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 $\eta^{\mu\nu} = \text{diag}(+1, -1)$ 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

$$S = \int d^2x \; i\bar{\psi}\not{\partial}\psi \tag{3.3}$$

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: $m\bar{\psi}\psi = m(\chi_+^\dagger\chi_- + \chi_-^\dagger\chi_+)$

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 $E = |p|$.

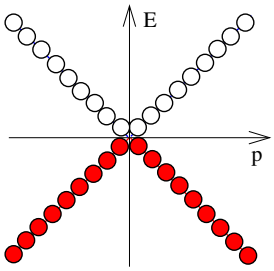


Figure 23:

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*.

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.

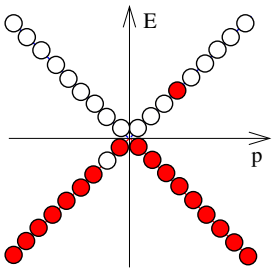


Figure 24:

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

$$S = \int d^2x \; i\bar{\psi}\not{D}\psi \tag{3.6}$$

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!

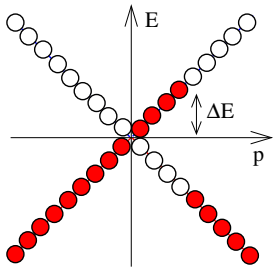


Figure 25:

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.

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.

Figure 26:

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

$$E^2 = eB(2n+1) + p_z^2 - 2eBS_z \qquad \text{with } n = 0, 1, 2, \ldots$$

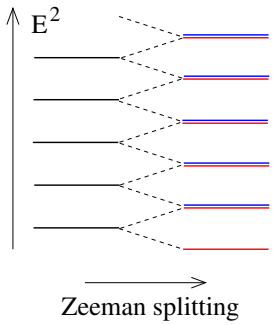


Figure 27:

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

$$S = A \int dz dt \; i\bar{\chi}_+(\partial_t - \partial_z)\chi_+$$

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

$$\begin{aligned} Z[K] &= \int \mathcal{D}\phi' \ \exp\left(-S[\phi] + \int d^d x \ K\phi\right) \exp\left(-\int d^d x \ \epsilon\,(\partial_\mu J^\mu - KX)\right) \\ &\approx \int \mathcal{D}\phi' \ \exp\left(-S[\phi] + \int d^d x \ K\phi\right) \left[1 - \int d^d x \ \epsilon\,(\partial_\mu J^\mu - KX)\right] \end{aligned} \tag{3.17}$$

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

$$\int \mathcal{D}\phi \ \exp\left(-S[\phi] + \int d^d x \ K\phi\right) \left[\int d^d x \ \epsilon(x)\,(\partial_\mu J^\mu - KX)\right] = 0$$

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,

$$\int \mathcal{D}\phi \ \exp\left(-S[\phi] + \int d^d x \ K\phi\right) (\partial_\mu J^\mu - K(x)X(\phi)) = 0$$

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

$$\partial_\mu \langle J^\mu(x)\phi(x^1)\dots\phi(x^n)\rangle = 0 \quad \text{for } x \neq x^i$$

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

$$j^\mu_A = i\bar{\psi}\gamma^\mu\gamma^5\psi$$

Repeating the rest of the path integral manipulations seems to tell us that the Ward identities hold with $\partial_\mu j^\mu_A = 0$. 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

$$\begin{aligned}\int d^4x\ \epsilon(x) \sum_n \bar{\phi}_n \gamma^5 \phi_n &= \lim_{\Lambda\to\infty} \int d^4x\ \epsilon(x) \sum_n \bar{\phi}_n \gamma^5 \phi_n\, e^{-\lambda_n^2/\Lambda^2} \\ &= \lim_{\Lambda\to\infty} \int d^4x\ \epsilon(x) \sum_n \bar{\phi}_n \gamma^5\, e^{-(i\not{D})^2/\Lambda^2} \phi_n \end{aligned} \tag{3.28}$$

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

$$\sum_n \bar{\phi}_n(x)\, \gamma^5 e^{+\not{D}^2/\Lambda^2} \phi_n(x) = \int \frac{d^4k}{(2\pi)^4}\, \mathrm{tr}\left(\gamma^5 e^{-ik\cdot x} e^{+\not{D}^2/\Lambda^2} e^{ik\cdot x}\right) \tag{3.29}$$

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

  $$\mathrm{tr}\,\gamma^5 = \mathrm{tr}\,\gamma^5\gamma^\mu\gamma^\nu = 0$$

  However, if we multiply all five (Euclidean) gamma matrices together we get the identity matrix. This is captured by the expression

  $$\mathrm{tr}\,\gamma^5\gamma^\mu\gamma^\nu\gamma^\rho\gamma^\sigma = 4\epsilon^{\mu\nu\rho\sigma}$$

  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 $e^{\not{D}^2/\Lambda^2}$. 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

$$\begin{aligned}\not{D}^2 = \gamma^\mu\gamma^\nu\mathcal{D}_\mu\mathcal{D}_\nu &= \frac{1}{2}\{\gamma^\mu,\gamma^\nu\}\mathcal{D}_\mu\mathcal{D}_\nu + \frac{1}{2}[\gamma^\mu,\gamma^\nu]\mathcal{D}_\mu\mathcal{D}_\nu \\ &= \mathcal{D}^2 + \frac{1}{4}[\gamma^\mu,\gamma^\nu][\mathcal{D}_\mu,\mathcal{D}_\nu] \\ &= \mathcal{D}^2 - \frac{ie}{2}\gamma^\mu\gamma^\nu F_{\mu\nu}\end{aligned}$$

The second is

$$e^{-ik\cdot x}\mathcal{D}_\mu e^{+ik\cdot x} = \mathcal{D}_\mu + ik_\mu$$

Combining these, we have

$$\begin{aligned}e^{-ik\cdot x}e^{\not{D}^2/\Lambda^2}e^{ik\cdot x} &= e^{-ik\cdot x}e^{\mathcal{D}^2/\Lambda^2 - \frac{ie}{2}\gamma^\mu\gamma^\nu F_{\mu\nu}/\Lambda^2}e^{ik\cdot x} \\ &= e^{(\mathcal{D}_\mu+ik_\mu)^2/\Lambda^2 - \frac{ie}{2}\gamma^\mu\gamma^\nu F_{\mu\nu}/\Lambda^2} \\ &= e^{(\mathcal{D}_\mu+ik_\mu)^2/\Lambda^2}\, e^{-\frac{ie}{2}\gamma^\mu\gamma^\nu F_{\mu\nu}/\Lambda^2}\, e^{\cdots}\ldots\end{aligned} \tag{3.30}$$

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

$$\begin{aligned}\sum_n \bar{\phi}_n\gamma^5\phi_n &= \lim_{\Lambda\to\infty}\int \frac{d^4k}{(2\pi)^4}\ \mathrm{tr}\left(\gamma^5 e^{-ik\cdot x}e^{+\not{\partial}^2/\Lambda^2}e^{ik\cdot x}\right)\\ &= \lim_{\Lambda\to\infty}\int \frac{d^4k}{(2\pi)^4}\ e^{-k^2/\Lambda^2}\ \left(\frac{e^2}{2}\epsilon^{\mu\nu\rho\sigma}F_{\mu\nu}F_{\rho\sigma}\frac{1}{\Lambda^4} + \ldots\right)\\ &= \frac{e^2}{32\pi^2}\epsilon^{\mu\nu\rho\sigma}F_{\mu\nu}F_{\rho\sigma}\end{aligned} \tag{3.32}$$

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,

$$S = \int d^4x\; i\bar{\psi}\, \not{\partial} \psi$$

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:

$$\Gamma^{\mu\nu\rho}(x_1, x_2, x_3) = \langle 0|T(j^\mu(x_1)\, j^\nu(x_2)\, j_A^\rho(x_3))|0\rangle$$

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_A^μ 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,

$$p_{1\mu}\Gamma^{\mu\nu\rho}(p_1, p_2, q) = 0 \tag{3.37}$$

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

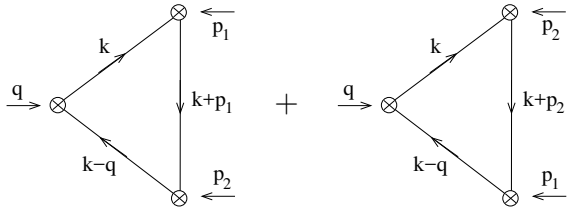
$$q_\rho \Gamma^{\mu\nu\rho}(p_1, p_2, q) = 0 \quad \Leftrightarrow \quad -(p_{1\rho} + p_{2\rho})\Gamma^{\mu\nu\rho}(p_1, p_2, q) = 0 \tag{3.38}$$

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,

$$-i\Gamma^{\mu\nu\rho}(p_1, p_2, q) = \quad + \tag{3.39}$$



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

$$-iq_\rho\Gamma^{\mu\nu\rho}(p_1, p_2, q) = i\int \frac{d^4k}{(2\pi)^4}\ \mathrm{tr}\left[\frac{1}{\not{k}}\left(\gamma^5(\not{k}-\not{q}) + \not{k}\gamma^5\right)\frac{1}{\not{k}-\not{q}}\gamma^\nu\frac{1}{\not{k}+\not{p}_1}\gamma^\mu\right] + \begin{pmatrix} p_1 \leftrightarrow p_2 \\ \mu \leftrightarrow \nu \end{pmatrix}$$

$$= i \int \frac{d^4k}{(2\pi)^4} \text{ tr} \left[\left(\frac{1}{\not{k}} \gamma^5 + \gamma^5 \frac{1}{\not{k} - \not{q}} \right) \gamma^\nu \frac{1}{\not{k} + \not{p}_1} \gamma^\mu \right. \\ \left. + \left(\frac{1}{\not{k}} \gamma^5 + \gamma^5 \frac{1}{\not{k} - \not{q}} \right) \gamma^\mu \frac{1}{\not{k} + \not{p}_2} \gamma^\nu \right]$$

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

$$\Delta_1^{\mu\nu} = i \int \frac{d^4k}{(2\pi)^4} \text{ tr} \left[\frac{1}{\not{k}} \gamma^5 \gamma^\nu \frac{1}{\not{k} + \not{p}_1} \gamma^\mu + \gamma^5 \frac{1}{\not{k} - \not{q}} \gamma^\mu \frac{1}{\not{k} + \not{p}_2} \gamma^\nu \right] \\ = i \int \frac{d^4k}{(2\pi)^4} \text{ tr} \left[\frac{1}{\not{k}} \gamma^5 \gamma^\nu \frac{1}{\not{k} + \not{p}_1} \gamma^\mu - \frac{1}{\not{k} + \not{p}_2} \gamma^5 \gamma^\nu \frac{1}{\not{k} - \not{q}} \gamma^\mu \right]$$

and

$$\Delta_2^{\mu\nu} = i \int \frac{d^4k}{(2\pi)^4} \text{ tr} \left[\gamma^5 \frac{1}{\not{k} - \not{q}} \gamma^\nu \frac{1}{\not{k} + \not{p}_1} \gamma^\mu + \frac{1}{\not{k}} \gamma^5 \gamma^\mu \frac{1}{\not{k} + \not{p}_2} \gamma^\nu \right] \\ = i \int \frac{d^4k}{(2\pi)^4} \text{ tr} \left[-\frac{1}{\not{k} + \not{p}_1} \gamma^5 \gamma^\mu \frac{1}{\not{k} - \not{q}} \gamma^\nu + \frac{1}{\not{k}} \gamma^5 \gamma^\mu \frac{1}{\not{k} + \not{p}_2} \gamma^\nu \right]$$

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

$$\tilde{\Delta} = -i \int \frac{d^4k}{(2\pi)^4} \left[a^\mu \partial_{k^\mu} f + \frac{1}{2} a^\mu a^\nu \partial_{k^\mu} \partial_{k^\nu} f + \dots \right]$$

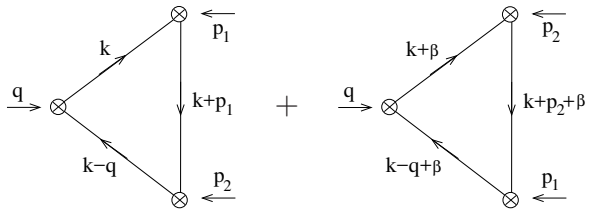
Each term above is a boundary term. Moreover, each term in the expansion is less and less divergent. If the original integral is only linearly divergent we need keep only the first of these terms. We have

$$\tilde{\Delta} = -i \int_{\mathbf{S}^3_\infty} \frac{d\hat{k}_\mu}{(2\pi)^4}\ a^\mu\, |k|^3 f(k) \tag{3.41}$$

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 configu-

ration. 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

$$\begin{aligned}\int \mathcal{D}\psi\mathcal{D}\bar{\psi} \; \exp\left(-\int d^4x \; i\bar{\psi}\not{D}\psi\right) &= \prod_n \int da_n d\bar{b}_n \; \exp\left(\sum_m \lambda_m \bar{b}_m a_m\right) \\ &= \prod_n \int da_n d\bar{b}_n \; \prod_m (1 + \lambda_m \bar{b}_m a_m)\end{aligned}$$

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

$$m(z) \to \begin{cases} +m & \text{as } z \to \infty \\ -m & \text{as } z \to -\infty \end{cases}$$

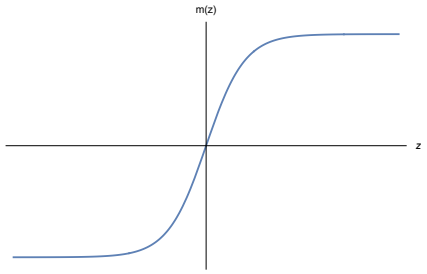


Figure 28:

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.

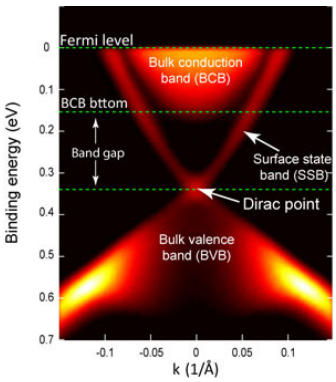


Figure 29:

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

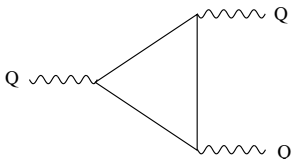


Figure 30:

$$\sum_{a=1}^{N} Q_a^3 = 0 \tag{3.57}$$

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

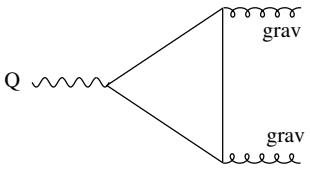


Figure 31:

$$\sum_{a=1}^{N} Q_a = 0 \tag{3.61}$$

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

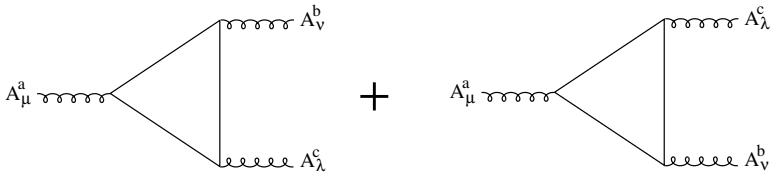
$$\{Q_a\} = \{1, 5, -7-8, 9\}$$

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 $\mathrm{tr}\, A = \mathrm{tr}\, A^T$. We learn that for any real or pseudoreal representation $\mathrm{tr}\, T^a\{T^b, T^c\} = 0$. 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 $\mathrm{tr}\, T^a\{T^b, T^c\}$ 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 $\begin{smallmatrix}\square\\\square\end{smallmatrix}$ representations are:

$$A(\square\square) = N + 4 \quad \text{and} \quad A(\begin{smallmatrix}\square\\\square\end{smallmatrix}) = N - 4$$

From this, we learn that we can construct a number of chiral gauge theories by taking, for $N \geq 5$,

$$G = SU(N) \text{ with a } \begin{smallmatrix}\square\\\square\end{smallmatrix} \text{ and } N-4\ \overline{\square} \text{ Weyl fermions}$$

where $\overline{\square}$ is shorthand for the anti-fundamental. Alternatively, we could have, for $N \geq 3$,

$$G = SU(N) \text{ with a } \square\square \text{ and } N+4\ \overline{\square}$$

or

$$G = SU(N) \text{ with a } \square\square, \text{ a } \begin{array}{c}\square\\\square\end{array} \text{ and } 2N\ \overline{\square}$$

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

$$\sigma_i = -\sum_{k=1}^{i-1} kq_k + \sum_{k=1}^{n-1}(n-k)q_k \qquad i = 1, \ldots, N. \tag{3.68}$$

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

$$Z = \int \mathcal{D}A \ \det\left(i\not{D}\,\frac{1+\gamma^5}{2}\right) \ \exp\left(-\int d^4x \ \frac{1}{2g^2}\,\mathrm{tr}\, F^{\mu\nu}F_{\mu\nu}\right) \tag{3.71}$$

As we discussed in Section 3.3.1, $i\not{D}$ is a Hermitan 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.

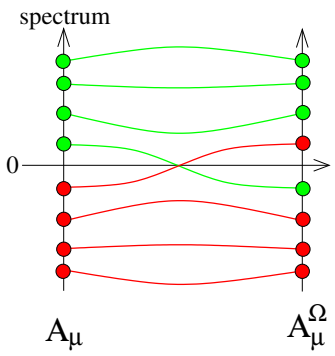


Figure 32:

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

$$\not{D}_5\Psi = \gamma^\tau \frac{\partial\Psi}{d\tau} + \not{D}\Psi$$

The operator $\not{D}_5$ 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 $\not{D}_5$ tells us that the determinant necessarily flips sign as in (3.73). Any zero mode of the $\not{D}$ 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 **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 $\sum_{\rm left} Y^3 - \sum_{\rm right} Y^3 = 0$. (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 $\sum_{\rm left} Y - \sum_{\rm right} Y = 0$ must vanish

$$U(1)\times{\rm gravity}^2: \quad \Big[2\times(-3)+6\Big]-\Big[-6+3\times 4+3\times(-2)\Big]=0$$

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,

$$q = l = x = 0 \quad \text{and} \quad u = -d \tag{3.81}$$

The second is, up to an overall rescaling, the charge assignment (3.77) seen in Nature,

$$x = 2l = -3(u + d) = -6q \quad \text{and} \quad u - d = \pm 6q$$

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

$$\begin{aligned}
\text{Left-Handed:}\quad & l_L:\ (\mathbf{2r},\mathbf{1})_{-N}\,,\quad q_L:\ (\mathbf{2r},\mathbf{N})_{+1}\\
\text{Right-Handed:}\quad & (e_\alpha)_R:\ (\mathbf{1},\mathbf{1})_{-2\alpha N}\,,\quad (\nu_\alpha)_R:\ (\mathbf{1},\mathbf{1})_{(2\alpha-2)N}\\
& (u_\alpha)_R:\ (\mathbf{1},\mathbf{N})_{1+(2\alpha-1)N}\,,\quad (d_\alpha)_R:\ (\mathbf{1},\mathbf{N})_{1-(2\alpha-1)N}
\end{aligned}$$

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.

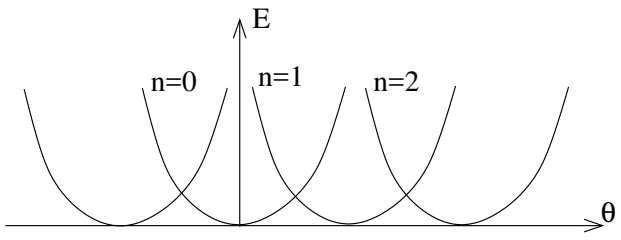


Figure 33: The energy spectrum for a particle moving around a solenoid.

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

$$U(1) \to \mathbf{Z}_2$$

This means that at $\theta = 0$ and $\theta = \pi$ we have two discrete symmetries: $T_\pi : x \to x + \pi$ and $P : x \to -x$.

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 $d^*J = 0$.

(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 $G = SU(N) \to \mathbf{Z}_N$, 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

$$\begin{aligned} \phi &\to \phi + N\alpha \quad \text{and} \quad A \to A + d\alpha \\ \hat{A} &\to \hat{A} + N\lambda \quad \text{and} \quad B \to B + d\lambda \end{aligned}$$

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

$$Z \to Z + pd\lambda \tag{3.96}$$

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].