -\frac{N}{2\lambda}\int d^4x\ {\rm tr}\, F^{\mu\nu}F_{\mu\nu} \tag{6.2}$$

This is the form we will work with.

6.2.1 The Topology of Feynman Diagrams

To proceed, we're going to look more closely at the Feynman diagrams that arise from the Yang-Mills action 6.2. We'll see that, in the 't Hooft limit $N \to \infty$, λ fixed, there is a rearrangement in the importance of various diagrams.

We will write down the Feynman rules for Yang-Mills. Each gluon field is an $N \times N$ matrix,

$$(A_\mu)^i{}_j \quad , \; i, j = 1, \ldots, N$$

The propagator has the index structure

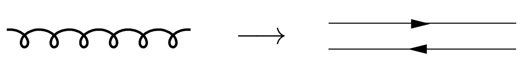
$$\langle A^i_{\mu\, j}(x)\, A^k_{\nu\, l}(y)\rangle = \Delta_{\mu\nu}(x-y)\, \left(\delta^i_l\, \delta^k_j - \frac{1}{N}\delta^i_j\, \delta^k_l\right)$$

where $\Delta_{\mu\nu}(x)$ is the usual photon propagator for a single gauge field. The $1/N$ term arises because we're working with traceless $SU(N)$ gauge fields, rather than $U(N)$ gauge fields. But clearly it is suppressed by $1/N$ and so, at leading order in $1/N$, we don't lose anything by dropping this term. We then have

$$\langle A^i_{\mu\, j}(x)\, A^k_{\nu\, l}(y)\rangle = \Delta_{\mu\nu}(x-y)\; \delta^i_l\, \delta^k_j$$

This means that we're really working with $U(N)$ gauge theory rather than $SU(N)$ gauge theory.

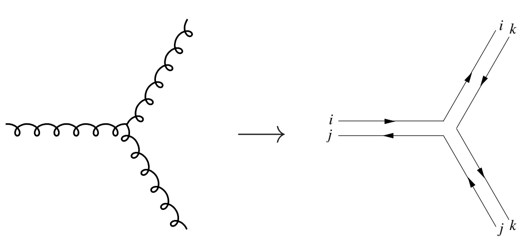
At this point, it is useful to introduce some new notation. The fact that the gauge field has two indices, i, j, suggests that we can represent it as two lines in a Feynman diagram rather than one. One of these lines represents the top index, which transforms in the $\bar{\mathbf{N}}$ representation; the other the bottom index which transforms in the $\mathbf{N}$ representation. Instead of the usual curly line notation for the gluon propagator, we have



$$\sim \quad \frac{\lambda}{N} \tag{6.3}$$

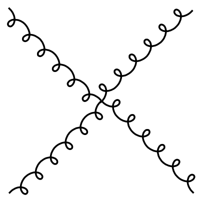
Note that each line comes with an arrow, and the arrows point in opposite ways. This reflects the fact that the upper and lower lines are associated to complex conjugate representations. The propagator scales as λ/N, as can be read off from the action (6.2).

Similarly, the cubic vertex that comes from expanding out the Yang-Mills action takes the form

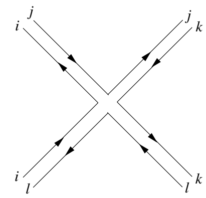


$$\sim \quad \frac{N}{\lambda}$$

where we've now included the $i, j, k = 1, \ldots, N$ indices to show how these must match up as we follow the arrows. (There is also a second diagram from the cubic vertex in which the arrows are reversed.) Similarly, the quartic coupling vertex becomes



$$\longrightarrow$$



$$\sim \quad \frac{N}{\lambda}$$

Each vertex comes with a factor of N/λ. This also follows from the action 6.2. The fact that the vertex comes with an inverse power of the coupling might be unfamiliar, but it is because of the way we chose to scale our fields. It will all come out in the wash, with the propagators compensating so that increasingly complicated diagrams are suppressed by powers of λ as expected. We'll see examples shortly.

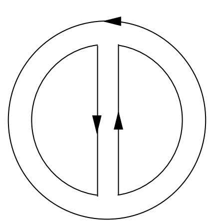
As we evaluate the various Feynman diagrams, we will now have a double expansion in both λ and in $1/N$. We'd like to understand how the diagrams arrange themselves. The general scaling will be

$$\text{diagram} \sim \left(\frac{\lambda}{N}\right)^{\#\text{propagators}} \left(\frac{N}{\lambda}\right)^{\#\text{vertices}} N^{\#\text{index contractions}} \tag{6.4}$$

where the index contractions come from the loops in the diagram. To see this more clearly, it's best to look at some examples.

Vacuum Bubbles

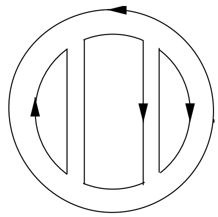
To understand the Feynman diagram expansion, let's start by considering the vacuum bubbles. The leading order contribution is a diagram which, in double line notation, looks like,



$$\sim \left(\frac{\lambda}{N}\right)^3 \left(\frac{N}{\lambda}\right)^2 N^3 \sim \lambda N^2 \tag{6.5}$$

Here the first two factors come from the 3 propagators and the 2 vertices in the diagram. The final factor is important: it comes from the fact that we have three contractions over the indices $i, j, k = 1, \ldots, N$. These are denoted by the three arrows in the diagram. Note that we get a contribution from the outside circle since we're dealing with vacuum bubbles.

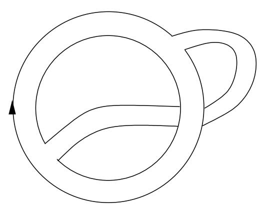
Similarly, at the next order in λ, we have the diagram



$$\sim \left(\frac{\lambda}{N}\right)^6 \left(\frac{N}{\lambda}\right)^4 N^4 \sim \lambda^2 N^2 \tag{6.6}$$

There are now four contractions over internal loops. This diagram has the same N^2 behaviour as our first one-loop diagram, but it is down in the expansion in 't Hooft coupling. It is easy to convince yourself that the two diagrams above give the leading contribution (in N) to the free energy, which scales as $\sim \mathcal{O}(N^2)$. This reflects the fact that Yang-Mills theory has N^2 degrees of freedom.

However, there is another diagram that we could have drawn. This has the same momentum structure as (6.5), but a different index structure. In double line notation it takes the form,



$$\sim \left(\frac{\lambda}{N}\right)^3 \left(\frac{N}{\lambda}\right)^2 N \sim \lambda \tag{6.7}$$

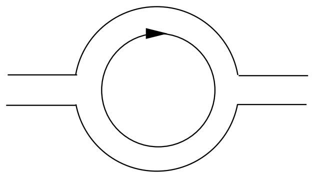
If you follow the loop around, you will find that there is now just a single contraction of the group indices. The result is a contribution to the vacuum energy which occurs at the same value of λ as (6.5), but is suppressed by $1/N^2$ relative to the first two diagrams. This means that in the limit $N \to \infty$, with λ fixed this diagram will be sub-dominant.

We see that, among all the possible Feynman diagrams, a subset dominate in the large N limit. The dominant diagrams are those which, like (6.5) and (6.6), can be drawn flat on a plane in the double line notation. These are referred to as *planar diagrams*. In contrast, diagrams like (6.7) need a third dimension to draw them. These non-planar diagrams are subleading.

The large N limit has seemed to simplify our task. We no longer need to sum over all Feynman diagrams; only the planar ones. This remains daunting. Nonetheless, as we will see below, this new structure does give us some insight into the strong coupling dynamics of non-Abelian gauge theory.

The Gluon Propagator

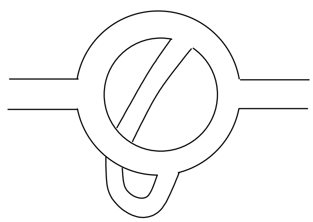
The ideas above don't just apply to the vacuum bubbles. A similar distinction holds for any Feynman diagram. We can, for example, consider the gluon propagator (6.3). A planar, one-loop correction is given by the diagram



$$\sim \left(\frac{\lambda}{N}\right)^4 \left(\frac{N}{\lambda}\right)^2 N \sim \frac{\lambda^2}{N}$$

Now we sum only over the indices on the internal loop, because we have fixed the external legs. We see that this again gives a contribution with the same $1/N$ scaling as the original propagator (6.3), but is down by a power of the 't Hooft coupling.

Meanwhile, the following two-loop, non-planar graph scales as



$$\sim \left(\frac{\lambda}{N}\right)^7 \left(\frac{N}{\lambda}\right)^4 \sim \frac{\lambda^3}{N^3}$$

and is suppressed by $1/N^2$ compared to the earlier contributions.

The Topology of Feynman Diagrams

Let's understand better how to order the different diagrams. We'll return to the vacuum diagrams. The key idea is that each of these can be inscribed on the surface of a two dimensional manifold of a given topology.

The planar diagrams can all be drawn on the surface of a sphere. This is because for any graph on a sphere, you can remove one of the faces and flatten out what's left to give the planar graph. The simplest example is the vacuum diagram (6.5) which sits nicely on the sphere as shown on the right.

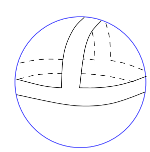
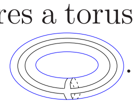


Figure 48:

In contrast, the non-planar diagrams must be drawn on higher genus surfaces. For example, the non-planar vacuum diagram (6.7) cannot be inscribed on a sphere, but requires a torus. It also requires more artistic skill than I can muster, but looks something like .

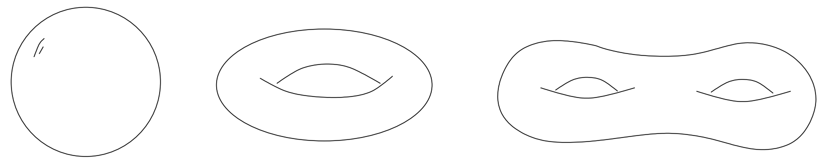


Figure 49: Examples of the simplest Riemann surfaces with $\chi = 2, 0$ and -2.

In general, the Feynman diagram tiles a two dimensional surface Σ. The map is

$$\begin{aligned} E &= \#\text{ of edges} = \#\text{ of propagators} \\ F &= \#\text{ of faces} = \#\text{ of index loops} \\ V &= \#\text{ of vertices} \end{aligned}$$

From (6.4), a given diagram then scales as

$$\text{diagram} \sim N^{F+V-E}\lambda^{E-V}$$

But there is a beautiful fact, due to Euler, which says that the following combination determines the topology of the Riemann surface

$$\chi(\Sigma) = F + V - E \tag{6.8}$$

The quantity $\chi(\Sigma)$ is called the *Euler character*. It is related to the number of handles H of the Riemann surface, also called the *genus*, by

$$\chi(\Sigma) = 2 - 2H \tag{6.9}$$

The simplest examples are shown in the figure. The sphere has $H = 0$ and $\chi = 2$; the torus has $H = 1$ and $\chi = 0$; the thing with two holes has $H = 2$ and $\chi = -2$. In this way, the large N expansion is a sum over Feynman diagrams, weighted by their topology

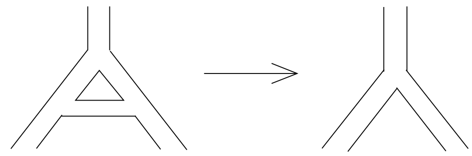
$$\text{diagram} \sim N^{\chi}\lambda^{E-V}$$

For each genus, the Riemann surface can be tiled in different ways by Feynman diagram webs, giving the expansion in the 't Hooft coupling. There is no topological interpretation of this exponent $V - E$. We'll shortly discuss the implication of this large N expansion.

The Euler Character

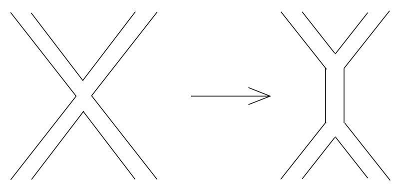
Before we proceed, it will be useful to get some intuition for why the Euler character (6.8) is a topological invariant, and why it is given by (6.9).

To see the former, it's best to play around a little bit by deforming various diagrams. The key manipulation is to take a face and shrink it to vanishing size. For example, we have



Under such a transformation, the number of faces shrinks by 1: $F \to F-1$. The number of vertices has also decreased, $V \to V-2$, as has the number of edges, $E \to E-3$. But the combination $\chi = F + V - E$ remains unchanged.

In all the examples above, we used only the cubic Yang-Mills vertex. Including the quartic vertex doesn't change the counting. This is because we can always split the quartic vertex into two cubic ones,



The left hand side has $V = 1$ and $E = 4$, which transforms into the right hand side with $V = 2$ and $E = 5$. We see neither χ, nor the power of λ depend on the kind of vertex that we use.

This should help explain why the Euler character does not vary under manipulations that make the diagram more and more complicated, but leave the underlying topology unchanged. For the sphere, the example we drew above shows that $\chi = 2$. For each extra handle, we can consider first cutting a hole in the surface. We do this by removing a face, leaving us with a boundary. To build a handle, we cut out two faces, each of which is an n-gon. This reduces the number of faces $F \to F - 2$. Now we glue the faces together by identifying the perimeters of the holes. This act reduces $E \to E - n$ and $V \to V - n$. But the net effect is that for each handle we add, $\chi \to \chi - 2$.

6.2.2 A Stringy Expansion of Yang-Mills

The large N limit of Yang-Mills has been repackaged as a sum over Riemann surfaces of different topologies. But this is the defining feature of weakly coupled string theory. This is discussed in much detail in the lectures on String Theory; here we'll just mention some pertinent facts.

In string theory, the sum over Riemann surfaces is weighted by the string coupling constant g_s. By analogy, we see that

$$g_s = \frac{1}{N}$$

But there are also differences. In string theory, the Riemann surfaces are smooth objects, which suffer quantum fluctuations governed by the inverse string tension α'. This is a quantity with dimension $[\alpha'] = -2$ and it is often written as $\alpha' = l_s^2$ with l_s the typical size of a string. The fluctuations of the Riemann surface are really governed by α'/L^2 where L is the spatial size of the background in which the string propagates.

In contrast, the Riemann surfaces that arise in the large N expansion are not smooth at all; they are tiled by Feynman diagrams and in the perturbative limit, $\lambda \ll 1$, the diagrams with the fewest vertices dominate. However, taken naively, it appears that in the opposite limit $\lambda \gg 1$, the diagrams with large numbers of vertices are important. With some imagination, these can be viewed as diagrams which finely cover the Riemann surface, so that it looks more and more like a classical geometry. This suggests that, in the 't Hooft limit, strongly coupled Yang-Mills may be a weakly coupled string theory in some background, with

$$\lambda^{-1} \sim \left(\frac{\alpha'}{L^2}\right)^{\#}$$

where I've admitted ignorance about the positive exponent #.

This is a bold idea. Weakly coupled string theory is a theory of quantum gravity, and gives rise to general relativity at long distances. If we can somehow make the idea above fly, then Yang-Mills theory would contain general relativity! But the strings and gravity would not live in the $d = 3 + 1$ dimensions of the Yang-Mills theory. Instead, we would find gravity in the "space in which the Feynman diagrams live", whatever that means.

So far, no one has made sense of these ideas for pure Yang-Mills. However, it is now understood how these ideas fit together in a very closely related theory called maximally supersymmetric (or $\mathcal{N} = 4$) Yang-Mills which is just $SU(N)$ Yang-Mills coupled to a bunch of adjoint scalars and fermions. In that case, the strongly coupled 't Hooft limit is indeed a theory of gravity in a $d = 9 + 1$ dimensional spacetime that has the form $AdS_5 \times \mathbf{S}^5$. The $d = 3 + 1$ dimensional world in which the Yang-Mills theory lives is the boundary of AdS_5. This remarkable connection goes by the name of the AdS/CFT correspondence or, more generally, gauge-gravity duality. It is a topic for another course.

It's an astonishing fact that, among the class of gauge theories in $d = 3+1$ dimensions, is a theory of quantum gravity in higher dimensional spacetime. It leaves us wondering just what else is hiding in the land of strongly coupled quantum field theories.

6.2.3 The Large N Limit is Classical

We can use the large N counting described above to understand the scaling of correlation functions.

In what follows, we consider gauge invariant operators which cannot be further decomposed into colour singlets. Since Yang-Mills has only adjoint fields, this means that we are interested in operators that have just a single trace. The simplest is

$$\mathcal{G}_{\mu\nu,\rho\sigma}(x) = \mathrm{tr}\, F_{\mu\nu} F_{\rho\sigma}(x)$$

There's a slew of further operators in which we add more powers of $F_{\mu\nu}$ inside the trace. However, it's important that the number of fields inside the trace is kept finite as we take $N \to \infty$, otherwise it will infect our N counting. This means, for example, that we can't discuss operators like $\det F_{\mu\nu} F^{\mu\nu}$. Of course, Yang-Mills also has non-local operators – the Wilson loops – and much of what we say will hold for them. But, for once, our main interest will be on the local, single trace operators.

We could also consider coupling our theory to adjoint matter, either scalars or fermions. Restricting to the adjoint representation means that these new fields are also $N \times N$ matrices, and the same $1/N$ counting that we developed above holds for their Feynman diagram expansion. This gives us the option to build more single trace operators, such as $\mathcal{G} = \mathrm{tr}(\phi^m)$ for a scalar ϕ, or combinations of scalars and field strengths. Once again, we insist only that the number of fields inside the trace does not scale with N.

We can compute correlation functions of any of these operators by adding sources in the usual way,

$$S_{YM} = N \int d^4x \ -\frac{1}{2\lambda} \mathrm{tr}\, F^{\mu\nu} F_{\mu\nu} + \ldots + J_a \mathcal{G}_a$$

where the $\ldots$ is any further adjoint matter that we've included, and where the operators $\mathcal{G}_a$ denote any single trace involving strings of the field strength, the other adjoint matter, or their derivatives. Note that we've scaled both fields and operators to keep an overall factor of N in front of the action. The connected correlation functions can be computed in the usual way by differentiating the partition function,

$$\langle \mathcal{G}_1 \ldots \mathcal{G}_p \rangle_c = \frac{1}{N^p} \frac{\delta}{\delta J_1} \cdots \frac{\delta}{\delta J_p} \ \log\, Z[J] \qquad (6.10)$$

where the subscript c is there to remind us that we're dealing with connected correlators. Because the action, including the source terms, has the form $S = N\,\mathrm{tr}\,(\text{something})$, our previous large N counting goes over unchanged, and the free energy is dominated by planar graphs at order $\log Z \sim N^2$. (This conclusion would no longer hold if we included multi-trace operators as sources, or if there were some other powers of N that had somehow snuck unseen into the action). We learn that connected correlation functions of single trace operators have the leading scaling

$$\langle \mathcal{G}_1 \ldots \mathcal{G}_p \rangle_c \sim N^{2-p} \tag{6.11}$$

where in this formula, and others below, we're ignoring the dependence on the 't Hooft coupling λ.

The simple formula (6.11) is telling us something interesting: the leading contribution to any correlation function comes from disconnected diagrams, rather than connected diagrams. For example, any two-point function has a connected piece $\langle \mathcal{G}\mathcal{G} \rangle \sim \langle \mathcal{G} \rangle \langle \mathcal{G} \rangle \sim N^2$. This should be contrasted with the connected piece which scales as $\langle \mathcal{G}\mathcal{G} \rangle_c \sim N^0$.

This means that the strict $N \to \infty$ limit of Yang-Mills is a free, classical theory. All correlation functions of single trace, gauge invariant operators factorise. Said slightly differently, quantum fluctuations are highly suppressed in the large N limit, with the variance of any gauge singlet operator $\mathcal{O}$ given by

$$(\Delta\mathcal{G})^2 = \langle \mathcal{G}\mathcal{G} \rangle - \langle \mathcal{G} \rangle \langle \mathcal{G} \rangle = \langle \mathcal{G}\mathcal{G} \rangle_c \sim N^0 \quad \Rightarrow \quad \frac{(\Delta\mathcal{G})^2}{\langle \mathcal{G} \rangle^2} \sim \frac{1}{N^2}$$

Usually when we hear the words "free, classical theory", we think "easy". That's not the case here. The large N limit is a theory of an infinite number of single trace operators $\mathcal{G}_a(x)$. If the theory is confining and has a mass gap, like Yang-Mills, each of these corresponds to a particle in the theory. (We will make this connection clearer below.) Or, to be more precise, each of the operators $\mathcal{G}(x)$ corresponds to some complicated linear combination of particles in the theory. After diagonalising the Hamiltonian, we will have a free theory of an infinite number of massive particles. Determining these masses is a difficult problem which remains unsolved.

The large N limit does not only hold for confining theories. For example, maximally supersymmetric Yang-Mills is a conformal field theory and does not confine. Now the goal in the large N limit is to diagonalise the dilatation operator to find the conformal dimensions of single trace operators. This is a difficult problem that is largely solved using techniques of integrability.

The fact that the large N limit is free leads to the concept of the *master field*. There should be a configuration of the gauge fields A_μ on which we can evaluate any correlation function to get the correct $N \to \infty$ answer. (If we add more adjoint matter fields, we would need to specify their value as well.) Once we have this master field, there is nothing left to do: no fluctuations, no integrations. We just evaluate. Furthermore, the master field should be translationally invariant so, at least in a suitable gauge, the A_μ are just constant. In other words, all of the information about Yang-Mils in the $N \to \infty$ limit is contained in four matrices, A_μ. The twist, of course, is that these are $\infty \times \infty$ matrices and, as a well known physicist is fond of saying, "you can hide a lot in a large N matrix". For pure Yang-Mills in $d = 3 + 1$ dimensions, no progress has been made in understanding the master field in decades. For maximally supersymmetric Yang-Mills, the master field should be equivalent to saying that the theory is really ten dimensional gravity in disguise.

6.2.4 Glueball Scattering and Decay

The strict $N \to \infty$ limit is free, with the degrees of freedom organised in single trace operators $\mathcal{G}(x)$. All of the difficulties of the strong coupling dynamics goes into diagonalising the Hamiltonian to determine masses (or scaling dimensions) of the corresponding states.

At large, but finite N, we introduce interactions between these degrees of freedom, which must scale as some power of $1/N$. Even though we can't solve the $N \to \infty$ limit, we can still get some useful intuition for the theory by looking at these interactions in a little more detail.

To see this, let's revert to pure Yang-Mills. We will assume that this theory confines in the large N limit. There is no reason to think this is not the case but it's important to stress that we can currently no more prove confinement in the large N limit than at finite N[11]. We consider the local glueball operators

$$\mathcal{G}(x) = \text{tr}\, F^m(x) \tag{6.12}$$

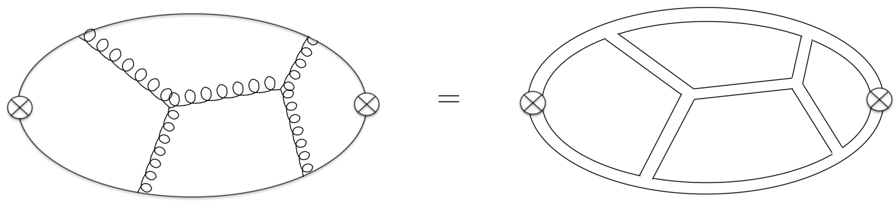
for some $m \geq 2$. We've ignored the Lorentz indices, which endow each operator with a certain spin. We could also include derivatives to increase the spin yet further.

[11]The Millennium Prize Problem requires that you prove confinement for *all* compact, simple non-Abelian gauge groups. This stipulation was put in place to avoid a scenario where confinement was proven only in the large N limit. Apparently, the authors of the problem originally meant to find a different phrasing, one that avoided the caveat of large N but would award a proof of confinement in, say, $SU(3)$ Yang-Mills. But they never got round to changing the wording. Like with all such prizes, if you're genuinely interested in the million dollars then you are probably in the wrong field.

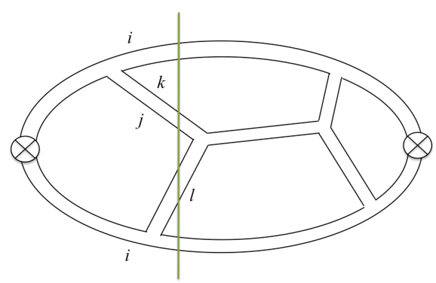
At large N, there is a connected component to the two point function which, with the normalisation (6.11), scales as

$$\langle \mathcal{G}(x)\, \mathcal{G}(0)\rangle_c \sim N^0$$

which means that $\mathcal{G}(x)$ creates a glueball state with amplitude of order 1. In terms of our original Feynman diagrams, this picks up contributions from very complicated processes, such as the one below



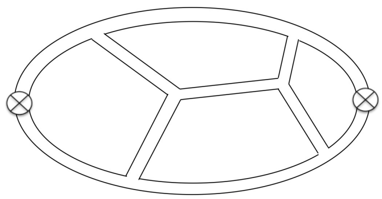
In the large N limit, this is converted into tree-level propagation of gauge singlet operators created by $\mathcal{G}(x)$. Importantly, the operator $\mathcal{G}(x)$ creates only single-particle states. To see this, we can cut the diagram to see the intermediate state, as shown below



We've now included $i, j = 1, \ldots, N$ indices to help keep track. To make something gauge invariant, we need to take the trace, which means combining each index with its partner. The only way to do this is to include all the internal legs together. This is the statement that the internal state corresponds to a single trace operator. In contrast, multi-particle states only propagate in non-planar diagrams where the internal lines can be combined into multi-trace colour singlets.

The fact that the single-trace operator $\mathcal{G}(x)$ creates single particle states also follows from the scaling of the correlation function (6.11). To see this, first suppose that the statement isn't true, and $\mathcal{G}$ creates a two particle state with amplitude order 1. Then one could construct a suitable correlation function which has the value $\langle \mathcal{G}\tilde{\mathcal{G}}\tilde{\mathcal{G}}\mathcal{G}\rangle \sim 1$, with the operators $\tilde{G}$ each interacting, with amplitude 1, with one of the the two intermediate particles. But we know from large N counting (6.11) that $\langle \mathcal{G}\tilde{\mathcal{G}}\tilde{\mathcal{G}}\mathcal{G}\rangle \sim 1/N^2$. (There is actually an implicit assumption here that there is no degeneracy of states at order N. But this is precisely the assumption of confinement.)

So we can think of any two-point function $\langle \mathcal{G}(x)\, \mathcal{G}(0)\rangle_c$ as the tree-level propagation of confined, single particle states. We are repackaging

$$\sum_{\text{planar graphs}}$$  $$= \sum_{\text{single particles}}$$ 

In general, the only singularities in tree-level graphs are poles. (This is to be contrasted with one-loop diagrams where we can have two-particle cuts, and higher loop diagrams with multi-particles cuts.) This means that there should be some expansion of the two-point function in momentum space as

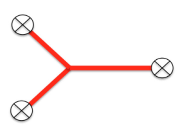
$$\langle \mathcal{G}(k)\, \mathcal{G}(-k)\rangle_c = \sum_n \frac{|a_n|^2}{k^2 - M_n^2} \tag{6.13}$$

where $a_n = \langle 0|\mathcal{G}|n\rangle$, with $|n\rangle$ the single particle state with mass M_n. But now there's something of a puzzle. At large k, Yang-Mills theory is asymptotically free, and we can compute this correlation function to find that it scales as

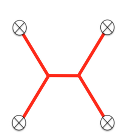
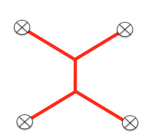
$$\langle \mathcal{G}(k)\, \mathcal{G}(-k)\rangle_c \to k^2 \log\, k^2$$

Yet naively the propagator (6.13) would appear to scale as $1/k^2$ for large momentum. The only way we can reproduce the expected log behaviour is if there are an infinite number of stable intermediate states $|n\rangle$, with an infinite tower of masses m_n. This coincides with our earlier expectations: as $N \to \infty$ Yang-Mills is a theory of an infinite number of free particles.

At large but finite N, there can no longer be an infinite tower of stable, massive particles. The heavy ones surely decay to the light ones. But this process is captured by the correlation functions of the schematic form

$$\langle \mathcal{G}\mathcal{G}\mathcal{G}\rangle \sim \sum$$  $$+ \ldots \sim \frac{1}{N}$$

which tells us that the amplitude for a glueball to decay to two glueballs scales as $1/N$, so their lifetime scales as N^2. Similarly, for scattering we can turn to the four-point function

$$\langle \mathcal{G}\mathcal{G}\mathcal{G}\mathcal{G}\rangle \sim \sum$$  $$+$$  $$+ \ldots \sim \frac{1}{N^2}$$

So the amplitude for gluon-gluon scattering scales as $1/N^2$.

6.2.5 Theta Dependence Revisited

We saw in Section 2.2 that Yang-Mills theory comes with an extra, topological parameter: the theta-term. How does this fare in the large N limit? The Lagrangian is

$$\begin{aligned}\mathcal{L}_{YM} &= -\frac{1}{2g^2}\mathrm{tr}\, F_{\mu\nu}F^{\mu\nu} + \frac{\theta}{16\pi^2}\mathrm{tr}\, F_{\mu\nu}{}^{\star}F^{\mu\nu} \\ &= N\left(-\frac{1}{2\lambda}\mathrm{tr}\, F_{\mu\nu}F^{\mu\nu} + \frac{\theta}{16\pi^2 N}\mathrm{tr}\, F_{\mu\nu}{}^{\star}F^{\mu\nu}\right)\end{aligned}$$

With the appropriate factor of N sitting outside the action, we see that we should keep θ/N fixed as we send $N \to \infty$. The first question that we should ask is: does the physics still depend on θ?

At first glance, it appears that the answer to this question should be no. The reasons for this are two-fold. At leading order in perturbation theory, none of the planar graphs appear to depend on θ. Moreover, the instanton effects which, at weak coupling, give us θ dependence now scale as $\sim e^{-8\pi^2/g^2} \sim e^{-8\pi^2 N/\lambda}$ and so are exponentially suppressed in the large N limit.

Although both of these arguments appear compelling, the conclusion is thought to be wrong. It is believed that, at leading order in the $1/N$ expansion, the physics continues to depend on θ (or, more precisely, on θ/N). Perhaps the simplest observable is the ground state energy, defined schematically in the Euclidean path integral as

$$e^{-V\,E(\theta)} = \int \mathcal{D}A\, \exp\left(-\int d^4x\; \mathcal{L}_{YM}\right) \tag{6.14}$$

where V is the spacetime volume. Recall that, in Euclidean space, the theta term weights the path integral as $e^{i\theta\nu}$ where ν is the topological winding of the configuration. The large N arguments that we've seen above tell us that $E \sim N^2$. It is believed that the θ dependence affects this quantity at leading order

$$E(\theta) = N^2 h\left(\frac{\theta}{N}\right) \tag{6.15}$$

for some function $h(x)$.

There are two main reasons for thinking that θ dependence survives in the large N limit. The first is that, in the presence of light quarks, the dependence can be seen in the chiral Lagrangian; we will describe this in Section 6.4. The second is that both the arguments we gave above also hold in toy models in two-dimensions (specifically the $\mathbf{CP}^N$ model that we will introduce in 7.3) where one can see that they lead to the wrong conclusion. The loophole lies in the first argument; at leading order in the $1/N$ expansion we must sum an infinite number of diagrams, and interesting things can happen for infinite series that don't arise for finite sums.

To make this more concrete, let's introduce the *topological susceptibility*,

$$\chi(k) = \int d^4x\, e^{ik\cdot x} \langle \mathrm{tr}\,(F_{\mu\nu}{}^\star F^{\mu\nu}(x))\, \mathrm{tr}\,(F_{\rho\sigma}{}^\star F^{\rho\sigma}(0))\rangle \tag{6.16}$$

(Not to be confused with the Euler character that we encountered earlier.) Roughly speaking, this tells us how the theory responds to changes in θ. In particular, the ground state energy $E(\theta)$ has the dependence

$$\frac{d^2E}{d\theta^2} = \left(\frac{1}{16\pi^2 N}\right)^2 \lim_{k\to 0} \chi(k) \tag{6.17}$$

We can compute contributions to $\chi(k)$ in perturbation theory. One finds that, at leading order in $1/N$, each individual diagram has $\chi(k) \to 0$ as $k \to 0$. Nonetheless, it is expected that the sum of all such diagrams does not vanish. No one has managed to perform this calculation explicitly in four-dimensional Yang-Mills theory. To see that such behaviour is indeed possible, you need only consider the series

$$f(k) = k^2 \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \log^n k^2 = k^2 \exp^{-\log k^2} = 1$$

The behaviour of the ground state energy (6.15) brings a new puzzle. The energy depends on θ/N, but must obey $E(\theta) = E(\theta + 2\pi)$. How can we reconcile these two properties? The accepted answer – and the one which is seen in the $\mathbf{CP}^N$ model – is that there is a level crossing in the ground state as θ is varied. This works as follows: at large N the theory is thought to have a large number of meta-stable, Lorentz-invariant states that differ in energy. There are order N such states and, in the k^{th}, the energy is given by

$$E_k(\theta) = N^2 h\left(\frac{\theta + 2\pi k}{N}\right)$$

The ground state energy is then

$$E(\theta) = \mathrm{min}_k\, E_k \tag{6.18}$$

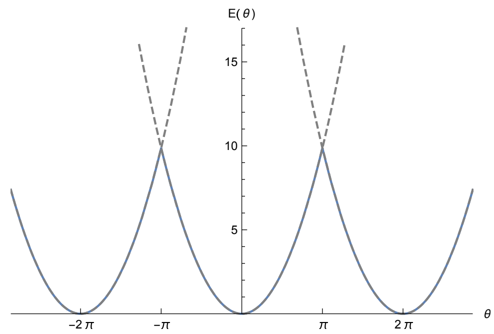


Figure 50: The vacuum energy as a function of θ.

We're left with a function which is periodic, but not smooth. In particular, when $\theta = \pi$ two levels cross.

What does the function $E(\theta)$ look like? First, we know that it has its minimum at $\theta = 0$. This is because the Euclidean path integral (6.14) is a sum over configurations weighted by $e^{i\theta}$. Only for $\theta = 0$ is this real and positive, hence maximising $e^{-VE(\theta)}$, and so minimising $E(\theta)$. Taylor expanding, we therefore expect that

$$E(\theta) = \min_k \frac{1}{2} C(\theta + 2\pi k)^2 + \mathcal{O}\left(\frac{1}{N}\right)$$

where $C = \chi(0)/(16\pi^2 N)^2$. This is shown in the figure.

A general value of θ explicitly breaks time-reversal or, equivalently, $\mathcal{CP}$. The two exceptions are $\theta = 0$ and $\theta = \pi$. (We explained why $\theta = \pi$ is time reversal invariant in Section 1.2.5.) But, at $\theta = \pi$, there are two degenerate ground states and time-reversal invariance maps one to the other. We learn that, at large N Yang-Mills, time-reversal invariance is spontaneously broken at $\theta = \pi$. This coincides with our conclusion from Section 3.6 using discrete anomalies.

6.3 Large N QCD

Our discussion in the previous section focussed purely on matrix valued fields. To get closer to QCD, we add quarks, as Dirac fermions in the fundamental representation.

We rescale the quark field $\psi \rightarrow \sqrt{N}\psi$, so that the action continues to have a factor of N sitting outside,

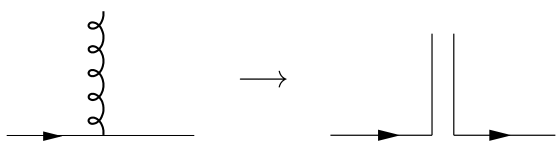
$$S_{QCD} = N \int d^4x \; -\frac{1}{2\lambda} \mathrm{tr}\, F^{\mu\nu} F_{\mu\nu} + i\bar{\psi} \not{D} \psi$$

We'll stick with just a single quark field for now, but everything that we say will go over for N_f flavours of quarks provided that we keep N_f fixed as $N \to \infty$.

The quark field carries just a single gauge index, ψ^i with $i = 1, \ldots, N$. Correspondingly, it is represented by just a single line in a Feynman diagram,

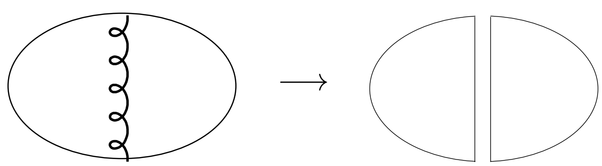
$$\sim \frac{\lambda}{N}$$

Meanwhile, the quark-gluon vertex is represented by



$$\sim N$$

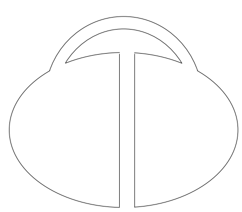
We can now repeat the large N counting that we saw previously. We can start by looking at contributions to the vacuum energy that include a quark loop. For example, we have



$$\sim \left(\frac{\lambda}{N}\right)^3 N^2\, N^2 \sim \lambda^3 N$$

where the first factor of N^2 comes from the two quark-gluon vertices, while the second factor comes from the index loops. We see that this is subleading compared to the pure glue vacuum diagrams which are $\sim N^2$. Including extra internal gluons, all planar diagrams with a single quark loop on the boundary will continue to scale as $\sim N$. This is the leading order contribution to the vacuum energy that includes quarks. This is simple to understand: the amplitude to create a quark is the same as the amplitude to create a gluon, but there are N^2 gluon degrees of freedom and only N quark degrees of freedom.

If the quark loop does not run around the boundary, the diagram is suppressed yet further. For example, consider the diagram



$$\sim \left(\frac{\lambda}{N}\right)^6 N^4\, N \sim \lambda^6 N^{-1}$$

Similarly, if we include internal quark lines in other Feynman diagrams, say the gluon propagator, we again get a suppression factor of $1/N$.

We can again interpret the large N Feynman diagrams in terms of 2d surfaces. However, now the surfaces are no longer closed. Instead, each quark loop should be thought of as the boundary of a hole on the Riemann surface. Each boundary increases the number of edges E by one, so a given Feynman diagram again scales as

$$\text{diagram} \sim N^{F+V-E}\,\lambda^{E-V} = N^{\chi}\lambda^{E-V}$$

which is the same result that we had before. But now the expression for the Euler character is

$$\chi = 2 - 2H - B$$

where B is the number of boundaries, or holes, in the surface.

In terms of string theory, the addition of quarks means that the large N limit includes open strings, with boundaries, as well as closed strings. This is closely related to the concept of D-branes in string theory.

6.3.1 Mesons

We can now rerun the arguments of Sections 6.2.3 and 6.2.4 for large N QCD. In addition to the glueball operators (6.12), we also have the meson operators

$$\mathcal{J}(x) = \sqrt{N}\,\bar{\psi}F^m\psi \tag{6.19}$$

where the F^m can denote any number of field strengths, derivatives and gamma matrices, so that $\mathcal{J}(x)$ is a local, gauge invariant operator that cannot be decomposed into smaller colour singlets.

Note that we've included an overall factor of $\sqrt{N}$ in (6.19). To see why this is, we compute correlation functions

$$\langle \mathcal{J}_1 \ldots \mathcal{J}_p \rangle_c \sim N^{1-p/2} \tag{6.20}$$

The first factor of N comes from the planar diagrams with a quark loop running along the boundary. The normalisation factor of $\sqrt{N}$ in (6.19) means that correlation function scale as $N^{-p/2}$ rather than as N^{-p}. This normalises the two-point function as $\langle \mathcal{J}\mathcal{J} \rangle_c \sim N^0$, so $\mathcal{J}$ creates a meson state with amplitude 1.

The same arguments that we used for pure Yang-Mills still apply here. The strict $N \to \infty$ limit is again a free theory, now including infinite towers of both glueball and meson states. In momentum space, the analog of the propagator (6.13) is

$$\langle \mathcal{J}(k)\, \mathcal{J}(-k) \rangle_c = \sum_n \frac{|b_n|^2}{k^2 - m_n^2} \tag{6.21}$$

where $b_n = \langle 0|\mathcal{J}|n\rangle$, with $|n\rangle$ the single particle meson state with mass m_n. As for glueballs, this expression is only compatible with the log behaviour of asymptotic freedom if there is an infinite tower of massive meson states.

At large N, the three point function of meson fields

$$\langle \mathcal{J}\mathcal{J}\mathcal{J} \rangle \sim \frac{1}{\sqrt{N}}$$

tells us that the amplitude for a meson to decay into two lighter mesons scales as $1/\sqrt{N}$. The lifetime of a meson is then typically of order N. They are shorter lived than the glueballs. Similarly, the four point function of meson fields is

$$\langle \mathcal{J}\mathcal{J}\mathcal{J}\mathcal{J} \rangle \sim \frac{1}{N}$$

The amplitude for meson-meson scattering scales as $1/N$.

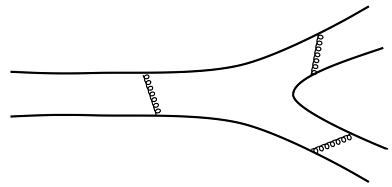
We can also compute correlation functions of both glueballs and mesons. At leading order, we have

$$\langle \mathcal{J}_1 \ldots \mathcal{J}_p\, \mathcal{G}_1 \ldots \mathcal{G}_q \rangle \sim N\, N^{-p/2}\, N^{-q}$$

This means that the two-point function $\langle \mathcal{J}\mathcal{G} \rangle \sim 1/\sqrt{N}$, so mesons and glueballs don't mix at large N, even if they share the same quantum numbers. (We had assumed this when talking separately about meson and glueballs above, so it's good to know it's true.) We can also extract the amplitude for a gluon to decay into two mesons which is $\langle \mathcal{G}\mathcal{J}\mathcal{J} \rangle \sim 1/N$, which is the same order as the decay into two gluons. Meanwhile, the amplitude for a meson to decay into two gluons is $\langle \mathcal{J}\mathcal{G}\mathcal{G} \rangle \sim 1/N^{3/2}$. We see that a gluon doesn't much mind who it decays into, while a meson greatly prefers decaying into other mesons.

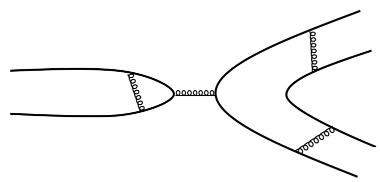
The OZI Rule

The large N approach helps explain a couple of phenomenological facts that had been previously observed to hold for QCD. In particular, note that the leading order meson decays have the form



$$\sim \frac{1}{\sqrt{N}}$$

In such a process, one of the original quarks ends up in each of the final decay products. In contrast, a process in which the two original quarks decay into pure glue which subsequently produces two further mesons, is suppressed by an extra factor of $1/N$,



$$\sim \frac{1}{N^{3/2}}$$

This suppression was observed experimentally in the early days of meson physics and goes by the name of the OZI rule (for Okubo, Zweig and Iizuka; it is also sometimes called the Zweig rule).

The standard example is the ϕ vector meson, which has quark content $s\bar{s}$. On energy considerations alone, one would have thought this would decay to $\pi^+\pi^-\pi^0$, none of which contain a strange quark. In reality, this decay is suppressed by QCD dynamics, and the ϕ meson decays primarily to K^+K^-, where the positively charged kaon has quark content $u\bar{s}$. This fact is clearest in the $1/N$ expansion.

The large N expansion also makes it clear that we don't expect to see meson bound states or, more generally, $\bar{q}q\bar{q}q$ states with four quarks. Such states are referred to as *exotics*. The amplitude for meson interactions scales as $1/N$, so such exotics certainly don't form in the large N limit. The lack of exotics in particle data book suggests that this suppression extends down to $N = 3$.

6.3.2 Baryons

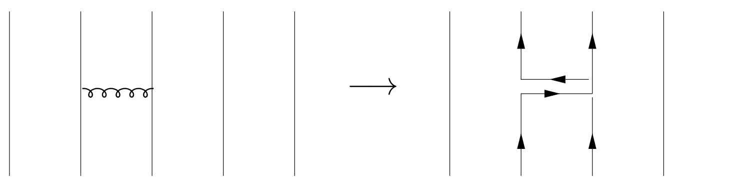
We now turn to baryons. These are a little more subtle because they contain N quarks, anti-symmetrised over the colour indices. Nonetheless, as first explained by Witten, they are naturally accommodated in the large N limit of QCD.

In what follows we will consider the large N limit with just a single flavour of quark, although it is not difficult to include $N_f > 1$ flavours. The baryon is then

$$B = \epsilon^{i_1 \ldots i_N} \psi_{i_1} \ldots \psi_{i_N} \tag{6.22}$$

This is the large N analog of, say, the Δ^{++} in QCD which contains three up quarks, or the Δ^- which contains three down quarks.

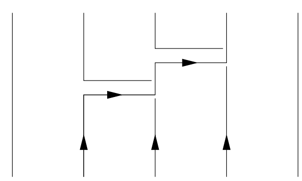
We can start by modelling these as N distinct quark lines. A gluon exchange between any pair of quarks is



$$\sim \frac{1}{N} \tag{6.23}$$

where we've been more careful in the second diagram in showing how the arrows flow. However, there $\frac{1}{2}N(N-1) \sim N^2$ different pairs of quarks, so the total amplitude for a gluon exchange within a baryon is order N.

There is a similar story for three body interactions. The gluon exchange is now



$$\sim \frac{1}{N^2} \tag{6.24}$$

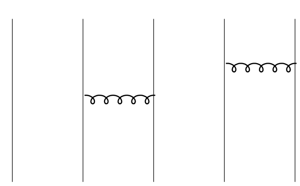
but there are order N^3 triplets of quarks, so again the total amplitude scales as N.

These simple arguments suggest that many-body interactions are all equally important, and contribute to the energy of the baryon at order N. It is therefore natural to guess that

$$M_{\text{baryon}} \sim N \tag{6.25}$$

This is perhaps not a surprise since the baryon contains N quarks, and is certainly to be expected in the non-relativistic quark model.

There's a calculation which may give you pause. Consider the the gluon exchange between two different pairs of quarks,



$$\sim \frac{1}{N^2} \tag{6.26}$$

But now there are $\sim N^4$ ways of picking two pairs of quarks, so it looks as if this contributes to the energy at order $N^4 \times N^{-2} \sim N^2$. It seems like we get increasingly divergent answers as we look at more and more disconnected pieces. In fact, this is the kind of behaviour that we would expect if the baryon mass scales as (6.25). The propagator for large times T then takes the form

$$e^{-iM_{\rm baryon}T} \approx 1 - iM_{\rm baryon}T - \frac{1}{2}M_{\rm baryon}^2 T^2 + \ldots$$

For the diagram (6.26), each of the gluons can be exchanged at any time and so it corresponds to the second order term in the expansion above which, we see, should indeed scale as $M_{\rm baryon}^2 \sim N^2$.

At this point, we could start to explore the interactions between baryons and mesons, and build towards a fuller phenomenology of QCD. However, we won't go in this direction. Instead, I will point out a nice connection between baryons in the large N expansion and another recurring topic from these lectures.

The Hartree Approximation

A particularly simple way to proceed is to assume that the quarks are non-relativistic. This is not particularly realistic for QCD, but it will provide a simple way to shine a light on the structure of the baryon. If each quark has mass m, we could try to model their physics inside a baryon by the following Hamiltonian

$$H = Nm + \frac{1}{2m}\sum_{i=1}^{N} p_i^2 + \frac{1}{2N}\sum_{i\neq j} V_2(x_{ij}) + \frac{1}{6N^2}\sum_{i\neq j\neq k} V_3(x_{ij}, x_{jk}) + \ldots$$

where $x_{ij} = x_i - x_j$ and the coefficients in front of the potentials are taken from (6.23) and (6.24). We should also include all multi-particle potentials. As we have seen, it is a mistake to think that these potentials are genuinely suppressed by the $1/N$ factors in the Hamiltonian: these are compensated by the sums over particles, so each term ends up of order N.

There is a straightforward variational approach to such many-body Hamiltonians called the Hartree approximation. It is the first port of call in atomic physics, when studying atoms with many electrons, and we met it in the lectures on Topics in Quantum Mechanics. The idea is to work with the ansatz for the ground state wavefunction given by

$$\psi(\mathbf{x}_1, \ldots, \mathbf{x}_N) = \prod_{i=1}^{N} \phi_0(\mathbf{x}_i)$$

Note that the quarks are fermions, but they have already been anti-symmetrised over the colour indices (6.22), so it is appropriate that the wavefunction for the remaining degrees of freedom is symmetric.

The Hartree ansatz neglects interactions between the quarks. Instead, it is a self-consistent approach in which each quark experiences a potential due to all the others. This approach becomes increasingly accurate as the number of particles becomes large, so it is particularly well suited to baryons in the large N limit.

Evaluating the Hamiltonian on the Hartree wavefunction gives

$$\langle\psi|H|\psi\rangle = N \left[m + \frac{1}{2m}\int d^3x\ |\phi(\mathbf{x})|^2 + \frac{1}{2}\int d^3x_1 d^3x_2\ V_2(x_{12})|\phi(\mathbf{x}_1)\phi(\mathbf{x}_2)|^2 \right.$$
$$\left. + \frac{1}{6}\int d^3x_1 d^3x_2 d^3x_3\ V_3(x_{12}, x_{23})|\phi(\mathbf{x}_1)\phi(\mathbf{x}_2)\phi(\mathbf{x}_3)|^2\right]$$

We then find the $\phi(\mathbf{x})$ which minimises this expression. This, obviously, is a hard problem. But fortunately it is not one we need to solve in order to extract the main lessons. These come simply from the fact that there is a factor of N outside the bracket, but nothing inside. This confirms our earlier conclusion (6.25) that the mass of the baryon indeed scales as $M_{\text{baryon}} \sim N$. But we also learn something new, because whatever function $\phi(\mathbf{x})$ ends up being, it certainly does not depend on N. This means that the size of the baryon – its spatial profile in $\phi(\mathbf{x})$ – is order 1.

The mass and size of the baryon are rather suggestive. Recall that the large N limit is a theory of weakly coupled gauge singlets, interacting with coupling $1/N$. This means that the mass of the baryon scales as the inverse coupling, N, with the size independent of the coupling. But this is the typical behaviour of solitons. For example, the 't Hooft Polyakov monopole that we met in Section 2.8 has a mass which scales as $1/g^2$ and a size which is independent of g^2. This strongly suggests that the baryon should emerge as a soliton in large N QCD.

We have, of course, already seen a context in which baryons emerge as solitons: they are the Skyrmions in the chiral Lagrangian that we met in Section 5.3. To my knowledge, this connection has not been fully explained.

Before we move on, there is one further twist to the "baryons as solitons" story. The mass of the baryon, N, is not quite like the mass of the monopole: it is proportional to the inverse coupling, rather than the square of the inverse coupling. Returning to the language of string theory that we introduced in Section 6.2.2, the mass of the baryon scales as

$$M_{\text{baryon}} \sim \frac{1}{g_s}$$

with $g_s = 1/N$ the string coupling constant. This suggests that baryons are a rather special kind of soliton: they are D-branes. These are objects in string theory on which strings can end, and have a number of magical properties. (You can read more about D-branes in the lectures on String Theory.) With its N constituent quarks, the baryon is indeed a vertex on which N QCD flux tubes can end.

6.4 The Chiral Lagrangian Revisited

In this section, we will see what becomes of the chiral Lagrangian at large N. Let's first recall the usual story: Yang-Mills coupled to N_f massless fermions has a classical global symmetry

$$G = U(N_f)_L \times U(N_f)_R \tag{6.27}$$

However, the anomaly means that $U(1)_A$ does not survive the quantisation process, leaving us just with $U(1)_V \times SU(N_f)_L \times SU(N_f)_R$. This is subsequently broken to $U(1)_V \times SU(N_f)_V$, and the resulting Goldstone modes are described by the chiral Lagrangian.

How does this story change at large N. The key lies in the anomaly, which is given by

$$\partial_\mu J^\mu_A = \frac{g^2 N_f}{8\pi^2} \mathrm{tr}\, F_{\mu\nu} {}^\star F^{\mu\nu} \tag{6.28}$$

In the large N limit, we send $g^2 \to 0$ keeping $\lambda = g^2 N$ fixed. This suggests that the anomaly is suppressed in the large N limit and the quantum theory enjoys the full chiral symmetry (6.27). This means that there is one further Goldstone mode that appears: the η' meson. In this section we will see how this plays out.

6.4.1 Including the η'

Our first steps are a straightforward generalisation of the chiral Lagangian derived in Section 5.2. The chiral condensate takes the form

$$\langle \bar{\psi}_{-\tilde{i}} \psi_{+\tilde{j}} \rangle = \sigma \Sigma_{\tilde{i}\tilde{j}}$$

but now with $\Sigma \in U(N_f)$ rather than $SU(N_f)$. (The ugly $\tilde{i}, \tilde{j} = 1, \ldots, N_f$ flavour indices are to ensure that we don't confuse them with the i, j colour indices we've used elsewhere in this Section.) As before, we promote the order parameter to a dynamical field, $\Sigma \to \Sigma(x)$, whose ripples describe our massless mesons, transforming under the chiral symmetry G as

$$\Sigma(x) \to L^\dagger \Sigma(x)\, R \tag{6.29}$$

with $L \times R \in G$. The overall phase of Σ is our new Goldstone boson, η',

$$\det \Sigma = e^{i\eta'/f_{\eta'}} \tag{6.30}$$

We would now like to construct the Lagrangian consistent with the chiral symmetry (6.29). Unlike in Section 5.2, we now have two different terms with two derivatives,

$$(\mathrm{tr}\,\Sigma^\dagger \partial_\mu \Sigma)^2 \quad \text{and} \quad \mathrm{tr}(\partial_\mu \Sigma^\dagger \partial^\mu \Sigma)^2 \tag{6.31}$$

The first term vanishes when $\Sigma \subset SU(N_f)$, but survives when $\Sigma \subset U(N_f)$. In other words, it provides a kinetic term only for η'. Meanwhile, the second term treats all Goldstone modes on the same footing.

Large N-ology tells us that all these mesons have the same properties and, in particular, to leading order in $1/N$ we have $f_{\eta'} = f_\pi$. This means that we need only the second kinetic term and the chiral Lagrangian takes the same form as (5.7),

$$\mathcal{L} = \frac{f_\pi^2}{4}\,\mathrm{tr}(\partial_\mu \Sigma^\dagger \partial^\mu \Sigma)^2$$

We can compute the expected scaling of f_π with N. Recall that the pion decay constant f_π is defined by (5.13)

$$\langle 0|J^a_{L\,\mu}(x)|\pi^b(p)\rangle = -i\frac{f_\pi}{2}\delta^{ab}\,p_\mu e^{-ix\cdot p}$$

with J_L a generator of the $SU(N_f)$ flavour current. At this point we need to be a little careful about normalisations. The current J above is defined with the usual kinetic kinetic term $\mathcal{L} \sim i\bar{\psi}\not{D}\psi$. Meanwhile, our large N counting used a different normalisation in which there was an overall factor of N outside the action. Chasing this through, means that the current J_L is related to the appropriate normalised large N current (6.19) by

$$J_L = \sqrt{N}\mathcal{J}_L$$

We can then use the general result (6.20) to find

$$\langle J_L J_L\rangle = \sum_n \langle 0|J_L|n\rangle\langle n|J_L|0\rangle \sim N \quad \Rightarrow \quad \langle 0|J_L|n\rangle \sim \sqrt{N}$$

This means that the pion decay constant scales as

$$f_\pi \sim \sqrt{N}$$

6.4.2 Rediscovering the Anomaly

So far, things are rather easy. Now we would like to consider what happens at the next order in $1/N$. Obviously, we could add the other kinetic term in (6.31), splitting $f_{\eta'}$ and f_π. This doesn't greatly change the physics and we will ignore this possibility below. Instead, there is a much more dramatic effect that we must take into account, because the anomaly now gives η' a mass. How do we describe that?

We can isolate η' by taking the determinant (6.30), and therefore introduce a mass term by

$$\mathcal{L} = \frac{f_\pi^2}{4}\,\mathrm{tr}(\partial_\mu\Sigma^\dagger\partial^\mu\Sigma)^2 - \frac{1}{2}f_\pi^2\, m_{\eta'}^2\,(-i\log\det\Sigma)^2$$

Here $m_{\eta'}^2$ is the mass which must vanish as $N\to\infty$. We will see shortly that $m_{\eta'}^2\sim 1/N$.

It is unusual to include a log term in an effective action. However, as we will now see, it captures a number of aspects of the anomaly. To illustrate this, let's first add masses for the other quarks. As we saw in Section 5.2.3, this is achieved by including the term

$$\mathcal{L} = \int d^4x\;\frac{f_\pi^2}{4}\,\mathrm{tr}\,(\partial^\mu\Sigma^\dagger\,\partial_\mu\Sigma) - \frac{\sigma}{2}\mathrm{tr}\,\left(M\Sigma+\Sigma^\dagger M^\dagger\right) - \frac{1}{2}f_\pi^2\,m_{\eta'}^2\,(-i\log\det\Sigma)^2$$

with M a complex mass matrix. By a suitable $SU(N_f)\times SU(N_f)$ rotation, we can always choose

$$M = e^{i\theta/N}\mathcal{M}$$

where $\mathcal{M}$ is diagonal, real and positive. This final phase can be removed by a $U(1)_A$ rotation, $\Sigma\to e^{-i\theta/N}\Sigma$ to make the mass real. But this now shows up in the mass term for the η',

$$\mathcal{L} = \int d^4x\;\frac{f_\pi^2}{4}\,\mathrm{tr}\,(\partial^\mu\Sigma^\dagger\,\partial_\mu\Sigma) - \frac{\sigma}{2}\mathrm{tr}\,\left(\mathcal{M}\Sigma+\Sigma^\dagger\mathcal{M}^\dagger\right) - \frac{1}{2}f_\pi^2\,m_{\eta'}^2\,(-i\log\det\Sigma-\theta)^2$$

However, we've played these games before: in Section 3.3.3, we saw that rotating the phase of the mass matrix in equivalent to introducing a theta angle. We conclude that this is how the QCD theta angle appears in the chiral Lagrangian.

We can now minimise this potential to find the ground state. With $\mathcal{M}$ diagonal, the ground state always takes the form

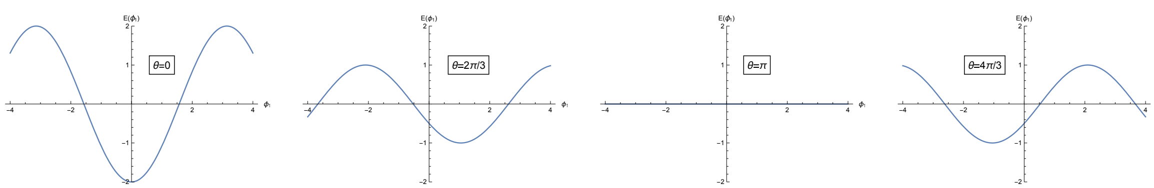
$$\Sigma = \mathrm{diag}\left(e^{i\phi_1},\ldots,e^{i\phi_{N_f}}\right)$$

The exact form depends in a fairly complicated manner on the choices of mass matrix $\mathcal{M}$ and theta angle. To proceed, we must make some assumptions. We will take m_η much bigger than all other masses, which means that we first impose the second term as a constraint

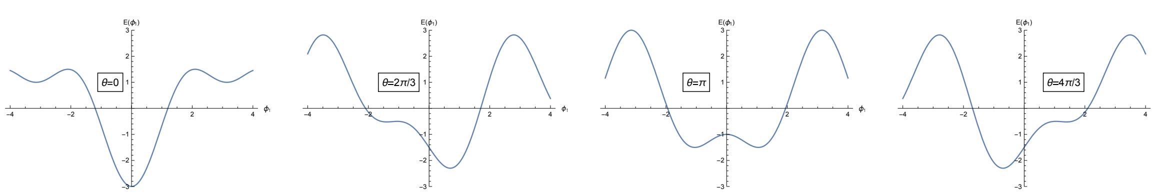
$$\sum_{i=1}^{N_f} \phi_i = \theta$$

We further look at the simplest case of a diagonal mass matrix: $\mathcal{M} = m\mathbf{1}_{N_f}$ with $m > 0$. We will then see how the ground states change as we vary θ.

For $\theta = 0$, the ground state sits at $\Sigma = 1$. Now we increase θ. What happens next differs slightly for $N_f = 2$ and $N_f > 2$. Let's start with $N_f = 2$. As we increase θ, the ground state moves to $\phi_1 > 0$ and the overall magnitude of the potential decreases. At $\theta \to \pi^-$, the ground state tends towards $\phi_1 = \pi/2$. At $\theta = \pi$ itself, the potential vanishes for all ϕ_1, which is symptomatic of a second order phase transition. If we now increase θ just a little more, the ground state jumps to $\phi_1 = -\pi/2$, before moving back towards $\phi_1 = 0$ as $\theta \to 2\pi$. The sequence is shown in the plots below for $\theta = 0, \frac{2\pi}{3}, \pi$ and $\frac{4\pi}{3}$



The fact that the potential vanishes when $\theta = \pi$ is special to $N_f = 2$. The story for $N_f \geq 3$ is similar, except that there are now just two degenerate vacua at $\theta = \pi$. This is characteristic of a first order phase transition. The potential for $N_f = 3$ for $\theta = 0, \frac{2\pi}{3}, \pi$ and $\frac{4\pi}{3}$ is shown below.



6.4.3 The Witten-Veneziano Formula

So far, we've happily incorporated the new η' Goldstone boson into our chiral Lagrangian. However, this brings something of a puzzle, which is to reconcile the following facts:

- The ground state energy is $E(\theta) \sim N^2$ and depends on θ.
- Quarks contribute to quantities such as $E(\theta)$ at order N.
- All θ dependence vanishes if we have a massless fermion.

These three facts seem incompatible. How can the $\sim N$ contribution from quarks cancel the $\sim N^2$ contribution from gluons to render $E(\theta)$ independent of θ?

To see how this might work, let's consider schematically the contribution to the susceptibility (6.16)

$$\chi(k) = \sum_{\text{glueballs}} \frac{N^2 a_n^2}{k^2 - M_n^2} + \sum_{\text{mesons}} \frac{N b_n^2}{k^2 - m_n^2}$$

where M_n are the masses of glueballs, m_n the masses of mesons, and a_n and b_n the amplitudes for $\text{tr}\, F_{\mu\nu}{}^\star F^{\mu\nu}$ to create these states from the vacuum,

$$\langle 0|\text{tr}\, F\,{}^\star F|n^{\text{th}} \text{ glueball}\rangle = N a_n \quad , \quad \langle 0|\text{tr}\, F\,{}^\star F|n^{\text{th}} \text{ meson}\rangle = \sqrt{N} b_n$$

We want the second term to cancel the first in the limit $k \rightarrow 0$. We can achieve this only if there is some meson whose mass scales as $m^2 \sim 1/N$. But this tallies with our discussion above; we expect that the η' becomes a genuine Goldstone boson in the large N limit. We're therefore led to the conclusion

$$\chi(0)\Big|_{\text{Yang-Mills}} = \frac{N b_{\eta'}^2}{m_{\eta'}^2} \tag{6.32}$$

But we can now use our anomaly equation (6.28) to write

$$\sqrt{N} b_{\eta'} = \langle 0|F\,{}^\star F|\eta'\rangle = \frac{8\pi^2 N}{\lambda N_f}\langle 0|\partial_\mu J_A^\mu|\eta'\rangle = \frac{8\pi^2 N}{\lambda N_f} p_\mu \langle 0|J_A^\mu|\eta'\rangle$$

But we know from our discussion of currents in the chiral Lagrangian (5.13) that $\langle 0|J_A^\mu|\eta'\rangle = -i\sqrt{N_f} f_\pi p^\mu$. (The factor of $\sqrt{N_f}$ here is a novel normalisation, but ensures that f_π is independent of N_f in the large N limit.) We therefore find that $\sqrt{N} b_{\eta'} = (8\pi^2 N/\sqrt{N_f}\lambda) f_\pi m_{\eta'}^2$. Inserting this into (6.32), and using (6.17), we have

$$m_{\eta'}^2 = \frac{4N_f}{f_\pi^2} \frac{d^2 E}{d\theta^2}\Big|_{\theta=0}$$

This is the *Witten-Veneziano formula*. Rather remarkably, it relates the mass of the η' meson to the vacuum energy $\chi(0)$ of large N, pure Yang-Mills theory without quarks.

It's worth pausing to see how the N scaling works in this formula. While $E(\theta) \sim N$, we expect that $d^2E/d\theta^2$ is of order 1. Meanwhile, $f_\pi \sim \sqrt{N}$. We then see that $m^2_{\eta'} \sim 1/N$ as anticipated previously.

We don't know how to measure the topological susceptibility $\chi(0)$ experimentally. Nonetheless, we can use the Witten-Veneziano formula, with $m_{\eta'} \approx 950$ and $f_\pi \approx 93$ MeV and $N_f = 3$ to get $d^2E/d\theta^2 \approx (150\text{MeV})^4$.

6.5 Further Reading

The large N expansion in Yang-Mills was introduced by 't Hooft in 1974 [97]. ('t Hooft was astonishingly productive in those years!) Although we didn't cover it in these lectures, 't Hooft quickly showed how these methods could be used to solve QCD in two dimensions, a theory that is now referred to as the 't Hooft model [98].

The discussion of baryons in the $1/N$ expansion is due to Witten [221], as is the $1/D$ expansion in atomic physics [223]. Witten goes on to apply the $1/D$ expansion to helium. It's clever, but also shows why chemists tend not to adopt this approach.

The fact that, despite all appearances, dependence on the θ angle survives in the large N limit was first emphasised by Witten in [220]. The large N limit of the chiral Lagrangian was constructed in [224, 41], and the Witten-Veneziano formula was introduced in [222, 199]. The symmetry breaking pattern needed for the chiral Lagrangian can be proven in the large N limit: this result is due to Coleman and Witten [31]. The idea that QCD at $\theta = \pi$ spontaneously breaks time reversal was pointed out pre-QCD and pre-theta by Dashen [38] and is sometimes referred to as the Dashen phase.

The tantalising connection between string theory and the large N expansion can be made explicit in a number of low dimensional examples; the lectures by Ginsparg and Moore are a good place to start [75]. In $d = 3+1$ dimensions, this relationship underlies the AdS/CFT correspondence [129].

Coleman's lectures remain the go-to place for a gentle introduction to the $1/N$ expansion [32]. Manohar has written an excellent review of the phenomenology of large N QCD [132]. Any number of reviews on the gauge/gravity duality also contain a discussion of $1/N$ and its relationship to string theory: I particularly like the lectures by McGreevy [135].

7. Quantum Field Theory on the Line

In this section, and the next, we describe the physics of relativistic quantum field theories that live in $d = 1 + 1$ and $d = 2 + 1$ dimensions.

There are several reasons to be interested in quantum field theories in lower dimensions. Perhaps most importantly, these field theories play important roles in condensed matter systems. However, it turns out that it is often easier to solve quantum field theories in lower dimensions. This makes them a testing ground where we can understand some of the subtleties of field theory and build some intuition for the kinds of issues that arise when the interactions between fields becomes strong.

As we go down in dimension, we find an increased richness in the interactions that a field theory can enjoy. More specifically, we find an increase in the number of relevant and marginally relevant interactions that theories admit. These are the terms that drive us from weakly coupled physics in the UV towards something more interesting in the IR. In $d = 3 + 1$, this can only be achieved by non-Abelian gauge fields. As we will see below, in lower dimensions we have other options. This means that Yang-Mills theory, which has dominated our lectures so far, becomes somewhat less prominent in the story of lower dimensional quantum field theories.

7.1 Electromagnetism in Two Dimensions

Maxwell theory in $d = 1 + 1$ dimensions is rather special. The gauge field is A_μ, with $\mu = 0, 1$ and the corresponding field strength has just a single component F_{01}. The action is given by

$$S = \int d^2x \ -\frac{1}{2e^2} F_{01} F^{01} - A_\mu j^\mu$$

where j^μ denotes the coupling to charged matter. Note that we have retained the notation of Yang-Mills theory where the coupling constant e^2 sits outside the action. With this convention, the matter is taken to have integer valued electric charge.

Electromagnetism in $d = 1 + 1$ dimensions has a number of properties that are rather different from its $d = 3 + 1$ dimensional counterpart. These occur both at the classical and quantum levels. Let's first look at some basic classical properties. The first difference comes in the pure Maxwell theory, which has equation of motion

$$\partial_0 F^{01} = \partial_1 F^{01} = 0 \tag{7.1}$$

We see that this allows only for a constant electric field. In particular, there are no electromagnetic wave solutions in $d = 1 + 1$ dimensions.

This is an important point and it's worth explaining from a slightly different perspective. In general d dimensional spacetime, the gauge field is A_μ with the index running over $\mu = 0, 1, \ldots, d-1$. However, not all of these components are physical. The standard way to isolate the physical degrees of freedom is to use the gauge symmetry $A_\mu \to A_\mu + \partial_\mu \omega$ to set $A_0 = 0$. This leaves us with only the spatial gauge fields $\vec{A}$. However, we still have to impose the equation of motion for A_0 which is solved by insisting that $\nabla \cdot \vec{A} = 0$. This projects out the longitudinal fluctuations of $\vec{A}$, leaving us just with the transverse modes. The upshot is that the gauge field in d dimensions carries $d-2$ physical degrees of freedom. When $d = 3+1$, these are the familiar two polarisation modes of the photon. However, in $d = 1+1$ dimensions, there are no transverse modes and the electromagnetic field has no propagating degrees of freedom.

Now let's look at what happens when we add matter. The classical equations of motion are

$$\frac{1}{e^2}\partial_\mu F^{\mu\nu} = j^\nu$$

We can consider placing a point charge q at the origin, so the equation that we have to solve is

$$\frac{1}{e^2}\partial_1 F^{01} = q\delta(x) \quad \Rightarrow \quad F^{01} = qe^2\theta(x) + \mathcal{E} \tag{7.2}$$

where $\theta(x)$ is the Heaviside step function ($\theta(x) = 0$ for $x < 0$ and $\theta(x) = 1$ for $x > 0$) and $\mathcal{E}$ is a constant background electric field which is typically fixed by the choice of electric field at spatial infinity. We see that the electric field emitted by a point charge in $d = 1+1$ dimensions is constant. (This is the same as the statement that a uniform surface charge in $d = 3+1$ dimensions gives rise to a constant electric field.)

The energy contained in the electric field is

$$H = \int dx\ \frac{1}{2e^2}F_{01}^2 \tag{7.3}$$

This means that a classical point charge in $d = 1+1$ dimensions costs infinite energy. The finite energy states must be charge neutral. To this end, consider a charge q at position $x = -L/2$ and a charge $-q$ at position $x = +L/2$. We have the equation of motion

$$\frac{1}{e^2}\partial_1 F^{01} = q\left[\delta(-L/2) - \delta(+L/2)\right] \quad \Rightarrow \quad F^{01} = \begin{cases} qe^2 & x \in (-L/2, +L/2) \\ 0 & \text{otherwise} \end{cases} \tag{7.4}$$

where we have chosen the integration constant $\mathcal{E}$ in (7.2) to ensure vanishing electric field at $x = \pm\infty$. The total energy (7.3) stored in the electric field is

$$H = \frac{q^2 e^2}{2} L$$

We see that the energy grows linearly with the separation. In other words, electric charges in $d = 1+1$ dimensions are classically confined. The reason is that the electric field is forced to form a flux tube, simply because it has nowhere else to go.

7.1.1 The Theta Angle

As we described above, pure Maxwell theory in $d = 1+1$ dimensions has no propagating, wave-like solutions. This does not, however, mean that the theory is completely devoid of content. The classical equations of motion (7.1) still allow for constant electric fields. As we now explain, this is enough to give rise to a Hilbert space in the quantum theory.

We also take this opportunity to add a new ingredient to pure Maxwell theory. This is a θ term, analogous to the θ terms which we met in four dimensional gauge theories in Sections 1.2 and 2.2. (In fact, such a term exists in any even spacetime dimension.) The action is

$$S = \int d^2x \; \left(\frac{1}{2e^2} F_{01}^2 + \frac{\theta}{2\pi} F_{01} \right) \tag{7.5}$$

Like its four-dimensional counterpart, the theta term is a total derivative and does not affect the classical equations of motion. Nonetheless, it does affect the quantum spectrum.

Our first task is to isolate the dynamical degrees of freedom in pure Maxwell theory. This is best illustrated by taking the theory to live on $\mathbf{R} \times \mathbf{S}^1$ where we take the spatial $\mathbf{S}^1$ to have radius R. Although the theory has no propagating degrees of freedom, there is a single physical mode which is spread all over the $\mathbf{S}^1$. It is known as the *zero mode*

$$\phi(t) = \int_0^{2\pi R} dx \; A_1(x,t) \tag{7.6}$$

The fact that $\phi(t)$ does not depend on space means that there is no sense in which it propagates. Said another way, this is just a single degree of freedom rather than the infinite number of degrees of freedom — one per spatial point — that are typically contained in a field theory.

The quantity $\phi(t)$ is gauge invariant and dimensionless. Importantly, it is also periodic. This arises from performing large gauge transformations of the kind that we met a number of times previously. These are single valued gauge transformations of the form $e^{i\omega(x)}$, but where ω is not single valued. Instead ω obeys

$$\omega(x = 2\pi R) = \omega(x = 0) + 2\pi n \quad \text{ for some } n \in \mathbf{Z}$$

The simplest such example, with $n = 1$, is just $\omega = x/R$. Under such a gauge transformation, we have

$$A_1 \to A_1 + \partial_x \omega = A_1 + \frac{1}{R}$$

Under this, or any gauge transformation with $n = 1$, the zero mode (7.6) transform as

$$\phi \to \phi + 2\pi$$

This is the statement that ϕ is periodic.

The dynamics of ϕ follows from the Lagrangian

$$L = \frac{1}{4\pi e^2 R}\,\dot{\phi}^2 + \frac{\theta}{2\pi}\dot{\phi}$$

As usual, the θ term does not affect the classical equations of motion, but it does affect the definition of the canonical momentum p, which is given by

$$p = \frac{1}{2\pi e^2 R}\dot{\phi} + \frac{\theta}{2\pi}$$

The Hamiltonian is then

$$H = \frac{1}{4\pi e^2 R}\,\dot{\phi}^2 = \pi e^2 R \left(p - \frac{\theta}{2\pi}\right)^2$$

This is precisely the problem of a particle moving on a circle in the presence of flux. We already met this in Section 2.2 as an analogy which captures some of the aspects of the four dimensional theta term. We also met it subsequently in Section 3.6 where we saw that it exhibits some interesting discrete anomaly when $\theta = \pi$; we won't need this fact in what follows.

A familiar theme now emerges: although the classical physics remains unchanged by θ, there is an important effect in the quantum physics. This arises because the wavefunctions ψ should be single valued. The energy eigenstates are $\psi_l = e^{il\phi}$ with $l \in \mathbf{Z}$. The spectrum is given by

$$H\psi_l = E_l\psi_l \quad \text{with} \quad E_l = \pi e^2 R\left(l - \frac{\theta}{2\pi}\right)^2$$

The spectrum is periodic in θ as expected. For $\theta \in (-\pi, \pi)$, the ground state is $l = 0$. For $\theta = \pm\pi$, there are two degenerate ground states, $l = 0$ and $l = \pm 1$. If we increase $\theta \to \theta + 2\pi$, then the spectrum remains the same, but all the states shift along by one. This is a phenomenon known as spectral flow.

7.1.2 The Theta Angle is a Background Electric Field

There is a particularly simple interpretation of the θ angle in two dimensions: it gives rise to a background electric field. We have already noticed that, classically, the equation of motion $\partial_1 F^{10} = 0$ allows for a constant background electric field. In $A_0 = 0$ gauge, this is given by

$$F_{01} = \frac{1}{2\pi R}\dot{\phi} = e^2\left(p - \frac{\theta}{2\pi}\right)$$

Evaluated on the state ψ_l, the electric field is given by

$$F_{01} = e^2\left(l - \frac{\theta}{2\pi}\right) \qquad l \in \mathbf{Z} \tag{7.7}$$

We see that the Hilbert space of pure Maxwell theory in $d = 1 + 1$ dimensions can be thought of as describing integrally spaced, constant electric fields, shifted by the θ angle.

The above analysis was all performed on a spatial circle of radius R. However, the ultimate quantisation of the electric field (7.7) is independent of this radius. Indeed, there is a particularly simple way to see that the θ angle gives rise to a background electric field if we work on spatial $\mathbf{R}$. We return to the action (7.5) which, noting that the θ term is a total derivative, we rewrite as

$$S = \int d^2x \; -\frac{1}{2e^2}F_{01}F^{01} + \frac{\theta}{2\pi}\oint dx^\mu A_\mu$$

where the contour integral should be taken around the boundary of spacetime. Written this way, it looks like the insertion of a Wilson line, with a particle of charge $\theta/2\pi$ at $x = -\infty$, together with a particle of charge $-\theta/2\pi$ at $x = +\infty$. As we saw in the classical analysis leading to (7.4), this results in an electric field $F_{01} = -\theta e^2/2\pi$. This agrees with the more careful quantum computation (7.7).

Our discussion above suggests that something interesting happens when $\theta = \pi$: there are two degenerate ground states. These are the states (7.7) with $l = 0$ and $l = +1$ which have $F_{01} = \pm e^2\theta/2\pi$. If we were to change θ slowly, passing through the value $\theta = \pi$, we jump discontinuously from the background field $F_{01} = -e^2/2$ to the background field $F_{01} = +e^2/2$. This is an example of a first order phase transition.

Our next task is to understand what happens to our theory when we include dynamical matter.

7.2 The Abelian-Higgs Model

In this section, we consider a $U(1)$ gauge theory coupled to a complex scalar field ϕ. The action is

$$S = \int d^2x \; \frac{1}{2e^2}F_{01}^2 + \frac{\theta}{2\pi}F_{01} + |\mathcal{D}_\mu\phi|^2 - m^2|\phi|^2 - \frac{\lambda}{2}|\phi|^4 \tag{7.8}$$

We take the scalar field to have charge $q = 1$, so that $\mathcal{D}_\mu\phi = \partial_\mu\phi - iA_\mu\phi$. In two-dimensions, the gauge coupling has scaling dimension $[e^2] = 2$. This means that electromagnetism will always be strongly coupled in the infra-red unless some other physics kicks in at a higher scale. It will be straightforward to understand the dynamics of the scalar when $|m^2| \gg e^2$, but harder in the regime $|m^2| \lesssim e^2$. In what follows, we will discuss the Abelian-Higgs model in two different semi-classical regimes: $m^2 \gg e^2$ and $m^2 \ll -e^2$.

$\underline{m^2 \gg e^2}$: For very large, positive m^2, quantization of the scalar field simply gives us particles and anti-particles, each of mass m and charge $q = \pm 1$. These particles then interact through the two-dimensional Coulomb force. We will call this the *Coulomb phase*.

To start our discussion, let's focus on the case $\theta = 0$. A particle of charge $q = 1$ gives rise to a constant electric field, $F_{01} = e^2$, which we take to be emitted to the right of the particle. If an anti-particle, with charge $q = -1$, sits at a distance L, as shown in the figure, then we are left with an energy $\mathcal{E}$ in the electric field given by

$$\mathcal{E} = \frac{e^2L}{2} \tag{7.9}$$

This linear growth in energy is the characteristic of confinement. We see that, in $d = 1 + 1$ dimensions, confinement occurs rather naturally, with the electric field automatically forming a flux tube. Indeed, in two dimensions, the Coulomb phase is the same thing as the confining phase.

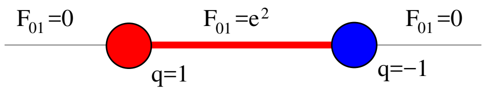


Figure 51: When $\theta = 0$, there is a confining string between particles and anti-particles

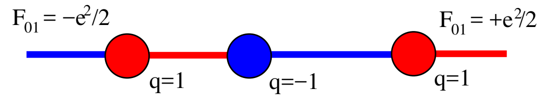


Figure 52: When $\theta = \pi$, the string tensions cancel on either side and alternating particles/anti-particles feel no long-distance force.

There is, however, a limit to how far this flux tube can stretch. If we attempt to separate a particle-anti-particle pair too far, then the energy stored in the string is greater than the energy required to create a particle-anti-particle pair, and we expect the string to break. This should happen for $e^2L/2 \gtrsim 2m$ or, $L \gtrsim 4m/e^2$. The upshot of this argument is that we expect the spectrum of the theory to consist of a tower of neutral meson-like states, each containing a particle and anti-particle. The low-lying modes of this spectrum can be easily computed using a non-relativistic Schrödinger equation, although we will not do so here[12].

We could also ask how the theory responds if we insert test charges of $q \notin \mathbf{Z}$. A particle-anti-particle pair will, once again, be confined by the electric field $F_{01} = qe^2$. However, the electric field cannot be removed by pair creation of ϕ particles, since these can only result in a change $\Delta F_{01} = e^2$. We learn that these test particles are confined no matter how far they are separated.

The story does not change much as we turn on θ, until we reach $\theta = \pi$. Now something more interesting can happen. Suppose that the electric field at $x \to -\infty$ is given by $F_{01} = -e^2/2$. The presence of a particle of charge q means that the electric field jumps to $F_{01} = +e^2/2$. Since its magnitude doesn't change, this particle is free to roam along the line. We can follow this by a chain of alternating particles and anti-particles, each of which is free to move at no extra cost of energy (ignoring any short distance forces between the particles). In this case, the particles are no longer confined, at least when placed with a particular ordering along the line.

$\underline{m^2 \ll -e^2}$: With a large negative mass-squared, the scalar condenses. The minimum of the classical potential lies at

$$|\phi|^2 = -\frac{m^2}{\lambda} \tag{7.10}$$

[12]See, for example, the discussion of the linear potential and Airy function in the lectures on Applications of Quantum Mechanics.

Our naive expectation is that we now lie in the Higgs phase, with the electric field screened and the charged particles free to roam at will. Rather strikingly, this naive expectation is completely wrong. Instead, it turns out that the physics in this regime is exactly the same as the physics when $m^2 \gg e^2$. As we now explain, this is due to a special property of Abelian gauge theories in two dimensions.

7.2.1 Vortices

The new ingredient is the existence of vortices. These are solutions to the equations of motion that exist when the theory is formulated in the Euclidean space. These same vortices were discussed in Section 2.5.2, where they arise as string-like solutions in $d = 3 + 1$ dimensions. In contrast, these same solutions will now be localised in spacetime; they play a role similar to the instantons discussed in Section 2.3 although, as we shall see, their effect is arguably more profound: they destroy the long-range order (7.10).

To see this, let's first formulate the action in Euclidean space. We write the action (7.8) as

$$S_E = \int d^2x\ \frac{1}{2e^2}F_{12}^2 + \frac{i\theta}{2\pi}F_{12} + |\mathcal{D}_i\phi|^2 + \frac{\lambda}{2}\left(|\phi|^2 - v^2\right)^2 \tag{7.11}$$

where now $i = 1, 2$. We have written the Higgs vev as $v^2 = -m^2/\lambda$. A finite action configuration requires $|\phi| \to v$ as $r \to \infty$. This provides us with some interesting topology: the asymptotic $\mathbf{S}_\infty$ of Euclidean spacetime is mapped into the $\mathbf{S}^1$ defined by $|\phi| = v$. Mathematically, this means that field configurations are characterised by $\Pi_1(\mathbf{S}^1) = \mathbf{Z}$, in which the phase of ϕ winds asymptotically. For example, we may take

$$\phi \to e^{in\theta}v \tag{7.12}$$

where θ is the polar coordinate on the spatial $\mathbf{R}^2$ (and not to be confused with the coefficient of the topological tern, which is also denoted as the theta angle). This is single valued for $n \in \mathbf{Z}$. This integer n is called the *winding*. Configurations with $n > 0$ are called *vortices*; those with $n < 0$ are *anti-vortices*.

However, a scalar that winds in this way has infinite action unless it is also accompanied by non-vanishing gauge field. This is because the gradient terms are given by

$$\int d^2x\ |\partial_i\phi|^2 = \int d\theta dr\ r\ \frac{1}{r^2}|\partial_\theta\phi|^2 + \ldots = 2\pi\int_0^\infty dr\ \frac{n^2}{r}|\phi|^2 + \ldots$$

which is logarithmically divergent. We see that the trouble arises because the gradient terms fall off too slowly, as $1/r$. To compensate for this, we must turn on a gauge field A_i, such that $\mathcal{D}_i\phi = \partial_i\phi - iA_i\phi$ falls off at a faster rate. For a configuration that winds as (7.12), this ensures that the gauge field must take the asymptotic form $A_\theta \rightarrow n/r$ which, in turn, tells us that vortices are accompanied by a quantised flux

$$\frac{1}{2\pi}\int d^2x\ F_{12} = \frac{1}{2\pi}\oint d\theta\ rA_\theta = n \tag{7.13}$$

One can construct solutions to the equations of motion with this asymptotic behaviour by working with an ansatz of the form $\phi(x) = e^{in\theta}g_n(r)$ and $A_\theta = nf_n(r)$, where the radial functions $g_n(r)$ and $f_n(r)$ satisfy second order differential equations subject to certain boundary conditions. The exact form of these solutions will not concern us here: all we need is the statement that solutions always exist for $n = \pm 1$. In this solution, the flux is restricted to a region of size $1/ev$, while the scalar field deviates from the vacuum over a region $1/\sqrt{\lambda}v$. We'll denote the vortex size, a, by the larger of these two scales,

$$a = \max\left(\frac{1}{ev}, \frac{1}{\sqrt{\lambda}v}\right)$$

We will also denote the real part of the action for a single, $n = \pm 1$, vortex as $S_{\rm vortex}$. Because the vortices come with flux (7.13), their contribution to the path integral will have the characteristic form

$$e^{-S_{\rm vortex}\pm i\theta}$$

where the $\pm$ sign distinguishes a vortex from an anti-vortex.

So much for solutions with $n = \pm 1$. What about vortices with higher winding? It turns out that solutions exist for higher n, but only when $\lambda < e^2$. Nonetheless, we shall not make use of these solutions. Instead, it will suffice to consider a dilute gas of $n = \pm 1$ vortices separated by distances $\gg a$.

Summing over Vortices

Let's start by computing the partition function,

$$Z[\theta] = \int \mathcal{D}A\,\mathcal{D}\phi\ \exp\left(-S_E[A,\phi]\right)$$

As always, the partition function depends on the parameters, or sources, of the action. As the notation suggests, we will be particularly interested in the dependence on the theta angle. In the semi-classical approximation, this path integral gets contributions from the (approximate) solutions of far-separated vortices and anti-vortices. The strategy for performing these kinds of calculations was sketched in Section 2.3.3 in the context of the double well potential in quantum mechanics. The contribution from a single vortex takes the schematic form

$$Z_{\text{vortex}}[\theta] = V\, K\; e^{-S_{\text{vortex}}+i\theta}$$

Here V denotes the volume of spacetime (which, of course, is really an area since we are in two dimensions). This factor comes from the fact that the vortex can sit anywhere. V is, of course, infinite if we work on $\mathbf{R}^2$ but it will prove useful to consider it finite for now. The factor K comes from computing the one-loop determinant contribution around the background of the vortex; it will depend on parameters such as e^2, v^2 and λ but its precise form will not be important for our needs. Finally, we have the characteristic exponential suppression of the vortex. Similarly, for an anti-vortex we have

$$Z_{\text{anti-vortex}}[\theta] = V\, K\, e^{-S_{\text{vortex}}-i\theta}$$

For our final expression, we sum over a dilute gas with all possible combinations of p vortices and $\bar{p}$ anti-vortices, to get

$$Z[\theta] = \sum_{p,\bar{p}} \frac{1}{p!\bar{p}!}\,(VK\,e^{-S_{\text{vortex}}})^{p+\bar{p}}\,e^{i(p-\bar{p})\theta} = \exp\Big(2VK\,e^{-S_{\text{vortex}}}\,\cos\theta\Big) \tag{7.14}$$

What physics can we extract from this? First, this result tells us how the ground state energy varies as a function of θ. For this, we need to recall the interpretation of the partition function as a propagator between states,

$$Z[\theta] = \langle\theta|e^{-HT}|\theta\rangle = \langle\theta|e^{-E_0T}|\theta\rangle$$

If we write $V = TR$, with T the size of the temporal direction, and R the radius of the spatial direction, then we find the ground state energy density

$$\frac{E_0(\theta)}{R} = -2K\,e^{-S_{\text{vortex}}}\,\cos\theta \tag{7.15}$$

We can also compute the expected value of the background electric field. This is

$$\langle F_{12}\rangle = -\frac{2\pi i}{V}\frac{\partial}{\partial\theta}\log\, Z[\theta] = 4\pi i K\,e^{-S_{\text{vortex}}}\,\sin\theta$$

The fact that the right-hand-side is imaginary should not concern us; after Wick rotating back to Lorentzian signature, we get the result

$$\langle F_{01}\rangle = 4\pi K\, e^{-S_{\rm vortex}}\,\sin\theta$$

We see that turning on a θ angle once again induces a background electric field. Admittedly, there are some differences from the case of pure electromagnetism (7.7) or, indeed, the case of $m^2 \gg e^2$. In particular, the electric field is maximum at $\theta = \pi/2$, rather than $\theta = \pi$.

Classically, the energy density in the electric field is proportional to F_{01}^2. Quantum mechanically, the energy density (7.15) is not proportional to $\langle F_{01}\rangle^2$; instead, it is proportional to $\langle F_{01}^2\rangle \sim \partial^2/\partial\theta^2 \log Z$. This is telling us that there are large fluctuations in the electric field. At $\theta = \pi$, it is these fluctuations which are contributing to the energy, even though $\langle F_{01}\rangle = 0$.

Note in particular, that when $\theta = \pi$, there is a change in the vacuum structure: when $m^2 \gg e^2$, there were two values for the electric field, $\langle F_{01}\rangle = \pm e^2/2$, while for $m^2 \ll e^2$ there is just one, $\langle F_{01}\rangle = 0$. This behaviour is characteristic of a phase transition and we will return to it shortly when we sketch the phase diagram of the theory.

7.2.2 The Wilson Loop

We can now address our main question of interest: when $m^2 \ll -e^2$, are charged particles screened, as one would expect in a Higgs phase? To answer this we use the Wilson loop, introduced in Section 2.5.3, describing the insertion of a particle with charge q, and an anti-particle with charge $-q$,

$$W[C] = \exp\left(iq\oint_C A\right) \tag{7.16}$$

Here C is the rectangular loop; the particle and anti-particle are separated by a spatial distance L, and propagate for time T'. We will take each of these distances to be much larger than the size of the vortices, so $L, T' \gg a$, but much smaller than the size of our universe, so $L \ll R$ and $T' \ll T$.

We would like to compute the expectation value of the Wilson loop,

$$\langle W[C]\rangle = \frac{1}{Z}\int \mathcal{D}A\,\mathcal{D}\phi\; W[C]\,\exp\left(-S_E[A,\phi]\right) \tag{7.17}$$

But this is particularly simple in the semi-classical approximation. First, we assume that we can divide all (anti) vortices into those inside the loop C, and those outside.

This ignores those vortices that happen to overlap with the curve C, but these should be negligible when C is large. In the semi-classical approximation, the expression (7.17) decomposes into two pieces; one from inside the loop and the other from outside the loop,

$$\int \mathcal{D}A\,\mathcal{D}\phi\; W[C]\; \exp\left(-S_E[A,\phi]\right) = \tilde{Z}_{\text{inside}}[\theta]\,\tilde{Z}_{\text{outside}}[\theta]$$

The contribution from outside the loop is given by our original expression for $Z[\theta]$ (7.14), but with the area of spacetime V reduced by the area of the loop,

$$\tilde{Z}_{\text{outside}}[\theta] = \exp\Big(2(V-LT')K\,e^{-S_{\text{vortex}}}\cos\theta\Big)$$

Meanwhile, the Wilson loop affects only the contribution $\tilde{Z}_{\text{inside}}$ from inside the loop. In a given background, the Wilson loop (7.16) simply counts the total winding number, $\nu = \#(\text{vortices}) - \#(\text{anti-vortices})$ in the loop.

$$W[C] = e^{2\pi i q\nu}$$

Comparing to the expression (7.14), we see that the Wilson loop effectively shifts the theta angle $\theta \to \theta + 2\pi q$. We therefore have

$$\tilde{Z}_{\text{inside}}[\theta] = \exp\Big(2LT'K\,e^{-S_{\text{vortex}}}\cos(\theta+2\pi q)\Big)$$

Combining these results, the expectation value of the Wilson loop becomes

$$\langle W[C]\rangle = \exp\Big(2LT'K\,e^{-S_{\text{vortex}}}\;[\cos(\theta+2\pi q)-\cos\theta]\Big)$$

Our task now is to interpret this result. First notice that, for $q \notin \mathbf{Z}$, the Wilson loop exhibits an area law, telling us that the charges are confined. The string tension is given by the energy density

$$\frac{E}{L} = 2K\,e^{-S_{\text{vortex}}}\;[\cos(\theta+2\pi q)-\cos\theta] \tag{7.18}$$

This is already surprising, since it disagrees with our naive expectation that all charges should be screened in the Higgs phase. Instead, charges $q \notin \mathbf{Z}$ are confined, just as they are in the Coulomb phase with $m^2 \gg e^2$. In contrast, the string tension vanishes for $q = 1$. But, this too, agrees with the Coulomb phase picture, where pair creation of ϕ particles results in the string breaking, and the test particles forming gauge neutral meson states.

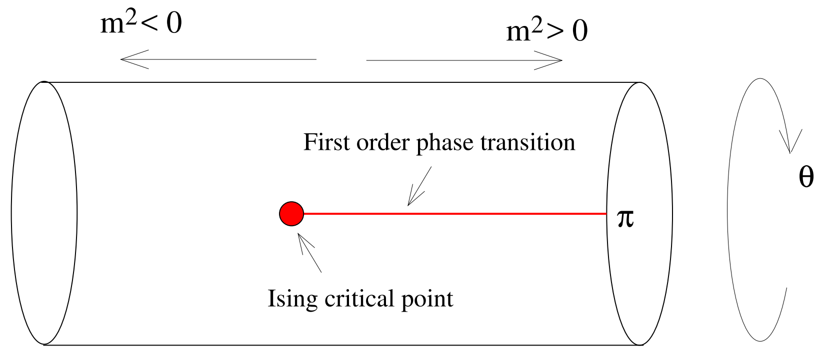


Figure 53: The phase diagram of the 2d Abelian Higgs model

We learn that, in the $d = 1+1$ Abelian Higgs model, there is no qualitative distinction between the behaviour of the theory at $m^2 \gg e^2$ and $m^2 \ll -e^2$. In both cases, the charged particles are confined. The only difference is a quantitative one: the string tension (7.18) is exponentially suppressed when $m^2 \ll -e^2$, compared to its value (7.9) when $m^2 \gg e^2$.

The Phase Diagram of the Abelian Higgs Model

The discussion above strongly suggests that there is no phase transition as we move from $m^2 \gg e^2$ to $m^2 \ll -e^2$: the would-be Higgs phase is washed away by vortices, leaving us only with the Coulomb phase.

However, there is one remaining subtlety, which occurs at $\theta = \pi$. As we saw above, there are two degenerate ground states, $\langle F_{01} \rangle = \pm e^2/2$ when $m^2 \gg e^2$, with a first order phase transition between them as we vary θ through π. In contrast, there is a unique ground state $\langle F_{01} \rangle = 0$ when $m^2 \ll -e^2$. This line of first order phase transitions must end somewhere. The simplest possibility is that it ends at a critical point at some value of the mass, presumably around $m^2 \sim -e^2$. Since the order parameter, F_{01}, is a parity-odd real scalar, it is natural to conjecture that this critical point is described by the $d = 2$ Ising CFT. The resulting phase diagram for the $d = 1 + 1$ Abelian Higgs model is shown in the figure.

(As an aside: The story above is similar, but ultimately different, from the story from the XY-model in $d = 1 + 1$ dimensions. This theory describes a complex scalar without the associated gauge field and was discussed in the lectures on Statistical Field Theory. Once again, vortices play an important role, but this time they induce the Kosterlitz-Thouless phase transition.)

7.3 The $\mathbf{CP}^{N-1}$ Model

We now turn to a theory that is closely related to the Abelian Higgs model. It consists of N complex scalars, ϕ_a, $a = 1, \ldots, N$, each coupled to a $U(1)$ gauge field with charge $q = +1$.

Our interest will lie in the theory where all scalars have negative m^2 so, following (7.11), we write the action in Euclidean space as

$$S = \int d^2x \; \frac{1}{2e^2} F_{12}^2 + \frac{\theta}{2\pi} F_{12} + \sum_{a=1}^{N} |\mathcal{D}_i \phi_a|^2 + \frac{\lambda}{2} \Big(\sum_{a=1}^{N} |\phi_a|^2 - v^2 \Big)^2 \qquad (7.19)$$

Note that our theory has a $SU(N)$ global symmetry, acting in the obvious way on the ϕ_a. This will be important below. As always, we would like to ask: what is the low-energy physics? This arises in the limit $e^2 \to \infty$ and $\lambda \to \infty$.

We can first look classically. At low-energies, the scalars sit in the minima of the potential,

$$\sum_{a=1}^{N} |\phi_a|^2 = v^2 \qquad (7.20)$$

This restricts the values of the complex ϕ fields to lie on a $\mathbf{S}^{2N-1}$ sphere of radius v. But we still have to divide out by gauge transformations. These identify configurations related by

$$\phi_a \to e^{i\alpha} \phi_a$$

We're left with scalar fields ϕ_a which parameterise the manifold,

$$\mathbf{S}^{2N-1}/U(1) = \mathbf{CP}^{N-1}$$

The manifold $\mathbf{CP}^{N-1}$ is known as *complex projective space*; it can be equivalently defined as the space of all complex lines in $\mathbf{C}^N$ which pass through the origin. $\mathbf{CP}^{N-1}$ has real dimension $2(N-1)$, or complex dimension $N-1$, and should be thought of as the complex analog of a round sphere, with the $SU(N)$ global symmetry descending to an isometry of $\mathbf{CP}^{N-1}$.

To proceed, we could choose to parameterise the ϕ_a by coordinates X^m on $\mathbf{CP}^{N-1}$. Plugging this back into our action would result in a non-linear sigma model of the kind

$$S = \int d^2x \; g_{mn}(X)\, \partial_i X^m \partial_i X^n \qquad (7.21)$$

where $g_{mn}(X)$ is the metric on $\mathbf{CP}^{N-1}$. (There is an additional term coming from the theta angle that we will discuss below.) For our purposes, however, it will prove more useful to work with the action (7.19); this form of the action is sometimes referred to as a *gauged linear sigma model*.

Classically, we learn that our $\mathbf{CP}^{N-1}$ model describes $N-1$, interacting, massless complex scalars. These are Goldstone modes. Indeed, picking a solution to (7.20) breaks the global $SU(N)$ symmetry to $SU(N-1)\times U(1)$, and the target space $\mathbf{CP}^{N-1}$ can equivalently be written as the coset space

$$\mathbf{CP}^{N-1} = \frac{SU(N)}{S[U(N-1)\times U(1)]}$$

The interactions between the Goldstone modes are determined by the coupling v^2, which is the size of $\mathbf{CP}^{N-1}$ or, more pertinently, the inverse curvature. This means that the theory is weakly coupled when $v^2 \gg 1$, and strongly coupled when $v^2 \ll 1$. However, as we should now expect: we don't get to choose, since quantum fluctuations will cause v^2 to change as we flow towards the infra-red. Do we flow to weak coupling or strong coupling? As we will see below, the answer is that we flow to strong coupling: the $\mathbf{CP}^{N-1}$ sigma model in two dimensions is asymptotically free.

7.3.1 A Mass Gap

Rather than compute the beta function for v^2, we will instead jump straight to figuring out the low-energy dynamics. This will give us the interesting information that we care about and, indirectly, also allow us to extract the beta function.

We're interested in the low-energy limit, $e^2, \lambda \to \infty$. We force the fields to live in the minima (7.20) by using a Lagrange multiplier constraint, and replace the action (7.19) with

$$S = \int d^2x \ \sum_{a=1}^{N} |\mathcal{D}_i\phi_a|^2 + i\sigma\Big(\sum_{a=1}^{N} |\phi_a|^2 - v^2\Big) + \frac{i\theta}{2\pi}F_{12} \tag{7.22}$$

where σ is now a dynamical field. Note that σ comes with a factor of i because we want it to impose the constraint (7.20) as a delta function. This will result in some strange looking factors of i in the effective potential below. However, upon Wick rotating back to Lorentzian signature, $\sigma \to i\sigma$ and everything looks nice and real again.

We have succeeded in writing the path integral so that the ϕ_a occur quadratically. They can now be happily integrated out, and we're left with the partition function,

$$Z = \int \mathcal{D}A\mathcal{D}\sigma\mathcal{D}\phi\mathcal{D}\phi^\star \ e^{-S} = \int \mathcal{D}A\mathcal{D}\sigma \ e^{-S_{\rm eff}}$$

with

$$S_{\text{eff}} = N\,\text{tr}\,\log\Big(-(\partial_i - iA_i)^2 + i\sigma\Big) - i\int d^2x\ \left(v^2\sigma - \frac{\theta}{2\pi}F_{12}\right) \tag{7.23}$$

The problem is that we're now left with a very complicated looking path integral over the auxiliary A and σ. In general, this is hard. However, some respite comes from the factor of N in front of the first term, which suggests that one can evaluate the integral using the saddle point in the limit $N \to \infty$. The is rather similar to the large N expansion that we met in Section 6 for Yang-Mills. It turns out, perhaps reasonably, that theories like the $\mathbf{CP}^{N-1}$ model, where the number of fields grows linearly with N, are much easier to deal with than Yang-Mills, where the number of fields grows as N^2.

To proceed, we will first restrict to configurations with $A_i = 0$, and extract an effective potential for the constant value of the auxiliary scalar σ. The trace above is an integral over momentum,

$$V_{\text{eff}}(\sigma) = N\int \frac{d^2k}{(2\pi)^2}\ \log(k^2 + i\sigma) - iv^2\sigma$$

The integral is divergent and requires us to introduce a UV cut-off Λ_{UV}. Performing the integral then gives

$$\begin{aligned} V_{\text{eff}}(\sigma) &= \frac{N}{4\pi}\left[i\sigma\log\left(\frac{i\sigma + \Lambda_{UV}^2}{i\sigma}\right) + \frac{1}{2}\Lambda^2\log\left(\frac{i\sigma + \Lambda_{UV}^2}{\Lambda_{UV}^2}\right)\right] - iv^2\sigma \\ &= \frac{N}{4\pi}i\sigma\left[1 - \log\left(\frac{i\sigma}{\Lambda_{UV}^2}\right)\right] - iv^2\sigma + \ldots \end{aligned} \tag{7.24}$$

where, to reach the second line, we've Taylor expanded in σ/Λ_{UV}^2, and the $\ldots$ include constant terms and terms which vanish as $\Lambda_{UV}^2 \to \infty$.

We still have to do the path integral over σ and that will, in general, be hard. However, the overall factor of N provides a glimmer of hope, because it means that the integral will be dominated by the saddle point in the $N \to \infty$ limit. This saddle point is given by

$$\begin{aligned} \frac{\partial V_{\text{eff}}}{\partial\sigma} = 0 \quad &\Rightarrow \quad \frac{N}{4\pi}\log\left(\frac{i\sigma}{\Lambda_{UV}^2}\right) = -v^2 \\ &\Rightarrow \quad i\sigma = \Lambda_{UV}^2\exp\left(-\frac{4\pi v^2}{N}\right) \end{aligned} \tag{7.25}$$

There are a number of different lessons to take from this. First, note that the $\mathbf{CP}^{N-1}$ model has undergone the phenomenon of dimensional transmutation that we saw in

Yang-Mills theory. The original Lagrangian (7.19) has only dimensionless parameters (at least, this is true after we have sent $e^2 \to \infty$). Nonetheless, the theory generates a physical dimensionful scale, arising from the UV cut-off Λ_{UV} in the partition function,

$$\Lambda_{CP^{N-1}} = \Lambda_{UV} \, \exp\left(-\frac{2\pi v^2}{N}\right) \tag{7.26}$$

The scale $\Lambda_{CP^{N-1}}$ is entirely analogous to Λ_{QCD} (2.59) that arises in Yang-Mills. While the cut-off Λ_{UV} is unphysical, the low-energy $\Lambda_{CP^{N-1}}$ is the scale at which interesting physical things can happen. This is sensible only because the dimensionless coupling v^2 runs under RG. In (7.26) the coupling should be thought of as being evaluated at the cut-off, $v^2 = v^2(\Lambda_{UV})$. More generally, the physical scale is written as

$$\Lambda_{CP^{N-1}} = \mu \, \exp\left(-\frac{2\pi v^2(\mu)}{N}\right)$$

From the requirement that this physical scale is invariant RG we can extract the beta-function for v^2,

$$\frac{d\Lambda_{CP^{N-1}}}{d\mu} = 0 \quad \Rightarrow \quad \mu \frac{dv^2}{d\mu} = \frac{N}{2\pi} \tag{7.27}$$

This tells us that v^2 gets smaller as we flow towards the IR (small μ). From our previous discussion, we know that this is the strong coupling limit of the $\mathbf{CP}^{N-1}$ model. In other words, this beta function tells us that, just like Yang-Mills, the $\mathbf{CP}^{N-1}$ model is strongly coupled in the IR, and asymptotically free in the UV.

Although the physics very much parallels that of Yang-Mills theory, it's worth pointing out the logic of our derivation is somewhat different. For Yang-Mills, we started off by computing the one-loop beta function and, from that, extracted the physical scale Λ_{QCD}. For the $\mathbf{CP}^{N-1}$ model, our discussion ran the other way round. Both are valid.

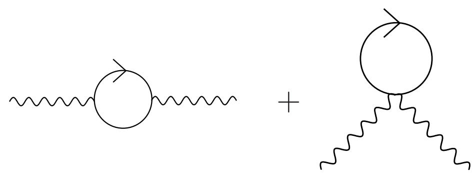
So far, we've figured out that there is a dynamically generated scale $\Lambda_{CP^{N-1}}$. But what happens at this scale? To see this, we need to note that, from (7.25), we have $i\sigma = \Lambda^2_{CP^{N-1}}$. But substituting this into (7.22), we see that an expectation value for σ acts as a mass term for our original fields ϕ_a. In other words, the 2d $\mathbf{CP}^{N-1}$ sigma model is not a theory of massless Goldstone modes at all! In the quantum theory, these massless modes pick up a mass given by $\Lambda_{CP^{N-1}}$. Moreover, the $SU(N)$

global symmetry is restored at low-energies. This is an example of the Mermin-Wagner theorem which states that there can be no Goldstone bosons in two dimensions[13].

Once again, we see the close analogy with Yang-Mills. Both theories appear massless but actually have a gap. The difference is that we can actually show this for the $\mathbf{CP}^{N-1}$ model.

7.3.2 Confinement

So far we have ignored the role of the gauge field in the effective action (7.23). At leading order, the effect of integrating out the scalars ϕ_a is captured by two Feynman diagrams:



These generate a Maxwell kinetic term

$$S_{\text{eff}} = -\frac{N}{48\pi\Lambda^2_{CP^{N-1}}}F_{\mu\nu}F^{\mu\nu}$$

Note that we started with a Maxwell term in our original action (7.19), but sent $e^2 \to \infty$. This was to no avail: we generate a new term at one-loop, now with a coefficient that is comparable to the mass gap in the theory.

The upshot of our discussion is that low-energy physics of the $\mathbf{CP}^{N-1}$ model is that of N massive scalars, each with mass $m = \Lambda_{CP^{N-1}}$, interacting through an unbroken $U(1)$ gauge field. As we saw in Section 7.1, electromagnetism gives rise to a linear, confining force between charged particles in two dimensions. The original scalars ϕ^a transform in the $\mathbf{N}$ of the $SU(N)$ global symmetry. We learn that not only are these now massive, but they are also confined. The physical spectrum of the theory consists of massive, $SU(N)$ singlets. These are mesons, constructed from ϕ and $\phi^\star$.

[13]We met another example of the Mermin-Wagner theorem in the lecture notes on Statistical Field Theory. There we discussed the $O(N)$ model, a non-linear sigma model with target space $\mathbf{S}^N$; it is the real version of the $\mathbf{CP}^{N-1}$ model. Indeed, the first two models in each class coincide at the bottom of the list, since $\mathbf{CP}^1 = \mathbf{S}^2$. After this, the models differ. In particular, the $\mathbf{CP}^{N-1}$ models have instantons for all N, while the $O(N)$ models do not for $N \geq 4$. Nonetheless, the two classes of models share the same fate. Both are gapped at low-energies.

7.3.3 Instantons

The low-energy physics of the $\mathbf{CP}^{N-1}$ model is very similar to that of the Abelian Higgs model that we met in Section 7.2. In both cases, the quantum theory eschews the Higgs phase, and the fundamental excitations are confined. Yet the way we reached these conclusions is rather different. For the Abelian Higgs model, we placed the blame firmly on the instantons (which we identified as vortices); for the $\mathbf{CP}^{N-1}$ model, we reached the same conclusion but using the large N expansion.

We could ask: are there instantons in the $\mathbf{CP}^{N-1}$ model? And, if so, what role do they play?

The answer to the first question is: yes, the $\mathbf{CP}^{N-1}$ model does have instantons. There are actually two different ways to see this. If we start with the gauged linear model (7.19), then the instantons again arise as vortices. (Vortices with more than one scalar field sometimes go by the unhelpful name of "semi-local vortices".) They are labelled by a winding number

$$n = \frac{1}{2\pi}\int d^2x\ F_{12} \tag{7.28}$$

Alternatively, if we work with the non-linear sigma-model (7.21), these instantons show up in a rather different guise. Here field configurations are a map from spatial $\mathbf{R}^2 \mapsto \mathbf{CP}^{N-1}$. However, we must first choose a point on the $\mathbf{CP}^{N-1}$ target space which is the vacuum. This choice breaks the $SU(N)$ symmetry down to $SU(N-1)\times U(1)$. The requirement that the fields asymptote to this vacuum point at spatial infinity means that field configurations are really a map from $\mathbf{S}^2 \mapsto \mathbf{CP}^{N-1}$, and these are characterised by the winding number

$$\Pi_2\left(\mathbf{CP}^{N-1}\right) = \mathbf{Z}$$

This winding is given by

$$n = \frac{1}{2\pi i v^2}\int d^2x\ \partial_\mu\epsilon_{\mu\nu}\left(\phi_a^\star\partial_\nu\phi_a\right)$$

One can show that this coincides with magnetic flux (7.28) using the equation of motion for A_μ from (7.19).

These instantons have a number of interesting properties. One can show that their action is given by

$$S_{\text{instanton}} = 2\pi v^2 \tag{7.29}$$

The scale invariance of the classical 2d sigma model means that the instantons cannot have a fixed size. Instead, like their Yang-Mills counterparts discussed in Section 2.3, they have a scaling modulus. There are also further moduli that describe how the instanton is oriented inside $\mathbf{CP}^{N-1}$. In all, the single instanton has $2N$ parameters, which decompose into two position moduli, a scaling modulus, and $2N-3$ orientational moduli.

We now come to the second question: what role do these instantons play in determining the low energy physics? For $N \geq 2$, the answer is: surprisingly little. This can be seen, for example, by comparing the mass scale (7.26) to the instanton action (7.29),

$$\Lambda_{CP^{N-1}} = \Lambda_{UV}\, e^{-S_{\text{instanton}}/N}$$

This factor of N is important: it is telling us that the instantons are not responsible for the mass gap in the $\mathbf{CP}^{N-1}$ model.

The issue here is that, as we have seen, the $\mathbf{CP}^{N-1}$ model is strongly coupled, and it is not appropriate to try to employ semi-classical techniques like instantons. Indeed, the existence of instantons hinges on the fact that we pick a vacuum state on $\mathbf{CP}^{N-1}$ which, in turn, spontaneously breaks the $SU(N)$ global symmetry. Yet, the large N expansion tells us that this is a red herring: in the quantum theory the $SU(N)$ symmetry is restored. The true ground state does not involve a preferred point on $\mathbf{CP}^{N-1}$, but rather a wavefunction that spreads over the whole space. As such, the role of instantons in this theory is limited when it comes to determining the infra-red physics. The same lesson is expected to hold in Yang-Mills.

The Theta Angle

So far we have not discussed the role of the theta angle in the $\mathbf{CP}^{N-1}$ model. There is something interesting here. For $N \geq 3$ (i.e. for $\mathbf{CP}^2$ or higher) it is thought that, while the theta angle affects the spectrum of the theory, it does not change the phase and the theory remains gapped for all θ. However, for $\mathbf{CP}^1$, something special happens. Here, the theory is thought to be gapped for all $\theta \neq \pi$. At $\theta = \pi$, the theory is expected to be gapless, with the low-energy physics described by an $SU(2)_1$ Wess-Zumino-Witten model. This is sometimes referred to as the *Haldane conjecture*.

7.4 Fermions in Two Dimensions

It's now time to look at fermions. In this section, we will describe a theory that consists only of interacting fermions. In $d = 3+1$ dimensions, such theories are not particularly interesting because the simplest interaction – a four fermion term – is irrelevant. This is no longer the case in $d = 1 + 1$ dimensions and, as we will see, even the simplest theories of interacting fermions are strongly interacting and, like the $\mathbf{CP}^{N-1}$ model above, share a number of surprising properties with QCD.

We start by reviewing some basic facts about fermions in $d = 1+1$ dimensions. The Clifford algebra $\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu}$ is satisfied by 2×2 matrices. Working in signature $\eta^{\mu\nu} = \mathrm{diag}(+1, -1)$, we take the gamma matrices to be

$$\gamma^0 = \sigma^1 \quad \text{and} \quad \gamma^1 = i\sigma^2 \quad \Rightarrow \quad \gamma^3 = -\gamma^0\gamma^1 = \sigma^3 \tag{7.30}$$

Here γ^3 plays the same role as γ^5 in $d = 3+1$ dimensions. It is an extra, anti-commuting matrix which can be used to decompose the two-component Dirac fermion as

$$\psi = \begin{pmatrix} \psi_+ \\ \psi_- \end{pmatrix}$$

Here $\psi_\pm$ are 2d Weyl spinors; they are eigenstates of γ^3.

Fermions in $d = 1 + 1$ dimensions have the special property that they can be both Weyl and Majorana at the same time. This follows because the chiral basis of gamma matrices (7.30) is also real. (In contrast, in $d = 3 + 1$ dimensions you can pick a real basis of gamma matrices but it is not chiral, or a chiral basis which is not real.) This means that we can decompose the Dirac fermion as $\psi = \chi_1 + i\chi_2$ and, moreover, decompose each Weyl fermion as $\psi_\pm = \chi_{1\pm} + i\chi_{2\pm}$. In what follows, we won't need this Majorana decomposition until section 7.4.2.

The action for a free Dirac fermion is

$$\begin{aligned} S &= \int d^2x \; i\bar{\psi}\,\partial\!\!\!/\psi - m\bar{\psi}\psi \qquad (7.31) \\ &= \int d^2x \; i\psi_+^\dagger\partial_-\psi_+ + i\psi_-^\dagger\partial_+\psi_- - m(\psi_-^\dagger\psi_+ + \psi_+^\dagger\psi_-) \end{aligned}$$

where we have introduced lightcone coordinates $x^\pm = t \pm x$ and $\partial_\pm = \partial_t \pm \partial_x$.

For a massless fermion, with $m = 0$, the two Weyl spinors decouple, with equations of motion

$$\partial_+\psi_- = 0 \quad \Rightarrow \quad \psi_- = \psi_-(x^-) \quad \text{and} \quad \partial_-\psi_+ = 0 \quad \Rightarrow \psi_+ = \psi_+(x^+)$$

We learn that the chiral fermion ψ_- is a function only of x^-. In other words, ψ_- is a right-moving fermion. Similarly, ψ_+ is a left-moving fermion. Since the fermions are massless, each moves at the speed of light.

In $d = 3+1$ dimensions, interactions between fermions are always mediated by gauge or scalar fields. In $d = 1+1$ dimensions we have a more direct possibility. The fermion field has dimension $[\psi] = 1/2$ which means that four fermion term $(\bar{\psi}\psi)^2$ is marginal. We can ask: how does this change the low-energy physics. In fact, as we discuss, there are two different ways of adding four fermion terms.

7.4.1 The Gross-Neveu Model

The Gross-Neveu model describes N, classically massless Dirac fermions, ψ_i, $i = 1, \ldots, N$, with a four fermi interaction. The action is given by

$$S = \int d^2x \; i\bar{\psi}_i \partial\!\!\!/ \psi_i + \frac{\lambda}{2N}(\bar{\psi}_i\psi_i)^2 \tag{7.32}$$

Here λ is a dimensionless coupling. We have included the factor of N in anticipation of the fact that we will solve this theory in the large N limit.

The action has a manifest $U(1)_V \times SU(N)$ flavour symmetry, under which the fermions transform as $\mathbf{N}_{+1}$ (meaning that they are in the fundamental $\mathbf{N}$ of $SU(N)$ and have charge $+1$ under $U(1)_V$). In fact, if we decompose each Dirac fermion into two Majorana fermions, the symmetry group is actually $O(2N)$ symmetry, and this will play a role shortly. There is also a discrete $\mathbf{Z}_2$ chiral symmetry,

$$\mathbf{Z}_2 : \;\; \psi_i \rightarrow \gamma^3\psi_i \tag{7.33}$$

Importantly, a would-be mass term is odd under this discrete chiral symmetry, $\bar{\psi}_i\psi_i \rightarrow -\bar{\psi}_i\psi_i$. This means that the existence of the $\mathbf{Z}_2$ symmetry would naively prohibit the generation of a mass. Our goal is to see how this plays out in the quantum theory.

It turns out that life is easier if we introduce an auxiliary scalar field, σ, and write the action as

$$S = \int d^2x \; i\bar{\psi}_i \partial\!\!\!/ \psi_i - \frac{N}{2\lambda}\sigma^2 + \sigma\bar{\psi}_i\psi_i \tag{7.34}$$

Although σ is dynamical, we do not include a kinetic term for it. We can integrate it out by imposing the equation of motion

$$\sigma = \frac{\lambda}{N}\bar{\psi}_i\psi_i$$

and we get back the original action (7.32). The new form of the action (7.34) is again invariant under the discrete chiral symmetry, but only if we take σ to be odd,

$$\mathbf{Z}_2 : \;\; \sigma \rightarrow -\sigma$$

The introduction of σ is reminiscent of the auxiliary field that we introduced in the $\mathbf{CP}^{N-1}$ model. Indeed, we will proceed by following the same strategy. We will integrate out the fields that we thought we cared about – in this case the fermions – and focus on the resulting effective dynamics for σ. We will see that this is sufficient to teach us the relevant physics.

Integrating out the fermions leaves behind the following effective action for σ,

$$S_{\text{eff}} = iN \log\det\left(i\not{\partial} + \sigma\right) - \int d^2x\ \frac{N}{2\lambda}\sigma^2$$

We can write the first term in more concrete form. First,

$$\det\left(i\not{\partial} + \sigma\right) = \det\left(i\not{\partial}\right)\ \det\left(1 - i\not{\partial}^{-1}\sigma\right)$$

and we neglect the factor $\det(i\not{\partial})$ on the grounds that it contributes an irrelevant constant. The next step is to deal with the gamma matrix structure in the second term. Using $\det\gamma^3 = -1$, we have

$$\det\left(1 - i\not{\partial}^{-1}\sigma\right) = \det\left(\gamma^3(1 - i\not{\partial}^{-1}\sigma)\gamma^3\right) = \det\left(1 + i\not{\partial}^{-1}\sigma\right)$$

Multiplying these together then gives

$$\det\left(1 - i\not{\partial}^{-1}\sigma\right) = \det{}^{1/2}\left(1 + (\not{\partial}^{-1}\sigma)^2\right) = \det{}^{1/2}\left(1 - \sigma\,\partial^{-2}\sigma\right)$$

where the argument in the final argument comes with a 2×2 unit matrix for the spinor indices. But this simply changes $\det^{1/2}$ back to det. Finally, we use $\log\det = \text{Tr}\log$ to write

$$S_{\text{eff}} = iN\,\text{Tr}\log\left(1 - \sigma\,\partial^{-2}\sigma\right) - \int d^2x\ \frac{N}{2\lambda}\sigma^2$$

This action doesn't look particularly appealing. But it has one important feature going for it, which is that it's proportional to N. This means that in the large N limit it can be evaluated using a saddle point. We look for solutions in which σ is constant. In this case, the annoying log factor can be replaced by a simple integral, leaving us with the effective potential for the scalar field. Rotating to Euclidean space, we have

$$V_{\text{eff}}(\sigma) = -N\int^{\Lambda_{UV}} \frac{d^2p}{(2\pi)^2}\log\left(1 + \frac{\sigma^2}{p^2}\right) + \frac{N}{2\lambda}\sigma^2$$

This is the same kind of integral that we met in (7.24) when solving the 2d $\mathbf{CP}^{N-1}$ model. The same method that we used previously now gives

$$V_{\text{eff}}(\sigma) = \frac{N}{4\pi}\sigma^2\left(\log\left(\frac{\sigma^2}{\Lambda_{UV}^2}\right) - 1\right) + \frac{N}{2\lambda}\sigma^2 \tag{7.35}$$

In the large N limit, the path integral is dominated by the minimum of the potential which sits at

$$\frac{\partial V_{\rm eff}}{\partial \sigma} = 0 \quad \Rightarrow \quad \sigma^2 = \Lambda_{UV}^2 e^{-2\pi/\lambda}$$

We learn that the σ field gets an expectation value. The theory was originally invariant under the discrete chiral symmetry, $\sigma \to -\sigma$, but this is spontaneously broken in the ground state: the theory must choose one of the two ground states $\sigma = \pm\Lambda_{UV}\, e^{-\pi/\lambda}$.

With the protective $\mathbf{Z}_2$ symmetry spontaneously broken, there is nothing to stop the fermions getting a mass. Indeed, substituting the expectation value of σ back into the action (7.34), we find that the mass is given by

$$m_{GN} = \Lambda_{UV} e^{-\pi/\lambda} \tag{7.36}$$

Once again we have the phenomenon of dimensional transmutation: the dimensionless coupling λ has combined with the UV cut-off to provide a physical mass scale of the theory. Once again, we thought that we started out with a theory of massless particles, but the interactions find an ingenious way to generate a mass.

Above we have phrased the physics in the terms of the effective potential. Another approach would be to compute one-loop contributions to the running of the coupling. We would have found that the theory is asymptotically free, with the beta function

$$\mu \frac{d\lambda(\mu)}{d\mu} = -\frac{\lambda^2}{\pi} \quad \Rightarrow \quad \frac{1}{\lambda(\mu)} = \frac{1}{\lambda_0} - \frac{1}{2\pi}\log\frac{\Lambda_{UV}^2}{\mu^2}$$

Phrased in this way, the physical mass is seen to be RG invariant, as it should be: $m_{GN} = \mu\, e^{-\pi/\lambda(\mu)}$.

7.4.2 Kinks in the Gross-Neveu Model

As we've seen, the Gross-Neveu model spontaneously breaks the $\mathbf{Z}_2$ symmetry. This means that the theory has two degenerate ground states, distinguished by the sign of $\sigma = \pm\Lambda_{UV} e^{-\pi/\lambda}$. This gives us a new state in the the theory: a kink which interpolates between the two ground states, so that the profile of $\sigma(x)$ obeys

$$\sigma \to \pm\Lambda_{UV} e^{-\pi/\lambda} \qquad \text{as } x \to \pm\infty$$

We would like to understand what properties these kinks have and, in particular, how they transform under the symmetries of the theory. The key to this is to see what happens to the original fermions in the presence of the kink.

The Dirac equation from (7.34) is

$$i\partial\!\!\!/\psi_i + \sigma\psi_i = 0$$

We'd like to solve this in the kink background. You might think that this is tricky because we haven't determined the profile $\sigma(x)$ of the kink. Fortunately, this isn't a problem, because the property that we need is robust and independent of the exact form of $\sigma(x)$: this is the existence of a fermi zero mode.

We met fermi zero modes on domain walls previously, both in our discussion of topological insulators in Section 3.3.4 and lattice gauge theory in Section 4.4.1. The analysis needed here is exactly the same, and we won't repeat it. But the upshot is that each fermion ψ_i has a single, complex fermi zero mode on the kink.

At this point, it is important to recall that our Dirac fermions can be decomposed into Majorana fermions, which we write as

$$\psi_i = \chi_i + i\chi_{i+N} \qquad i = 1, \ldots, N$$

The existence of Majorana fermions means that the global symmetry of the Gross-Neveu model is $O(2N)$ rather than $U(N)$. Each of these Majorana fermions gives rise to a single, Majorana (i.e. real) fermi zero mode on the kink which we will denote as b_i. These obey the commutation relations,

$$\{b_i, b_j\} = 2\delta_{ij} \qquad\qquad i = 1, \ldots, 2N \tag{7.37}$$

To convince yourself that these are the right commutation relations, we can pair the Majorana modes back into their complex counterparts $c_i = \frac{1}{\sqrt{2}}(b_i + ib_{i+N})$, with $i = 1, \ldots, N$ which, from (7.37), obey the usual Grassmann creation and annihilation commutation relations $\{c_i, c_j\} = 0$ and $\{c_i, c_j^\dagger\} = 2\delta_{ij}$

The commutation relations (7.37) are familiar: they are simply the Clifford algebra in $D = 2N$ dimensions. This has a representation in terms of $2^N \times 2^N$ dimensionsal matrices. Said in a different way, the Majorana zero modes ensure that the Hilbert space of kink excitations has dimension 2^N.

This 2^N dimensional Hilbert space does not form an irreducible representation of the $O(2N)$ symmetry group. Instead, it decomposes into two chiral spinors. We achieve this by introducing the "γ^5" matrix, $\gamma^5 = ib_1 \ldots b_{2N}$ which obeys $\{\gamma^5, b_i\} = 0$ and $(\gamma^5)^2 = 1$. The two reducible representations are distinguished by the eigenvalue under $\gamma^5 = \pm 1$, and have dimension 2^{N-1}.

The upshot of this analysis is rather nice. We started with Majorana fermions transforming in the $2N$-dimensional vector representation of $O(2N)$. But the interactions generate new solitonic states. These are kinks which transform in the left and right-handed spinor representations of $O(2N)$. This can be thought of as a version of "charge fractionalisation".

Our results in this section used the large N approximation to determine the fate of the Gross-Neveu model. One might wonder if the kinks survive to small N. It turns out that for $N > 2$, both kinks and fermions exist in the spectrum. But, perhaps counterintuitively, when $N = 2$ only kinks, in the spinor representation of $O(4)$, survive; the original fermions no longer exist. For $N = 1$, the Gross-Neveu model coincides with the Thirring model and turns out to be free. We will discuss this case in Section 7.5.

An Odd Number of Majorana Fermions

So far, our discussion of the Gross-Neveu model has focussed on N Dirac fermions or, equivalently, $2N$ Majorana fermions. But there's nothing to stop us writing down the action for an odd number of Majorana fermions χ_i,

$$S = \int d^2x \; i\bar{\chi}_i \partial\!\!\!/ \chi_i - \frac{\tilde{N}}{4\lambda}\sigma^2 + \sigma\bar{\chi}_i\chi_i$$

where the summation is over $i = 1, \ldots, \tilde{N}$. When $\tilde{N} = 2N$, this reduces to our previous action (7.34) in terms of Dirac fermions. When $\tilde{N}$ is odd, our previous analysis goes through unchanged, and we again find that the $\mathbf{Z}_2$ is spontaneously broken, resulting in two degenerate ground states. The only novel question is: what becomes of the kinks?

The Majorana zero modes again give rise to a Clifford algebra (7.37), but this time it's a Clifford algebra in $D = \tilde{N}$ dimensions, with $\tilde{N}$ odd. There is a single reducible representation which has dimension $2^{(\tilde{N}-1)/2}$, and one might think this is the Hilbert space of the kinks. However, there is another discrete symmetry that we have to take into account. This is $\chi_i \to -\chi_i$ which is part of the $O(\tilde{N})$ group, but not $SO(\tilde{N})$. To implement this, we introduce the fermion parity operator $(-1)^F$ which obeys

$$(-1)^F \chi_i \, (-1)^F = -\chi_i \quad \Rightarrow \quad \{(-1)^F, b_i\} = 0$$

When $\tilde{N} = 2N$ is even, the operator γ^5 can be identified with $(-1)^F$. But when $\tilde{N}$ is odd, there is no action of $(-1)^F$ on a single irreducible representation of the Clifford algebra. Instead, we need two irreducible representations: one with $(-1)^F = +1$ and one with $(-1)^F = -1$. This means that for $\tilde{N}$ odd, we again have two irreducible representations of $O(\tilde{N})$, and the total number of kink states is $2 \times 2^{(\tilde{N}-1)/2}$.

7.4.3 The Chiral Gross-Neveu Model

There is a variant on the Gross-Neveu model that introduces yet another ingredient into the mix. First, consider the action of the axial symmetry

$$U(1)_A: \ \psi \rightarrow e^{i\alpha\gamma^3}\psi$$

There are two real, fermion bilinears that we can introduce: $\bar{\psi}\psi$ and $i\bar{\psi}\gamma^3\psi$. Neither of them is invariant under the axial symmetry. Instead, each rotates into the other. We can form the complex combination $\bar{\psi}\psi + \bar{\psi}\gamma^3\psi$, and this transforms as

$$\begin{aligned} U(1)_A: \quad & \bar{\psi}\psi + \bar{\psi}\gamma^3\psi \rightarrow \ e^{2i\alpha}\left(\bar{\psi}\psi + \bar{\psi}\gamma^3\psi\right) \\ & \bar{\psi}\psi - \bar{\psi}\gamma^3\psi \rightarrow \ e^{-2i\alpha}\left(\bar{\psi}\psi - \bar{\psi}\gamma^3\psi\right) \end{aligned}$$

This transformation motivates us to consider the following theory of N massless, interacting Dirac fermions,

$$S_{\chi GN} = \int d^2x \ i\bar{\psi}_i \partial\!\!\!/ \psi_i + \frac{\lambda}{2N}\left((\bar{\psi}_i\psi_i)^2 - (\bar{\psi}_i\gamma^3\psi_i)^2\right) \tag{7.38}$$

The advantage of this set-up is that the theory is protected from generating a mass term by the continuous $U(1)_A$ chiral symmetry, as opposed to the discrete $\mathbf{Z}_2$ chiral symmetry of the original Gross-Neveu model.

This is an important distinction. We saw above that the discrete $\mathbf{Z}_2$ symmetry proved ineffectual at protecting the Gross-Neveu model from developing a gap because it was spontaneously broken. However, there is a general theorem, due to Mermin and Wagner, that says it is not possible to spontaneously break continuous symmetries in $d = 1 + 1$ quantum field theory. We met this theorem in the lectures on Statistical Field Theory; its essence is that infra-red fluctuations of fields always destroy any long range order.

Given this theorem, you might think that the existence of a continuous chiral symmetry would be much more powerful and protect the fermions from developing a gap. You would be wrong. As we now show, the Mermin-Wagner theorem not withstanding, the chiral Gross-Neveu model (7.38) also generates a gap at low energies.

To see this, we use the same trick as before but this time introduce two auxiliary fields, σ and π. The action (7.38) can be written as

$$S = \int d^2x \ i\bar{\psi}_i \partial\!\!\!/ \psi_i - \frac{N}{2\lambda}(\sigma^2 + \pi^2) + \bar{\psi}_i(\sigma + i\pi\gamma^3)\psi_i \tag{7.39}$$

The equation of motion for σ and π then tell us that

$$\sigma \pm i\pi = \frac{\lambda}{N}\,\bar{\psi}_i(1 \mp \gamma^3)\psi_i \tag{7.40}$$

The action (7.39) remains invariant under $U(1)_A$ provided that the auxiliary scalars transform as

$$U(1)_A : \sigma + i\pi \to e^{-2i\alpha}(\sigma + i\pi) \tag{7.41}$$

Evaluating the fermion determinant in the same way as before, we find

$$\det\left(1 - i\,\partial\!\!\!/^{-1}(\sigma + i\pi\gamma^3)\right) = \det{}^{1/2}\left(1 + (\partial_\mu^{-1}\sigma)^2 + (\partial_\mu^{-1}\pi)^2\right)$$

Viewing both σ and π as constants, we're then left with the effective potential,

$$V_{\rm eff}(\sigma,\pi) = \frac{N}{4\pi}(\sigma^2+\pi^2)\left(\log\left(\frac{\sigma^2+\pi^2}{\Lambda_{UV}^2}\right) - 1\right) + \frac{N}{2\lambda}(\sigma^2+\pi^2)$$

This is identical to the potential (7.35) for the original Gross-Neveu model, but with σ^2 replaced with $\sigma^2+\pi^2$. Note, in particular, that the potential is invariant under the $U(1)_A$ action (7.41) as it should be.

What do we do with this potential? Because we're in $d = 1+1$ dimensions, we should be a little careful. We parameterise the complex scalar field as

$$\sigma + i\pi = \rho e^{i\theta}$$

The minimum of the potential sits at

$$\rho = m_{GN}$$

where m_{GN} is the same dynamically generated mass scale (7.36) that we saw in the previous model. This is already sufficient to tell us that the fermions generate a mass.

The care is needed when we come to the angular field mode $\theta(x)$. This transforms as $\theta \to \theta + \alpha$ under the $U(1)_A$ symmetry. If we were in a higher dimension, we would argue that $\theta(x)$ should take some fixed value in the ground state, breaking the $U(1)_A$ symmetry. In such a situation, we would identify the Goldstone boson as θ, which necessarily remains gapless.

However, in $d = 1{+}1$ dimensions the story is a little different. As we mentioned above, the Mermin-Wagner theorem tells us that there are no Goldstone modes. Instead, the ground state wavefunction is closer in spirit to quantum mechanics, spreading over all values of θ. This is a topic that we discussed in some detail in the lectures on Statistical Field Theory in the context of the Kosterlitz-Thouless phase transition. We will recount the important facts here. The key result is that while θ does indeed remain massless, it is not a Goldstone boson. This is not merely a matter of terminology: the physics differs.

First, we need to work a little harder in expanding the effective action. The potential is

$$S_{\text{eff}} = -iN\,\text{Tr}\log\left(i\,\partial\!\!\!/ + \rho e^{i\theta\gamma^3}\right) - \frac{N}{2\lambda}\int d^2x\;\rho^2$$

It's no longer sufficient to focus on constant values of σ and π since the resulting potential will not depend on θ. Instead, we need to consider slowly varying θ. The leading term in the effective action is the obvious one:

$$S_{\text{eff}} = \int d^2x\;\frac{N\rho^2}{4\pi}(\partial_\mu\theta)^2$$

This theory is less trivial than it looks! Because θ is a periodic variable $\theta \in [0, 2\pi)$, a so-called compact boson, the overall normalisation factor $N/4\pi$ is meaningful and will show up in correlation functions. We will need to study such theories in some detail in Section 7.5, but for now a quick and simple computation of the 2-point correlators will suffice. If θ was a normal scalar field in $d = 1 + 1$ dimensions, we would have

$$\langle\theta(x)\,\theta(0)\rangle = -N\log\left(\Lambda_{UV}|x|\right) \tag{7.42}$$

However, because it's a compact boson we should really work with the single-valued operator $e^{i\theta}$. The appropriate correlation function then follows from Wick's theorem, together with the result (7.42),

$$\langle e^{i\theta(x)}\,e^{-i\theta(0)}\rangle = e^{\langle(\theta(x)\theta(0))\rangle} \sim \frac{1}{|x|^{1/N}}$$

We see that in the strict $N \to \infty$ limit, the theory exhibits the long range order expected from spontaneous symmetry breaking. Indeed, there is a loophole in the Mermin-Wagner theorem and it breaks down in theories with an infinite number of fields. However, for any large, but finite N, we find "quasi-long range order", with correlation functions dropping off very slowly.

This translates directly into correlation functions between fermion bilinears. Using (7.40), we again see the phenomenon of quasi-long range order,

$$\langle \bar{\psi}(1-\gamma^3)\psi(x)\ \bar{\psi}(1+\gamma^3)\psi(0)\rangle \sim \frac{1}{|x|^{1/N}}$$

The upshot is that, once again, an interacting quantum field theory (7.38) has found a way to generate a mass. This time, the fermions get mass but the chiral $U(1)_A$ symmetry remains unbroken.

7.4.4 Back to Basics: Quantising Fermions in 2d

Given that we've just used path integral techniques to solve a theory of strongly interacting fermions, what we're about to do next may seem a little odd. We will return to the free fermion and solve it using canonical quantisation.

This is the kind of calculation that we did in our first course in Quantum Field Theory, and you may reasonably wonder why we're bothering to do it again now that we're grown up. The reason is that it will prove an important warm-up for the following section where we discuss bosonization.

We introduced the action for a massless fermion in (7.31). A two-component Dirac fermion can be decomposed into Weyl fermions $\psi^T = (\psi_+, \psi_-)$, in terms of which the action is

$$S = \int d^2x\ i\psi_-^\dagger(\partial_t + \partial_x)\psi_- + i\psi_+^\dagger(\partial_t - \partial_x)\psi_+$$

The two Weyl fermions $\psi_\pm$ are independent. This means that there are two conserved quantities: these are the vector and axial currents and will be particularly important in what follows. The vector current is

$$j_V^\mu = \bar{\psi}\gamma^\mu\psi \tag{7.43}$$

while a massless fermion also has a conserved axial current given by

$$j_{\rm A}^\mu = \bar{\psi}\gamma^\mu\gamma^3\psi \tag{7.44}$$

From these we can construct two conserved charges, Q_V and Q_A.

The Weyl fermion ψ_- is right-moving, and quantisation of this field will lead to particles with momentum $p > 0$. Similarly, the quantisation of ψ_+ will lead to particles with momentum $p < 0$. The mode expansion of the operators in the Schrödinger picture follows the familiar story described in the lectures on Quantum Field Theory

$$\psi_-(x) = \int_0^\infty \frac{dp}{2\pi}\,\left(b_{-\,p}e^{ipx} + c^\dagger_{-\,p}e^{-ipx}\right) \tag{7.45}$$

$$\psi_+(x) = \int_{-\infty}^0 \frac{dp}{2\pi}\,\left(b_{+\,p}e^{ipx} + c^\dagger_{+\,p}e^{-ipx}\right) \tag{7.46}$$

with the creation and annihilation operators obeying the standard anti-commutation relations $\{b_{\pm\,p}, b^\dagger_{\pm\,q}\} = \{c_{\pm\,p}, c^\dagger_{\pm\,q}\} = 2\pi\,\delta(p-q)$. The vacuum is defined by $b_{\pm\,p}|0\rangle = c_{\pm\,p}|0\rangle = 0$, and the operators $b^\dagger_{\pm\,p}$ and $c^\dagger_{\pm\,p}$ then create particles and anti-particles respectively.

It will turn out that we will need to be careful about various UV issues. For this reason, we work instead with the mode expansion

$$\psi_-(x) = \int_0^\infty \frac{dp}{2\pi}\,\left(b_{-\,p}e^{ipx} + c^\dagger_{-\,p}e^{-ipx}\right)e^{-p/2\Lambda}$$

$$\psi_+(x) = \int_{-\infty}^0 \frac{dp}{2\pi}\,\left(b_{+\,p}e^{ipx} + c^\dagger_{+\,p}e^{-ipx}\right)e^{-|p|/2\Lambda}$$

where Λ is a UV cut-of scale. In what follows, all integrals will be over the full range of $\mathbf{R}$ unless otherwise stated. We also introduce the UV length scale

$$\epsilon = \frac{1}{\Lambda}$$

We can then compute the two-point functions in position space. For example, we have

$$\begin{aligned}\langle\psi_-(x)\psi_-^\dagger(y)\rangle &= \int_0^\infty \frac{dp\,dq}{(2\pi)^2}\,\langle\, b_{-q}\,b^\dagger_{-\,p}\,\rangle\, e^{iqx-ipy}e^{-(|p|+|q|)/2\Lambda}\\ &= \int_0^\infty \frac{dp}{2\pi}\,e^{ip(x-y)}e^{-|p|/\Lambda}\\ &= \frac{i}{2\pi}\frac{1}{(x-y)+i\epsilon}\end{aligned} \tag{7.47}$$

You can also check that $\langle\psi_-^\dagger(x)\psi_-(y)\rangle = \langle\psi_-(x)\psi_-^\dagger(y)\rangle$. In particular, if we combine these results we have

$$\begin{aligned}\langle\,\{\psi_-(x),\psi_-^\dagger(y)\}\,\rangle &= \frac{i}{2\pi}\left(\frac{1}{(x-y)+i\epsilon} + \frac{1}{-(x-y)+i\epsilon}\right)\\ &= \frac{1}{\pi}\frac{\epsilon}{(x-y)^2+\epsilon^2}\ \longrightarrow\ \delta(x-y)\ \text{as}\ \epsilon\to 0\end{aligned}$$

in agreement with the standard anti-commutation relations between fermions. Similarly,

$$\langle \psi_+(x)\psi_+^\dagger(y)\rangle = -\frac{i}{2\pi}\frac{1}{(x-y)-i\epsilon} \tag{7.48}$$

and $\langle \psi_+^\dagger(x)\psi_+(y)\rangle = \langle \psi_+(x)\psi_+^\dagger(y)\rangle$.

The expressions (7.47) and (7.48) are the key bits of information that we need to take forward into the next section where we discuss bosonization.

7.5 Bosonization in Two Dimensions

There is something rather wonderful about fermions in two dimensions: they can be rewritten in terms of bosons! The purpose of this section is to explain how on earth this is possible.

At first sight, this is a surprise. After all, the difference between bosons and fermions is one of the most fundamental things we learn as undergraduates. However, there are reasons to suspect this difference is not so stark in $d = 1+1$ dimensions. First, the spin statistics theorem tells us that bosons have integer spin and fermions half-integer spin. Yet in one spatial dimension there is no meaning to rotation and, correspondingly, no meaning to spin. Relatedly, if we want to exchange two particles on a line, we can only do so by moving them past each other. This is in contrast to higher dimensions where particle positions can be exchanged, while keeping them separated by arbitrarily large distances. This simple observation suggests that interactions will be as important as statistics when particles are confined to live on a line.

To begin, we will show that a free massless Dirac fermion in $d = 1+1$ is equivalent to a free massless, real scalar field ϕ. Even for free fields, this is a rather remarkable claim. The Hilbert space of a single bosonic oscillator looks nothing the Hilbert space of a single fermionic oscillator, yet we claim that the theories in $d = 1+1$ not only have the same Hilbert space (at least after we include a subtle $\mathbf{Z}_2$ issue), but also the same spectrum. Furthermore, for any operator that we can construct out of fermions, there is a corresponding operator made from bosons. Here we will focus on these operators and show that the correlation functions of the fermionic theory coincide with those of the bosonic theory.

The Compact Boson

The bosonic theory that we will focus on is deceptively simple. It is the theory of a massless, real scalar field ϕ. We write its action as

$$S = \int d^2x \, \frac{\beta^2}{2} (\partial_\mu \phi)^2 \tag{7.49}$$

However, there is one difference with a usual scalar field: we will take our scalar ϕ to be periodic, taking values in the range

$$\phi \in [0, 2\pi) \tag{7.50}$$

We refer to this as a compact boson. The dimensionless parameter β is called the radius of the boson. (String theorists would usually define $R^2 = 2\pi\beta^2 l_s^2$ and call R the radius. Here l_s is the string length and which gives R dimension -1. Furthermore, it's not uncommon to work in conventions with $l_s^2 = 2$, in which case $R^2 = 4\pi\beta^2$.)

Usually, the overall coefficient of the kinetic term does not affect the physics, since it can always be absorbed into a redefinition of the field. But, in the present context, we can't absorb β without changing the periodicity of ϕ. This leads us to suspect that the simple action (7.49) describes a different theory for each choice of β^2, a suspicion that we will confirm below. We will see that there is one special choice of β^2 for which the compact boson coincides with the free fermion. (Spoiler: it's $\beta^2 = 1/4\pi$.)

What are the implications of having a compact boson? The first thing to notice is that we can't add terms like ϕ^2 or ϕ^4 to the action, since these don't respect the periodicity. Instead we should add terms like $\cos\phi$ and $\sin\phi$. Equivalently, the field ϕ is not really a well defined operator. We should instead focus on operators like $e^{i\phi}$ which, again, respect the periodicity. These are sometimes referred to as *vertex operators*, following their role in String Theory. Our task below will be to compute correlation functions of the vertex operators $e^{i\phi}$.

Now let's turn to the conserved currents of the theory (7.49). The action is invariant under the symmetry $\phi \to \phi + \text{constant}$. The associated current is

$$j^\mu_{\text{shift}} = \beta^2 \partial^\mu \phi$$

Clearly the equation of motion, $\partial^2\phi = 0$, ensures that j^μ_{shift} is conserved. The corresponding Noether charge is Q_{shift}, under which the operator $e^{i\phi}$ has charge $+1$.

However, in two dimensions a massless scalar also enjoys another conserved current,

$$j^\mu_{\rm wind} = \frac{1}{2\pi}\epsilon^{\mu\nu}\partial_\nu\phi$$

which is conserved by dint of the epsilon symbol; we don't need to invoke the equation of motion. To see the associated conserved quantity, it is useful to put the theory on a spatial circle of radius R. The charge associated to $j^\mu_{\rm wind}$ is then

$$Q_{\rm wind} = \int_0^{2\pi R} dx\ j^0_{\rm wind} = \frac{1}{2\pi}\int_0^{2\pi R} dx\ \partial_x\phi$$

The conserved charge $Q_{\rm wind}$ is the number of times that $\phi \in [0,2\pi)$ winds around its range as we go around the spatial circle. It is a topological charge. The existence of two, independent $U(1)$ global symmetries is reminiscent of the vector and axial symmetries of the massless fermion. We'll make this connection more precise shortly.

7.5.1 T-Duality

There is an alternative description of the compact boson in terms of a dual scalar. To realise this, we take the original action (7.49) and think of $\partial\phi$ as the variable, rather than ϕ. We can do this, only if we also impose an appropriate Bianchi identity. We might naively think that the Bianchi identity is $\partial_\mu(\epsilon^{\mu\nu}\partial_\nu\phi) = 0$, but in fact this is too strong since it kills all winding. Instead, we want

$$\frac{1}{2\pi}\int d^2x\ \partial_\mu(\epsilon^{\mu\nu}\partial_\nu\phi) = \frac{1}{2\pi}\oint dx^\mu\ \partial_\mu\phi \in \mathbf{Z} \tag{7.51}$$

To impose this, we introduce a second compact boson

$$\tilde{\phi} \in [0,2\pi)$$

and consider the action

$$S = \int d^2x\ \frac{1}{2}\beta^2(\partial_\mu\phi)^2 - \frac{1}{2\pi}\epsilon^{\mu\nu}\partial_\mu\phi\,\partial_\nu\tilde{\phi}$$

Integrating out $\tilde{\phi}$ in the partition function imposes the condition (7.51) and takes us back to the original action (7.49). Alternatively, we can integrate out $\partial\phi$. Completing the square, we have

$$S = \int d^2x\ \frac{1}{2}\beta^2\left(\partial^\mu\phi - \frac{1}{2\pi\beta^2}\epsilon^{\mu\nu}\partial_\nu\tilde{\phi}\right)^2 + \frac{1}{2}\frac{1}{4\pi^2\beta^2}(\partial\tilde{\phi})^2 \tag{7.52}$$

This then gives an equivalent theory in terms of the dual scalar,

$$S = \int d^2x\ \frac{1}{2}\tilde{\beta}^2(\partial_\mu\tilde{\phi})^2 \quad \text{with} \quad \tilde{\beta}^2 = \frac{1}{4\pi^2\beta^2} \tag{7.53}$$

The theory (7.53) is entirely equivalent to our original theory (7.49). This is referred to as *T-duality.*

T-duality is particularly striking in the context of string theory. There, the compact boson ϕ is interpreted as a compact direction of spacetime in which the string can move. In the usual conventions of string theory, the radius of this circle is taken to be $R = \sqrt{2\pi}\beta l_s$ with l_s the string length. T-duality says that, as far as the string is concerned, the physics is exactly the same if we instead take a spacetime with a compact circle of radius $\tilde{R} = l_s^2/R$. In other words, very big circles are the same as very small circles. You can read more about this interpretation in the lecture notes on String Theory.

How is this possible? The key is the relation between ϕ and $\tilde{\phi}$, which can be found inside the squared brackets in (7.52),

$$\partial_\mu \phi = \frac{1}{2\pi\beta^2}\epsilon_{\mu\nu}\partial^\nu\tilde{\phi} \tag{7.54}$$

This clearly relates the momentum current for ϕ to the winding current for $\tilde{\phi}$, and vice versa. What looks like momentum modes in one description becomes winding modes in the other. In particular, $e^{i\tilde{\phi}}$ carries charge $+1$ under $Q_{\rm wind}$.

Although the transformation between ϕ and $\tilde{\phi}$ is simple, it is also non-local. If we try to solve for $\tilde{\phi}$ in terms of ϕ, we must integrate. We'll see this clearly in (7.56) below.

Chiral Bosons

In what follows, it will be useful to introduce chiral bosons, which are either purely left moving or purely right moving. The equation of motion $\partial^2\phi = 0$ can be solved by

$$\phi = \phi_-(x^-) + \phi_+(x^+)$$

where $x^\pm = t \pm x$. In fact, the decomposition isn't quite as clean because there is also a zero mode which does not naturally divide between the two. We will ignore this fact here.

These chiral bosons give us a novel perspective on the dual scalar. The relation (7.54) is solved by writing

$$\tilde{\phi} = 2\pi\beta^2(\phi_- - \phi_+)$$

We can then express the chiral bosons in terms of the scalar and its dual by

$$\phi_\mp(x,t) = \frac{1}{2}\left[\phi(x,t) \pm \frac{1}{2\pi\beta^2}\tilde{\phi}(x,t)\right] \tag{7.55}$$

Indeed, we can check that

$$\begin{aligned}\partial_x\phi_- &= \frac{1}{2}\left(\partial_x\phi + \frac{1}{2\pi\beta^2}\partial_x\tilde{\phi}\right) \\ &= \frac{1}{2}\left(\partial^t\tilde{\phi} - \frac{1}{2\pi\beta^2}\partial_t\phi\right) = -\partial_t\phi_- \quad \Rightarrow \quad \partial_+\phi_- = 0\end{aligned}$$

as required.

7.5.2 Canonical Quantisation of the Boson

Let's now consider what happens when we quantise the boson. Let's start by ignoring the the fact that ϕ is compact: we'll then reinstate this condition later when we discuss the viable operators in the theory. In the Schrödinger picture, we expand the operator $\phi(x)$ in Fourier modes, following the usual story in Quantum Field Theory

$$\phi(x) = \frac{1}{\beta}\int \frac{dp}{2\pi}\frac{1}{\sqrt{2|p|}}\left(a_p e^{ipx} + a_p^\dagger e^{-ipx}\right)e^{-|p|/2\Lambda}$$

Classically, the momentum is $\pi = \beta^2\dot{\phi}$. In the Schrödinger picture, this is written as the operator

$$\pi(x) = -i\beta\int \frac{dp}{2\pi}\sqrt{\frac{|p|}{2}}\left(a_p e^{ipx} - a_p^\dagger e^{-ipx}\right)e^{-|p|/2\Lambda}$$

We've introduced a UV cut-off Λ in these expressions. We'll see the utility of this shortly. As for fermions, we also introduce the UV length scale $\epsilon = 1/\Lambda$. Using the usual commutation relations among the creation and annihilation operators $[a_p, a_q^\dagger] = 2\pi\,\delta(p-q)$, we have

$$[\phi(x), \pi(y)] = \frac{i}{\pi}\frac{\epsilon}{(x-y)^2+\epsilon^2} \;\longrightarrow\; i\delta(x-y) \text{ as } \epsilon\to 0$$

How do we construct the quantum operator for the chiral boson (7.55)? The dual scalar obeys $\partial_x\tilde{\phi}/2\pi = -\beta^2\dot{\phi} = -\pi(x)$. We can then write down a quantum operator in the Schrödinger picture, by integrating the momentum thus:

$$\phi_\pm(x) = \frac{1}{2}\left[\phi(x) \pm \frac{1}{\beta^2}\int_{-\infty}^{x} dx'\;\pi(x')\right] \tag{7.56}$$

Here we see what we promised earlier: the chiral bosons $\phi_\pm(x)$ are inherently non-local objects: they requires knowledge of the profile of the field everywhere to the left of the point x. To check that these are indeed the right objects, we can work in our mode expansion. We have

$$\begin{aligned}\phi_-(x) &= \frac{1}{2\beta}\int\frac{dp}{2\pi}\sqrt{\frac{1}{2|p|}}\left(1+\frac{|p|}{p}\right)\left(a_p\,e^{ipx}+a_p^\dagger e^{-ipx}\right)e^{-|p|/2\Lambda}\\ &= \frac{1}{2\beta}\int_0^\infty\frac{dp}{2\pi}\,\sqrt{\frac{2}{|p|}}\left(a_p\,e^{ipx}+a_p^\dagger e^{-ipx}\right)e^{-p/2\Lambda}\end{aligned}$$

which picks up contributions only from the right-moving, $p>0$ modes. This is reminiscent of the expansion (7.45) for the Weyl fermion ψ_-. Similarly,

$$\phi_+(x) = \frac{1}{2\beta}\int_{-\infty}^0\frac{dp}{2\pi}\,\sqrt{\frac{2}{|p|}}\left(a_p\,e^{ipx}+a_p^\dagger e^{-ipx}\right)e^{-|p|/2\Lambda}$$

which picks up contributions only from left-moving, $p<0$ modes. This is reminiscent of the expansion (7.46) for ψ_+.

The commutation relations of $\phi_\pm$ are easily computed. We have

$$\begin{aligned}[\phi_\pm(x),\phi_\pm(y)] &= \pm\frac{1}{4\beta^2}\int_{-\infty}^y dy'\;[\phi(x),\pi(y')]\pm\frac{1}{4\beta^2}\int_{-\infty}^x dx'\;[\pi(x'),\phi(y)]\\ &= \mp\frac{i}{4\beta^2}\mathrm{sign}(x-y)\end{aligned}\tag{7.57}$$

Again, we see the non-locality of chiral bosons in their commutation relation. The operators fail to commute no matter how far separated. Meanwhile,

$$[\phi_+(x),\phi_-(y)] = -\frac{i}{4\beta^2}\tag{7.58}$$

This latter commutation relation is telling us that, in contrast to the Weyl fermions, the left and right moving scalars have not fully decoupled. The culprit is the zero momentum mode of the scalar, which is shared by both ϕ_+ and ϕ_-. This zero mode is an important subtlety in a number of applications, but we will not treat it properly here. A slightly better treatment can be found in the lectures on String Theory.

Before we proceed, we need one more computation under our belts. This is the Green's functions for the chiral bosons $\langle\phi_\pm(x)\phi_\pm(y)\rangle$. This is straightforward. To avoid UV divergences, we first subtract the constant term and define

$$G_\pm(x,y) = \langle\,\phi_\pm(x)\phi_\pm(y)\,\rangle - \langle\,\phi_\pm(0)^2\,\rangle$$

We then have

$$\begin{aligned} G_-(x,y) &= \frac{1}{4\beta^2}\int_0^\infty \frac{dp\,dq}{(2\pi)^2}\,\frac{2}{\sqrt{pq}}\langle\, a_p a_q^\dagger\,\rangle\,\left(e^{ipx-iqy}-1\right)e^{-(p+q)/2\Lambda} \\ &= \frac{1}{4\beta^2}\int_0^\infty \frac{dp}{2\pi}\frac{2}{p}\left(e^{ip(x-y)}-1\right)e^{-p/\Lambda} \\ &= \frac{1}{4\pi\beta^2}\log\left(\frac{\epsilon}{\epsilon - i(x-y)}\right) \end{aligned}$$

Note that $G_-(x,x)=0$, as it should. Meanwhile, at large distances the Green's function exhibits a logarithmic divergence. This infra-red behaviour is characteristic of massless scalar fields in two dimensions. Similarly, we have

$$G_+(x,y) = \frac{1}{4\pi\beta^2}\log\left(\frac{\epsilon}{\epsilon + i(x-y)}\right)$$

The Correlators

Finally, we have the tools to compute correlation functions in this theory. But the question that we should first ask is: what are the operators? The first point to note is that ϕ is not a good operator, because the classical field is not single valued. The same is true of the dual $\tilde{\phi}$. Instead, we must work with derivatives such as $\partial\phi$ or with so-called *vertex operators* of the form

$$e^{i\phi} =: e^{i\phi}:$$

where, as usual, normal ordering means all annihilation operators are moved to the right. Whenever we write an operator like $e^{i\phi}$ or $\cos\phi$, we will always mean the normal ordered version of these operators. In subsequent equations, we will keep punctuation to a minimum and usually won't explicitly write the : :.

In what follows, we will compute correlation functions of the form

$$\langle e^{i\phi_-(x)}e^{-i\phi_-(y)}\rangle \quad \text{and} \quad \langle e^{i\phi_+(x)}e^{-i\phi_+(y)}\rangle$$

In the next section we will then compare these with expressions involving fermions. At the same time, we will look a little more closely at the conditions for $e^{i\phi_\pm}$ to be consistent with the periodicity of ϕ.

To compute these expressions, we need to think more carefully about what the normal ordering means. For this, we will need the usual BCH identity,

$$e^A e^B = e^{A+B}e^{+\frac{1}{2}[A,B]} = e^B e^A e^{[A,B]}$$

where the higher order terms vanish whenever $[A, B]$ is a constant. We apply this to the operators $A = \alpha a + \alpha' a^\dagger$ and $B = \beta a + \beta' a^\dagger$. We have

$$
\begin{aligned}
: e^A :: e^B : &= e^{\alpha' a^\dagger} e^{\alpha a} e^{\beta' a^\dagger} e^{\beta a} \\
&= e^{\alpha' a^\dagger} e^{\beta' a^\dagger} e^{\alpha a} e^{\beta a} e^{\alpha \beta'} \\
&= : e^{A+B} : \, e^{\langle AB \rangle}
\end{aligned}
$$

Applying this to the vertex operators $e^{i\phi}$, which are nothing more than exponentials of many creation and annihilation operators, we have

$$\langle e^{i\phi_-(x)} e^{-i\phi_-(y)} \rangle = \langle e^{i\phi_-(x) - i\phi_-(y)} \rangle \, e^{G_-(x,y)}$$

But the correlation function on the right-hand side is of a normal ordered operator and this is simply $\langle : e^{i\phi_-(x)-i\phi_-(y)} : \rangle = 1$, since only the 1 in the Taylor expansion of the exponential contributes. We're left with

$$\langle e^{i\phi_-(x)} e^{-i\phi_-(y)} \rangle = e^{G_-(x,y)} = \left(\frac{\epsilon}{\epsilon - i(x-y)} \right)^{1/4\pi\beta^2} \tag{7.59}$$

Similarly

$$\langle e^{i\phi_+(x)} e^{-i\phi_+(y)} \rangle = e^{G_+(x,y)} = \left(\frac{\epsilon}{\epsilon + i(x-y)} \right)^{1/4\pi\beta^2} \tag{7.60}$$

Note that the correlation functions depend in an interesting way on the radius of the compact boson β^2. This confirms a statement that we made at the beginning of this section: the radius of the boson β^2 is a genuine parameter of the theory. In the language of conformal field theory, we would say that the operator $e^{i\phi_\pm}$ has dimension $1/8\pi\beta^2$.

7.5.3 The Bosonization Dictionary

The hard work is now behind us. Looking at the correlation functions (7.59) and (7.60), it is clear that they take a particularly simple form if we choose the radius of the boson to be

$$\beta^2 = \frac{1}{4\pi}$$

We can then compare correlation functions for right-moving fermions (7.47) and bosons (7.59),

$$\langle \psi_-(x) \psi_-^\dagger(y) \rangle = \frac{i}{2\pi} \frac{1}{(x-y) + i\epsilon} \quad \text{and} \quad \langle e^{i\phi_-(x)} e^{-i\phi_-(y)} \rangle = \frac{i\epsilon}{(x-y) + i\epsilon}$$

This tells us that we should identify

$$\psi_-(x) \;\longleftrightarrow\; \sqrt{\frac{1}{2\pi\epsilon}}\, e^{i\phi_-(x)} \tag{7.61}$$

where, recall, $\Lambda = 1/\epsilon$ is our UV cut-off. Similarly, comparing the correlation functions for left-moving operators, we have the map

$$\psi_+(x) \;\longleftrightarrow\; \sqrt{\frac{1}{2\pi\epsilon}}\, e^{-i\phi_+(x)} \tag{7.62}$$

We can also develop the map between composite operators. The simplest is the quadratic, mass term for fermions

$$\bar{\psi}\psi = \psi_-^\dagger(x)\psi_+(x) + \psi_+^\dagger(x)\psi_-(x) \;\longleftrightarrow\; \frac{1}{2\pi\epsilon}\left(e^{-i\phi_-(x)}e^{-i\phi_+(x)} + e^{i\phi_+(x)}e^{i\phi_-(x)}\right)$$

At this point, we just need to use the standard BCH identity, $e^A e^B = e^{A+B}e^{-[A,B]/2}$. Using the commutation relation (7.58), we have

$$\bar{\psi}\psi \;\longleftrightarrow\; -\frac{1}{2\pi\epsilon}\left(e^{-i\phi(x)} + e^{i\phi(x)}\right) = -\frac{1}{\pi\epsilon}\cos\phi \tag{7.63}$$

Similarly, the chiral mass term

$$i\bar{\psi}\gamma^3\psi \;\longleftrightarrow\; -\frac{1}{\pi\epsilon}\sin\phi$$

These will be important in the next section when we will understand better how to think of massive fermions in the bosonic language.

Matching Currents

Bosonization is a kind of duality, in which two seemingly different theories secretly describe the same physics. In any such duality, the most important objects to match on both sides are the conserved currents. We will see how this pans out in the present context.

The vector (7.43) and axial (7.44) currents are, like the mass term, composite, quadratic operators. For example,

$$j_V^0 = -\psi^\dagger\psi = -(\psi_-^\dagger\psi_- + \psi_+^\dagger\psi_+) \quad \text{and} \quad j_V^1 = -\psi^\dagger\sigma^3\psi = -\psi_-^\dagger\psi_- + \psi_+^\dagger\psi_+$$

However, it turns out that we need to be a little more careful in defining these operators. We do this through point splitting. For example, consider

$$\begin{aligned}
\psi_-^\dagger\psi_- &:= \lim_{y\to x}\ \psi_-^\dagger(x)\psi_-(y) \\
&\longleftrightarrow \lim_{y\to x}\ \frac{1}{2\pi\epsilon}e^{-i\phi_-(x)}e^{i\phi_-(y)} \\
&= \lim_{y\to x}\ \frac{1}{2\pi\epsilon}e^{-i(\phi_-(x)+\phi_-(y))}e^{G_-(x,y)} \\
&= \lim_{x\to y}\ \frac{1}{2\pi\epsilon}\left(1-i(x-y)\frac{\partial\phi_-(x)}{\partial x}+\ldots\right)\frac{\epsilon}{\epsilon-i(x-y)} \\
&= \frac{1}{2\pi}\frac{\partial\phi_-(x)}{\partial x}+\lim_{x\to y}\left(\frac{i}{2\pi(x-y)}\right)
\end{aligned}$$

Note that this expression comes with an infinite, constant term. We can remove this simply by normal ordering the fermionic operator. Identical calculations also hold for $\psi_+^\dagger\psi_+$, leaving us with the map

$$:\psi_\pm^\dagger\psi_\pm:\ \longleftrightarrow\ \frac{1}{2\pi}\frac{\partial\phi_\pm}{\partial x}$$

From this we can read off the map between currents,

$$j_V^0\ \longleftrightarrow\ \frac{1}{2\pi}\frac{\partial(\phi_-+\phi_+)}{\partial x}=\frac{1}{2\pi}\frac{\partial\phi}{\partial x}$$

and

$$j_V^1\ \longleftrightarrow\ \frac{1}{2\pi}\frac{\partial(\phi_--\phi_+)}{\partial x}=-\frac{1}{2\pi}\frac{1}{\beta^2}\pi(x)$$

Recalling that the classical momentum is $\pi=\beta^2\dot\phi$, we identify $j_V^1\longleftrightarrow-\dot\phi/2\pi$. In other words, we learn that the vector current of fermions is related to the topological current in the bosonic language

$$j_V^\mu\ \longleftrightarrow\ -j^\mu_{\rm wind}=-\frac{1}{2\pi}\epsilon^{\mu\nu}\partial_\nu\phi \tag{7.64}$$

Similarly,

$$j_A^\mu\ \longleftrightarrow\ 2j^\mu_{\rm shift}=2\beta^2\partial^\mu\phi \tag{7.65}$$

The methods that we've described above can be used to find the map between all other operators in the theory. For our purposes, the basic dictionary described above will suffice.

7.5.4 The Allowed Operators: Is the Boson Really a Fermion?

We have seen that, when $\beta^2 = 1/4\pi$, the operators $e^{i\phi_\pm}$ can be identified with free fermions through the map (7.61) and (7.62). But there is one subtlety that we didn't address: are the operators $e^{i\phi_\pm}$ compatible with the periodicity of ϕ?[14]

Because $\phi \in [0, 2\pi)$, the operator $e^{i\phi}$ is perfectly fine, as indeed is $e^{in\phi}$ for any $n \in \mathbf{Z}$. The dual scalar, defined by (7.54), also has periodicity $\tilde{\phi} \in [0, 2\pi)$, so that $e^{i\tilde{\phi}}$ is also fine. In general, we can have any operator of the form $e^{in\phi + iw\tilde{\phi}}$ with $n, w \in \mathbf{Z}$. For a general value of β^2, this means that the allowed operators are

$$e^{in\phi + iw\tilde{\phi}} = e^{i(n+2\pi\beta^2 w)\phi_-}\, e^{i(n-2\pi\beta^2 w)\phi_+}$$

Restricting to $\beta^2 = 1/4\pi$, we have

$$e^{in\phi + iw\tilde{\phi}} = e^{i(n+w/2)\phi_-}\, e^{i(n-w/2)\phi_+}$$

To get a purely chiral operator we could, for example, set $n = 1$ and $w = \pm 2$. But this leaves us with $e^{2i\phi_\pm}$, rather than $e^{i\phi_\pm}$. This is rather disconcerting, since it means that the operators $e^{i\phi_\pm}$ are not in the spectrum of the theory because they are incompatible with the periodicity of ϕ and $\tilde{\phi}$. Yet these are precisely the operators that we want to identify with a single fermion. What's going on?!

The answer is that the compact boson is not actually equivalent to a theory of a free fermion. Instead, it is equivalent to a theory of a fermion coupled to a $\mathbf{Z}_2$ gauge symmetry, acting as

$$\mathbf{Z}_2 : \psi \mapsto -\psi \tag{7.66}$$

This eliminates the single fermion from the spectrum, but leaves us with the composite operators $\psi\psi$ and $\bar{\psi}\psi$.

The need to couple the free fermion to a $\mathbf{Z}_2$ gauge field shows up in another way which we briefly describe here. If the two theories are equivalent, then their partition functions should coincide. It is straightforward to compute the partition function for the compact boson on a torus $\mathbf{T}^2$. It agrees with that of a free fermion only if we sum over both periodic and anti-periodic boundary conditions on the torus. (These are usually referred to as Ramond and Neveu-Schwarz sectors respectively.) The fact that we need to sum over both boundary conditions is another way of saying that the fermion is coupled to a $\mathbf{Z}_2$ gauge field, ensuring that configurations related by (7.66) are physically identified.

[14] I'm grateful to Carl Turner for explaining this to me.

7.5.5 Massive Thirring = Sine-Gordon

Having spent all this time developing the bosonization dictionary, we can now use it in anger. As we will see, the nice thing about the bosonization map is that it very often takes a strongly coupled theory and rewrites it in terms of a weakly coupled theory using the other variables.

Let's go back to the free theory of a compact scalar,

$$S = \int d^2x \; \frac{\beta^2}{2}(\partial_\mu \phi)^2$$

We know that for the specific value $\beta^2 = 1/4\pi$, this is equivalent to a free, massless Dirac fermion. But what about the other values of β^2? This is easy to answer using our bosonization dictionary. We split the kinetic term up as

$$\beta^2 = \frac{1}{4\pi} + \frac{g}{2\pi^2}$$

and think of the second piece, proportional to g, as a bosonic current-current interaction,

$$j^\mu_{\text{wind}}\, j_{\text{wind}\,\mu} = -\frac{1}{4\pi^2}(\partial_\mu \phi)^2$$

Adding such a current is straightforward for the boson: it just shifts the coefficient of the kinetic term away from the magic value. Written in terms of the fermion, it must again be a current-current interaction, this time of the form

$$j^\mu_V j_{V\mu} = (\bar\psi\gamma^\mu\psi)(\bar\psi\gamma_\mu\psi)$$

This is referred to as a *Thirring interaction*. Rather surprisingly, we learn that a general, free compact boson corresponds to an interacting fermion.

More generally, we can consider the massive, interacting Thirring model, with action

$$S = \int d^2x \; i\bar\psi \not{\partial}\psi - m\bar\psi\psi - g(\bar\psi\gamma^\mu\psi)(\bar\psi\gamma_\mu\psi) \tag{7.67}$$

Bosonization maps this into a compact boson with a potential, known as the *Sine-Gordon model*,

$$S = \int d^2x \; \frac{\beta^2}{2}(\partial_\mu\phi)^2 + \frac{m}{\pi\epsilon}\cos\phi$$

Note that the action includes an explicit mention of the UV cut-off $\Lambda = 1/\epsilon$. The potential $V(\phi) \sim -\cos\phi$ has its minimum at $\phi = 0$ and so, indeed, would seem to give a mass to ϕ as required.

There are a couple of cute subtleties that we learn from the bosonization map. First, we usually think about adding interaction terms to the Hamiltonian which are positive definite. For our fermionic theory, the requirement is slightly different. We must have $\beta^2 > 0$ on the bosonic side but, in terms of fermions, this translates to

$$g > -\frac{\pi}{2}$$

We learn that we can suffer a negative contribution to the Hamiltonian, as long as it's not too negative.

Second, we expect that the role of m is to make the excitation massive on both sides. But that's not quite true. Recall that the two-point correlators (7.59) and (7.60) allow us to read off the dimension of the vertex operators $e^{i\phi_\pm}$ or, equivalently, the dimension of the fermion. This dimension is $1/8\pi\beta^2$. It means that the $\cos\phi$ potential for the boson (or, equivalently, the mass term for the fermion) is relevant only if

$$\frac{1}{4\pi\beta^2} < 2 \quad \Rightarrow \quad \beta^2 > \frac{1}{8\pi} \quad \Rightarrow \quad g > -\frac{\pi}{4}$$

In other words, for $-\pi/2 < g < -\pi/4$, the mass term is an irrelevant operator and the massive Thirring model describes a massless theory in the infra-red!

Fermion = Kink

It will pay to look a little more closely at what becomes of a single, massive fermion. The answer to this follows from looking at the map between currents (7.64). A single fermion carries charge $Q_V = \int dx\; j^0_V = 1$. Correspondingly, it corresponds to a state in the bosonic theory with charge

$$Q_{\rm wind} = \frac{1}{2\pi}\int dx\; \partial_x\phi = -1$$

It is straightforward to find a classical configuration that carries this charge. The minima of the potential $V(\phi) \sim -\cos\phi$ lie at $\phi = 2\pi n$. We simply need to take a a configuration that interpolates between two minima, say from $\phi = 2\pi$ at $x \to -\infty$ to $\phi = 0$ at $x \to +\infty$. We learn that the fermion is identified with a kink in the Sine-Gordon model.

We can explore this kink in more detail. The classical energy of any configuration in the Sine-Gordon model can be written, up to an unimportant constant, as

$$\mathcal{E} = \int dx\; \frac{\beta^2}{2}\phi'^{\,2} + \frac{2m}{\pi\epsilon}\sin^2(\phi/2)$$

We can rewrite this using the *Bogomolnyi trick*, in which we complete the square thus:

$$\mathcal{E} = \int dx\; \frac{\beta^2}{2}\left(\phi' \pm \sqrt{\frac{4m}{\beta^2\pi\epsilon}}\sin(\phi/2)\right)^2 \mp \sqrt{\frac{4m\beta^2}{\pi\epsilon}}\phi'\;\sin(\phi/2) \tag{7.68}$$

The first term is a total square, and hence positive definite. The second term is a total derivative. This ensures that we can bound the energy of any configuration in terms of the end points

$$\mathcal{E} \geq 4\sqrt{\frac{m\beta^2}{\pi\epsilon}}\left|\Big[\cos(\phi/2)\Big]^{+\infty}_{-\infty}\right|$$

For a kink that interpolates between neighbouring minima, we have

$$\mathcal{E}_{\rm kink} \geq 8\sqrt{\frac{m\beta^2}{\pi\epsilon}}$$

with equality if the Bogomolnyi equations are satisfied, which can be found in the total square in (7.68),

$$\phi' = \pm\sqrt{\frac{4m}{\beta^2\pi\epsilon}}\sin(\phi/2)$$

These equations aren't quite satisfactory, since they still include the UV cut-off ϵ. This arises here because we're using an unholy combination of classical and quantum analysis. Still, there's a simple way to fix it. For $g = 0$ or, equivalently, $\beta^2 = 1/4\pi$, the Sine-Gordon model describes a free fermion. Here, the mass of the Bogomolnyi kink is

$$\mathcal{E}_{\rm kink} = \frac{4}{\pi}\sqrt{\frac{m}{\epsilon}} \tag{7.69}$$

which suggests that we should take the $\epsilon = 16/m\pi^2$ if we want the semi-classical analysis of the Sine-Gordon model to reproduce the mass m of the fermion.

There is a more general lesson lurking here. Bosonization provides us with a duality between two different theories, in which the elementary excitation of one theory is mapped into a soliton of the other. This, it turns out, is a characteristic signature of dualities in different dimensions. (We will meet an example in 3d where particles are mapped to vortices in Section 8.2.) Often these other dualities are not well understood. Two dimensional bosonization provides a useful grounding, where the map between the two theories can be performed explicitly.

7.5.6 QED$_2$: The Schwinger Model

The *Schwinger model* is the name given to QED in two dimensions: it consists of a single Dirac fermion, coupled to a $U(1)$ gauge field. The action is

$$S = \int d^2x \; \frac{1}{2e^2} F_{01}^2 + \frac{\theta}{2\pi} F_{01} + i\bar{\psi} \not{D} \psi - im\bar{\psi}\psi$$

As we have seen in Sections 7.1 and 7.2, Maxwell theory is strongly coupled in two dimensions, and electric charges confine. When the fermion is very heavy, $m^2 \gg e^2$, we can use standard perturbative techniques to solve the model. In contrast, when the fermions are light the theory is strongly coupled and we must look elsewhere for help. Fortunately, as we now see, bosonization will do the job for us.

The coupling between the fermion and the gauge field is buried in the covariant derivative: $\not{D}\psi = \not{\partial}\psi - iA_\mu \gamma^\mu \psi$. As usual, the gauge field couples to the fermion current, as $A_\mu j_V^\mu$. This makes it straightforward to write down the bosonised version,

$$\begin{aligned} S &= \int d^2x \; \frac{1}{2e^2} F_{01}^2 + \frac{\theta}{2\pi} F_{01} + \frac{1}{8\pi}(\partial_\mu \phi)^2 + \frac{1}{2\pi} A_\mu \epsilon^{\mu\nu} \partial_\nu \phi + \frac{m}{\pi\epsilon} \cos\phi \\ &= \int d^2x \; \frac{1}{2e^2} F_{01}^2 + \frac{1}{2\pi}(\theta + \phi) F_{01} + \frac{1}{8\pi}(\partial_\mu \phi)^2 + \frac{m}{\pi\epsilon} \cos\phi \end{aligned} \tag{7.70}$$

where the second line follows after an integration by parts. Already here, there's something rather nice. Suppose that the mass $m = 0$. The equation of motion for ϕ is then

$$\frac{1}{4\pi} \partial^2 \phi = \frac{1}{2\pi} F_{01}$$

But we know from our bosonization formula (7.65) that the axial current is $j_A^\mu = -\partial^\mu \phi / 2\pi$, so we can write this as

$$\partial_\mu j_A^\mu = \frac{1}{\pi} F_{01}$$

But this agrees with our earlier derivation (3.36) of the anomaly in two dimensions. Previously the anomaly was a subtle quantum effect; after bosonization, it simply becomes the equation of motion.

Meanwhile, the equation of motion for the gauge field includes

$$\partial_x \left(\frac{1}{e^2} F_{01} + \frac{1}{2\pi} \phi \right) = 0 \quad \Rightarrow \quad F_{01} = -\frac{e^2}{2\pi} \phi$$

where the second condition comes from requiring that this combination vanishes at infinity. This is reminiscent of our result in Section 7.1 where we found that the theta angle gives rise to a background magnetic field (7.7). However, once again, we find this result simply from the classical equation of motion, without the need to invoke any quantisation. A more careful analysis, along the lines of Section 7.1 shows that

$$F_{01} = -\frac{e^2}{2\pi}(\theta + \phi)$$

which seems very reasonable given the action (7.70). (Note: in Section 7.1, we denoted the Wilson line as ϕ; this is not to be confused with the bosonized fermion ϕ we are working with here.)

To answer further questions, note that the gauge field A_μ only appears in the field strength in (7.70). If we take the theory to sit on a line, so that there is no quantisation condition on F_{01}, we can integrate out the gauge field to get

$$S = \int d^2x \; \frac{1}{8\pi}(\partial_\mu \phi)^2 + \frac{m}{\pi\epsilon}\cos\phi - \frac{e^2}{8\pi^2}(\theta + \phi)^2$$

Note that we have now lost the periodicity in ϕ. (This is restored on a compact space where $\int F_{01} \in 2\pi\mathbf{Z}$. In this case, the potential gets replaced by $\min_n (\theta+\phi+2\pi n)^2$. We encountered similar periodic, but non-smooth potentials in our study of 4d Yang-Mills theory at large N in (6.18).)

There are a number of things we can now look at. First, suppose that our original fermions were massless, with $m = 0$. Note that we can now absorb the theta angle simply by rescaling $\phi \to \phi - \theta$. This is to be expected: as discussed in Section 3.3.3, the chiral anomaly means that the theta angle is always redundant in the presence of massless fermions. We're left simply with a real scalar field whose mass is

$$\text{mass}^2 = \frac{e^2}{\pi}$$

We learn that the massless Schwinger model is not, in fact, massless. It has a gap.

Let's now turn on the fermion mass m. The minima of the potential now sit at

$$\sin\phi = -\frac{e^2\epsilon}{4\pi m}(\theta + \phi) \tag{7.71}$$

For large m, this has many solutions but, at least when $\theta \neq \pi$, there is only a unique ground state. There is now no kink solution that interpolates between neighbouring minima because the minima are no longer degenerate. This reflects the physics of confinement that we saw in Section 7.1: a single fermion costs infinite energy due to the resulting flux tube which stretches to infinity. The finite energy excitations are mesons, bound states of fermions and anti-fermions. One may use the bosonized action above to study these in the limit of small mass.

Something interesting happens when $\theta = \pi$. This is simplest to see if we shift $\hat{\phi} = \phi - \pi$. The minima of (7.71) then sit at

$$\sin\hat{\phi} = \frac{e^2\epsilon}{4\pi m}\hat{\phi} \tag{7.72}$$

This can be solved graphically. When $m \gg e^2\epsilon$, there are many solutions. The obvious one at $\hat{\phi} = 0$ is actually a local maxima of the potential. There are then two degenerate minima. This is what we expect from our discussion in 7.1: integrating out the very heavy fermion leaves us with pure Maxwell theory at $\theta = \pi$, and we know that this has two degenerate ground states.

Now we can decrease the mass. The number of solutions to (7.72) starts to decrease and for $m \ll e^2\epsilon$, we have just a single ground state at $\hat{\phi} = 0$. The critical point happens at $4\pi m = e^2\epsilon$, when the two degenerate minima merge into a single one. But this is a very familiar phase transition: it is described by the Ising critical point. We learn that as we vary the mass at $\theta = \pi$, the Schwinger model becomes gapless and is described by the 2d Ising CFT. Note that this is exactly the same behaviour that we saw for the Abelian Higgs model in Section 7.2.

7.6 Non-Abelian Bosonization

Consider N, massless Dirac fermions, ψ_i with $i = 1, \ldots, N$. Decomposing each into a Weyl fermion, the action is

$$S = \int d^2x \; i\psi^\dagger_{-i}\partial_+\psi_{-i} + i\psi^\dagger_{+i}\partial_-\psi_{+i} \tag{7.73}$$

where $\partial_\pm = \partial_t \pm \partial_x$. We clearly have a $U(N)\times U(N)$ chiral symmetry, which rotates the left- and right-handed fermion separately. In fact, in two dimensions each Weyl fermion can be further split into two Majorana-Weyl fermions. This follows from the fact that we can choose a basis of gamma matrices (7.30) that are both in the chiral basis and real. The upshot is that the free fermions (7.73) actually have an $O(2N) \times O(2N)$ chiral symmetry.

But what becomes of this symmetry on the bosonic side? We have N compact, real bosons ϕ_i. Because these are compact, there is not even an $O(N)$ symmetry that rotates them. (This is the statement that $\mathbf{R}^N$ has a $O(N)$ symmetry acting on it, but the torus $\mathbf{T}^N$ does not.) Instead, all we have is the Cartan subalgebra $U(1)^N$, together with the corresponding action on the dual scalars.

What to make of this? One might think that it's no biggie: after all, the bosonic theory should presumably have the enlarged symmetry since its equivalent to its fermionic cousin. But it would be nice to make this manifest. And, fortunately, there is a beautiful way to do so, as first explained by Witten.

Here we will bosonize, keeping the $U(N) \times U(N)$ symmetry manifest, although a similar method works for the $O(2N) \times O(2N)$ chiral symmetry too. Let's start by looking at the currents. The overall $U(1) \times U(1)$ takes a similar form to the previous section, but we write this as

$$j_- = 2\psi^\dagger_{-i}\psi_{-i} \quad \text{and} \quad j_+ = 2\psi^\dagger_{+i}\psi_{+i}$$

These are the components of the vector and axial current written in the lightcone coordinates $x^\pm = t \pm x$. But now we also have the non-Abelian flavour symmetries, with the corresponding $SU(N)$ currents,

$$J^a_- = 2\psi^\dagger_{-i}T^a_{ij}\psi_{-j} \quad \text{and} \quad J^a_+ = 2\psi^\dagger_{+i}T^a_{ij}\psi_{+j}$$

where T^a_{ij} are the generators of $su(N)$. The equations of motion for the fermions ensure that the currents obey

$$\partial_+ j_- = \partial_- j_+ = 0 \quad \text{and} \quad \partial_+ J^a_- = \partial_- J^a_+ = 0$$

We would like to ask: can we write down a bosonic model that has the same currents? Rather than jumping immediately to the model, we're first going to write down an ansatz for the form of the currents, and then see if we can come up with an action which reproduces this.

We've already seen how to do this for the $U(1)$ currents: we simply write them in terms of a compact boson ϕ. In lightcone coordinates, this becomes

$$j_- = \frac{1}{2\pi}\partial_-\phi \quad \text{and} \quad j_+ = -\frac{1}{2\pi}\partial_+\phi$$

What's the analog expression for the non-Abelian currents? Here's a guess. First let's write the Abelian currents in a way that highlights their $U(1)$-ness. We define $\tilde{g} = e^{i\phi} \in U(1)$. Then we can write

$$j_- = -\frac{i}{2\pi}\tilde{g}^{-1}\partial_-\tilde{g} \quad \text{and} \quad j_+ = \frac{i}{2\pi}\tilde{g}^{-1}\partial_+\tilde{g} \tag{7.74}$$

This is now something that we can hope to generalise. We introduce the group-valued field

$$g(x,t) \in SU(N)$$

We then define the currents

$$J_- = -\frac{i}{2\pi} g^{-1}\partial_- g \quad \text{and} \quad J_+ = \frac{i}{2\pi}(\partial_+ g)g^{-1} \tag{7.75}$$

Note that the ordering of g and g^{-1} matters in these expressions and differs from what we might naively have written down simply by copying (7.74). The reason for the choice above is that we want these currents to obey conservation laws

$$\partial_+ J_- = \partial_- J_+ = 0 \tag{7.76}$$

Happily, the ordering in (7.75) means that the first of these conservation laws implies the second,

$$\begin{aligned} \partial_+ J_- = 0 \quad \Rightarrow \quad (\partial_+ g^{-1})\partial_- g + g^{-1}\partial_+\partial_- g &= 0 \\ g(\partial_+ g^{-1})\partial_- g + \partial_+\partial_- g &= 0 \\ \partial_+ g(\partial_- g^{-1})g + \partial_+\partial_- g &= 0 \\ \partial_+ g\partial_- g^{-1} + (\partial_+\partial_- g)g^{-1} &= 0 \quad \Rightarrow \quad \partial_- J_+ = 0 \end{aligned} \tag{7.77}$$

Had we chosen a different order of g and g^{-1} in (7.75) then the conservation laws (7.76) turn out to be inconsistent with each other.

Now we've got a good candidate for the currents (7.75), we want to write down an action for g whose dynamics implies their conservation. In fact, given the group structure, we are pretty restricted in what we can write down. If we want an action with two derivatives, then there is a unique choice,

$$S = \int d^2x \; \frac{1}{4\lambda^2} \; \text{tr}\left(\partial_\mu g \, \partial^\mu g^{-1}\right) \tag{7.78}$$

for some dimensionless coupling λ^2. We have met this structure before: it is identical to the chiral Lagrangian (5.7) that we used to describe pions in QCD. This is a non-linear sigma model, whose target space is the group manifold $SU(N)$. In two-dimensions, the sigma-models whose target spaces are group manifolds are sometimes referred to as *principal chiral models*.

The action (7.78) enjoys two global symmetries, in which we act by an $SU(N)$ transformation on either the left or right,

$$g \to Lg \quad \text{or} \quad g \to gR \;, \quad L, R \in SU(N)$$

This gives rise to two currents $J_L^\mu \sim (\partial^\mu g)g^{-1}$ and $J_R^\mu \sim (\partial^\mu g^{-1})g$. (We computed these currents in the context of the chiral Lagrangian in (5.11) and (5.12).) These indeed take a similar form to our chiral currents J_- and J_+ defined in (7.75), which is encouraging. However, closer inspection tells us that things aren't quite as straightforward. The equation of motion from (7.78) implies that $\partial_\mu J_L^\mu = \partial_\mu J_R^\mu = 0$, but this not the same thing as what we wanted in (7.76). We learn that the symmetry structure of the bosonic model (7.78) differs from that of N free fermions.

There is also a dynamical reason why the sigma model (7.78) cannot describe free fermions: it is asymptotically free. The coupling $\lambda^2(\mu)$ runs with scale μ and its one-loop beta function can be shown to be

$$\mu \frac{d\lambda^2}{d\mu} = -(N-2)\frac{\lambda^2}{4\pi}$$

This is similar to the behaviour of the $\mathbf{CP}^{N-1}$ model that we met in Section 7.3. (It is even more similar to the behaviour of the $O(N)$ models in two dimensions that we met in the lectures on Statistical Field Theory.) In the infra-red, the non-linear sigma model (7.78) is expected to flow to a gapped phase.

7.6.1 The Wess-Zumino-Witten Term

The simple sigma-model (7.78) does not have the right properties to describe free fermions. However, it is possible to modify this theory to give us what we want. The modification is a little subtle, but it's a subtlety that we have met before: the extra term cannot be written as an integral over 2d spacetime, but instead only over a 3d spacetime. Such terms are called *Wess-Zumino-Witten* terms, and we saw an example in Section 5.5 in the context of the chiral Lagrangian for QCD.

Things are simplest if we work in the Euclidean path integral and take our spacetime to be $\mathbf{S}^2$. We introduce a three-dimensional ball, D, such that $\partial D = \mathbf{S}^2$. We extend the fields $g(x,t)$ over $\mathbf{S}^2$ to $g(y)$, where y are coordinates on the ball D. We then consider the modified action,

$$S = \int d^2x \, \frac{1}{4\lambda^2} \, \mathrm{tr}\left(\partial_\mu g \, \partial^\mu g^{-1}\right) + k \int_D d^3y \; \omega \tag{7.79}$$

where

$$\omega = \frac{i}{24\pi}\epsilon^{\mu\nu\rho}\,\mathrm{tr}\left(g^{-1}\frac{\partial g}{\partial y^\mu}g^{-1}\frac{\partial g}{\partial y^\nu}g^{-1}\frac{\partial g}{\partial y^\rho}\right)$$

This has a very similar structure to the five-dimensional WZW term (5.35) that we introduced in Section 5.5.

Just as in the 4d story, there is an ambiguity in our choice of 3d-dimensional ball D with $\partial D = \mathbf{S}^2$. We could just as well take a ball D', also with $\partial D' = \mathbf{S}^2$ but with the opposite orientation. The now-familiar topological quantisation conditions tell us that

$$\exp\left(ik\int_D d^3y\;\omega\right) = \exp\left(-ik\int_{D'} d^3y\;\omega\right) \quad\Rightarrow\quad \exp\left(ik\int_{\mathbf{S}^3} d^3y\;\omega\right) = 1$$

where we have stitched together the two three-balls to make the three-sphere $\mathbf{S}^3 = D\cup D'$. The integrand provides a map from $\mathbf{S}^3$ to the group manifold $SU(N)$ with fields $g(y)$. But, as we saw in the context of instantons in Section 2.3, these maps are characterised by the homotopy group

$$\Pi_3(SU(N)) = \mathbf{Z} \quad \text{for } N \geq 3$$

It turns out that, for configurations with winding n, the WZW term evaluates to $\int_{\mathbf{S}^3} d^3y\;\omega = 2\pi n$. This quantisation condition then tells us that the coefficient of the WZW term must be an integer

$$k \in \mathbf{Z}$$

We refer to this integer as the *level*.

The effect of the WZW term in two dimensions is, in many ways, much more dramatic than that of its four dimensional counterpart. In 4d, we had to look at rather specific scattering processes, or baryons, to see the implications of the WZW term. In contrast, in 2d the presence of the WZW term affects even the phase of the theory. To see this, we can look again at the beta function for λ^2. At one-loop, one finds that it picks up an extra term, given by

$$\mu\frac{d\lambda^2}{d\mu} = -(N-2)\frac{\lambda^2}{4\pi}\left[1-\left(\frac{\lambda^2 k}{4\pi}\right)^2\right]$$

We see that there is now a fixed point of the RG equation, at

$$\lambda^2 = \frac{4\pi}{|k|} \tag{7.80}$$

Here the theory is described by a gapless CFT, known as the $SU(N)_k$ WZW theory. It is completely solvable using various CFT techniques, although we will not discuss these here. Since our one-loop computation is valid for $\lambda^2 \ll 1$, we can trust the existence of this fixed point only when $k \gg 1$ and the theory remains weakly coupled. Nonetheless, the fixed point is known to persist for all $k \in \mathbf{Z}$.

At the fixed points, something nice happens with the currents. The classical equation of motion of the action (7.79) is

$$\frac{1}{2\lambda^2}\partial_\mu(g^{-1}\partial_\mu g) - \frac{k}{8\pi}\epsilon_{\mu\nu}\partial_\mu(g^{-1}\partial_\nu g) = 0$$

In lightcone coordinates, with metric $\eta_{+-} = 1$, this reads

$$\left(\frac{1}{2\lambda^2} + \frac{k}{8\pi}\right)\partial_-(g^{-1}\partial_+ g) + \left(\frac{1}{2\lambda^2} - \frac{k}{8\pi}\right)\partial_+(g^{-1}\partial_- g) = 0$$

At the fixed point (7.80), one of these terms vanishes. Which one depends on the sign of k. For $k > 0$, we're left with

$$\partial_-(g^{-1}\partial_+ g) = 0$$

which is precisely the condition $\partial_- J_+ = 0$ that we wanted for the chiral current (7.76). The other condition $\partial_+ J_- = 0$ then follows automatically, as shown in (7.77).

We've found that, for each N, there is a set of conformal field theories, labelled by $k \in \mathbf{Z}$. That's nice but which, if any, describe N free fermions? The answer to this comes from looking more closely at the algebra obeyed by the $SU(N)$ currents. We won't give details of the calculation here, and instead just sketch the basic facts. The $SU(N)$ currents turn out to obey an extension of the usual $su(N)$ Lie algebra, with an extra term referred to as a central charge,

$$[J^a_\pm(x), J^b_\pm(y)] = if^{abc}J^c_\pm(x)\delta(x-y) \pm \frac{ik}{4\pi}\delta^{ab}\delta'(x-y)$$

with f^{abc} the structure constants of $su(N)$ and $\delta'(x)$ the derivative of the delta function. This is known as a *Kac-Moody algebra*, and its properties are well studied. It is known that the algebra has unitary representations only if $k \in \mathbf{Z}$, a fact which sits well with our realisation as currents in the WZW model.

One can also compute the same algebra for N free Dirac fermions. Here the computation is somewhat simpler and follows from the usual commutation relations for free fermions. One finds the Kac-Moody algebra above, but with the specific value $k = 1$.

We learn that we can bosonize N free Dirac fermions to an $SU(N)$ WZW model at level $k = 1$, together with a compact boson ϕ to describe the $U(1)$ currents. In other words, the following action

$$S = \int d^2x \; \frac{1}{8\pi}(\partial_\mu\phi)^2 + \frac{1}{16\pi}\mathrm{tr}\left(\partial_\mu g\, \partial^\mu g^{-1}\right) + \int_D d^3y \; \omega$$

is, despite appearances, N free Dirac fermions in disguise.

7.7 Further Reading

Quantum field theories in low dimensions were originally studied by particle physicists. They were viewed as toy models, in which some of the more outlandish behaviour of quantum field theory, such as confinement, or a dynamically generated mass, could be viewed in a tractable setting, giving comfort in a time of confusion. Later it was realised that many of these quantum field theories have direct application to condensed matter systems.

This programme was initiated by Schwinger who, in 1962, studied massless QED in $d = 1 + 1$ [174], in what is probably the first time that a strongly interacting quantum field theory was solved. This is a model which trivially confines and, somewhat less trivially, exhibits a mass gap. In these lectures, we solved it using bosonization techniques. Schwinger used operator methods. One conclusion that he took from this study was that thinking in terms of elementary particles can be misleading in strongly interacting field theories:

> "This line of thought emphasizes that the question "Which particles are fundamental?" is incorrectly formulated. One should ask "What are the fundamental fields?"."

The massive Schwinger model was revisited by Coleman and collaborators in the 1970s to better understand both confinement and the role played by the theta angle in two dimensions [28, 30]. The full phase structure of the theory, including the critical point at $\theta = \pi$, was described in [179].

Gross and Neveu introduced their models of N interacting fermions in 1974 [85]. Their goal was to test drive an asymptotically free theory which exhibits a dynamically generated mass scale as well as, in this case, dynamical spontaneous symmetry breaking. Witten later determined the spectrum of kinks [218] and showed how to reconcile the apparent breaking of the $U(1)$ chiral symmetry [219] with the lack of Goldstone bosons in two dimensions in [134, 26].

The role of instantons in determining the phase structure of the two-dimensional Abelian-Higgs model was first discussed by Callan, Dashen and Gross in [24]. One might have thought that this was a warm-up to understanding the vacuum structure of four-dimensional gauge theories, but in fact it was a warm-down to check that their earlier 4d analysis was sensible. The full phase diagram, including the critical point at $\theta = \pi$, was described in the appendix of Witten's $\mathbf{CP}^N$ paper [220]. A more modern perspective on this critical point was discussed in [123].

The $\mathbf{CP}^N$ model was proposed in 1978 [50, 81]. It was quickly noticed that it shares a number of properties with Yang-Mills, including asymptotic freedom, instantons and a large N expansion. It was first solved at large N by D'Adda, Lüscher and Di Vecchia [36]. Soon after, Witten studied the interplay between instantons, the theta term and the large N expansion, and argued that this provided a useful analogy for Yang-Mills in four dimensions [220]. The fact that the $\mathbf{CP}^1$ model at $\theta = \pi$ is a gapless theory was first conjectured by Haldane in [87]

In the high energy literature, bosonization was introduced by Sidney Coleman [29]. In the condensed matter literature, related results were derived slightly earlier by Luther and Peschel [128], and also by Mattis. Coleman ends his paper with the typically charming admission "Schroer has also pointed out that many of the results obtained here are in close correspondence with the results of [...] Luther and collaborators. Luther and I are in total agreement with Schroer on this point; we are also united in our embarrassment that we were incapable to reaching this conclusion unprompted. (Our offices are on the same corridor.)" The non-local relationship between fermions and bosons was discovered soon after by Mandelstam [131]. An earlier, lattice version of this relationship can be found in the Jordan-Wigner transformation. Finally, the non-Abelian bosonization is due to Witten in the beautiful paper [227].

There are a number of excellent reviews on bosonization, including [177, 178]

These lectures note do not discuss conformal field theories in $d = 1 + 1$ dimensions. This is a vast topic that deserves its own course. An introduction to the very basics can be found in the lectures on string theory [192]; an introduction to more than the basics can be found in the lectures by Ginsparg [74]; and a fuller treatment can be found in the big yellow book [40].

8. Quantum Field Theory on the Plane

In this section, we step up a dimension. We will discuss quantum field theories in $d = 2+1$ dimensions. Like their $d = 1+1$ dimensional counterparts, these theories have application in various condensed matter systems. However, they also give us further insight into the kinds of phases that can arise in quantum field theory.

8.1 Electromagnetism in Three Dimensions

We start with Maxwell theory in $d = 2+1$. The gauge field is A_μ, with $\mu = 0, 1, 2$. The corresponding field strength describes a single magnetic field $B = F_{12}$, and two electric fields $E_i = F_{0i}$. We work with the usual action,

$$S_{\text{Maxwell}} = \int d^3x \; -\frac{1}{4e^2} F_{\mu\nu}F^{\mu\nu} - A_\mu j^\mu \tag{8.1}$$

The gauge coupling has dimension $[e^2] = 1$. This is important. It means that $U(1)$ gauge theories in $d = 2+1$ dimensions coupled to matter are strongly coupled in the infra-red. In this regard, these theories differ from electromagnetism in $d = 3+1$.

We can start by thinking classically. The Maxwell equations are

$$\frac{1}{e^2}\partial_\mu F^{\mu\nu} = j^\nu$$

Suppose that we put a test charge Q at the origin. The Maxwell equations reduce to

$$\nabla^2 A_0 = Q\,\delta^2(\mathbf{x})$$

which has the solution

$$A_0 = \frac{Q}{2\pi}\log\left(\frac{r}{r_0}\right) + \text{constant}$$

for some arbitrary r_0. We learn that the potential energy $V(r)$ between two charges, Q and $-Q$, separated by a distance r, increases logarithmically

$$V(r) = \frac{Q^2}{2\pi}\log\left(\frac{r}{r_0}\right) + \text{constant} \tag{8.2}$$

This is a form of confinement, but it's an extremely mild form of confinement as the log function grows very slowly. For obvious reasons, it's usually referred to as *log confinement*.

In the absence of matter, we can look for propagating degrees of freedom of the gauge field itself. As explained in the previous section, we expect the gauge field to have a single, propagating polarisation state in $d = 2 + 1$ dimensions.

8.1.1 Monopole Operators

Something special happens for $U(1)$ gauge theories in $d = 2 + 1$ dimensions: they automatically come with an associated global $U(1)$ symmetry that we will call $U(1)_{\rm top}$, the "top" for "topological". The associated current is

$$J^\mu_{\rm top} = \frac{1}{4\pi}\epsilon^{\mu\nu\rho}F_{\nu\rho} \tag{8.3}$$

which obeys the conservation condition $\partial_\mu J^\mu_{\rm top} = 0$ by the Bianchi identity on $F_{\mu\nu}$. The associated conserved quantity is simply the magnetic flux

$$Q_{\rm top} = \int d^2x\ J^0_{\rm top} = \frac{1}{2\pi}\int d^2x\ B$$

In quantum field theory, symmetries act on local operators. The operators that transform under $U(1)_{\rm top}$ are not the usual fields of the theory. Rather, they are *disorder operators*, entirely analogous to the 't Hooft lines that we introduced in Section 2.6. In the present context, they are referred to as *monopole operators*.

We work in Euclidean space. A monopole operator $\mathcal{M}(x)$ inserted at a point $x \in \mathbf{R}^3$ is defined in the path integral by requiring that we integrate over field configurations in which there is a Dirac monopole inserted inserted at the point x. This means that, for an $\mathbf{S}^2$ surrounding x, we have

$$\frac{1}{4\pi}\int_{\mathbf{S}^2} d^2S_\mu\ \epsilon^{\mu\nu\rho}F_{\nu\rho} = 1 \tag{8.4}$$

This operator creates a single unit of magnetic flux so that, in the presence of $\mathcal{M}(x)$, the topological current is no longer conserved; instead it has a source

$$\partial_\mu J^\mu_{\rm top} = \delta^3(x) \tag{8.5}$$

Equivalently, the monopole operator is charged under $U(1)_{\rm top}$ so that

$$U(1)_{\rm top} : \mathcal{M}(x) \mapsto e^{i\alpha}\mathcal{M}(x) \tag{8.6}$$

The definition of monopole operators given above is somewhat abstract. As we will now see, in certain phases of the theory it is possible to give a more concrete definition.

Consider free Maxwell theory. Alternatively, consider $U(1)$ gauge theory coupled to charged fields with masses $m \gg e^2$. In both cases, the theory lies in the *Coulomb phase*, meaning that low energy spectrum contains just a single, free massless photon. The partition function is particularly straightforward; ignoring gauge fixing terms, we have

$$Z = \int \mathcal{D}A_\mu \; \exp\left(- \int d^3x \; \frac{1}{4e^2} F_{\mu\nu}F^{\mu\nu} \right)$$

Because the action depends only on $F_{\mu\nu}$, and not explicitly on A_μ, we can choose instead to integrate over the field strength. However, we shouldn't integrate over all field strengths; in the absence of monopole operators, we should integrate only over those that satisfy the Bianchi identity $\epsilon^{\mu\nu\rho}\partial_\mu F_{\nu\rho} = 0$. We can do this by introducing a Lagrange multiplier field $\sigma(x)$,

$$Z = \int \mathcal{D}F_{\mu\nu}\mathcal{D}\sigma \; \exp\left(- \int d^3x \; \frac{1}{4e^2} F_{\mu\nu}F^{\mu\nu} + \frac{i}{4\pi}\sigma\epsilon^{\mu\nu\rho}\partial_\mu F_{\nu\rho} \right) \tag{8.7}$$

If the field strength obeys the Dirac quantisation condition, then σ has periodicity 2π. But in this formulation, it is particularly straightforward to implement a monopole operator. We simply add to the path integral

$$\mathcal{M}(x) \sim e^{i\sigma(x)} \tag{8.8}$$

This ensures that the topological current has a source (8.5) or, equivalently, inserts a monopole at x.

We can now go one step further, and integrate out the field strength $F_{\mu\nu}$. We're left with an effective action for the Lagrange multiplier field $\sigma(x)$ which, in this context, is usually referred to as the *dual photon*. We're left with the effective action,

$$S_{\text{eff}} = \int d^3x \; \frac{e^2}{8\pi^2} \, \partial_\mu\sigma\partial^\mu\sigma \tag{8.9}$$

Clearly this describes a single, propagating degree of freedom. But this is what we expect for a photon in $d = 2 + 1$ which has just a single polarisation state.

In this formulation, the global symmetry $U(1)_{\text{top}}$ is manifest, and is given by

$$U(1)_{\text{top}} : \sigma \mapsto \sigma + \alpha \tag{8.10}$$

This agrees with our expected symmetry transformation (8.6) given the identification (8.8). The associated current can be read off from (8.9); it is

$$J^\mu_{\text{top}} = \frac{e^2}{(2\pi)^2} \, \partial^\mu\sigma \tag{8.11}$$

There's one, nice twist to this story. The theory (8.9) has a degeneracy of ground states, given by constant $\sigma \in [0, 2\pi)$. These degenerate ground states reflects the fact that if we place some magnetic flux in the Coulomb phase then it spreads out. In any of these ground states, the global symmetry $U(1)_{\rm top}$ acts like (8.10) and so is spontaneously broken. The associated Goldstone boson is simply σ itself. But this is equivalent to the original photon. We have the chain of ideas

$$\begin{aligned} \text{Coulomb Phase : } \quad \text{Unbroken } U(1)_{\rm gauge} \quad &\Leftrightarrow \quad \text{Spontaneously Broken } U(1)_{\rm top} \\ &\Leftrightarrow \quad \text{Goldstone Mode} = \text{Photon} \end{aligned}$$

A related set of ideas also holds in higher dimensions, but now with the $U(1)_{\rm top}$ a generalised symmetry, which acts on higher dimensional objects, as we discussed in Section 3.6.2. $d = 2+1$ dimensions is special because the disorder operator $\mathcal{M}(x)$ is a local operator, ensuring that $U(1)_{\rm top}$ is a standard global symmetry, rather than the less familiar generalised symmetry.

8.2 The Abelian-Higgs Model

We can get some more intuition for the role of monopole operators, and 3d gauge theories in general, by looking at the Abelian-Higgs model. This is a $U(1)$ gauge theory coupled to a scalar field ϕ which we take to have charge 1. The action is

$$S_{AH} = \int d^3x \; -\frac{1}{4e^2} F_{\mu\nu}^2 + |\mathcal{D}_\mu \phi|^2 - m^2 |\phi|^2 - \frac{\lambda}{2}|\phi|^4 \tag{8.12}$$

We will look at what happens to this theory as we vary the mass m^2 from positive to negative. This is a game that we've already played in both $d = 3+1$ dimensions (in Section 2.5.2) and in $d = 1+1$ dimensions (in Section 7.2). In both cases, the interesting physics came from vortices in the $m^2 < 0$ phase, and the same will be true here.

When the mass is small, $|m| \ll e^2$, the theory is strongly coupled in the infra-red. It is difficult to get a handle on the physics here, although we will ultimately be able to understand what happens. In contrast, when $|m| \gg e^2$, we can first understand the dynamics of the scalar in a regime where the gauge field is weakly coupled, and then figure out what's left. We first look at these two phases.

$\underline{m^2 \gg e^4}$: When $m^2 > 0$ we can simply integrate out the scalar, to leave ourselves with free Maxwell theory below the scale of m^2. This is the gapless Coulomb phase, in which we have an unbroken $U(1)$ gauge symmetry. As we explained above, this means that the global symmetry $U(1)_{\rm top}$ is spontaneously broken. The Goldstone mode is the photon.

There are also massive, charged excitations in this phase that come from the ϕ field. They interact through the Coulomb force which means that charges of opposite sign experience a logarithmically confining potential (8.2).

<u>$m^2 \ll -e^4$:</u> This is the *Higgs phase*. The scalar condenses,

$$|\phi|^2 = -\frac{m^2}{\lambda}$$

giving the photon a mass. This phase is gapped. The $U(1)_{\rm gauge}$ symmetry is spontaneously broken. But now the global topological symmetry $U(1)_{\rm top}$ is unbroken.

The finite energy states of the theory which carry non-vanishing $Q_{\rm top}$ charge are the vortices. We discussed these in detail in both $d = 3+1$ dimensions where the vortices are strings (see section 2.5.2) and in $d = 1+1$ dimensions where the vortices are instantons (see Section 7.2). In $d = 2+1$, vortices are particle-like excitations. They are classical configurations in which the phase of ϕ winds asymptotically in the spatial plane $\mathbf{R}^2$. They have finite energy, and finite quantised magnetic flux

$$\oint dx^i \ \partial_i\phi = \frac{1}{2\pi}\int d^2x \ B = Q_{\rm top} \in \mathbf{Z}$$

This is what monopole operators do in the Higgs phase: they create vortices. The upshot is that we can characterise the Higgs phase of the theory as

$$\begin{aligned} \text{Higgs Phase}: \quad \text{Spontaneously broken } U(1)_{\rm gauge} \quad &\Leftrightarrow \quad \text{Unbroken } U(1)_{\rm top} \\ &\Leftrightarrow \quad \text{Charged Excitation} = \text{Vortex} \end{aligned}$$

<u>$m^2 = 0$:</u> In $d = 2+1$, the two phases at $m^2 > 0$ and $m^2 < 0$ are clearly different since they have a different global symmetry $U(1)_{\rm top}$. (This is in contrast to the story in $d = 1+1$ where vortices are instantons and blur the distinction between the two phases.)

We can ask: what happens as we dial m^2 from positive to negative. We expect a phase transition to occur at some point, which we heuristically refer to as $m^2 = 0$. (In practice, this point can be shifted away from zero.) Is this a first order phase transition, or second order? If second order, what universality class does the theory lie in? Because the theory is strongly coupled in the regime $|m| \lesssim e^2$ it is difficult to perform any quantitative calculations to answer this question. Instead, we will guess.

To guide our guess, we use the symmetries of the problem. Since we have identified the global $U(1)_{\rm top}$ symmetry as distinguishing phases, it seems reasonable to postulate that the phase transition lies in the same universality class as other theories governed by a $U(1)$ global symmetry. This turns out to be true, and underlies a rather beautiful feature of 3d gauge theories known as particle-vortex duality.

8.2.1 Particle-Vortex Duality

In quantum field theories, there are very often two kinds of particle excitations that can appear. The first kind is the familiar excitation that we get when we quantise a local field. This is that kind that we learned about in our Quantum Field Theory course. The second kind we've seen a number of times in these lectures: they are solitons.

Despite the fact that these two kinds of particles arise in different ways, there is really little difference between them in the quantum theory. In particular, both are described as states in the Fock space. Typically at weak coupling, the solitons are much heavier than the "elementary particles", but that's more a limitation of our need to work at weak coupling. It may be – and often is – that as we move into strongly coupled regimes, the solitons become light.

This opens up an intriguing possibility. Is it possible to write down a different quantum field theory in which the roles of solitons and elementary particles are reversed? These two quantum field theories would describe the same physics, but what appears as a soliton in one would appear as an elementary particle in the other, and vice versa. This is referred to a *duality*.

In fact, we've already met a simple example of a duality in these lectures. In Section 7.5, we used bosonization to demonstrate the equivalence between a massive fermion and the Sine-Gordon model. The elementary fermion arises as a kink in the Sine-Gordon model.

Typically, dualities get harder to construct with any conviction as the number of dimensions increases. There are wonderful examples of dualities in $d = 3 + 1$, which exchange electric and magnetic excitations, but they need supersymmetry to keep control over the dynamics and so are beyond the scope of these lectures. However, things are somewhat easier in $d = 2 + 1$. Here we do have examples of dualities. In contrast to the bosonization story of Section 7.5, we are unable to prove the $d = 2 + 1$ dualities from first principles, but nonetheless have convincing evidence that they are true. We will see a number of these dualities as we proceed.

As we've seen above, in $d = 2+1$ dimensions the appropriate solitons are vortices. We will now propose a second theory, whose classical dynamics is different from the Abelian-Higgs model (8.12), but whose quantum dynamics is argued to be identical. The vortices in one theory are identified with the elementary particles of the other. For this reason, the claimed equivalence of the two theories is referred to as *particle-vortex duality*.

The XY-Model

The theory which is claimed to be dual to the 3d Abelian-Higgs model is simply a theory of a complex scalar field $\tilde{\phi}$, without any gauge field,

$$S_{XY} = \int d^3x \; |\partial_\mu \tilde{\phi}|^2 - \tilde{m}^2 |\tilde{\phi}|^2 - \frac{\tilde{\lambda}}{2} |\tilde{\phi}|^4 \tag{8.13}$$

This is known as the *XY-Model*. At first glance, the physics of this model is rather different from the Abelian-Higgs model. Indeed, it appears to have fewer degrees of freedom because it is missing the gauge field. Nonetheless, as we now explain, they describe the same physics, albeit in a non-obvious and interesting way.

Let's first address the issue of degrees of freedom. The XY-model clearly has two degrees of freedom in the UV where it is weakly coupled. But the Abelian-Higgs model has the same number: the gauge redundancy removes one degree of freedom from ϕ, but this is replenished by the single polarization state of the photon. We learn an interesting lesson: gauging a $U(1)$ symmetry in $d = 2+1$ changes the dynamics, but does not change the overall number of degrees of freedom. This will be important in later developments.

We can also match the symmetries between the XY-model and the Abelian-Higgs model. The XY-model clearly has a $U(1)$ global symmetry which rotates the phase of $\tilde{\phi}$. The associated current is

$$J^\mu_{XY} = i \left(\tilde{\phi}^\dagger \partial^\mu \tilde{\phi} - (\partial^\mu \tilde{\phi}^\dagger) \tilde{\phi} \right)$$

The Abelian-Higgs model also has a single global symmetry that we called $U(1)_{\text{top}}$. You might worry that the Abelian-Higgs model also has a gauge symmetry, which is clearly not shared by the XY-model. But, as we have stressed many times, gauge symmetries are not symmetries at all, but redundancies. This gives another important lesson: there is no need for gauge symmetries to match on both sides of a duality.

We can now look at how the physics of the XY-model changes as we vary the mass:

$\underline{\tilde{m}^2 > 0}$: This is a gapped phase. The $\tilde{\phi}$ excitations are massive and carry charge under the unbroken $U(1)$ global symmetry. We see that, at least with broad brush, this looks similar to the the Higgs phase of the Abelian-Higgs model, in which the $U(1)_{\rm top}$ symmetry was unbroken. In that case, the vortices carried charge under $U(1)_{\rm top}$.

$\underline{\tilde{m}^2 < 0}$: In this phase, $\tilde{\phi}$ gets a vacuum expectation value and the $U(1)$ global symmetry is broken. We can write $\tilde{\phi} = \rho e^{i\sigma}$. The fluctuations of ρ are massive, while the σ field is massless: it is the Goldstone mode for the broken $U(1)$. Notice that we've given this field the same name as the dual photon in the Abelian-Higgs model. This is not a coincidence.

Again, with broad brush this looks similar to the gapless Coulomb phase of the Abelian-Higgs model. However, the Coulomb phase was also characterised by the existence of massive, charged ϕ excitations that were logarithmically confined. Can we see similar excitations in the XY-model? The answer is yes.

The ordered phase of the XY-model also has vortices. As before, these arise from the phase of $\tilde{\phi}$ winding asymptotically, but now there is no gauge field to cancel the log divergence in their energy,

$$\int d^2x\; |\partial_i\tilde{\phi}|^2 = \int d\theta dr\; r\;\frac{1}{r^2}|\partial_\theta\tilde{\phi}|^2 + \ldots = 2\pi\int_0^\infty dr\;\frac{n^2}{r}|\tilde{\phi}|^2 + \ldots$$

The energy of a single vortex is logarithmically divergent. But this divergence can be cancelled by placing an anti-vortex at some distance r. It's not hard to convince yourself that the logarithm reappears in the potential energy between the vortex and anti-vortex, which scales as

$$V = \frac{1}{2\pi}\log\left(\frac{r}{r_0}\right)$$

for some cut-off r_0. In other words, the vortices are logarithmically confined. This, of course, is the same behaviour exhibited by charged particles in 3d electromagnetism.

$\underline{\tilde{m}^2 = 0}$: Lying between the two phases above is a critical point. Once again, we are being a little careless in describing this as sitting at $\tilde{m} = 0$; strictly, you should tune both $\tilde{m}$ and the other parameters to hit the critical point.

This time, the physics of the critical point is well understood: this is the XY Wilson-Fisher fixed point. We studied this in some detail in the lectures on Statistical Field Theory using the epsilon expansion.

The essence of particle-vortex duality is the claim that the Abelian-Higgs model also flows to the XY Wilson-Fisher fixed point at $m = 0$. This claim can be traced back to work of Peskin in the 1970s, but was brought to prominence by Dasgupta and Halperin in the early 1980s. Given the similarity in their phase structure, this would seem to be a reasonable claim. There is currently no proof of the duality, but there is now convincing numerical evidence that it is true.

The Duality Dictionary

The key to particle-vortex duality is really the idea of universality: the two theories (8.12) and (8.13) share the same critical point. We can then attempt to map the operators of the two theories at the critical point. We have only an incomplete dictionary at the moment, but our discussion above allows us to start to fill in some entries. For example, we have seen how the currents match on both sides

$$J^\mu_{\rm top} \quad \longleftrightarrow \quad J^\mu_{XY}$$

With two theories flowing to the same critical point, we can now turn on relevant operators in each. As long as we turn on the same relevant operator, we are guaranteed that the theories coincide in the neighbourhood of the fixed points. We have seen above how this plays out: when the scalar condenses in one theory, it matches the phase in which the scalar is not condensed in the other. Roughly speaking, we have

$$m^2 \approx -\tilde{m}^2$$

Alternatively, we can write this in terms of the relevant operators at the critical point as

$$|\phi|^2 \quad \longleftrightarrow \quad -|\tilde{\phi}|^2 \tag{8.14}$$

although since the critical points are strongly coupled, this relation is likely to have corrections, with operators on both sides mixing with others.

Far from the critical point, we have seen that the theories have the same qualitative features. In particular, the duality inherits its name from the map between massive excitations,

$$\begin{array}{rcl} \text{gauge vortex} & \longleftrightarrow & \tilde{\phi} \text{ excitation} \\ \phi \text{ excitation} & \longleftrightarrow & \text{global vortex} \end{array}$$

Only the first of these describes a map between finite energy excitations. In this case, it is better to phrase the map in terms of local operators, rather than solitons: the essence of particle-vortex duality is that the monopole operator on one side is a traditional field in the Lagrangian on the other,

$$\mathcal{M}(x) \quad \longleftrightarrow \quad \tilde{\phi}(x) \tag{8.15}$$

We could ask: do the interactions between these massive excitations agree in detail? The answer is most likely no. One could add irrelevant operators to both the Abelian-Higgs model and the XY model which will affect the interactions between these massive particles. We would have to work much harder to get quantitative agreement away from the critical point. For what it's worth, it is possible to do this matching in certain supersymmetric versions of the duality. Here, particle-vortex duality is referred to as *3d mirror symmetry*.

The View from Statistical Physics

The claim of particle-vortex duality offers a very clear experimental prediction. Although we have phrased our discussion in the context of physics in $d = 2+1$ dimensions, everything goes through in the the Euclidean $d = 3+0$ world. Here, the theories (8.12) and (8.13) can be viewed as statistical field theories, with the path integral describing thermal rather than quantum fluctuations. More details can be found in the lecture notes on Statistical Field Theory.

In this context, the 3d XY-model (8.13) governs the phase transition of a number of systems, including the superfluid transition of liquid helium. Similarly, the 3d Abelian-Higgs model (8.12) governs the superconducting phase transition, with the field strength F_{ij}, $i, j = 1, 2, 3$ describing the fluctuating magnetic field.

In both cases, the mass^2 term determines the deviation from the critical temperature T_c at which the phase transition occurs. But that makes the map (8.14) between the masses rather surprising. It means that the duality maps the high temperature phase of the superfluid to the low temperature phase of the superconductor, and vice versa.

The claim that both theories share a critical point then becomes the claim that the two phase transitions have the same critical exponents. Experimentally, however, this claim is incorrect: the two phase transitions are *not* the same. While the superfluid transition exhibits the XY Wilson-Fisher exponents, the superconducting transition has mean field exponents. It would seem that particle-vortex duality has been ruled out experimentally!

In fact this is too quick. Recall that the XY-model has two critical points. The mean field critical point is unstable, with $|\phi|^4$ a relevant operator that drives the theory to the Wilson-Fisher point. The same should be true of the Abelian-Higgs model. It is thought that the mean field exponents seen in the superconducting transition reflect the fact that the experiments haven't got close enough to the true critical point, and are instead probing the unstable mean field point. Calculations suggest that one would start to see Wilson-Fisher critical exponents in the superconducting transition only at $T - T_c \sim 10^{-9}$ K. Such a level of precision is not technologically feasible.

But this brings its own issues. It appears that we have a system in Nature which is fine-tuned. The natural scale of the superconducting phase transition is $T_c \sim 10$ K or so. In the experiments, we tune the coefficient of $|\phi|^2$ by hand to hit the critical temperature. But why is the coefficient of the $|\phi|^4$ relevant operator so small that it only shows up when $T - T_c \sim 10^{-9}$ K? This is similar to the famous hierarchy problem in the Standard Model, where again the coefficient of a relevant operator appears to be fine-tuned.

Particle physicists have sleepless nights over fine tuning, and desperately search for an explanation. In large part, this is because of experience with RG in statistical physics, where any fine-tuning seen in Nature must also have an explanation. In the case of superconductors, the apparent fine tuning is understood: it arises because the underlying scalar field ϕ is not fundamental, but instead comprises of a Cooper pair of electrons. (The analogous possibility for the Higgs fine tuning goes by the name of technicolour.) A full explanation would take us too far from the purpose of these lectures, but this suffices to ensure that the smallness of the $|\phi|^4$ relevant operator seen in the superconducting transition is technically natural.

8.3 Confinement in $d = 2 + 1$ Electromagnetism

We've seen that classical electromagnetism in $d = 2 + 1$ dimensions confines particles, but only weakly with a log potential

$$V(r) = \frac{Q^2}{2\pi} \log\left(\frac{r}{r_0}\right)$$

There is, however, an important effect in the quantum theory that turns the logarithmic confining potential into a more powerful linearly confining potential. This effect, first discovered by Polyakov, is due to instantons.

We've met instantons in $d = 3 + 1$ Yang-Mills theory in Section 2.3, and again in the $d = 1 + 1$ Abelian-Higgs model in Section 7.2. In the latter case, vortices play the role of instantons. Now that we are living in $d = 2+1$ dimensions, the instantons should be objects localised in three Euclidean dimensions. But these are very familiar: they are magnetic monopoles.

We've already introduced the idea of monopole operators in Section 8.1. These can be thought of as Dirac monopoles at a point. They are not quite what we want for the present purposes. As a starting point for a semi-classical calculation, we would like the monopoles to be smooth configurations with finite action. But we've seen such objects before: we can use the 't Hooft-Polyakov monopole described in Section 2.8.

Recall that the 't Hooft-Polyakov monopoles arise in an $SU(2)$ gauge theory (or, more generally, any non-Abelian gauge theory) broken down to its Cartan subalgebra. To achieve this, we couple the $SU(2)$ gauge theory to a real, adjoint scalar ϕ and work with the action

$$S = \int d^3x \ -\frac{1}{2g^2}\mathrm{tr}\, F_{\mu\nu}F^{\mu\nu} + \frac{1}{g^2}\mathrm{tr}\,(\mathcal{D}_\mu\phi)^2 - \frac{\lambda}{4}\left(\mathrm{tr}\,\phi^2 - \frac{v^2}{2}\right)^2 \tag{8.16}$$

The ground state of the system has, up to a gauge transformation, $\phi = v\sigma^3$, and breaks the gauge symmetry

$$SU(2) \to U(1)$$

At low energies, the spectrum contains just a single massless photon and looks like pure electromagnetism. In addition, there is a neutral scalar with mass $\sim \sqrt{\lambda}gv$ and a charged W-boson of mass $\sim v$.

In this way, we can view the model as $U(1)$ gauge theory, with a UV cut-off at the scale v. The dimensionless gauge coupling constant is g^2/v and to trust any semi-classical calculation, we must take $g^2/v \ll 1$.

8.3.1 Monopoles as Instantons

Our main reason for introducing the action (8.16) is that, in Euclidean spacetime, it admits smooth monopole solutions. These are the 't Hooft-Polyakov monopoles that we introduced in Section 2.8, but now localised in Euclidean spacetime meaning that they play the role of instantons, rather than particles. Here we recount the basics.

The existence of the monopoles can be traced to topology. Any finite action configuration must obey $\mathrm{tr}\,\phi^2 \to v^2/2$ as $x \to \infty$. This defines a sphere $\mathbf{S}^2$ in field space, so all finite action configurations are classified by a winding number $\Pi_2(\mathbf{S}^2) = \mathbf{Z}$, defined as

$$n = \frac{1}{8\pi v^3}\int_{\mathbf{S}^2_\infty} d^2S_i\ \epsilon^{ijk}\epsilon_{abc}\phi^a\partial_j\phi^b\partial_k\phi^c \in \mathbf{Z} \tag{8.17}$$

However, winding comes at a cost. Any purely scalar configuration that winds has linearly divergent action. This can be compensated by turning on a gauge field and this, in turn, endows the soliton with magnetic charge in the unbroken $U(1) \subset SU(2)$, (2.91),

$$m = -\frac{1}{v}\int d^2S_i\ \frac{1}{2}\epsilon^{ijk}\,\mathrm{tr}\,(F_{jk}\phi) = 4\pi n$$

The solution for a single monopole, with winding $n = 1$, has asymptotic form

$$\phi^a \to v\frac{x^a}{r} \quad \text{and} \quad A_i^a \to -\epsilon_{aij}\frac{x^j}{r^2} \quad \text{as } x \to \infty$$

The action of this configuration is finite, and given by

$$S_{\text{mono}} = \frac{8\pi v}{g^2}\, f(\lambda g^2)$$

with $f(\lambda g^2)$ a monotonically increasing function. It has the property that $f(0) = 1$, so that the action above coincides with that of a BPS monopole (2.93) when $\lambda = 0$.

We're used to the idea that finite action configurations in Euclidean space tunnel between different vacua of the theory. But what vacua does the monopole tunnel between? Clearly, it changes the magnetic flux $\Phi = \int d^2x\ B$ on a spatial slice. If we were living on a compact space, this would change the energy of a state, which is given by

$$\Delta E = \int d^2x\ \frac{1}{2}B^2 \sim \frac{1}{2}\,\text{Area}\left(\frac{\Phi}{\text{Area}}\right)^2$$

with "Area" the area of a spatial slice. However, as the area tends to infinity, the flux is suitably diluted and the cost in energy is vanishingly small. These are the different vacua that the monopoles tunnel between.

A Dilute Gas of Monopoles and Anti-Monopoles

With our monopole solution in hand, we can use it as the starting point for a semi-classical evaluation of the path integral. We should be getting used to this by now, and we follow the structure of the calculation laid out in Section 2.3, and again in Section 7.2.

One key step in the calculation is to invoke the use of a dilute gas of instantons. In the present case, this means we treat configurations of widely separated monopoles and anti-monopoles, with magnetic charges $m_i = \pm 4\pi$, as saddle points in the path integral. In the previous situations, we argued that the action of a dilute gas of N (anti)-instantons was roughly $S \approx NS_{\rm inst}$, reflecting the fact that these are approximate solutions when the objects are far separated.

For monopoles, however, we should treat this step more carefully. Viewed as particles in $d = 3+1$ dimensions, we know that the energy will pick up contributions from the long range Coulomb forces between the monopoles. This translates into a contribution to the action in our context. If a monopole of charge $m_i = \pm 4\pi$ sits at position X_i, the total action will be

$$S = S_{\rm mono} \sum_i \left(\frac{m_i}{4\pi}\right)^2 + \frac{1}{4\pi g^2} \sum_{i<j} \frac{m_i m_j}{|X_i - X_j|}$$

where the second term reflects the long range Coulomb interaction.

We evaluate the path integral by summing over these dilute gas configurations, containing N constituents of either type. This results in the expression,

$$Z = \sum_{N=0}^{\infty} \sum_{m_i = \pm 4\pi} \frac{1}{N!} \left(Ke^{-S_{\rm mono}}\right)^N \int \prod_{i=1}^N d^3X_i \ \exp\left(-\frac{1}{8\pi g^2} \sum_{i\neq j} \frac{m_i m_j}{|X^i - X^j|}\right) \quad (8.18)$$

Here K is the usual contribution from one-loop determinants and Jacobian factors. We could compute it, but it does not give any qualitatively new insights into the physics so we will not. The second factor in the expression above is the novelty. When the instantons are non-interacting, this just gives a power of V^N to the path integral, with V the spacetime volume. Now that we have long range interactions between the instantons, we must work a little harder.

There is a useful way to rewrite the final expression. We use the fact that the $1/r$ factor also arises in the Green's function of the Laplacian in three dimensions. In general, for a scalar field $\sigma(x)$, and any fixed function $f(x)$, we have

$$\int \mathcal{D}\sigma \ \exp\left(-\int d^3x \ \frac{1}{2}(\partial_\mu \sigma)^2 + f(x)\sigma(x)\right) \sim \exp\left(\frac{1}{8\pi}\int d^3x d^3y \ \frac{f(x)f(y)}{|x-y|}\right)$$

Using this, we rewrite the sum over the Coulomb gas in (8.18) as a path integral

$$\exp\left(-\frac{1}{8\pi g^2}\sum_{i\neq j}\frac{m_i m_j}{|X^i - X^j|}\right) = \int \mathcal{D}\sigma \exp\left(-\int d^3x \ \frac{g^2}{8\pi^2}(\partial_\mu\sigma)^2 + \frac{i}{2\pi}\sum_i m_i \sigma\, \delta^3(x - X_i)\right)$$

(We used a very similar trick in the lectures on Statistical Field Theory when treating the 2d Coulomb gas in the XY model.)

In fact, we've met this field $\sigma(x)$ before: it is precisely the dual photon that we introduced in Section 8.1. To see this, note that the coupling to the magnetic charge above coincides with the coupling in (8.7).

Continuing with our calculation, the partition function becomes

$$\begin{aligned}
Z = \int \mathcal{D}\sigma\ \exp\left(-\int d^3x\ \frac{g^2}{8\pi^2}(\partial_\mu\sigma)^2\right)\ &\sum_{N=0}^{\infty}\frac{(Ke^{-S_{\rm mono}})^N}{N!} \\
&\times \int\prod_{i=1}^{N} d^3X_i \sum_{m_i=\pm 4\pi} e^{-\frac{i}{2\pi}\sum_i m_i\sigma(X_i)} \\
= \int \mathcal{D}\sigma\ \exp\left(-\int d^3x\ \frac{g^2}{8\pi^2}(\partial_\mu\sigma)^2\right)\ &\sum_{N=0}^{\infty}\frac{1}{N!}\left(Ke^{-S_{\rm mono}}\int d^3x\cos(2\sigma(x))\right)^N \\
= \int \mathcal{D}\sigma\ \exp\left(-\int d^3x\ \frac{g^2}{8\pi^2}(\partial_\mu\sigma)^2 - Ke^{-S_{\rm mono}}\cos(2\sigma)\right) & \qquad (8.19)
\end{aligned}$$

We can now see the net effect of the instantons: they have generated a potential for the dual photon σ. Expanding about the minimum at $\sigma = 0$, we find that the dual photon has acquired a mass,

$$m^2_{\rm photon} = \frac{16\pi^2 Ke^{-S_{\rm mono}}}{g^2}$$

On dimensional grounds, the one-loop determinants and Jacobian factors that we lumped into the constant K must have dimension $[K] = 3$. For small λ, it turns out to scale as $K \sim v^{7/2}/g$. At weak coupling $g^2/v \ll 1$ and $S_{\rm mono} \gg 1$, where our semi-classical analysis is valid, we find that the mass of the dual photon is exponentially smaller than all other scales in the game. This means that we can read off the effective action from (8.19)

$$S_{\rm eff} = \int d^3x\ \frac{g^2}{8\pi^2}(\partial_\mu\sigma)^2 + Ke^{-S_{\rm mono}}\cos(2\sigma) \qquad (8.20)$$

We recognise this as the Sine-Gordon model that we met in $d = 1+1$ dimensions in Section 7.5.5. Now it arises as the effective, low-energy description of a gauge theory in $d = 2+1$ dimensions.

8.3.2 Confinement

What does it mean for the dual photon to get a mass? To answer this, we can see how the ground state responds to various provocations.

First, let's try to turn on an electric field in the ground state, say $F_{01} \neq 0$. To understand what this means in terms of the dual photon, we need to relate $F_{\mu\nu}$ with σ. We can do this by comparing our expressions for the topological current (8.3) and (8.11),

$$J^\mu_{\rm top} = \frac{1}{4\pi}\epsilon^{\mu\nu\rho}F_{\nu\rho} = \frac{e^2}{(2\pi)^2}\partial^\mu\sigma$$

We find that an electric field corresponds to

$$F_{01} = \frac{e^2}{2\pi}\,\partial_2\sigma$$

However, the configuration $\partial_2\sigma =$ constant does not obey the equations of motion of our effective action (8.20). This means that the vacuum does not support a constant, background electric field. Instead, solutions to the equations of motion with $\partial_2\sigma \neq 0$ are kinks, or domain walls, in which σ interpolates from, say, $\sigma = 0$ as $x_2 \to -\infty$, to $\sigma = \pi$ as $x_2 \to +\infty$. We already met these kinks in Section 7.5.5 when discussing the Sine-Gordon model in $d = 1+1$ dimensions. In the present context, the domain walls are string-like configurations stretched in the x^1 direction, with width $\sim 1/m_{\rm photon}$ in the x^2 direction, and tension,

$$\gamma = \frac{4}{\pi}\sqrt{2Kg^2e^{-S_{\rm mono}}}$$

a result which follows from translating our earlier result (7.69). (Up until now, we've always referred to the string tension as σ. Obviously that's a bad choice for our current discussion.)

The domain wall, or string, is a collimated flux tube of electric field $F_{01} \neq 0$. This is the expected behaviour of a gauge theory that is linearly confining. In other words, the classical log potential (8.2) of 3d gauge theories has been replaced with a more severe,

$$V(r) = \gamma r$$

We could explicitly compute the Wilson loop in this framework and confirm that it does indeed exhibit an area law.

We have seen that 3d electromagnetism exhibits linear confinement due to instantons which, in this context, are monopoles. It is crucial that these monopoles have a finite action, which we achieved by embedding the theory in a non-Abelian gauge group. If we introduce other UV completions of the theory, with a finite cut-off, Λ_{UV}, these too will have monopoles, typically with action $S_{\rm mono} \sim \Lambda_{UV}/g^2$. (Lattice gauge theory provides a good example of this.) These too will then exhibit linear confinement.

8.4 Chern-Simons Theory

Gauge theories in $d = 2 + 1$ dimensions admit a rather special interaction that does not have a counterpart in even spacetime dimensions. This is the famous *Chern-Simons* interaction. It plays a key role in many areas of theoretical and mathematical physics, from the physics of the quantum Hall effect, to the mathematics of the knot invariants. Many details on the former application can be found in the lecture notes on the Quantum Hall Effect.

For $U(1)$ gauge theory, the Chern-Simons term takes the form

$$S_{CS} = \frac{k}{4\pi} \int d^3x \; \epsilon^{\mu\nu\rho} A_\mu \partial_\nu A_\rho \tag{8.21}$$

We could consider this term on its own, or in conjunction with the Maxwell action (8.1). In either case, the dimensionless coefficient k is known as the *level*. We can write down similar terms in any odd spacetime dimension; we briefly met the $d = 4+1$ dimensional version in Section 4.4.1.

Let's start by studying the symmetries of the Chern-Simons action. It is Lorentz invariant, courtesy of the $\epsilon^{\mu\nu\rho}$ invariant tensor. At an operational level, the existence of this tensor means that the term is exclusive to $d = 2 + 1$ dimensions. However, this same $\epsilon^{\mu\nu\rho}$ tensor means that the Chern-Simons interaction breaks both parity and time-reversal invariance. Here we focus on parity. In even dimensions we can always take parity to act as $\mathbf{x} \mapsto -\mathbf{x}$ (see, for example, (1.25)). But, in odd dimensions, this coincides with a rotation. We should instead take parity to flip the sign of just a single spatial coordinate,

$$x^0 \to x^0 \quad , \quad x^1 \to -x^1 \quad , \quad x^2 \to x^2 \tag{8.22}$$

and, correspondingly, $A_0 \to A_0$, $A_1 \to -A_1$ and $A_2 \to A_2$. This means that, as advertised, the Chern-Simons action is odd under parity.

8.4.1 Quantisation of the Chern-Simons level

At first glance, it's not obvious that the Chern-Simons term is gauge invariant since it depends explicitly on A_μ. However, under a gauge transformation, $A_\mu \to A_\mu + \partial_\mu \omega$, we have

$$S_{CS} \to S_{CS} + \frac{k}{4\pi} \int d^3x \; \partial_\mu \left(\omega \epsilon^{\mu\nu\rho} \partial_\nu A_\rho \right)$$

The change is a total derivative. In many situations we can simply throw this total derivative away and the Chern-Simons term is gauge invariant. However, there are some situations where the total derivative does not vanish. As we will now show, in these cases the Chern-Simons partition function is gauge invariant provided that

$$k \in \mathbf{Z} \tag{8.23}$$

For Abelian Chern-Simons theories, it's a little subtle to see the requirement (8.23) since it only shows up in the presence of magnetic flux. (This is to be contrasted with the situation for non-Abelian Chern-Simons theories described in Section 8.4.3 where one can see the analogous quantisation condition around the vacuum state.)

Perhaps the simplest way is to consider the theory on Euclidean spacetime $\mathbf{S}^1 \times \mathbf{S}^2$. We then add a single unit of magnetic flux through the $\mathbf{S}^2$. As we've seen many times in these lectures, if we take the gauge group to be compact $U(1)$, the flux is quantised, in the minimal unit

$$\frac{1}{2\pi}\int_{\mathbf{S}^2} F_{12} = 1 \tag{8.24}$$

We then consider large gauge transformations of this background that wind around the $\mathbf{S}^1$. We denote the radius of this $\mathbf{S}^1$ as R, and parameterise it by the coordinate $x^0 \in [0, 2\pi R)$. Consider a gauge transformation $A_\mu \to A_\mu + \partial_\mu \omega$ which winds around the $\mathbf{S}^1$, with

$$\omega = \frac{x^0}{R} \tag{8.25}$$

Under such a transformation, any matter field ϕ with charge $q \in \mathbf{Z}$ remains single valued, since $\phi \to e^{iq\tau/R}\phi$. Even in the absence of charged matter, the statement that we're working with a compact $U(1)$ gauge group, rather than a non-compact $\mathbf{R}$ gauge group, means that the theory admits fluxes (8.24) and gauge transformations (8.25).

Under the gauge transformation (8.25), we have

$$A_0 \to A_0 + \frac{1}{R} \tag{8.26}$$

This means that the zero mode of A_0 is a periodic variable, with periodicity $1/R$. (We came to the same conclusion in Section 7.1 where we discussed two dimensional electromagnetism on a spatial circle.)

We can now see what becomes of our Chern-Simons action under such a gauge transformation. Evaluated on a configuration with constant A_0, we have

$$S_{CS} = \frac{k}{4\pi}\int d^3x \; A_0 F_{12} + A_1 F_{20} + A_2 F_{01}$$

Now it's tempting to throw away the last two terms when evaluating this on our background. But we should be careful as it's a topologically non-trivial configuration. We can safely set all terms with ∂_0 to zero, but integrating by parts on the spatial derivatives we get an extra factor of 2,

$$S_{CS} = \frac{k}{2\pi} \int d^3x \; A_0 F_{12} \tag{8.27}$$

Evaluated on the flux (8.24), with constant $A_0 = a$, we have

$$S_{CS} = 2\pi k R a$$

And under the gauge transformation (8.26), we have

$$S_{CS} \to S_{CS} + 2\pi k$$

The Chern-Simons action is not gauge invariant. But all is not lost. The partition function depends only on $e^{iS_{CS}}$ and this remains gauge invariant provided $k \in \mathbf{Z}$, which is our claimed result. This last part of the argument is exactly the same as the one we met in Section 2.1.3 when we discussed Chern-Simons terms in quantum mechanics, and in a number of other places when we've discussed WZW terms.

Chern-Simons Theories and Spinors

There are further subtleties associated to the factor of 2 above, which we flag up here. A better way to think about the Chern-Simons theory on a 3-dimensional manifold M, is by viewing this as the boundary of 4-dimensional manifold X. The story is simplest in the language of forms, where we have

$$S_{CS}[A;X] = \frac{k}{4\pi} \int_{M=\partial X} A \wedge dA = \frac{k}{4\pi} \int_X F \wedge F$$

The fact that the Chern-Simons term is related to the 4-dimensional θ term was anticipated in (1.12). Written in this way, the Chern-Simons term is clearly gauge invariant since it depends only on F and not A. Our worry, however, has transmuted to the question of whether it depends on the choice of 4-manifold X. How can we be sure that we get the same answer if we chose a different 4-manifold X' which also has boundary $\partial X' = M$? The difference between the two answers involves the integral over the compact manifold $Y = X \cup X'$, formed by gluing together X and X' along their common boundary,

$$S_{CS}[A;X] - S_{CS}[A;X'] = \frac{k}{4\pi} \int_Y F \wedge F$$

We're safe provided that this difference is 2π times an integer, since then the partition function, which depends on $e^{iS_{CS}}$, is independent of the choice of X. Clearly this requires

$$\frac{1}{2}\int_Y \frac{F}{2\pi}\wedge\frac{F}{2\pi}\in\mathbf{Z} \tag{8.28}$$

So is this true? Well, actually no. Or, at least, not always! It turns out that (8.28) is true only if the 4-manifold Y admits spinors or, more precisely, admits a mathematical object called a *spin structure* which tells you whether or not a fermion picks up a minus sign when it is transported around a loop. Any manifold that admits such a spin structure is called a *spin manifold*. And (8.28) holds whenever Y is a spin manifold.

For example, $Y=\mathbf{T}^4$, $Y=\mathbf{S}^2\times\mathbf{S}^2$ and $Y=\mathbf{S}^4$ are all spin manifolds. In these cases (8.28) holds. To give you some sense of how this works, suppose that we take $Y=\mathbf{S}^2\times\mathbf{S}^2$. Dirac quantisation means that the flux through each of the spheres must be a multiple of 2π. If we take $F=F_1+F_2$, with F_n giving flux through the n^{th} 2-sphere, then

$$\frac{1}{2}\int_{\mathbf{S}^2\times\mathbf{S}^2}\frac{F}{2\pi}\wedge\frac{F}{2\pi}=\int_{\mathbf{S}^2}F_1\int_{\mathbf{S}^2}F_2\in\mathbf{Z}$$

with the factor of 2 coming from the cross-term.

However, there are 4-manifolds Y which do not admit a spin structure. The simplest example is $Y=\mathbf{CP}^2$. In this case, $\int_Y(F/2\pi)\wedge(F/2\pi)$ is an integer, not an even integer.

The upshot of this is that the Chern-Simons level k for a $U(1)$ gauge group can be integer valued provided that the theory admits fermions. But, otherwise, must be an even integer. The simple "integrate by parts to get an extra factor of 2" prescription that we used to get (8.27) sweeps all of these subtleties under the rug.

8.4.2 A Topological Phase of Matter

So what is the physics of Chern-Simons theory? Despite the simplicity of the action, the physics is remarkably subtle. Let's start with the basics. We'll take the $d=2+1$ dimensional gauge field to be governed by

$$S=S_{\text{Maxwell}}+S_{CS}=\int d^3x\;-\frac{1}{4e^2}F_{\mu\nu}F^{\mu\nu}+\frac{k}{4\pi}\epsilon^{\mu\nu\rho}A_\mu\partial_\nu A_\rho$$

We can start by gaining some intuition from the classical equation of motion,

$$\partial_\mu F^{\mu\nu}+\frac{ke^2}{4\pi}\epsilon^{\nu\rho\sigma}F_{\rho\sigma}=0 \tag{8.29}$$

In terms of the electric field $E_i = F_{0i}$ and the magnetic field $B = F_{12}$, Gauss' law becomes

$$\partial_i E_i = \frac{ke^2}{2\pi} B \tag{8.30}$$

which tells us that a magnetic field acts as a source for the electric field. This simple observation will underlie much of the physics of Section 8.6 where we discuss bosonization in 3d.

What are the propagating excitations of the equations of motion (8.29)? Taking one further derivative of the equations of motion, we can decouple electric and magnetic fields to show that each component obeys the massive wave equation,

$$\partial^2 E_i - \left(\frac{ke^2}{2\pi}\right)^2 E_i = \partial^2 B - \left(\frac{ke^2}{2\pi}\right)^2 B = 0$$

(To do this, it's perhaps simplest to first define the field $G^\mu = \epsilon^{\mu\nu\rho} F_{\nu\rho}$ and show that G^μ obeys the massive wave equation.) We see that, at least classically, the excitations do not propagate at the speed of light. Instead, they are exponentially damped. In the quantum theory, this means that we have a theory of massive excitations. The mass of the photon is

$$m_{CS} = \frac{ke^2}{2\pi}$$

Yet again, we find ourselves in a situation with a massive gauge boson. How should we think of this phase?

We've already met other situations in $d = 2+1$ dimensions where the photon gets a mass. There is the confining phase, driven by instantons, that we saw in Section 8.3, in which the Wilson loop has an area law. And there is, of course, the Higgs phase in which a charged scalar field condenses and the Wilson line has a perimeter law. It turns out that the Chern-Simons phase differs from both of these. Instead, it is a novel phase of matter, referred to as a *topological phase*.

Topological phases of matter are subtle. They typically have interesting things going on at energies $E \ll m_{CS}$ way below the gap, even though there are no physical excitations beyond the vacuum. We'll explain below what these interesting things are.

Chern-Simons Terms are Topological

Before we address the novel physics of Chern-Simons theory, we first point out an important property of the Chern-Simons action (8.21): it doesn't depend on the metric of the background spacetime manifold. It depends only on the topology of the manifold. To see this, let's first look at the Maxwell action for comparison. If we were to couple this to a background metric $g_{\mu\nu}$, the action becomes

$$S_{\text{Maxwell}} = \int d^3x\sqrt{-g}\ \ -\frac{1}{4e^2}\, g^{\mu\rho}g^{\nu\sigma}F_{\mu\nu}F_{\rho\sigma}$$

We see that the metric plays two roles: first, it is needed to raise the indices when contracting $f_{\mu\nu}f^{\mu\nu}$; second it provides a measure $\sqrt{-g}$ (the volume form) which allows us to integrate in a diffeomorphism invariant way. Recall from our first lectures on *Quantum Field Theory* that this allows us to quickly construct the stress-tensor of the theory by differentiating with respect to the metric,

$$T^{\mu\nu} = \frac{2}{\sqrt{-g}}\,\frac{\partial\mathcal{L}}{\partial g_{\mu\nu}}\bigg|_{g_{\mu\nu}=\eta_{\mu\nu}}$$

In contrast, we have no need to introduce a metric when generalising (8.21) to curved spacetime. This is best stated in the language of differential geometry: $A\wedge dA$ is a 3-form, and we can quite happily integrate this over any three-dimensional manifold

$$S_{CS} = \frac{k}{4\pi}\int A\wedge dA$$

This means that pure Chern-Simons theory knows nothing about length scales. In particular, the Wilson loop can exhibit neither area nor perimeter law, since both of these are statements about lengths. Moreover, pure Chern-Simons theory has vanishing stress tensor.

Chern-Simons Theory on a Torus

If Chern-Simons theory has vanishing stress tensor, and no physical excitations, then what can it possibly do? The answer is that the theory responds to low-energy probes in interesting ways.

Here is a simple, yet dramatic way to probe the theory. We will place it on a spatial 2-dimensional manifold Σ. As we have seen, Chern-Simons theory knows nothing about the metric on Σ. However, as we now show, it does know about the topology and responds accordingly.

For pure Chern-Simons theory (or, equivalently, the $e^2\to\infty$ limit of Maxwell-Chern-Simons theory), Gauss' law (8.30) becomes

$$F_{12} = 0$$

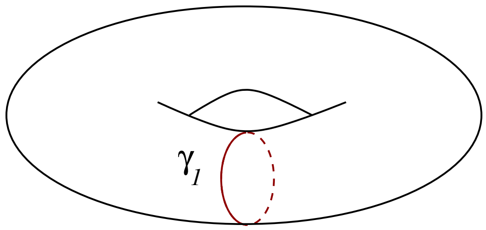


Figure 54:

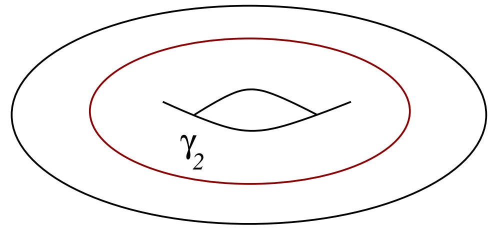


Figure 55:

Although this equation is very simple, it can still have interesting solutions if the background has some non-trivial topology. These are called, for obvious reason, *flat connections*. It's simple to see that such solutions exist on the torus $\Sigma = \mathbf{T}^2$, where one example is to simply set each A_i to be constant. Our first task is to find a gauge-invariant way to parameterise this space of solutions.

We'll denote the radii of the two circles of the torus $\mathbf{T}^2 = \mathbf{S}^1 \times \mathbf{S}^1$ as R_1 and R_2. We'll denote two corresponding non-contractible curves shown in the figure as γ_1 and γ_2. The simplest way to build a gauge invariant object from a gauge connection is to integrate

$$w_i = \oint_{\gamma_i} dx^j \, A_j$$

This is invariant under most gauge transformations, but not those that wind around the circle. By the same kind of arguments that led us to (8.26), we can always construct gauge transformations which shift $A_j \rightarrow A_j + 1/R_j$, and hence $w_i \rightarrow w_i + 2\pi$. The correct gauge invariant objects to parameterise the solutions are therefore the Wilson loops

$$W_i = \exp\left(i \oint_{\gamma_i} A_j dx^j\right) = e^{iw_i}$$

Because the Chern-Simons theory is first order in time derivatives, these Wilson loops are really parameterising the phase space of solutions, rather than the configuration space. Moreover, because the Wilson loops are complex numbers of unit modulus, the phase space is compact. On general grounds, we expect that when we quantise a compact phase space, we get a finite-dimensional Hilbert space. (We met an example of this in Section 2.1.3 when first describing Wilson lines.) Our next task is to understand how to quantise the space of flat connections.

The canonical commutation relations can be read off from the Chern-Simons action (8.21)

$$[A_1(x), A_2(x')] = \frac{i}{k}\,\delta^2(x - x') \quad \Rightarrow \quad [w_1, w_2] = \frac{2\pi i}{k}$$

The algebraic relation obeyed by the Wilson loops then follows from the usual Baker-Campbell-Hausdorff formula,

$$e^{iw_1}e^{iw_2} = e^{[w_1,w_2]/2}e^{i(w_1+w_2)}$$

which tells us that

$$W_1W_2 = e^{2\pi i/k}\,W_2W_1 \tag{8.31}$$

But such an algebra of operators can't be realised on a single vacuum state. This immediately tells us that the ground state must be degenerate. The smallest representation of (8.31) has dimension k, with the action

$$W_1|n\rangle = e^{2\pi ni/k}|n\rangle \quad \text{and} \quad W_2|n\rangle = |n+1\rangle$$

We have seen that on a torus $\Sigma = \mathbf{T}^2$, an Abelian Chern-Simons theory has k degenerate ground states. The generalisation of this argument to a genus-g Riemann surface tells us that the ground state must have degeneracy k^g. Notice that we don't have to say anything about the shape or sizes of these manifolds. The number of ground states depends only on the topology. This is an example of *topological order*.

8.4.3 Non-Abelian Chern-Simons Theories

We've not had much to say about non-Abelian gauge theories in low dimensions. This is not because they're boring, but simply because there is enough to keep us busy elsewhere. Here we make an exception and give a brief description of non-Abelian Chern-Simons theory.

Like Yang-Mills, Chern-Simons is based on a Lie algebra valued gauge connection A_μ. The non-Abelian Chern-Simons action is

$$S_{CS} = \frac{k}{4\pi}\int d^3x\ \epsilon^{\mu\nu\rho}\,\mathrm{tr}\left(A_\mu\partial_\nu A_\rho - \frac{2i}{3}A_\mu A_\nu A_\rho\right) \tag{8.32}$$

We've met this term before: the theta term in $d = 3+1$ dimensions can be written as a derivative of the Chern-Simons term (2.24). (It also arose in the same context when discussing canonical quantisation of Yang-Mills (2.35).) Chern-Simons theories with gauge group G and level k are sometimes denoted as G_k.

Once again, we will find that the level must be integer, $k \in \mathbf{Z}$. This time, however, the computation is more direct than in the Abelian case. Under a gauge transformation, we have

$$A_\mu \to \Omega^{-1} A_\mu \Omega + i\Omega^{-1}\partial_\mu \Omega$$

with $\Omega \in G$. The field strength transforms as $F_{\mu\nu} \to \Omega^{-1} F_{\mu\nu} \Omega$. A simple calculation shows that the Chern-Simons action changes as

$$S_{CS} \to S_{CS} + \frac{k}{4\pi}\int d^3x \left\{ i\epsilon^{\mu\nu\rho}\partial_\nu \mathrm{tr}\left(\partial_\mu\Omega\,\Omega^{-1}A_\rho\right) + \frac{1}{3}\epsilon^{\mu\nu\rho}\mathrm{tr}\,\left((\Omega^{-1}\partial_\mu\Omega)(\Omega^{-1}\partial_\nu\Omega)(\Omega^{-1}\partial_\rho\Omega)\right)\right\}$$

The first term is a total derivative. The same kind of term arose in Abelian Chern-Simons theories. However, the second term is novel to non-Abelian gauge theories, and this is where the quantisation requirement now comes from. In fact, we have seen this calculation before in Section 2.2.2 when discussing the theta angle in $d = 3 + 1$ Yang-Mills. On a spacetime manifold $\mathbf{S}^3$ (or on $\mathbf{R}^3$ with the requirement that gauge transformations asymptote to the same value at infinity), gauge transformations are characterised by the homotopy group $\Pi_3(SU(N)) \cong \mathbf{Z}$. The winding is counted by the function

$$n(\Omega) = \frac{1}{24\pi^2}\int_{\mathbf{S}^3} d^3S\ \epsilon^{\mu\nu\rho}\mathrm{tr}\left(\Omega^{-1}\partial_\mu\Omega\,\Omega^{-1}\partial_\nu\Omega\,\Omega^{-1}\partial_\rho\Omega\right) \quad \in \mathbf{Z} \tag{8.33}$$

We recognise this as the final term that appears in the variation of the Chern-Simons action. This means that the Chern-Simons action is not invariant under these large gauge transformations; it changes as

$$S_{CS} \to S_{CS} + \frac{k}{12\pi}\,24\pi^2\,n(\Omega) = S_{CS} + 2\pi k\,n(\Omega)$$

Insisting that the path integral, with its weighing $e^{iS_{CS}}$ is gauge invariant then gives us immediately our quantisation condition $k \in \mathbf{Z}$.

Wilson Loops

We have so far avoided talking about Wilson lines in Chern-Simons theories. There is rather a lot to say. We will not describe this in detail here, but just sketch the key idea.

In $d = 3$ Euclidean spacetime dimensions, a Wilson loop can get tangled. Mathematicians call closed curves in three dimensions *knots*, and there has been a great deal of effort in trying to classify the ways in which they can get tangled. It turns out that

Chern-Simons theories provide one of the most powerful tools. For a given knot C, we can compute the Wilson loop $\langle W[C]\rangle$. In Chern-Simons theory the Wilson loop exhibits neither an area law, nor a perimeter law. Instead, it depends on the details of the topology of the knot C. For each gauge group G, the Wilson loop gives a topological invariant which is a polynomial (roughly in $q = e^{2\pi i/k}$.) In simple cases, these topological invariants coincide with ones already understood by mathematicians (such as the Jones polynomial), but they also offer a large number of generalisations. Edward Witten was awarded the Fields medal, in large part for understanding this connection.

8.5 Fermions and Chern-Simons Terms

There is an intricate interplay between fermions in $d = 2 + 1$ dimensions and Chern-Simons terms.

In signature $\eta^{\mu\nu} = \text{diag}(+1, -1, -1)$, the Clifford algebra $\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu}$ is satisfied by the 2×2 gamma matrices,

$$\gamma^0 = \sigma^2 \quad , \quad \gamma^1 = i\sigma^1 \quad , \quad \gamma^3 = i\sigma^3$$

The Dirac spinor is then a two-component complex object. In odd spacetime dimensions, there is no "γ^5" matrix and, correspondingly, no Weyl fermions. In $d = 2+1$, we can take the gamma matrices as above to be purely imaginary, which means that we can have Majorana fermions. However, we won't have a need for this real representation in what follows.

It will prove useful to understand the action of parity on fermions. As we saw in (8.22), in three dimensions parity acts as

$$x^0 \to x^0 \quad , \quad x^1 \to -x^1 \quad , \quad x^2 \to x^2$$

The Dirac action is then invariant if we take parity to act as

$$\mathcal{P} : \ \psi \mapsto \gamma^1 \psi \tag{8.34}$$

But this means that the fermion mass term necessarily breaks parity,

$$\mathcal{P} : \bar{\psi}\psi = \psi^\dagger \gamma^0 \psi \mapsto -\bar{\psi}\psi$$

where, to see this, you need to remember that $(\gamma^1)^\dagger = -\gamma^1$ and $(\gamma^1)^2 = -1$.

This is different from what happens in $d = 3 + 1$ dimensions or, indeed, in any even spacetime dimension. There parity flips the sign of all spatial dimensions and, correspondingly, the Dirac action is invariant if we take $\mathcal{P} : \psi \mapsto \gamma^0\psi$. This means that in even spacetime dimensions, $\bar{\psi}\psi$ is even under parity; in odd spacetime dimensions $\bar{\psi}\psi$ is odd.

We can understand why this is by counting degrees of freedom. In $d = 3 + 1$ dimensions, the Dirac spinor has 4 components. When we quantise a massive fermion, we get two particle states – spin up and spin down – and the same anti-particle states. But a Dirac fermion in $d = 2+1$ dimensions has only two components, and so we must have half the number of particle states of the $d = 3 + 1$ theory. The pair that we keep is dictated by the sign of the mass, and by $\mathcal{CPT}$ invariance: if we have a particle with spin, or angular momentum, $+\frac{1}{2}$, the theory must also include an anti-particle of spin $-\frac{1}{2}$. But this necessarily breaks parity: the theory has a particle of spin $+\frac{1}{2}$ but no particle of spin $-\frac{1}{2}$.

8.5.1 Integrating out Massive Fermions

Let us take a single Dirac fermion, of mass m, coupled to a $U(1)$ gauge field A_μ. The action is

$$S = \int d^3x \; i\bar{\psi}\not{D}\psi + m\bar{\psi}\psi$$

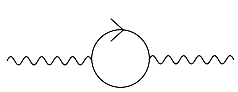
If we care about physics at energies below the fermion mass m, we can integrate out the fermion. We work in Euclidean space. The fermion then gives a contribution to the low-energy effective action for the gauge field,

$$S_{\text{eff}} = \log\det\left(i\not{D} + m\right) = \operatorname{Tr}\log\left(i\not{\partial} + \gamma^\mu A_\mu + m\right)$$

We expand this as,

$$S_{\text{eff}} = \operatorname{Tr}\log\left(i\not{\partial} + m\right) + \operatorname{Tr}\log\left(\frac{1}{i\not{\partial}+m}\gamma^\mu A_\mu\right) - \frac{1}{2}\operatorname{Tr}\log\left(\frac{1}{i\not{\partial}+m}\gamma^\mu A_\mu \frac{1}{i\not{\partial}+m}\gamma^\nu A_\nu\right) + \ldots$$

The first term is an overall constant, and the second term cannot lead to anything gauge invariant. But the third term holds something interesting. If we give the background field A_μ momentum p, then the trace over momenta corresponds to the diagram,



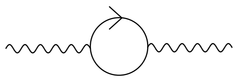
$$\begin{aligned} &= \frac{1}{2}A_\mu(-p)A_\nu(p)\int \frac{d^3k}{(2\pi)^3}\;\operatorname{tr}\left(\frac{1}{(\not{p}+\not{k})+m}\gamma^\mu\frac{1}{\not{k}+m}\gamma^\nu\right) \\ &= \frac{1}{2}A_\mu(-p)A_\nu(p)\int \frac{d^3k}{(2\pi)^3}\;\operatorname{tr}\left(\frac{\not{p}+\not{k}-m}{(p+k)^2+m^2}\gamma^\mu\frac{\not{k}-m}{k^2+m^2}\gamma^\nu\right)\end{aligned}$$

where we've used the fact that, after the Wick rotation, each gamma matrix squares to -1.

The trace picks out the non-vanishing gamma matrix structure. There will be a contribution to the Maxwell term; that doesn't interest us here. Instead, we care about the term we get when three gamma matrices are multiplied together. The trace structure gives

$$\mathrm{tr}\,\gamma^\rho\gamma^\mu\gamma^\nu = -2\epsilon^{\mu\nu\rho}$$

The resulting term is



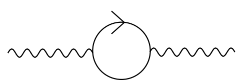
$$= \epsilon^{\mu\nu\rho} A_\mu(-p) A_\nu(p)\, p_\rho \int \frac{d^3k}{(2\pi)^3}\ \frac{m}{((p+k)^2+m^2)(k^2+m^2)}$$

We're interested in this integral in the infra-red limit, $p \to 0$, where it is given by

$$\int \frac{d^3k}{(2\pi)^3}\ \frac{m}{(k^2+m^2)^2} = \frac{1}{2\pi^2}\int_0^\infty dk\ \frac{mk^2}{(k^2+m^2)^2} = \frac{m}{8\pi|m|}$$

Putting this together, the 1-loop diagram gives

$$\lim_{p\to 0}$$



$$= \frac{1}{8\pi}\frac{m}{|m|}\epsilon^{\mu\nu\rho} A_\mu(-p) A_\nu(p)\, p_\rho$$

Back in real space, this gives us the leading term to the low energy effective action

$$S_{\text{eff}} = \frac{i}{4\pi}\frac{\mathrm{sign}(m)}{2}\int d^3x\ \epsilon^{\mu\nu\rho} A_\mu \partial_\nu A_\rho \tag{8.35}$$

There are a number of interesting things to point out about this result. First, the effective action comes with a power of i; this is expected for the Chern-Simons term in Euclidean space, and follows from Wick rotating terms with an ϵ symbol.

Second, and more surprisingly, the fermion does not decouple in the limit $m \to \infty$. After integrating out a massive field, one typically generates terms in the effective action that scale as a power of $1/m$. Not so for the Chern-Simons term: it is proportional to the sign of the mass. This behaviour holds for fermions in any odd spacetime dimensions; we met a similar example in $d = 4+1$ when discussing anomaly inflow in Section 4.4.2.

Finally, and most importantly, the effective action (8.35) is not gauge invariant! It is a Chern-Simons term (8.21) with level $k = \pm\frac{1}{2}$. Yet, we saw in the previous section, that the Chern-Simons term is only gauge invariant for $k \in \mathbf{Z}$. With $k = \pm\frac{1}{2}$, the sign of the partition function can flip under gauge transformations.

What are we to make of this? It appears that a single massive Dirac fermion, coupled to a $U(1)$ gauge field, is inconsistent. This is very much reminiscent of the gauge anomalies that we met in $d = 3+1$ dimensions in Section 3. However, we shouldn't be too hasty. After all, anomalies in $d = 3+1$ dimensions were strictly related to massless fermions, and here we're dealing with a massive fermion. What's going on?

Indeed, we were sloppy in how we dealt with UV divergences in the calculation above. They do not arise in the calculation of the Chern-Simons term, but they will surely be important if we compute other quantities and, as in any quantum field theory, we need a way to regulate them. To achieve this, we introduce a Pauli-Villars regulator field, together with suitable counterterms. We take the Pauli-Villars field to have real mass $\Lambda_{UV} > 0$. The regulated Dirac determinant is then

$$\frac{\det(i\not{D} + m)}{\det(i\not{D} + \Lambda_{UV})}$$

This gives two contributions to the Chern-Simons term; one from our fermion, and one from the regulator. The effective action for the gauge field then becomes

$$\begin{aligned} S_{\text{eff}} &= \log\left(\frac{\det(i\not{D} + m)}{\det(i\not{D} + \Lambda_{UV})}\right) \\ &= \frac{i}{8\pi}\left(\text{sign}(m) - 1\right)\int d^3x\ \epsilon^{\mu\nu\rho}A_\mu\partial_\nu A_\rho \end{aligned}$$

which vanishes when $m > 0$ but gives a Chern-Simons term of level $k = -1$ when $m < 0$. In other words, when the regulated fermion determinant is defined more carefully, there is no problem with gauge invariance.

The resulting situation is notationally inconvenient. Usually we would like to write down an action as shorthand for a quantum field theory, even though we know that to fully define the theory really requires a statement about how we regulate. The issue above means that the sign of the mass of the Pauli-Villars regulator matters in a crucial fashion. To avoid this, we are often sloppy and pretend that we've already integrated out the Pauli-Villars field to generate a bare Chern-Simons term with level $k = -\frac{1}{2}$ in the action.

More generally, we can couple N_f Dirac fermions to a $U(1)$ gauge field with the leading terms in the action given by

$$S = \int d^3x\ -\frac{1}{4e^2}F_{\mu\nu}F^{\mu\nu} + \frac{k}{4\pi}\epsilon^{\mu\nu\rho}A_\mu\partial_\nu A_\rho + \sum_{i=1}^{N_f} i\bar{\psi}_i\not{D}\psi_i + m_i\bar{\psi}_i\psi_i$$

Using the convention that the Chern-Simons term already includes the contributions from Pauli-Villars fields, gauge invariance requires

$$k + \frac{N_f}{2} \in \mathbf{Z}$$

This interplay between the level k and the number of fermions is sometimes referred to as the *parity anomaly*. It's not a great name since the theory with fermion masses is not parity invariant to begin with.

8.5.2 Massless Fermions and the Parity Anomaly

We can gain a slightly different perspective on the ideas above by considering a massless fermion coupled to a $U(1)$ gauge field, A_μ. The action is now

$$S = \int d^3x \; i\bar{\psi}\not{D}\psi$$

The transformation (8.34) ensures that the classical action is invariant under parity, provided that we also act with $A_1 \to -A_1$.

The classical action is invariant under parity. But what about the partition function? To answer this, we must make sense of the determinant of the Dirac operator,

$$Z[A] = \det\left(i\not{D}\right)$$

As above, we work in Euclidean space. The Dirac operator is Hermitian, which means that it has real eigenvalues,

$$i\not{D}\phi_n = \lambda_n\phi_n \quad \lambda_n \in \mathbf{R}$$

So formally we can write

$$Z = \prod_n \lambda_n$$

Of course, this formula is divergent and so we must work to make sense of it. For now, we would like to ask the following question: what is the sign of $\det(i\not{D})$? Roughly speaking, this depends on whether there are an even or odd number of negative eigenvalues. But, as there are an infinite number, it is not clear how to count them.

Why do we care so much about the sign? The problem comes if we try to reconcile a given sign with the requirements of gauge invariance. Suppose that we start with some gauge configuration $A_\mu^\star$ and decide that $\det(i\not{D})$ has a specific sign. Then it better be the case that, for any gauge configuration A_μ^ω, related to $A_\mu^\star$ by a gauge transformation, the sign of $\det(i\not{D})$ remain the same.

At this point, the discussion may be ringing bells. It is entirely analogous to the $SU(2)$ anomaly that we described in Section 3.4.3. We proceed in a very similar way. Consider the 1-parameter family of gauge configurations,

$$A_\mu(s;x) = (1-s)A_\mu^\star(x) + sA_\mu^\omega(x) \tag{8.36}$$

This has the property that it interpolates from $A_\mu^\star$ when $s=0$ to A_μ^ω when $s=1$. The question that we would like to answer is: how many eigenvalues pass through zero and change sign as we vary $s \in [0,1]$? To answer this, we can consider the gauge configuration $A_\mu(s;x)$ in (8.36) to live on the four manifold $I \times \mathbf{R}^3$, where I is the interval parameterised by $0 \le s \le 1$.

The number of times that an eigenvalue crosses zero is given by the index of the Dirac operator. This is the object that we introduced in Section 3.3.1 where, on a closed four manifold, the Atiyah-Singer index theorem allowed us to write

$$\text{Index}(i\not{D}_{4d}) = \frac{1}{32\pi^2}\int_{I\times\mathbf{R}^3} d^4x\ \epsilon^{\mu\nu\rho\sigma}F_{\mu\nu}F_{\rho\sigma}$$

In 4d, the index counts the difference between the number of left-handed and right-handed zero modes. For our purposes, it tells us the difference between the number of eigenvalues that switch from positive to negative, and those which switch from negative to positive. In other words, under the gauge transformation $A_\mu^0 \rightarrow A_\mu^\omega$, the partition function of the massless fermion changes as

$$Z \rightarrow Z\,(-1)^{\text{Index}(i\not{D}_{4d})}$$

There is no reason for this index to be even. We see, once again, that without regularisation the sign of the partition function can change under a suitable gauge transformation.

What happens if we now include a regulator? In mathematics, a suitably regulated sum of the signs of the eigenvalues of $i\not{D}$ is known as the Atiyah-Patodi-Singer eta-invariant. It is defined by

$$\eta(A) = \lim_{\epsilon\rightarrow 0^+}\sum_n e^{-\epsilon\,\lambda_n^2}\,\text{sign}(\lambda_n)$$

We then define a regulated version of the fermion partition function as

$$Z = |\det(i\not{D})|\,e^{-i\pi\eta(A)}$$

The η invariant depends on the background gauge field A. The Atiyah-Patodi-Singer index theorem provides an expression for η in terms of the gauge field. If we restrict to the generic situation where the gauge field has no zero modes, then one can show that

$$\pi\,\eta(A) = \frac{1}{4\pi}\int d^3x\ \epsilon^{\mu\nu\rho}A_\mu\partial_\nu A_\rho$$

This reproduces the expression that we found previously from the Pauli-Villars regularisation. In general, the eta-invariant is the more mathematically rigorous way to describe what's happening as it allows one to track what happens as eigenvalues pass through zero.

8.6 3d Bosonization

In two spacetime dimensions, there is not much of a distinction between bosons and fermions. The map between them is known as bosonization and was described in Section 7.5.

In three spacetime dimensions, bosons are not the same as fermions. We can tell which one we have in the same way as we would in four dimensions. Given a pair of particles we can rotate them by 180°, keeping them well separated. The wavefunction for a pair of bosons will come back to itself, while the wavefunction for a pair of fermions comes back with a minus sign.

Nonetheless, it is possible to use Chern-Simons terms to change statistics of an excitation from a boson to a fermion. This process is referred to as *3d bosonization*.

8.6.1 Flux Attachment

To get a feel for what's going on, it's useful to first revert to some non-relativistic physics. Consider Chern-Simons theory coupled to a current J^μ

$$S = \int d^3x\ \frac{k}{4\pi}\epsilon^{\mu\nu\rho}A_\mu\partial_\nu A_\rho - A_\mu J^\mu \tag{8.37}$$

We can insert a test particle of unit charge by taking $J^0 = \delta^2(\mathbf{x})$. How does the gauge field respond? Gauss' law tells us that the charged particle is accompanied by a fractional magnetic flux,

$$\frac{1}{2\pi}B = \frac{1}{k}\,\delta^2(\mathbf{x}) \tag{8.38}$$

This is referred to as *flux attachment*.

Now consider two such particles. We will exchange them to determine their quantum statistics. The wavefunction will pick up a factor of ± 1 depending on whether the original particles were fermions. However, there is a second contribution to the phase of the wavefunction that comes from the Aharonov-Bohm effect.

Recall that a particle of charge q moving around a flux Φ picks up a phase $e^{iq\Phi}$. But because of flux attachment (8.38), the particles carry both charge $q = 1$ and flux $\Phi = 2\pi/k$. If we move one particle all the way around another, we will get a phase $e^{iq\Phi}$. But the statistical phase is defined by exchanging particles, which consists of only half an orbit (followed by a translation which contributes no phase). So, after exchange, the expected statistical phase is

$$\pm\, e^{iq\Phi/2} = \pm e^{i\pi/k}$$

where we take the $+$ sign if our original particles are bosons and the $-$ sign if they were fermions. We see that the effect of the Chern-Simons term is to transmute the quantum statistics of the particles. In particular, if we take a Chern-Simons term at level $k = \pm 1$, what were bosons become fermions and vice versa. Once again, we see that the topological nature of the Chern-Simons term endows it with seemingly magic infra-red properties: it can change the behaviour of far separated particles, even though it has no propagating degrees of freedom.

For $|k| > 1$, the particles are neither bosons nor fermions. Instead they carry fractional quantum statistics. Such particles are called *anyons* and are allowed only in $d = 2 + 1$ dimensions. You can read more about them in the lecture notes on the Quantum Hall Effect where they play a prominent role.

A Famously Fiddly Factor of 2

The calculation above contains an annoying factor of 2 that we've swept under the rug. Here's the issue. As the charge q in the first particle moved around the flux Φ in the second, we picked up a phase $e^{iq\Phi}$. But you might think that the flux Φ of the first particle also moved around the charge q of the second. So surely this should give another factor of $e^{iq\Phi}$. Right? Well, no. To see why, it's best to just do the calculation.

For generality, let's take N particles sitting at positions $\mathbf{x}_a(t)$ which, as the notation shows, we allow to change with time. The charge density and currents are

$$J^0(\mathbf{x}, t) = \sum_{a=1}^{N} \delta^2(\mathbf{x} - \mathbf{x}_a(t)) \quad \text{and} \quad \mathbf{J}(\mathbf{x}, t) = \sum_{a=1}^{N} \dot{\mathbf{x}}_a\, \delta^2(\mathbf{x} - \mathbf{x}_a(t))$$

The equation of motion from (8.37) is

$$\frac{1}{2\pi}F_{\mu\nu} = \frac{1}{k}\epsilon_{\mu\nu\rho}J^\rho$$

and can be easily solved even in this general case. We work in Coulomb gauge with $A_0 = 0$ and $\nabla \cdot \mathbf{A} = 0$. The solution is then

$$A_i(\mathbf{x},t) = \frac{1}{k}\sum_{a=1}^{N} \epsilon^{ij}\,\frac{x^j - x_a^j(t)}{|\mathbf{x}-\mathbf{x}_a(t)|^2} \tag{8.39}$$

This follows from the standard methods that we know from our Electromagnetism lectures, but this time using the Green's function for the Laplacian in two dimensions: $\nabla^2 \log|\mathbf{x}-\mathbf{y}| = 2\pi\delta^2(\mathbf{x}-\mathbf{y})$. This solution is again the statement that each particle carries flux $1/k$. However, we can also use this solution directly to compute the phase change when one particle – say, the first one – is transported along a curve C. It is simply

$$\exp\left(i\oint_C \mathbf{A}\cdot d\mathbf{x}_1\right)$$

If the curve C encloses one other particle, the resulting phase change can be computed to be $e^{2\pi i/m}$. As before, if we exchange two particles, we get half this phase, or $e^{i\pi/m}$. This, of course, is the same result we got above.

8.6.2 A Bosonization Duality

The discussion above shows that Chern-Simons terms can turn bosons into fermions and vice-versa. However, it holds only for massive particles, and cannot be easily generalised to massless particles, let alone to relativistic quantum field theories. Nonetheless, it is suggestive that it may be possible to write down a quantum field theory of bosons coupled to Chern-Simons terms that has a dual interpretation in terms of fermions. As we now explain, it is thought that this is indeed the case.

Before we proceed, we're going to make a small change in notation. In what follows, there will be lots of $U(1)$ gauge fields floating around. Some of them will be dynamical, while others will be background gauge fields that we couple to currents. To distinguish between these, we use the following convention: dynamical gauge fields will be written in lower case, e.g. a_μ. Meanwhile, background gauge fields will be written in upper case, e.g. A_μ.

This convention differs from what we've used throughout these lectures, where we typically refer to all gauge fields, dynamical or background, as A_μ. It is, however, a standard convention in condensed matter physics where the true electromagnetic gauge field A_μ is typically a background field, describing electric or magnetic fields that the experimenter has chosen to turn on. In contrast, 3d dynamical gauge fields a_μ are always emergent excitations, arising from some collective behaviour of strongly coupled electrons.

Consider the following theory, that we refer to as Theory A: a complex scalar field coupled to a $U(1)$ gauge field, with Chern-Simons term at level $k = 1$,

$$S_A[\phi, a] = \int d^3x \ -\frac{1}{4e^2} f_{\mu\nu} f^{\mu\nu} + \frac{1}{4\pi}\epsilon^{\mu\nu\rho} a_\mu \partial_\nu a_\rho + |\mathcal{D}_\mu \phi|^2 - m^2|\phi|^2 - \frac{\lambda}{2}|\phi|^4 \tag{8.40}$$

This is the Abelian Higgs model (8.12), but with the addition of a Chern-Simons term. Just as before, it is straightforward to analyse in the limits $m^2 \gg e^2$ and $m^2 \ll -e^2$ where it is a theory of weakly interacting massive particles. But we'd like to understand what happens in the strongly coupled regime. We will argue below that as we vary the m^2 from positive to negative, there is a unique second order phase transition, roughly at $m = 0$. You can think of this gapless theory as the XY critical point, coupled to a Chern-Simons gauge field $U(1)_1$. Below, we will conjecture an alternative, and somewhat simpler, description.

In the infra-red limit $e^2 \to \infty$, the Gauss' law constraint gives rise to the local flux attachment condition,

$$\frac{f_{12}}{2\pi} + \rho_{\text{scalar}} = 0 \tag{8.41}$$

where ρ_{scalar} is the charge density of the scalar field ϕ. In the non-relativistic setting – which can be invoked when $m^2 \gg e^2$ – we viewed this as attaching flux to every scalar excitation and saw that, for $k = 1$, this turns a boson into a fermion. In the relativistic setting, it turns out to be more appropriate to think of attaching a scalar to every flux.

To see this, first note that the theory has a conserved global symmetry, with the topological current (8.3)

$$j^\mu_{\text{top}} = \frac{1}{2\pi}\epsilon^{\mu\nu\rho}\partial_\nu a_\rho \tag{8.42}$$

We know from our earlier discussion in Section 8.1 that the local operators which carry charge under this current are monopole operators $\mathcal{M}(x)$, which insert magnetic flux at a point. The flux attachment (8.41) is telling us that, in the presence of a Chern-Simons term, these monopole operators are not gauge invariant. We can make them gauge invariant only by dressing them with some scalar charge $\rho_{\rm scalar}$. Schematically, we refer to the gauge invariant composite operator as $\mathcal{M}\phi$.

How do we do this less schematically? The right way to proceed is to solve the equation of motion for the scalar in the presence of a Dirac monopole. We then treat each mode quantum mechanically: the flux attachment condition (8.41) tells us that we should excite a single mode. The monopole operator with the lowest dimension will correspond to exciting the lowest energy scalar mode.

We won't go through this full calculation. However, the key physics can be seen from a simple calculation that we did back in Section 1.1: a charged particle moving in a minimal Dirac monopole receives a shift of $\hbar/2$ to its angular momentum. (See, in particular, equation (1.9).) This means that exciting any bosonic mode will shift the angular momentum of the monopole to become 1/2-integer. But, in a relativistic theory, the spin-statistics relation must hold. If our gauge invariant monopole operator $\mathcal{M}\phi$ has spin 1/2, then it must also be a fermion.

We see that this argument leads to the same result as before: a bosonic theory coupled to a $U(1)$ Chern-Simons gauge field at level $k = 1$ is really a theory of fermions. The obvious question is: what theory of fermions?

It is conjectured that, close to the critical point, the bosonic theory (8.40) is really just a free Dirac fermion! In other words, it can be equivalently described as

$$S_B[\psi] = \int d^3x \; i\bar{\psi}\,\partial\!\!\!/\psi - m'\bar{\psi}\psi \tag{8.43}$$

The map is very similar to that of particle-vortex duality that we saw in Section 8.2.1. In particular, the fermion is described by the dressed monopole operator in Theory A,

$$\mathcal{M}\phi \quad \longleftrightarrow \quad \psi$$

while the $U(1)$ currents map between themselves

$$j^\mu_{\rm top} = \frac{1}{2\pi}\epsilon^{\mu\nu\rho}\partial_\nu a_\rho \quad \longleftrightarrow \quad j^\mu = \bar{\psi}\gamma^\mu\psi \tag{8.44}$$

Checking the Topological Phases

Let's now look for some evidence that this claimed duality is correct. In the case of particle-vortex duality, we checked that the theories looked similar in the weakly coupled regimes $|m^2| \gg e^2$. We can try to do something similar here.

This is simplest for Theory B. To study the relevant physics, we couple the current (8.44) to a background gauge field A_μ. The partition function for each theory then depends on this background field. For Theory B it is

$$Z_B[A] = \int \mathcal{D}\psi \;\exp\left(iS_B[\psi] + i\int d^3x\; j^\mu A_\mu - \frac{1}{2}\frac{1}{4\pi}\epsilon^{\mu\nu\rho}A_\nu\partial_\nu A_\rho\right)$$

Note that we are using the convention described in Section 8.5, in which the half-integer Chern-Simons term arising from the Pauli-Villars regulator field is shown explicitly in the action. We have chosen to add this term with level $k = -1/2$.

When the fermions are massive, $m' \neq 0$, we can integrate them out and generate an effective theory for the background fields A_μ. The lowest dimension term is a Chern-Simons interaction for A_μ,

$$Z[A] = \exp\left(i\frac{\tilde{k}}{4\pi}\epsilon^{\mu\nu\rho}A_\mu\partial_\nu A_\rho + \ldots\right) \tag{8.45}$$

From our discussion in Section 8.5, we know that after integrating out the massive fermion ψ the Chern-Simons level for the background gauge field will be

$$\tilde{k} = \frac{1}{2}\left(-1 + \mathrm{sign}(m')\right) = \begin{cases} 0 & m' > 0 \\ -1 & m' < 0 \end{cases}$$

It may seem odd to write down an action for background fields which don't fluctuate, but there's important information in the coefficient $\tilde{k}$; it is the Hall conductivity of the topological gapped phase. This follows by using the partition function $Z[A]$ to compute the response of the current j^μ to a background electric field

$$\langle j^\mu(x)\rangle = -i\frac{\delta \log Z[A]}{\delta A_\mu(x)} \quad\Rightarrow\quad \langle j_i\rangle = -\frac{\tilde{k}}{2\pi}\epsilon_{ij}E_j$$

You can read (a lot) more about the Hall conductivity in the lectures on the Quantum Hall Effect.

We would like to see how this effect is encoded in the bosonic Theory A. We couple the background gauge field A_μ to the topological current (8.42) to get the partition function

$$Z_A[A] = \int \mathcal{D}\phi\mathcal{D}a \;\exp\left(iS_A[\phi, a] + i\frac{1}{2\pi}\epsilon^{\mu\nu\rho}A_\mu\partial_\nu a_\rho\right)$$

where we're neglecting gauge fixing terms. This time we only have a scalar field, which does not shift the level of the Chern-Simons term when integrated out. Nonetheless, we can still reproduce the result (8.46) for the Hall conductivity. To see how this works, let's start with the mass $m^2 \gg e^2$ where, at low energies, the scalar field simply decouples, leaving us with the effective action

$$S_{\rm eff}[a,A] = \int d^3x \; \frac{1}{4\pi}\epsilon^{\mu\nu\rho}a_\mu\partial_\nu a_\rho + \frac{1}{2\pi}\epsilon^{\mu\nu\rho}A_\mu\partial_\nu a_\rho$$

The equation of motion for the dynamical gauge field a is simply $a = -A$. Substituting this back in, given the effective action (8.45) with $\tilde{k} = -1$.

What happens when $m^2 \ll -e^2$? In this case the scalar field condenses and the dynamical gauge field a becomes gapped. This extra term kills the Hall conductivity, leaving us with (8.45) with $\tilde{k} = 0$. We see that the scalar field does reproduce the topological phases of the the fermion theory as promised. This requires the map,

$$m^2 \quad \longleftrightarrow \quad -m' \quad \Rightarrow \quad \phi^\dagger\phi \quad \longleftrightarrow \quad -\bar{\psi}\psi$$

The agreement between the topological phases is promising, but a long way from demonstrating the claimed duality between Theory A (8.40) and the free fermion (8.43). There are a number of other routes which lead us to the duality (including large N methods, holography, lattice constructions and supersymmetry) but we will not discuss them here. Instead we will assume that bosonization duality holds and ask: what can we do with it?

8.6.3 The Beginning of a Duality Web

We will now show how, starting from the bosonization duality, we can derive further equivalences between quantum field theories. First, some conventions. We will revert to form notation for the gauge fields, and write the Chern-Simons terms as

$$\frac{1}{4\pi}\epsilon^{\mu\nu\rho}a_\mu\partial_\nu a_\rho = \frac{1}{4\pi}ada$$
$$\frac{1}{2\pi}\epsilon^{\mu\nu\rho}A_\mu\partial_\nu a_\rho = \frac{1}{2\pi}Ada = -\frac{1}{2\pi}adA$$

(The final equality really holds only when integrated over spacetime, and after discarding a total derivative.) Both of these are correctly normalised as explained in Section 8.4: they can be added to the action only with integer-valued coefficients. We will denote the gauge field under which matter is charged by adding a subscript to the covariant derivative like this,

$$\mathcal{D}_a\phi = \partial\phi - ia\phi$$

The spacetime index on the derivatives will be suppressed. In what follows, the distinction between dynamical gauge fields and background gauge fields will be crucial. As we mentioned previously, they are distinguished by case. Lower case gauge fields, $a, b, c, \ldots$ will always be dynamical; upper case gauge fields $A, B, C, \ldots$ will always be background.

In this notation, we write the 3d bosonization duality that we described above as an equivalence between two theories

$$|\mathcal{D}_b\phi|^2 - |\phi|^4 + \frac{1}{4\pi}bdb + \frac{1}{2\pi}Adb \quad \longleftrightarrow \quad i\bar{\psi}\not{D}_A\psi - \frac{1}{2}\frac{1}{4\pi}AdA \tag{8.46}$$

Much of this expression is shorthand. First, we have set the mass terms to zero on both sides. This really means that we tune to the critical point. On the fermionic side this is obvious, but the scalar side includes a $|\phi|^4$ term which is taken to mean that we flow to the Wilson-Fisher fixed point of the theory, rather than the free fixed point. Of course, we don't literally get to the Wilson-Fisher by simply setting $m^2 = 0$; instead we must tune m^2, or more generally the coefficient of the relevant operator, as we flow to the IR to hit the critical point. All of this is buried in the notation above.

Second, we reiterate that the scalar ϕ in the above expression is charged under a dynamical gauge field, which we have called b to prepare us for some manipulations ahead. This means that we integrate over (gauge equivalent) configurations of b in the path integral. In contrast, the fermion ψ is charged under the background field A. We can read off the duality map (8.44) between currents by seeing which terms on both side are coupled to A. Finally, we've omitted nearly all the details of the regularisation of the field theory, with one exception: the level $-1/2$ Chern-Simons term on the right-hand-side can be thought of as coming from integrating out a Pauli-Villars regulator. This was explained in Section 8.5. (A warning: some places in the literature adopt a different convention where this level $-1/2$ Chern-Simons term remains hidden in the regulator.)

At this point we start to play with these two theories. Both sides of the duality (8.46) have a background $U(1)$ gauge field A. The key idea is to promote this to a dynamical gauge field. This is misleadingly easy in our notation: we simply write a instead of A. As we explained in Section 8.1, gauging a $U(1)$ symmetry in $d = 2 + 1$ results in a new global symmetry,

$$j^\mu = \frac{1}{2\pi}\epsilon^{\mu\nu\rho}\partial_\nu a_\rho$$

We couple this to a background gauge field C. This means that we add $\frac{1}{2\pi}AdC$ to both sides of (8.46), and then make $A \to a$ dynamical. This results in a new duality,

$$|\mathcal{D}_b\phi|^2 - |\phi|^4 + \frac{1}{4\pi}bdb + \frac{1}{2\pi}adb + \frac{1}{2\pi}adC \quad \longleftrightarrow \quad i\bar{\psi}\not{D}_a\psi - \frac{1}{2}\frac{1}{4\pi}ada + \frac{1}{2\pi}adC$$

The number of gauge fields on the left-hand side are proliferating. But, at this point, something nice happens: the gauge field a only appears linearly in the action. This means that it acts as a Lagrange multiplier, setting $db = -dC$. But, this, in turn, freezes the first dynamical gauge field b to be equal, up to gauge connection, to the new background field $-C$. The upshot is that we end up with a scalar field theory with no dynamical gauge fields at all, and the duality

$$|\mathcal{D}_C\phi|^2 - |\phi|^4 + \frac{1}{4\pi}CdC \quad \longleftrightarrow \quad i\bar{\psi}\not{D}_a\psi - \frac{1}{2}\frac{1}{4\pi}ada + \frac{1}{2\pi}adC \tag{8.47}$$

This is a new equivalence between two, seemingly very different looking, theories. The left-hand-side is something very familiar: it is the XY Wilson-Fisher fixed point. In contrast, the right-hand side is a strongly coupled $U(1)$ gauge theory. The claim is that these two fixed points are the same, so

$$\text{XY Wilson-Fisher} \quad \longleftrightarrow \quad U(1)_{-1/2} \text{ coupled to a Dirac fermion}$$

From our first bosonization duality, we have derived another. Similarly, we can go in reverse: starting from the equality of partition functions (8.47), it is not hard to derive the original (8.46).

We can continue in this vein, adding different matter fields and gauging global symmetries, to derive an infinite number of dualities between different 3d Abelian theories with Chern-Simons terms. This is referred to as the *duality web*. Below we give just a handful of interesting examples.

8.6.4 Particle-Vortex Duality Revisited

Our second bosonization duality (8.47) includes a Chern-Simons coupling for the background field C on the left-hand-side. Since we don't integrate over the background field, there is nothing to stop us taking this term onto the other side of the equation. We will also take this opportunity to rename some of the variables. The duality (8.47) is equivalent to

$$|\mathcal{D}_A\phi|^2 - |\phi|^4 \quad \longleftrightarrow \quad i\bar{\psi}\not{D}_b\psi - \frac{1}{2}\frac{1}{4\pi}bdb + \frac{1}{2\pi}bdA - \frac{1}{4\pi}AdA \tag{8.48}$$

Having moved the background Chern-Simons term to the other side, we now play the same game as before: we add a term $\frac{1}{2\pi}AdC$, and then again promote A to a dynamical field, $A \to a$. We now have

$$|\mathcal{D}_a\phi|^2 - |\phi|^4 + \frac{1}{2\pi}adC \quad \longleftrightarrow \quad i\bar{\psi}\not{D}_b\psi - \frac{1}{2}\frac{1}{4\pi}bdb + \frac{1}{2\pi}bda - \frac{1}{4\pi}ada + \frac{1}{2\pi}adC$$

Again, there's a lot of gauge fields on the right-hand-side. Now a does not appear linearly as a Lagrange multiplier, but quadratically. Still, it is begging to be integrated out by imposing the equation of motion $a = b + C$, leaving us with

$$|\mathcal{D}_a\phi|^2 - |\phi|^4 + \frac{1}{2\pi}adC \quad \longleftrightarrow \quad i\bar{\psi}\not{D}_b\psi + \frac{1}{2}\frac{1}{4\pi}bdb + \frac{1}{2\pi}bdC + \frac{1}{4\pi}CdC \tag{8.49}$$

This is still a bosonization duality, relating a scalar theory to a fermionic theory. But the right-hand-side is very nearly the same expression that we started with in (8.48), but with one important difference: two of the Chern-Simons terms have their sign flipped. In fact, we we send $C \to -C$, all of the Chern-Simons terms have their sign flipped. In other words, this partition function describes the time reversal of the theory in (8.48).

As we have seen, Chern-Simons terms break time reversal, so one would not naively expect that $U(1)_{1/2}$ coupled to a Dirac fermion is time reversal invariant. However, if we take the time reversal of the duality (8.48), we have

$$|\mathcal{D}_{-C}\phi|^2 - |\phi|^4 \quad \longleftrightarrow \quad i\bar{\psi}\not{D}_b\psi + \frac{1}{2}\frac{1}{4\pi}bdb - \frac{1}{2\pi}bd(-C) + \frac{1}{4\pi}CdC \tag{8.50}$$

By charge conjugation we can replace $\mathcal{D}_{-C}\phi \to \mathcal{D}_C\phi$. The left-hand-side is once again the XY critical point. It is clearly time-reversal invariant. The duality tells us that $U(1)_{1/2}$ coupled to a massless fermion must be secretly time reversal invariant: it must emerge as a discrete symmetry of the quantum theory.

Combining (8.49) together with (8.50) gives us yet another duality. It is

$$|\mathcal{D}_a\phi|^2 - |\phi|^4 + \frac{1}{2\pi}adC \quad \longleftrightarrow \quad |\mathcal{D}_C\phi|^2 - |\phi|^4$$

But this is precisely the statement of particle vortex duality that we discussed in Section 8.2.1: the left-hand-side is the Abelian Higgs model while the right-hand-side is the XY model. We learn that particle-vortex duality = bosonization^2.

8.6.5 Fermionic Particle-Vortex Duality

Above we have managed to use 3d bosonization to derive a duality between purely bosonic theories. We might ask: can we do something similar to derive a duality between purely fermionic theories? The answer is yes. But, there will be a new subtlety that we have to address.

We can see this subtlety by retracing the steps above. To derive bosonic particle-vortex duality, we started with the bosonization dual (8.47), moved the background Chern-Simons term to the other side, and then promoted the background gauge field to a dynamical one. To derive a fermionic particle-vortex duality, it is natural to attempt the same manoeuvres for our original bosonization duality (8.46),

$$|\mathcal{D}_b\phi|^2 - |\phi|^4 + \frac{1}{4\pi}bdb + \frac{1}{2\pi}Adb \quad \longleftrightarrow \quad i\bar{\psi}\not{D}_A\psi - \frac{1}{2}\frac{1}{4\pi}AdA \tag{8.51}$$

But we immediately run into a stumbling block: we can't move the background Chern-Simons term to the other side because it is half-integer valued. It is needed on the right-hand-side to ensure that the fermion partition function is gauge invariant.

To get around this, we will stipulate that the background gauge field A only admits flux quantised as

$$\frac{1}{2\pi}\int dA \in 2\mathbf{Z}$$

This is twice the usual requirement. We can then write

$$A = 2C$$

with C a background gauge field whose flux is correctly quantised. The duality (8.51) is then

$$|\mathcal{D}_b\phi|^2 - |\phi|^4 + \frac{1}{4\pi}bdb + \frac{2}{2\pi}Cdb \quad \longleftrightarrow \quad i\bar{\psi}\not{D}_{2C}\psi - \frac{2}{4\pi}CdC \tag{8.52}$$

All Chern-Simons terms are now properly quantised. But the fermion on the right-hand-side has charge 2 under the gauge field C. If we give a fermion of charge q a mass m and integrate it out, it will generate a Chern-Simons term with level $\frac{1}{2}q^2\text{sign}(m)$. (This follows from the fact that the one-loop diagram in Section 8.5 has two insertions of the photon-fermion vertex.) So integrating out a fermion of charge 2 generates an integer-valued Chern-Simons level and there is no problem with the parity anomaly.

Now let us play games with this theory. We will move the CdC background Chern-Simons term to the other side, add $\frac{1}{2\pi}BdC$ to both sides, and finally breathe life into C to make it dynamical, $C \to a$. We have

$$|\mathcal{D}_b\phi|^2 - |\phi|^4 + \frac{1}{4\pi}bdb + \frac{2}{2\pi}adb + \frac{2}{4\pi}ada + \frac{1}{2\pi}adB \quad \longleftrightarrow \quad i\bar{\psi}\not{D}_{2a}\psi + \frac{1}{2\pi}adB$$

The mess of mixed Chern-Simons terms on the left-hand-side is easily dealt with: we simply define the new linear combination

$$\hat{a} = a + b$$

Then we find

$$|\mathcal{D}_b\phi|^2 - |\phi|^4 - \frac{1}{4\pi}bdb - \frac{1}{2\pi}bdB + \frac{2}{4\pi}\hat{a}d\hat{a} + \frac{1}{2\pi}\hat{a}dB \quad \longleftrightarrow \quad i\bar{\psi}\not{D}_{2a}\psi + \frac{1}{2\pi}adB$$

But the first four terms in this expression – those which involve ϕ and b — coincide with the time-reversal of the left-hand-side of (8.51). We can then use the duality (8.51) to replace them, leaving us with the promised fermion-fermion duality,

$$i\bar{\psi}\not{D}_A\psi + \frac{1}{2}\frac{1}{4\pi}AdA + \frac{2}{4\pi}\hat{a}d\hat{a} + \frac{1}{2\pi}\hat{a}dA \quad \longleftrightarrow \quad i\bar{\psi}\not{D}_{2a}\psi + \frac{1}{2\pi}adA$$

where we've taken this opportunity to rename the background field A.

What is this final expression telling us? The right-hand-side is a $U(1)$ gauge theory coupled to a single Dirac fermion of charge 2. The left-hand-side is very almost a free fermion. But it also includes a decoupled topological theory, $U(1)_2$, described by the dynamical gauge field $\hat{a}$. We learn that

$$U(1) \text{ with Dirac fermion of charge 2} \quad \longleftrightarrow \quad \text{Free Dirac fermion} + U(1)_2$$

This is the fermionic version of particle-vortex duality, with the monopole operators of the gauge theory identified with the fermion. A closely related duality was first suggested by Son in the context of the half-filled Landau level. It has also been invoked in the context of topological insulators.

8.7 Further Reading

Quantum field theories in $d = 2 + 1$ dimensions have a rather special relation to the real world because, after a Wick rotation, many of them (but not all of them!) can be viewed as statistical field theories in $d = 3 + 0$ dimensions, where they describe systems near critical points. For example, ϕ^4 scalar field theory in $d = 3$ dimensions describes the water boiling in your kettle. (Admittedly, you might need to put a fairly tight lid on the kettle.)

From the high energy perspective, $d = 2 + 1$ dimensions offer another arena to study questions about gauge theories that seemed too challenging in $d = 3 + 1$. Polyakov's demonstration of confinement [159, 160], driven by the proliferation of instantons (monopoles), was a highlight in this regard. Similarly, particle vortex duality was first introduced by Peskin [152], in an attempt to see whether a similar duality in $d = 3 + 1$ could help explain confinement. This was subsequently rediscovered in the condensed matter community by Dasgupta and Halperin, who also performed numerics to find convincing evidence of a second order phase transition [37]. Both of these papers originally expressed the duality in terms of lattice theories; the continuum version that we described here was first proposed in [61].

Chern-Simons theory was introduced by Deser, Jackiw and Templeton [42, 43], initially as a surprising, gauge invariant mechanism to give the three dimensional photon a mass. The depth of the theory became apparent with Witten's Fields medal winning work on knot invariants [228], and the connection to WZW models [53]. The interplay between massive fermions and Chern-Simons terms was discovered in [149] and [168, 169]; a more modern perspective was provided by Witten in [230]. A very clear discussion of the properties of Chern-Simons theories can be found in the lectures by Dunne [49]. You can read more about the subtleties related to the quantisation of Abelian Chern-Simons theories in the appendices of [175] and [176]

The story of 3d bosonization has a long and complicated history. The idea that one can use Chern-Simons terms to transmute the statistics of non-relativistic particles from bosons to fermions was pointed out by Wilczek and Zee [211]. Polyakov was the first to conjecture that there might be a relativistic version of bosonization, but he missed the need to bosonize at the Wilson-Fisher fixed point [161]. The full story came by bringing together a wonderfully diverse set of ideas from both high energy and condensed matter physics. These include dualities in supersymmetric theories [110], large N bosonization and its relation to holography [76, 5, 6, 7], and physics associated to superfluids [13], the half-filled Landau level [186] and topological insulators [200, 136]. The web of dualities among Abelian gauge theories, relating bosonization and particle-vortex duality, was first described in [119, 176].

References

[1] L. F. Abbott, "*Introduction to the Background Field Method*," Acta Phys. Polon. B **13**, 33 (1982).

[2] S. Adler, "*Axial Vector Vertex in Spinor Electrodynamics*", Phys. Rev. 177, 2426 (1969)

[3] S. L. Adler and W. A. Bardeen, "*Absence of higher order corrections in the anomalous axial vector divergence equation*," Phys. Rev. **182**, 1517 (1969).

[4] O. Aharony, N. Seiberg and Y. Tachikawa, "*Reading between the lines of four-dimensional gauge theories*," JHEP **1308**, 115 (2013) [arXiv:1305.0318 [hep-th]].

[5] O. Aharony, G. Gur-Ari and R. Yacoby, "*d=3 Bosonic Vector Models Coupled to Chern-Simons Gauge Theories*," JHEP **1203**, 037 (2012) [arXiv:1110.4382 [hep-th]].

[6] O. Aharony, G. Gur-Ari and R. Yacoby, "*Correlation Functions of Large N Chern-Simons-Matter Theories and Bosonization in Three Dimensions*," JHEP **1212**, 028 (2012) [arXiv:1207.4593 [hep-th]].

[7] O. Aharony, "*Baryons, monopoles and dualities in Chern-Simons-matter theories*," JHEP **1602**, 093 (2016) [arXiv:1512.00161 [hep-th]].

[8] R. Balian, J. M. Drouffe and C. Itzykson, "*Gauge Fields on a Lattice. 1. General Outlook*," Phys. Rev. D **10**, 3376 (1974); "*Gauge Fields on a Lattice. 2. Gauge Invariant Ising Model*," Phys. Rev. D **11**, 2098 (1975); "*Gauge Fields on a Lattice. 3. Strong Coupling Expansions and Transition Points*," Phys. Rev. D **11**, 2104 (1975) Erratum: [Phys. Rev. D **19**, 2514 (1979)].

[9] T. Banks and E. Rabinovici, "*Finite Temperature Behavior of the Lattice Abelian Higgs Model*," Nucl. Phys. B **160**, 349 (1979).

[10] T. Banks and A. Zaks, "*On the Phase Structure of Vector-Like Gauge Theories with Massless Fermions*," Nucl. Phys. B **196**, 189 (1982).

[11] T. Banks and N. Seiberg, "*Symmetries and Strings in Field Theory and Gravity*," Phys. Rev. D **83**, 084019 (2011) [arXiv:1011.5120 [hep-th]].

[12] W. A. Bardeen, "*Anomalous Ward identities in spinor field theories*," Phys. Rev. **184**, 1848 (1969).

[13] M. Barkeshli and J. McGreevy, "*Continuous transition between fractional quantum Hall and superfluid states,*" Phys. Rev. B **89**, no. 23, 235116 (2014) [arXiv:1201.4393 [cond-mat.str-el]].

[14] A. A. Belavin, A. M. Polyakov, A. S. Schwartz and Y. S. Tyupkin, "*Pseudoparticle Solutions of the Yang-Mills Equations*," Phys. Lett. B **59**, 85 (1975)

[15] J. S. Bell, "*On the Einstein-Podolsky-Rosen paradox*," Physics Physique Fizika **1**, 195 (1964).

[16] J. S. Bell and R. Jackiw, "*A PCAC Puzzle: $\pi^- \to 2\gamma$ in the σ-model*", Nuovo Cim. A60, 47 (1969)

[17] B. A. Bernevig, "*Topological Insulators and Topological Superconductors*", Princeton University Press (2013)

[18] A. Bilal, "*Lectures on Anomalies*," arXiv:0802.0634 [hep-th].

[19] J. D. Bjorken, "*Asymptotic Sum Rules at Infinite Momentum*," Phys. Rev. **179**, 1547 (1969).

[20] E. B. Bogomolny, "*Stability of Classical Solutions*," Sov. J. Nucl. Phys. **24**, 449 (1976)

[21] C. Bouchiat, J. Iliopoulos and P. Meyer, "*An Anomaly Free Version of Weinberg's Model*," Phys. Lett. **38B**, 519 (1972).

[22] C. G. Callan, Jr., S. R. Coleman, J. Wess and B. Zumino, "*Structure of phenomenological Lagrangians. 2.*," Phys. Rev. **177**, 2247 (1969).

[23] C. G. Callan, Jr., R. F. Dashen and D. J. Gross, "*The Structure of the Gauge Theory Vacuum*," Phys. Lett. B **63**, 334 (1976)

[24] C. G. Callan, Jr., R. F. Dashen and D. J. Gross, "*Instantons and Massless Fermions in Two-Dimensions*," Phys. Rev. D **16** (1977) 2526.

[25] C. G. Callan, Jr. and J. A. Harvey, "*Anomalies and Fermion Zero Modes on Strings and Domain Walls*," Nucl. Phys. B **250**, 427 (1985).

[26] S. R. Coleman, "*There are no Goldstone bosons in two-dimensions*," Commun. Math. Phys. **31**, 259 (1973).

[27] S. R. Coleman and E. J. Weinberg, "*Radiative Corrections as the Origin of Spontaneous Symmetry Breaking*," Phys. Rev. D **7**, 1888 (1973).

[28] S. R. Coleman, R. Jackiw and L. Susskind, "*Charge Shielding and Quark Confinement in the Massive Schwinger Model*," Annals Phys. **93**, 267 (1975).

[29] S. R. Coleman, "*The Quantum Sine-Gordon Equation as the Massive Thirring Model*," Phys. Rev. D **11**, 2088 (1975).

[30] S. R. Coleman, "*More About the Massive Schwinger Model*," Annals Phys. **101**, 239 (1976).

[31] S. R. Coleman and E. Witten, "*Chiral Symmetry Breakdown in Large N Chromodynamics*," Phys. Rev. Lett. **45**, 100 (1980).

[32] S. Coleman, "*Aspects of Symmetry : Selected Erice Lectures*,", Cambridge University Press

[33] M. Creutz, "*Monte Carlo Study of Quantized SU(2) Gauge Theory*," Phys. Rev. D **21**, 2308 (1980).

[34] M. Creutz, "*Quarks, Gluons and Lattices*", Cambridge University Press

[35] M. Creutz, "*Why rooting fails*," PoS Lattice **2007**, 007 (2007) [arXiv:0708.1295 [hep-lat]].

[36] A. D'Adda, M. Luscher and P. Di Vecchia, "*A 1/N Expandable Series of Nonlinear Sigma Models with Instantons,*" Nucl. Phys. B **146**, 63 (1978).

[37] C. Dasgupta and B. I. Halperin, "*Phase Transition in a Lattice Model of Superconductivity*," Phys. Rev. Lett. **47**, 1556 (1981).

[38] R. F. Dashen, "*Some features of chiral symmetry breaking*," Phys. Rev. D **3**, 1879 (1971).

[39] R. Delbourgo and A. Salam, "*The gravitational correction to PCAC*," Phys. Lett. **40B**, 381 (1972). doi:10.1016/0370-2693(72)90825-8

[40] P. Di Francesco, P. Mathieu and D. Senechal, "*Conformal Field Theory*,", Springer

[41] P. Di Vecchia and G. Veneziano, "*Chiral Dynamics in the Large n Limit*," Nucl. Phys. B **171**, 253 (1980).

[42] S. Deser, R. Jackiw and S. Templeton, "*Topologically Massive Gauge Theories*," Annals Phys. **140**, 372 (1982)

[43] S. Deser, R. Jackiw and S. Templeton, "*Three-Dimensional Massive Gauge Theories*," Phys. Rev. Lett. **48**, 975 (1982).

[44] P. A. M. Dirac, "*Quantised Singularities in the Electromagnetic Field*", Proc. R. Soc. Lond. A, 1931.

[45] P. A. M. Dirac "*The Theory of Magnetic Poles*", Physical Review 75, 7 (1948)

[46] P. A. M. Dirac, letter to Abdus Salam 1981, published in N. Craigie, P. Goddard and W. Nahm, "*Monopoles in Quantum Field Theory*", Proc. of Monopole Meeting, Trieste, Italy, 1981.

[47] N. Dombey, "*An Interview with Murray Gell-Mann on June 5 2001*", arXiv:2007.12615

[48] S. D. Drell, M. Weinstein and S. Yankielowicz, "*Strong Coupling Field Theories. 2. Fermions and Gauge Fields on a Lattice*," Phys. Rev. D **14**, 1627 (1976).

[49] G. V. Dunne, "*Aspects of Chern-Simons theory*," hep-th/9902115.

[50] H. Eichenherr, "*SU(N) Invariant Nonlinear Sigma Models*," Nucl. Phys. B **146**, 215 (1978) Erratum: [Nucl. Phys. B **155**, 544 (1979)].

[51] E. Eichten and J. Preskill, "*Chiral Gauge Theories on the Lattice*," Nucl. Phys. B **268**, 179 (1986).

[52] S. Elitzur, "*Impossibility of Spontaneously Breaking Local Symmetries*," Phys. Rev. D **12**, 3978 (1975).

[53] S. Elitzur, G. W. Moore, A. Schwimmer and N. Seiberg, "*Remarks on the Canonical Quantization of the Chern-Simons-Witten Theory*," Nucl. Phys. B **326**, 108 (1989).

[54] A. M. Essin, Essin, J. E. Moore and D. Vanderbilt, "*Magnetoelectric polarizability and axion electrodynamics in crystalline insulators*", arXiv:0810.2998 [cond-mat], Phys. Rev. Lett. 102, 146805 (2009)

[55] G. Farmelo, "*The Strangest Man: The Hidden Life of Paul Dirac, Quantum Genius*", Faber and Faber, 2009

[56] R. P. Feynman, "*Very high-energy collisions of hadrons*," Phys. Rev. Lett. **23**, 1415 (1969).

[57] R. P. Feynman, "*The behavior of hadron collisions at extreme energies*," Conf. Proc. C **690905**, 237 (1969).

[58] L. Fidkowski and A. Kitaev, "*The effects of interactions on the topological classification of free fermion systems,*" Phys. Rev. B **81**, 134509 (2010) [arXiv:0904.2197 [cond-mat.str-el]].

[59] L. Fidkowski and A. Kitaev,, "*Topological phases of fermions in one dimension*", Phys. Rev. B 83, 075103 (2011), [arXiv:1008.4138 [cond-mat.str-el]].

[60] D. Finkelstein and J. Rubinstein, "*Connection between spin, statistics, and kinks*," J. Math. Phys. **9**, 1762 (1968).

[61] D. H. Lee and M. P. A. Fisher, "*Correspondence between two-dimensional bosons and a bulk superconductor in a magnetic field*" Phys. Rev. B 39, 2756 (1989).

[62] V. Fock, "*Über die invariante Form der Wellen- und der Bewegungsgleichungen für einen geladenen Massenpunkt*", Zeit. für Physik 39, 226-232. (1926)

[63] E. H. Fradkin and S. H. Shenker, "*Phase Diagrams of Lattice Gauge Theories with Higgs Fields*," Phys. Rev. D **19**, 3682 (1979).

[64] Y. Frishman, A. Schwimmer, T. Banks and S. Yankielowicz, "*The Axial Anomaly and the Bound State Spectrum in Confining Theories*," Nucl. Phys. B **177**, 157 (1981).

[65] H. Fritzsch, M. Gell-Mann and H. Leutwyler, "*Advantages of the Color Octet Gluon Picture*," Phys. Lett. **47B**, 365 (1973).

[66] L. Fu, C.L. Kane, E.J. Mele, "*Topological Insulators in Three Dimensions*", Phys. Rev. Lett. 98, 106803 (2007)

[67] L. Fu and C. L. Kane, "*Topological Insulators with Inversion Symmetry*", cond-mat/0611341, Phys. Rev. B 76, 045302 (2007)

[68] K. Fujikawa, "*Path Integral Measure for Gauge Invariant Fermion Theories*," Phys. Rev. Lett. **42**, 1195 (1979).

[69] K. Fujikawa, "*Path Integral for Gauge Theories with Fermions*," Phys. Rev. D **21**, 2848 (1980) Erratum: [Phys. Rev. D **22**, 1499 (1980)].

[70] D. Gaiotto, A. Kapustin, N. Seiberg and B. Willett, "*Generalized Global Symmetries*," JHEP **1502**, 172 (2015) [arXiv:1412.5148 [hep-th]].

[71] D. Gaiotto, A. Kapustin, Z. Komargodski and N. Seiberg, "*Theta, Time Reversal, and Temperature*," JHEP **1705**, 091 (2017) [arXiv:1703.00501 [hep-th]].

[72] M. Gell-Mann and M. Levy, "*The axial vector current in beta decay*," Nuovo Cim. **16**, 705 (1960).

[73] P. H. Ginsparg and K. G. Wilson, "*A Remnant of Chiral Symmetry on the Lattice*," Phys. Rev. D **25**, 2649 (1982).

[74] P. H. Ginsparg, "*Applied Conformal Field Theory*," hep-th/9108028.

[75] P. H. Ginsparg and G. W. Moore, "*Lectures on 2-D gravity and 2-D string theory,*", hep-th/9304011.

[76] S. Giombi, S. Minwalla, S. Prakash, S. P. Trivedi, S. R. Wadia and X. Yin, "*Chern-Simons Theory with Vector Fermion Matter*," Eur. Phys. J. C **72**, 2112 (2012) [arXiv:1110.4386 [hep-th]].

[77] S. L. Glashow, "*Partial Symmetries of Weak Interactions*," Nucl. Phys. **22**, 579 (1961).

[78] J. Goldstone, "*Field Theories with Superconductor Solutions*," Nuovo Cim. **19**, 154 (1961).

[79] J. Goldstone, A. Salam and S. Weinberg, "*Broken Symmetries*," Phys. Rev. **127**, 965 (1962).

[80] P. Goddard, J. Nuyts and D. I. Olive, "*Gauge Theories and Magnetic Charge*," Nucl. Phys. B **125**, 1 (1977).

[81] V. L. Golo and A. M. Perelomov, "*Solution of the Duality Equations for the Two-Dimensional SU(N) Invariant Chiral Model*," Phys. Lett. **79B**, 112 (1978).

[82] D. Gross and R. Jackiw, "*Effect of Anomalies on Quasirenormalisable Theories*", Phys. Rev. D6, 477 (1972)

[83] D. J. Gross and F. Wilczek, "*Ultraviolet Behavior of Nonabelian Gauge Theories*," Phys. Rev. Lett. **30**, 1343 (1973).

[84] D. J. Gross and F. Wilczek, "*Asymptotically Free Gauge Theories I*" Phys. Rev. D **8**, 3633 (1973).

[85] D. J. Gross and A. Neveu, "*Dynamical Symmetry Breaking in Asymptotically Free Field Theories*," Phys. Rev. D **10**, 3235 (1974).

[86] D. J. Gross, "*Twenty five years of asymptotic freedom*," Nucl. Phys. Proc. Suppl. **74**, 426 (1999) [hep-th/9809060].

[87] F. D. M. Haldane, "*Continuum dynamics of the 1-D Heisenberg antiferromagnetic identification with the O(3) nonlinear sigma model*," Phys. Lett. A **93**, 464 (1983).

[88] D. Haldane, "*Model for a Quantum Hall Effect without Landau Levels: Condensed Matter Realisation of the Parity Anomaly*", Phys. Rev. Lett. 61, 2015 (1988).

[89] J. A. Harvey, "*Magnetic monopoles, duality and supersymmetry*," In *Trieste 1995, High energy physics and cosmology* 66-125 [hep-th/9603086].

[90] J. A. Harvey, "*TASI 2003 lectures on anomalies*," hep-th/0509097.

[91] M. Z. Hasan, C. L. Kane, "*Topological Insulators*", arXiv:1002.3895, Rev. Mod. Phys. 82: 3045 (2010)

[92] P. Hasenfratz and G. 't Hooft, "*A Fermion-Boson Puzzle in a Gauge Theory*," Phys. Rev. Lett. **36**, 1119 (1976).

[93] G. 't Hooft, "*Renormalization of Massless Yang-Mills Fields*," Nucl. Phys. B **33**, 173 (1971).

[94] G. 't Hooft, "*Renormalizable Lagrangians for Massive Yang-Mills Fields*," Nucl. Phys. B **35**, 167 (1971).

[95] G. 't Hooft and M. Veltman "*Regularization and Renormalisation of Gauge Fields*", Nucl. Phys. B44, 189 (1972)

[96] G. 't Hooft, "*The Renormalisation Procedure for Yang-Mills Fields*", PhD Thesis Utrecht University (1972)

[97] G. 't Hooft, "*A Planar Diagram Theory for Strong Interactions*," Nucl. Phys. B **72**, 461 (1974).

[98] G. 't Hooft, "*A Two-Dimensional Model for Mesons*," Nucl. Phys. B **75**, 461 (1974). doi:10.1016/0550-3213(74)90088-1

[99] G. 't Hooft, "*Magnetic Monopoles in Unified Gauge Theories*," Nucl. Phys. B **79**, 276 (1974).

[100] G. 't Hooft, "*Quarks and Gauge Fields*", Proceedings of the Colloqium on "Recent Progress in Lagrangian Field Theory and Applications", Marseille, June 24-28, 1974 Available here.

[101] G. 't Hooft, "*Symmetry Breaking Through Bell-Jackiw Anomalies*," Phys. Rev. Lett. **37**, 8 (1976).

[102] G. 't Hooft, "*Computation of the Quantum Effects Due to a Four-Dimensional Pseudoparticle*," Phys. Rev. D **14**, 3432 (1976) Erratum: [Phys. Rev. D **18**, 2199 (1978)].

[103] G. 't Hooft, "*On the Phase Transition Towards Permanent Quark Confinement*," Nucl. Phys. B **138**, 1 (1978).

[104] G. 't Hooft, "*A Property of Electric and Magnetic Flux in Non-Abelian Gauge Theories*," Nucl. Phys. B **153**, 141 (1979).

[105] G. 't Hooft, "*Naturalness, chiral symmetry, and spontaneous chiral symmetry breaking*," NATO Sci. Ser. B **59**, 135 (1980).

[106] G. 't Hooft, "*Topology of the Gauge Condition and New Confinement Phases in Non-Abelian Gauge Theories*," Nucl. Phys. B **190**, 455 (1981).

[107] G. 't Hooft, "*When was asymptotic freedom discovered? or the rehabilitation of quantum field theory*," Nucl. Phys. Proc. Suppl. **74**, 413 (1999) [hep-th/9808154].

[108] G. 't Hooft ed. "*50 years of Yang-Mills theory*," Hackensack, USA: World Scientific (2005) 487 p

[109] D. Hsieh, D. Qian, L. Wray, Y. Xia, Y.S. Hor, R.J. Cava, M.Z. Hasan, "*A topological Dirac insulator in a quantum spin Hall phase (experimental realization of a 3D Topological Insulator)*", arXiv:0910.2420 [cond-mat], Nature 452, 970 (2008).

[110] K. A. Intriligator and N. Seiberg, "*Mirror symmetry in three-dimensional gauge theories*," Phys. Lett. B **387**, 513 (1996) [hep-th/9607207].

[111] R. Jackiw and C. Rebbi, "*Solitons with Fermion Number 1/2*," Phys. Rev. D **13**, 3398 (1976).

[112] R. Jackiw and C. Rebbi, "*Spin from Isospin in a Gauge Theory*," Phys. Rev. Lett. **36**, 1116 (1976).

[113] R. Jackiw and C. Rebbi, "*Vacuum Periodicity in a Yang-Mills Quantum Theory*," Phys. Rev. Lett. **37**, 172 (1976).

[114] R. Jackiw and C. Rebbi, "*Spinor Analysis of Yang-Mills Theory*," Phys. Rev. D **16**, 1052 (1977).

[115] R. Jackiw, "*Topological Investigations in Quantum Gauge Theories*", in S. B. Treiman, E. Witten, R. Jackiw and B. Zumino, "*Current Algebra And Anomalies*," Singapore, Singapore: World Scientific (1985) 537p

[116] J. D. Jackson and L. B. Okun, "*Historical Roots of Gauge Invariance*", hep-ph/0012061 Rev. Mod. Phys. 73:663-680 (2001)

[117] D. B. Kaplan, "*A Method for simulating chiral fermions on the lattice*," Phys. Lett. B **288**, 342 (1992) [hep-lat/9206013].

[118] D. B. Kaplan, "*Chiral Symmetry and Lattice Fermions*," arXiv:0912.2560 [hep-lat].

[119] A. Karch and D. Tong, "*Particle-Vortex Duality from 3d Bosonization*," Phys. Rev. X **6**, no. 3, 031043 (2016) [arXiv:1606.01893 [hep-th]].

[120] A. Kapustin, "*Wilson-'t Hooft operators in four-dimensional gauge theories and S-duality*," Phys. Rev. D **74**, 025005 (2006) [hep-th/0501015].

[121] Y. Kikukawa and T. Noguchi, "*Low-energy effective action of domain wall fermion and the Ginsparg-Wilson relation*," Nucl. Phys. Proc. Suppl. **83**, 630 (2000) [hep-lat/9902022].

[122] O. Klein, "*On the Theory of Charged Fields*", Written for a conference in Conference in Kazimierz, Poland 1938.

[123] Z. Komargodski, A. Sharon, R. Thorngren and X. Zhou, "*Comments on Abelian Higgs Models and Persistent Order*," arXiv:1705.04786 [hep-th].

[124] J. B. Kogut and L. Susskind, "*Hamiltonian Formulation of Wilson's Lattice Gauge Theories*," Phys. Rev. D **11**, 395 (1975).

[125] J. Kondo, "*Resistance Minimum in Dilute Magnetic Alloys*," Prog. Theor. Phys. **32**, no. 1, 37 (1964).

[126] M. Lüscher, "*Exact chiral symmetry on the lattice and the Ginsparg-Wilson relation,*" Phys. Lett. B **428**, 342 (1998) [hep-lat/9802011].

[127] M. Luscher, "*Chiral gauge theories revisited*," Subnucl. Ser. **38**, 41 (2002) [hep-th/0102028].

[128] A. Luther and I. Peschel, "*Single-particle states, Kohn anomaly, and pairing fluctuations in one dimension*", Phys. Rev. B9, 7 (1974)

[129] J. M. Maldacena, "*The Large N limit of superconformal field theories and supergravity,*" Int. J. Theor. Phys. **38**, 1113 (1999) [Adv. Theor. Math. Phys. **2**, 231 (1998)] [hep-th/9711200].

[130] S. Mandelstam, "*Vortices and Quark Confinement in Nonabelian Gauge Theories*," Phys. Rept. **23**, 245 (1976).

[131] S. Mandelstam, "*Soliton Operators for the Quantized Sine-Gordon Equation*," Phys. Rev. D **11**, 3026 (1975).

[132] A. V. Manohar, "*Large N QCD*," hep-ph/9802419.

[133] N. S. Manton and P. Sutcliffe, "Topological solitons,", Cambridge University Press (2004)

[134] N. D. Mermin and H. Wagner, "*Absence of ferromagnetism or antiferromagnetism in one-dimensional or two-dimensional isotropic Heisenberg models*," Phys. Rev. Lett. **17**, 1133 (1966).

[135] J. Mcgreevy, "*Applied String Theory*", https://mcgreevy.physics.ucsd.edu/fall08/index.html

[136] M. A. Metlitski and A. Vishwanath, "*Particle-vortex duality of two-dimensional Dirac fermion from electric-magnetic duality of three-dimensional topological insulators*," Phys. Rev. B **93**, no. 24, 245151 (2016) [arXiv:1505.05142 [cond-mat.str-el]].

[137] M. A. Metlitski, "*S-duality of $u(1)$ gauge theory with $\theta = \pi$ on non-orientable manifolds: Applications to topological insulators and superconductors*," arXiv:1510.05663 [hep-th].

[138] G. Moore, "*Informal Lectures on Lattice Gauge Theory*", https://theorie.ikp.physik.tu-darmstadt.de/qcd/moore/latt_lectures.pdf

[139] J. E. Moore and L. Balents, "*Topological invariants of time-reversal-invariant band structures*", cond-mat/0607314, Phys. Rev. B 75, 121306(R) (2007)

[140] Y. Nambu, "*Axial vector current conservation in weak interactions*," Phys. Rev. Lett. **4**, 380 (1960).

[141] Y. Nambu and G. Jona-Lasinio, "*Dynamical Model of Elementary Particles Based on an Analogy with Superconductivity. 1.*," Phys. Rev. **122**, 345 (1961).

[142] Y. Nambu, "*Strings, Monopoles and Gauge Fields*," Phys. Rev. D **10**, 4262 (1974).

[143] H. Neuberger, "*Exactly massless quarks on the lattice*," Phys. Lett. B **417**, 141 (1998) [hep-lat/9707022].

[144] H. Neuberger, "*More about exactly massless quarks on the lattice*," Phys. Lett. B **427**, 353 (1998) [hep-lat/9801031].

[145] H. B. Nielsen and P. Olesen, "*Vortex Line Models for Dual Strings*," Nucl. Phys. B **61**, 45 (1973).

[146] H. B. Nielsen and M. Ninomiya, "*Absence of Neutrinos on a Lattice. 1. Proof by Homotopy Theory*," Nucl. Phys. B **185**, 20 (1981) Erratum: [Nucl. Phys. B **195**, 541 (1982)].

[147] H. B. Nielsen and M. Ninomiya, "*Absence of Neutrinos on a Lattice. 2. Intuitive Topological Proof*," Nucl. Phys. B **193**, 173 (1981).

[148] N. K. Nielsen, "*Asymptotic Freedom as a Spin Effect*" Am. J. Phys 49;12 1171 (1981)

[149] A. J. Niemi and G. W. Semenoff, "Âxial Anomaly Induced Fermion Fractionization and Effective Gauge Theory Actions in Odd Dimensional Space-Times," Phys. Rev. Lett. **51**, 2077 (1983). doi:10.1103/PhysRevLett.51.2077

[150] L. O'Raifeartaigh, "*The Dawning of Gauge Theory*", Princeton University Press (1997)

[151] W. Pauli, letter to A. Pais entitled "*Meson-Nucleon Interaction and Differential Geometry*", 1953

[152] M. E. Peskin, "*Mandelstam 't Hooft Duality in Abelian Lattice Models*," Annals Phys. **113**, 122 (1978).

[153] M. E. Peskin, "*Chiral Symmetry And Chiral Symmetry Breaking*," SLAC-PUB-3021. pdf here.

[154] M. E. Peskin and D. V. Schroeder, "*An Introduction to quantum field theory*," CRC Press

[155] H. Poincaré, "*Remarques sur une expérience de M. Birkeland*, Compt. Rend. Acad. Sci. Paris 123 (1896), 530-533.

[156] H. D. Politzer, "*Reliable Perturbative Results for Strong Interactions?*," Phys. Rev. Lett. **30**, 1346 (1973).

[157] H. D. Politzer, "*The dilemma of attribution*," Nobel Lecture, Proc. Nat. Acad. Sci. **102**, 7789 (2005) [Int. J. Mod. Phys. A **20**, 5741 (2005)]

[158] A. M. Polyakov, "*Particle Spectrum in the Quantum Field Theory*," JETP Lett. **20**, 194 (1974) [Pisma Zh. Eksp. Teor. Fiz. **20**, 430 (1974)].

[159] A. M. Polyakov, "*Compact Gauge Fields and the Infrared Catastrophe*," Phys. Lett. B **59**, 82 (1975)

[160] A. M. Polyakov, "*Quark Confinement and Topology of Gauge Groups*," Nucl. Phys. B **120**, 429 (1977).

[161] A. M. Polyakov, "*Fermi-Bose Transmutations Induced by Gauge Fields*," Mod. Phys. Lett. A **3**, 325 (1988).

[162] M. K. Prasad and C. M. Sommerfield, "*An Exact Classical Solution for the 't Hooft Monopole and the Julia-Zee Dyon*," Phys. Rev. Lett. **35**, 760 (1975).

[163] J. Preskill and S. Weinberg, "*Decoupling Constraints on Massless Composite Particles*," Phys. Rev. D **24** (1981), 1059

[164] J. Preskill, "*Magnetic Monopoles*," Ann. Rev. Nucl. Part. Sci. **34**, 461 (1984).

[165] X-L. Qi, T. Hughes and S-C. Zhang, "*Topological Field Theory of Time-Reversal Invariant Insulators*", arXiv:0802.3537 [cond-mat], Phys. Rev. B78 195424 (2008)

[166] X-L. Qi, R. Li, J. Zang, S-C. Zhang, "*Seeing the magnetic monopole through the mirror of topological surface states*", arXiv:0811.1303 [cond-mat] Science 323, 1184 (2009)

[167] X-L Qi and S-C Zhang, "*Topological insulators and superconductors*", arXiv:1008.2026 [cond-mat], Rev. Mod. Phys. 83, 1057-1110 (2011)

[168] A. N. Redlich, "*Gauge Noninvariance and Parity Violation of Three-Dimensional Fermions*," Phys. Rev. Lett. **52**, 18 (1984).

[169] A. N. Redlich, "*Parity Violation and Gauge Noninvariance of the Effective Gauge Field Action in Three-Dimensions*," Phys. Rev. D **29**, 2366 (1984).

[170] R. Roy, "*Three dimensional topological invariants for time reversal invariant Hamiltonians and the three dimensional quantum spin Hall effect*", cond-mat/0607531, Phys. Rev. B 79, 195322 (2009)

[171] M.N. Saha, "*The Origin of Mass in Neutrons and Protons*", Ind. Jour. Phys., 10, 141, 1936

[172] M.N. Saha, "*Note on Dirac's Theory of Magnetic Poles*", Phys. Rev. **95**1968 (1949)

[173] S. Scherer and M. R. Schindler, "*A Chiral perturbation theory primer*," hep-ph/0505265.

[174] J. S. Schwinger, "*Gauge Invariance and Mass. 2.*," Phys. Rev. **128**, 2425 (1962).

[175] N. Seiberg and E. Witten, "*Gapped Boundary Phases of Topological Insulators via Weak Coupling*," PTEP **2016**, no. 12, 12C101 (2016) [arXiv:1602.04251 [cond-mat.str-el]].

[176] N. Seiberg, T. Senthil, C. Wang and E. Witten, "*A Duality Web in 2+1 Dimensions and Condensed Matter Physics*," Annals Phys. **374**, 395 (2016) [arXiv:1606.01989 [hep-th]].

[177] D. Senechal, "*An Introduction to bosonization*," cond-mat/9908262.

[178] R. Shankar, "*Bosonization: How to make it work for you in condensed matter*," Acta Phys. Polon. B **26**, 1835 (1995).

[179] R. Shankar and G. Murthy, "*Deconfinement in d=1: A Closer look*," Phys. Rev. B **72**, 224414 (2005) [cond-mat/0508242].

[180] S. R. Sharpe, "*Rooted staggered fermions: Good, bad or ugly?*," PoS LAT **2006**, 022 (2006) [hep-lat/0610094].

[181] R. Shaw, Cambridge PhD Thesis.

[182] M. Shifman, "*Instantons in Gauge Theories*", World Scientific (1994)

[183] Y. M. Shnir, "*Magnetic Monopoles*, Springer (2005)

[184] T. H. R. Skyrme, "*A Nonlinear field theory*," Proc. Roy. Soc. Lond. A **260**, 127 (1961).

[185] T. H. R. Skyrme, "*A Unified Field Theory of Mesons and Baryons*," Nucl. Phys. **31**, 556 (1962).

[186] D. T. Son, "*Is the Composite Fermion a Dirac Particle?*," Phys. Rev. X **5**, no. 3, 031027 (2015) [arXiv:1502.03446 [cond-mat.mes-hall]].

[187] J. Steinberger, "*On the Use of subtraction fields and the lifetimes of some types of meson decay*," Phys. Rev. **76**, 1180 (1949).

[188] L. Susskind, "*Lattice Fermions*," Phys. Rev. D **16**, 3031 (1977).

[189] S. Raby, S. Dimopoulos and L. Susskind, "*Tumbling Gauge Theories*," Nucl. Phys. B **169**, 373 (1980).

[190] D. G. Sutherland, "*Current algebra and some nonstrong mesonic decays*," Nucl. Phys. B **2**, 433 (1967).

[191] D. Tong, "*TASI Lectures on Instantons*", hep-th/0509216 http://www.damtp.cam.ac.uk/user/tong/tasi.html

[192] D. Tong, "*Lectures on String Theory*", arXiv:0908.0333 http://www.damtp.cam.ac.uk/user/tong/string.html

[193] D. Tong, "*Lectures on the Quantum Hall Effect*", arXiv:1606.06687 http://www.damtp.cam.ac.uk/user/tong/qhe.html

[194] D. J. Thouless, M. Kohmoto, M. P. Nightingale, and M. den Nijs, "*Quantized Hall Conductance in a Two-Dimensional Periodic Potential*", Phys. Rev. Lett. 49, 405 (1982).

[195] C. Vafa and E. Witten, "*Restrictions on Symmetry Breaking in Vector-Like Gauge Theories*," Nucl. Phys. B **234**, 173 (1984).

[196] C. Vafa and E. Witten, "*Parity Conservation in QCD*," Phys. Rev. Lett. **53**, 535 (1984).

[197] S, Vandoren and P. van Nieuwenhuizen, "*Lectures on Instantons*", arXiv:0802.1862 [hep-th]

[198] M. Veltman "*Theoretical aspects of high energy neutrino interactions*", Proc. Roy. Soc. A 301, 1465

[199] G. Veneziano, "*U(1) Without Instantons*," Nucl. Phys. B **159**, 213 (1979).

[200] C. Wang and T. Senthil, "*Half-filled Landau level, topological insulator surfaces, and three-dimensional quantum spin liquids*," Phys. Rev. B **93**, no. 8, 085110 (2016) [arXiv:1507.08290 [cond-mat.str-el]].

[201] F. J. Wegner, "*Duality in Generalized Ising Models and Phase Transitions Without Local Order Parameters,*" J. Math. Phys. **12**, 2259 (1971).

[202] X. G. Wen, "*A lattice non-perturbative definition of an SO(10) chiral gauge theory and its induced standard model*," Chin. Phys. Lett. **30**, 111101 (2013) [arXiv:1305.1045 [hep-lat]].

[203] H. Weyl, "*Gruppentheorie und Quantenmechanik*, Leipzig, Hirzel (1928)

[204] H. Weyl, "*"Gravitation and the electron*", Proc. Nat. Acad. Sci. 15, 323 (1929)

[205] S. Weinberg, "*A Model of Leptons*," Phys. Rev. Lett. **19**, 1264 (1967).

[206] S. Weinberg, "*Nonlinear realizations of chiral symmetry*," Phys. Rev. **166**, 1568 (1968). doi:10.1103/PhysRev.166.1568

[207] S. Weinberg, "*Nonabelian Gauge Theories of the Strong Interactions*," Phys. Rev. Lett. **31**, 494 (1973).

[208] S. Weinberg, "*Current algebra and gauge theories. 2. NonAbelian gluons*," Phys. Rev. D **8**, 4482 (1973).

[209] S. Weinberg, "*Effective Field Theory, Past and Future*," PoS CD **09**, 001 (2009) [arXiv:0908.1964 [hep-th]].

[210] D. Weingarten, "*Mass Inequalities for QCD,*" Phys. Rev. Lett. **51** (1983), 1830

[211] F. Wilczek and A. Zee, "*Linking Numbers, Spin, and Statistics of Solitons*," Phys. Rev. Lett. **51**, 2250 (1983).

[212] F. Wilczek, "*Two Applications of Axion Electrodynamics*," Phys. Rev. Lett. **58**, 1799 (1987).

[213] H. Wilson, "*Note on Dirac's Theory of Magnetic Poles*", Phys. Rev. 75, 309 (1949)

[214] K. G. Wilson, "*Confinement of Quarks*," Phys. Rev. D **10**, 2445 (1974).

[215] K. G. Wilson, "*The Renormalization Group: Critical Phenomena and the Kondo Problem*," Rev. Mod. Phys. **47**, 773 (1975).

[216] E. Witten, "*Dyons of Charge $e\theta/2\pi$*," Phys. Lett. B **86**, 283 (1979)

[217] E. Witten and D. I. Olive, "*Supersymmetry Algebras That Include Topological Charges*," Phys. Lett. **78B**, 97 (1978).

[218] E. Witten, "*Some Properties of the $(\bar{\psi}\psi)^2$ Model in Two-Dimensions*," Nucl. Phys. B **142**, 285 (1978).

[219] E. Witten, "*Chiral Symmetry, the 1/N Expansion, and the SU(N) Thirring Model*," Nucl. Phys. B **145**, 110 (1978).

[220] E. Witten, "*Instantons, the Quark Model, and the 1/n Expansion*," Nucl. Phys. B **149**, 285 (1979).

[221] E. Witten, "*Baryons in the 1/N Expansion*," Nucl. Phys. B **160**, 57 (1979).

[222] E. Witten, "*Current Algebra Theorems for the U(1) Goldstone Boson*," Nucl. Phys. B **156**, 269 (1979).

[223] E. Witten, "*The 1/N Expansion In Atomic And Particle Physics*," NATO Sci. Ser. B **59**, 403 (1980).

[224] E. Witten, "*Large N Chiral Dynamics*," Annals Phys. **128**, 363 (1980).

[225] E. Witten, "*An SU(2) Anomaly*," Phys. Lett. B **117**, 324 (1982)

[226] E. Witten, "*Current Algebra, Baryons, and Quark Confinement*," Nucl. Phys. B **223**, 433 (1983).

[227] E. Witten, "*Nonabelian Bosonization in Two-Dimensions*," Commun. Math. Phys. **92**, 455 (1984).

[228] E. Witten, "*Quantum Field Theory and the Jones Polynomial*," Commun. Math. Phys. **121**, 351 (1989).

[229] E. Witten, "*Three Lectures on Topological Phases*" arXiv:1510.07698, La Rivista del Nuovo Cimento, 39 (2016) 313-370

[230] E. Witten, "*Fermion Path Integrals And Topological Phases*", arXiv:1508.04715, Rev. Mod. Phys. 88, 35001 (2016)

[231] T. T. Wu and C. N. Yang, "*Concept of Nonintegrable Phase Factors and Global Formulation of Gauge Fields*," Phys. Rev. D **12**, 3845 (1975).

[232] C. N. Yang and R. L. Mills, "*Conservation of Isotopic Spin and Isotopic Gauge Invariance*", Phys. Rev. 96 (1954) 191-195

[233] K. Yee, "*Numerical solution of initial boundary value problems involving maxwell's equations in isotropic media*", IEEE Transactions on Antennas and Propagation, Vol 14, 3, May 1966.

[234] Y. You, Y. BenTov and C. Xu, "*Interacting Topological Superconductors and possible Origin of* $16n$ *Chiral Fermions in the Standard Model*," arXiv:1402.4151 [cond-mat.str-el].

[235] D. Zwanziger, "*Quantum field theory of particles with both electric and magnetic charges*," Phys. Rev. **176**, 1489 (1968).