Preprint typeset in JHEP style - HYPER VERSION 2018

Gauge Theory

David Tong

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

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

Contents

For Mum.

Acknowledgements, Caveats and Apologies

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

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

Pre-Requisites

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

Recommended Books and Resources

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

- John Preskill, Lectures on Quantum Field Theory

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

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

- Sidney Coleman, *Aspects of Symmetry*

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

- Alexander Polyakov, *Gauge, Fields and Strings*

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

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

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

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

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

- Eduardo Fradkin, *Field Theories in Condensed Matter Physics*
- Shankar, *Quantum Field Theory and Condensed Matter*

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

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

0. Introduction

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

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

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

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

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

is a theme that becomes particularly subtle when we include fermions in the mix, and see how they intertwine with the gauge fields.

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

1. Topics in Electromagnetism

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

1.1 Magnetic Monopoles

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

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

Here g is called the *magnetic charge*.

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

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

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

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

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

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

1.1.1 Dirac Quantisation

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

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

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

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

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

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

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

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

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

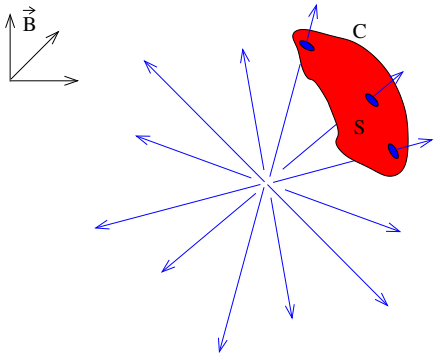

Figure 1: Integrating over S...

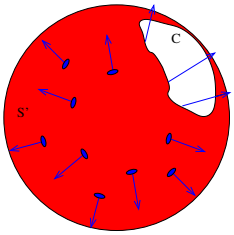

Figure 2: ...or over S'.

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

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

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

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

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

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

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

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

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

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

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

1.1.2 A Patchwork of Gauge Fields

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

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

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

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

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

The resulting magnetic field is

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

Substituting in (1.5) gives

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

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

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

Now consider a different gauge connection

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

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

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

$$A_\phi^N = A_\phi^S + \frac{1}{r\sin\theta}\,\partial_\phi\omega \quad \text{where} \quad \omega = \frac{g\phi}{2\pi} \tag{1.8}$$

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

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

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

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

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

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

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

1.1.3 Monopoles and Angular Momentum

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

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

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

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

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

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

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

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

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

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

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

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

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

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

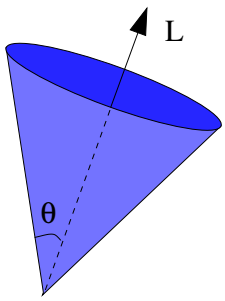

Figure 3:

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

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

Once again, we find the Dirac quantisation condition.

On Bosons and Fermions

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

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

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

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

1.2 The Theta Term

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

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

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

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

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

We start by defining the dual tensor

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

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

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

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

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

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

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

Axion Electrodynamics

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

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

The equations of motion from this action read

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

where

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

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

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

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

1.2.1 The Topological Insulator

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

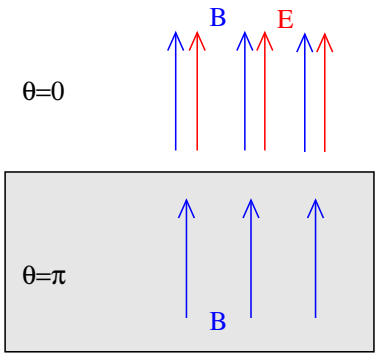

Figure 4: Applying a magnetic field

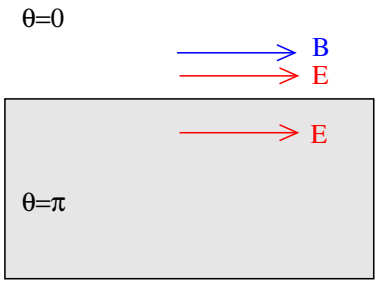

Figure 5: Applying an electric field

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

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

$$\sigma = \int_{-\epsilon}^{\epsilon} d^2x \ \rho = \alpha c B$$

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

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

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

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

The creation of a two-dimensional current which lies perpendicular to an applied electric field is called the *Hall effect*. The coefficient of proportionality is known as the *Hall conductivity* and there is a long and beautiful story about how it takes certain

very special values which are rational multiples of $e^2/2\pi\hbar$. (More details can be found in the lecture notes on the Quantum Hall Effect.) In the present example (1.14), the Hall conductivity is

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

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

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

Continuity Conditions

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

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

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

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

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

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

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

Now we can use the standard arguments (involving Gaussian pillboxes and line integrals) that tell us $\mathbf{B}$ perpendicular to a surface and $\mathbf{E}$ tangential to a surface are

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

necessarily continuous. This means that if we introduce the normal vector to the surface $\hat{\mathbf{n}}$, then

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

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

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

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

1.2.2 A Mirage Monopole

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

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

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

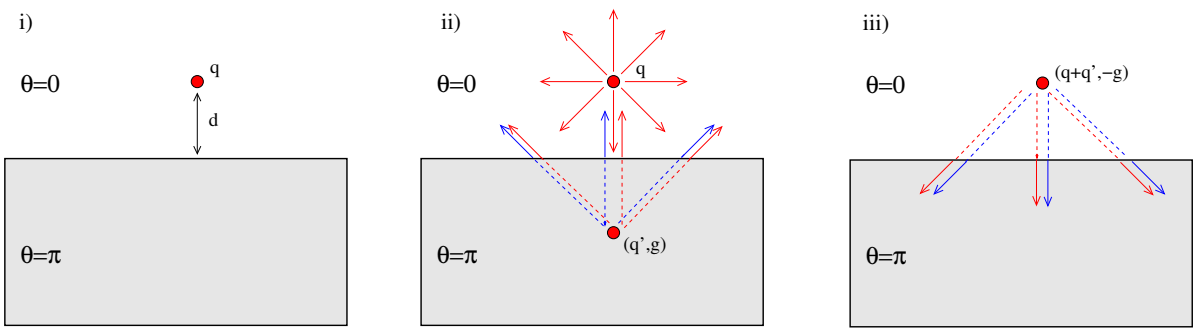

Figure 6: i) An electric charge q placed near a topological insulator. ii) The resulting electric and magnetic field lines as seen outside. iii) The field lines as seen inside.

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

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

and

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

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

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

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

and

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

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

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

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

and

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

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

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

Similarly, the magnetic field tangent to the interface is

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

and

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

which gives us

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

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

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

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

1.2.3 The Witten Effect

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

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

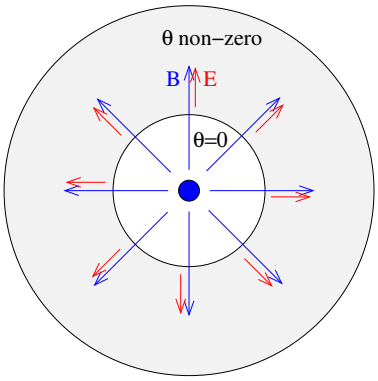

Figure 7:

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

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

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

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

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

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

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

by a time-varying electric field which lies parallel to $\mathbf{B}$. At the end of this process, the final electric field will be

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

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

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

1.2.4 Why θ is Periodic

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

We can work in a gauge where $A_1 = 0$, so that $B = \partial_1 A_2$. If we want a uniform, constant B then we need to write $A_2 = Bx^1$. This isn't single valued. However, that

needn't be a problem because, as we've seen above, A_2 is actually a periodic variable with periodicity $\hbar/eR$. This means that we're perfectly at liberty to write $A_2 = Bx_1$, but only if this has the correct period (1.22). This holds provided

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

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

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

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

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

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

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

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

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

which is our promised result (1.20)

The above explanation was rather laboured. It's pretty straightforward to generalise it to non-constant $\mathbf{E}$ and $\mathbf{B}$ fields. If you're mathematically inclined, it is the statement that the second Chern number of a $U(1)$ bundle is integer valued and, as we have seen above, is actually equal to the product of two first Chern numbers. Finally note that,

although we took Euclidean spacetime to be a torus T^4, the end result does not depend on the volume of the torus which is set by R. Nonetheless, the introduction of the torus was crucial in our argument: we needed the circles of T^4 to exploit the fact that $\Pi_1(U(1)) \cong \mathbf{Z}$. We will see another derivation of this when we come to discuss the anomaly in Section 3.3.1.

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

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

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

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

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

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

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

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

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

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

and

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

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

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

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

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

1.3 Further Reading

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

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

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

These lectures are concerned with the theoretical structure of gauge theories. It is a subject whose history is inextricably bound with experimental discoveries in particle

physics and the development of the Standard Model. Our understanding of gauge theory took place slowly, over many decades, and involved many hundreds, if not thousands, of physicists.

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

Gauge Symmetry

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

Monopoles

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

It seems that Dirac held anti-matter and magnetic monopoles, both predictions made within a few pages of each other, on similar footing [55]. He returned to the subject of monopoles only once, in 1948, elaborating on the concept of the "Dirac string" [45]. But the spectacular experimental discovery of anti-matter in 1932, followed by a long,

fruitless wait in the search for monopoles, left Dirac disillusioned. Fifty years after his first paper, and seemingly unconvinced by theoretical arguments (like "monopoles are heavy"), he wrote in a letter to Salam [46],

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

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

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

The extension of the Dirac quantisation condition to dyons was made by Zwanziger in 1968 [234], while the idea of patching gauge fields is due to Wu and Yang [231]

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

Topological Insulators

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

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

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

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