2. Yang-Mills Theory

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

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

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

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

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

2.1 Introducing Yang-Mills

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.1.1 The Action

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

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

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

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

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

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

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

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

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

and noting that this obeys the identity

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

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

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

A Rescaling

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

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

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

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

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

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

In what follows, we will use the normalisation (2.8). This is the more useful choice in the quantum theory, where $S_{\rm YM}$ sits exponentiated in the partition function. One way to see this is to note that g^2 sits in the same place as $\hbar$ in the partition function. This already suggests that $g^2 \to 0$ will be a classical limit. Heuristically you should think that, for g^2 small, one pays a large price for field configurations that do not minimize the action; in this way, the path integral is dominated by the classical configurations. In contrast, when $g^2 \to \infty$, the Yang-Mills action disappears completely. This is the strong coupling regime, where all field configurations are unsuppressed and contribute equally to the path integral.

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

2.1.2 Gauge Symmetry

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

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

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

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

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

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

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

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

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

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

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

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

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

The redundancy inherent in the gauge symmetry means that only gauge independent quantities should be considered physical. These are the things that do not depend on our underlying choice of description. In general relativity, we would call such objects "coordinate independent", and it's not a bad metaphor to have in mind for Yang-Mills. It's worth pointing out that in Yang-Mills theory, the "electric field" $E_i = F_{0i}$ and the "magnetic field" $B_i = -\frac{1}{2}\epsilon_{ijk}F_{jk}$ are not gauge invariant as they transform as (2.12). This, of course, is in contrast to electromagnetism where electric and magnetic fields are physical objects. Instead, if we want to construct gauge invariant quantities we should work with traces such as $\mathrm{tr}\, F_{\mu\nu}F_{\rho\sigma}$ or the Wilson lines that we will describe below. (Note that, for simple gauge groups such as $SU(N)$, the trace of a single field strength vanishes: $\mathrm{tr}\, F_{\mu\nu} = 0$.)

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

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

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

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

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

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

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

2.1.3 Wilson Lines and Wilson Loops

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

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

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

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

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

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

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

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

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

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

where

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

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

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

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

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

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

Quantising the Colour Degree of Freedom

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

- $\kappa = 0$: The Hilbert space consists of a single state, $|0\rangle$. This is equivalent to putting a particle in the trivial representation of the gauge group.

- $\kappa = 1$: The Hilbert space consists of N states, $w_i^\dagger|0\rangle$. This describes a particle transforming in the fundamental representation of the $SU(N)$ gauge group.

- $\kappa = 2$: The Hilbert space consists of $\frac{1}{2}N(N+1)$ states, $w_i^\dagger w_j^\dagger|0\rangle$, transforming in the symmetric representation of the gauge group.

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

The Path Integral over the Colour Degrees of Freedom

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

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

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

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

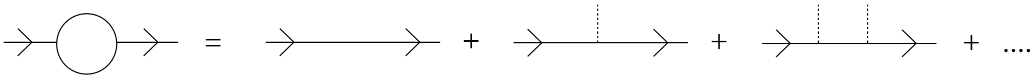
under a gauge transformation which seems bad. However, in the quantum theory it's not the action S_w that we have to worry about but e^{iS_w} because this is what appears in the path integral. And e^{iS_w} is gauge invariant provided that $\kappa \in \mathbf{Z}$.

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

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

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

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



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

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

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

2.2 The Theta Term

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

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

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

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

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

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

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

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

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

  where

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

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

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

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

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

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

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

2.2.1 Canonical Quantisation of Yang-Mills

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

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

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

$$A_0 = 0$$

With this ansatz, the Lagrangian becomes

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

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

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

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

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

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

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

From this we can build the Hamiltonian

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

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

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

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

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

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

Building the Hilbert Space

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

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

$$\mathcal{D}_iE_i\,|\text{phys}\rangle = 0 \qquad (2.29)$$

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

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

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

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

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

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

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

2.2.2 The Wavefunction and the Chern-Simons Functional

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

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

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

Meanwhile, the Schrödinger equation is

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

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

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

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

Now consider the following state

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

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

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

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

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

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

which gives us the following relation,

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

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

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

Is the Chern-Simons Functional Gauge Invariant?

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

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

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

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

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

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

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

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

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

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

Such maps fall into disjoint classes. This arises because the gauge transformations can "wind" around the spatial $\mathbf{S}^3$, in such a way that one gauge transformation cannot be continuously transformed into another. We'll meet this kind of idea a lot throughout these lectures. Such maps are characterised by homotopy theory. In general, we will be interested in the different classes of maps from spheres $\mathbf{S}^n$ into some space X. Two maps are said to be homotopic if they can be continuously deformed into each other.

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

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

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

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

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

An Example: $SU(2)$

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

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

$$\Omega_n(\mathbf{x}) = \exp\left(i\omega(r)\frac{\sigma_i \hat{x}^i}{2}\right) = \cos\left(\frac{\omega}{2}\right) + i \sin\left(\frac{\omega}{2}\right)\sigma^i \cdot \hat{x}^i \tag{2.37}$$

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

$$\omega(r) = \begin{cases} 0 & r = 0 \\ 4\pi n & r = \infty \end{cases}$$

Note that whenever ω is a multiple of 4π then $\Omega = e^{2\pi i \sigma_i \hat{x}^i} = 1$. This means that as we move out radially from the origin, the gauge transformation (2.37) is equal to the identity n times, starting at the origin and then on successive spheres $\mathbf{S}^2$ before it reaches the identity the final time at infinity $\mathbf{S}^2_\infty$. If we calculate the winding (2.36) of this map, we find

$$n(\Omega_n) = n$$

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

The Chern-Simons Functional is not Gauge Invariant!

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

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

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

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

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

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

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

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

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

2.2.3 Analogies From Quantum Mechanics

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

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

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

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

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

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

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

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

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

The Theta Angle as a "Hidden" Parameter

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

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

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

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

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

The eigenstates of (2.40) with this requirement are

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

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

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

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

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

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

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

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

Another Analogy: Bloch Waves

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

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

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

$$\mathbf{A} = iV\boldsymbol{\nabla}V^{-1} \qquad (2.42)$$

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

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

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

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

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

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

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

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

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

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

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

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

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

2.3 Instantons

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

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

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

where

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

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

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

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

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

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

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

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

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

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

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

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

2.3.1 The Self-Dual Yang-Mills Equations

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

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

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

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

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

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

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

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

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

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

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

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

A Single Instanton in $SU(2)$

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

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

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

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

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

Here the $\eta^i_{\mu\nu}$ are usually referred to as 't Hooft matrices. They are three 4×4 matrices which provide an irreducible representation of the $su(2)$ Lie algebra. They are given by

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

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

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

You can check that the associated field strength is

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

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

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

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

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

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

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

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

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

General Instanton Solutions

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

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

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

What about solutions with higher ν? There is a beautiful story here. It turns out that such solutions exist and have $4N\nu$ collective coordinates. Among these solutions are configurations which look like ν well separated instantons, each with $4N$ collective coordinates describing its position, scale size and orientation. However, as the instantons overlap this interpretation breaks down.

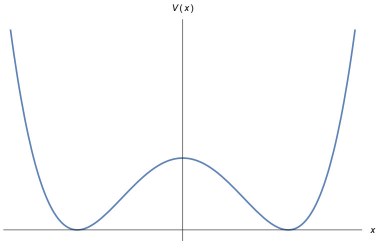


Figure 8: The double well

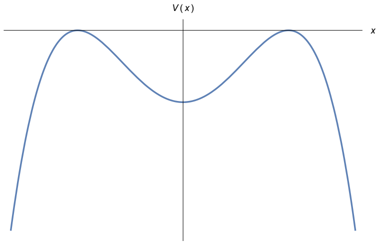


Figure 9: The upside down double well

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

2.3.2 Tunnelling: Another Quantum Mechanics Analogy

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

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

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

$$\tau = it$$

After this Wick rotation, the action

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

turns into the Euclidean action

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

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

As an example, consider the quartic potential

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

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

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

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

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

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

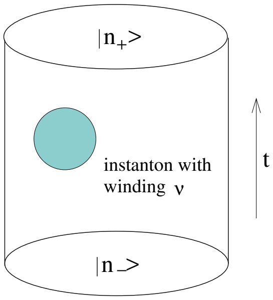


Figure 10:

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

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

2.3.3 Instanton Contributions to the Path Integral

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

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

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

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

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

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

and expand the Euclidean action as

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Back to Yang-Mills Instantons

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

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

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

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

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

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

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

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

2.4 The Flow to Strong Coupling

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Asymptotic freedom means that Yang-Mills theory is simple to understand at high energies, or short distance scales. Here it is a theory of massless, interacting gluon fields whose dynamics are well described by the classical equations of motion, together with quantum corrections which can be computed using perturbation methods. In particular, our discussion of instantons in Section 2.3 is valid at short distance scales. However, it becomes much harder to understand what is going on at large distances where the coupling gets strong. Indeed, the beta function (2.58) is valid only when $g^2(\mu) \ll 1$. This equation therefore predicts its own demise at large distance scales.

We can estimate the distance scale at which we think we will run into trouble. Taking the one-loop beta function at face value, we can ask: at what scale does $g^2(\mu)$ diverge? This happens at a finite energy

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

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

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

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

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

Viewed naively, there's something very surprising about the emergence of the scale Λ_{QCD}. This is because classical Yang-Mills has no dimensionful parameter. Yet the quantum theory has a physical scale, Λ_{QCD}. It seems that the quantum theory has generated a scale out of thin air, a phenomenon which goes by the name of *dimensional transmutation*. In fact, as the definition (2.59) makes clear, there is no mystery about this. Quantum field theories are not defined only by their classical action alone, but also by the cut-off Λ_{UV}. Although we might like to think of this cut-off as merely a crutch, and not something physical, this is misleading. It is not something we can do without. And it this cut-off which evolves to the physical scale Λ_{QCD}.

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

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

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

2.4.1 Anti-Screening and Paramagnetism

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

In condensed matter physics, materials are not boring passive objects. They contain mobile electrons, and atoms with a flexible structure, both of which can respond to any external perturbation, such as applied electric or magnetic fields. One consequence of this is an effect known as *screening*. In an insulator, screening occurs because an applied electric field will polarise the atoms which, in turn, generate a counteracting electric field. One usually describes this by introducing the electric displacement $\mathbf{D}$, related to the electric field through

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

where the permittivity $\epsilon = \epsilon_0(1 + \chi_e)$ with χ_e the electrical susceptibility. For all materials, $\chi_e > 0$. This ensures that the effect of the polarisation is always to reduce the electric field, never to enhance it. You can read more about this in Section 7 of the lecture notes on Electromagnetism.

(As an aside: In a metal, with mobile electrons, there is a much stronger screening effect which turns the Coulomb force into an exponentially suppressed Debye-Hückel, or Yukawa, force. This was described in the final section of the notes on Electromagnetism, but is not the relevant effect here.)

What does this have to do with quantum field theory? In quantum field theory, the vacuum is not a passive boring object. It contains quantum fields which can respond to any external perturbation. In this way, quantum field theories are very much like condensed matter systems. A good example comes from QED. There the one-loop beta function is positive and, at distances smaller than the Compton wavelength of the electron, the gauge coupling runs as

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

This tells us that the charge of the electron gets effectively smaller as we look at larger distance scales. This can be understood in very much the same spirit as condensed matter systems. In the presence of an external charge, electron-positron pairs will polarize the vacuum, as shown in the figure, with the positive charges clustering closer to the external charge. This cloud of electron-positron pairs shields the original charge, so that it appears reduced to someone sitting far away.

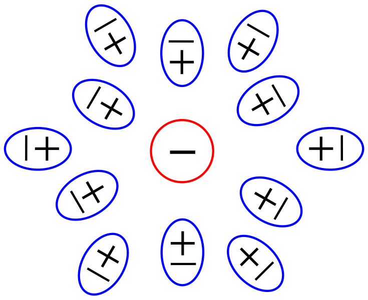


Figure 11:

The screening story above makes sense for QED. But what about QCD? The negative beta function tells us that the effective charge is now getting larger at long distances, rather than smaller. In other words, the Yang-Mills vacuum does not screen charge: it anti-screens. From a condensed matter perspective, this is unusual. As we mentioned above, materials always have $\chi_e > 0$ ensuring that the electric field is screened, rather than anti-screened.

However, there's another way to view the underlying physics. We can instead think about magnetic screening. Recall that in a material, an applied magnetic field induces dipole moments and these, in turn, give rise to a magnetisation. The resulting magnetising field $\mathbf{H}$ is defined in terms of the applied magnetic field as

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

with the permeability $\mu = \mu_0(1 + \chi_m)$. Here χ_m is the magnetic susceptibility and, in contrast to the electric susceptibility, can take either sign. The sign of χ_m determines the magnetisation of the material, which is given by $\mathbf{M} = \chi_m \mathbf{H}$. For $-1 < \chi_m < 0$, the magnetisation points in the opposite direction to the applied magnetic field. Such materials are called *diamagnets*. (A perfect diamagnet has $\chi_m = -1$. This is what happens in a superconductor.) In contrast, when $\chi_m > 1$, the magnetisation points in the same direction as the applied magnetic field. Such materials are called *paramagnets*.

In quantum field theory, polarisation effects can also make the vacuum either diamagnetic or paramagnetic. Except now there is a new ingredient which does not show up in real world materials discussed above: relativity! This means that the product must be

$$\epsilon \mu = 1$$

because "1" is the speed of light. In other words, a relativistic diamagnetic material will have $\mu < 1$ and $\epsilon > 1$ and so exhibit screening. But a relativistic paramagnetic material will have $\mu > 1$ and $\epsilon < 1$ and so exhibit anti-screening. Phrased in this way, the existence of an anti-screening vacuum is much less surprising: it follows simply from paramagnetism combined with relativity.

For free, non-relativistic fermions, we calculated the magnetic susceptibility in the lectures on Statistical Physics when we discussed Fermi surfaces. In that context, we found two distinct contributions to the magnetisation. Landau diamagnetism arose because electrons form Landau levels. Meanwhile, Pauli paramagnetism is due to the spin of the electron. These two effects have the same scaling but different numerical coefficients and one finds that the paramagnetism wins.

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

2.4.2 Computing the Beta Function

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

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

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

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

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

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

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

Gauge Fixing and Ghosts

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

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

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

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

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

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

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

We add to our action the gauge-fixing term

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

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

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

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

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

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

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

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

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

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

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

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

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

One-Loop Determinants

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

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

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

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

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

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

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

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

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

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

<u>The Ghost Contribution</u>

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

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

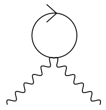
where the subscripts keep track of how many $\bar{A}_\mu$ terms each operator has,

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

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

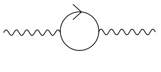
$$\begin{aligned}\text{Tr}\log\Delta_{\text{ghost}} &= \text{Tr}\log\left(-\partial^2+\Delta_1+\Delta_2\right)\\ &= \text{Tr}\log(-\partial^2) + \text{Tr}\log\left(1+(-\partial^2)^{-1}(\Delta_1+\Delta_2)\right)\\ &= \text{Tr}\log(-\partial^2) + \text{Tr}\left((-\partial^2)^{-1}(\Delta_1+\Delta_2)\right)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

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



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

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



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

with

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

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

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

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

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

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

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

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

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

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

The Gauge Contribution

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

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

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

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

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

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

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

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

The sum then gives the contribution to the effective action,

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

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

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

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

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

2.5 Electric Probes

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

2.5.1 Coulomb vs Confining

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

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

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

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

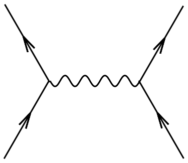


Figure 12:

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

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

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

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

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

An Aside on Group Theory

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

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

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

We then have

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

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

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

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

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

The two are related by

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

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

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

while the symmetric $\square\square$ and anti-symmetric $\substack{\square\\\square}$ representations have

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

Non-Abelian Coulomb Force

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

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

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

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

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

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

Perhaps surprisingly, this is a repulsive force.

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

$$\mathbf{N} \otimes \mathbf{N} = \square\square \oplus \substack{\square\\\square}$$

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

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

and

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

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

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

The Confining Force

In the previous section, we stated (but didn't prove!) that Yang-Mills has a mass gap. This means that, at distances $\gg 1/\Lambda_{QCD}$, the force will be due to the exchange of massive particles rather than massless particles. In many situations, the exchange of massive particles results in an exponentially suppressed Yukawa force, of the form $V(r) \sim e^{-mr}/r$, and you might have reasonably thought this would be the case for Yang-Mills. You would have been wrong.

Let's again consider a quark and an anti-quark, in the $\mathbf{N}$ and $\bar{\mathbf{N}}$ representations respectively. The energy between the two turns out to grow linearly with distance

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

for some value σ that has dimensions of energy per length. For reasons that we will explain shortly, it is often referred to as the *string tension*. On dimensional grounds, we must have $\sigma \sim \Lambda_{QCD}^2$ since there is no other scale in the game.

For two quarks, the result is even more dramatic. Now the tensor product of the two representations does not include a singlet (at least this is true for $SU(N)$ with $N \geq 3$). The energy between the two quarks turns out to be infinite. This is a general property of quantum Yang-Mills: the only finite energy states are gauge singlets. The theory is said to be *confining*: an individual quark cannot survive on its own, but is forced to enjoy the company of friends.

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

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

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

2.5.2 An Analogy: Flux Lines in a Superconductor

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

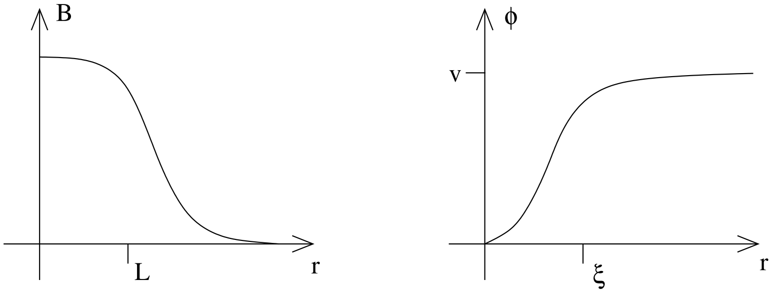


Figure 13: The profile for the magnetic field and Higgs field in a vortex.

We'll consider radially symmetric scalar profiles of the form

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

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

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

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

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

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

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

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

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

Type I, Type II and Bogomonlyi

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

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

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

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

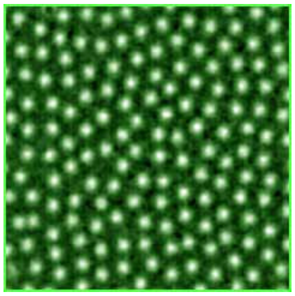


Figure 14: An Abrikosov lattice in a Type II superconductor.

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

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

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

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

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

[4] This picture is taken from P. Goa et al, Supercond. Sci. Technol. 14, 729 (2001). A nice gallery of vortex lattices can be found here.

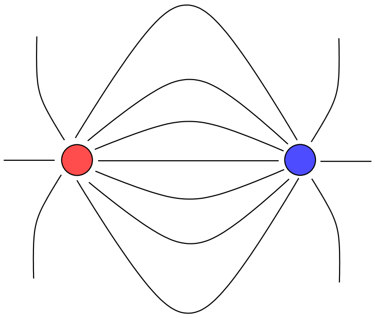


Figure 15: The flux lines for a monopole and anti-monopole in vacuum.

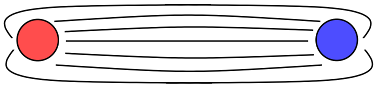


Figure 16: The same flux lines in a superconductor.

where we have related the flux to the winding using (2.71). This is nice. In a sector with winding $n > 0$, there is a minimum energy bound. Moreover, we can saturate this bound by requiring that the quantities in the squares vanish,

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

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

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

The Confinement of Monopoles

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

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

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

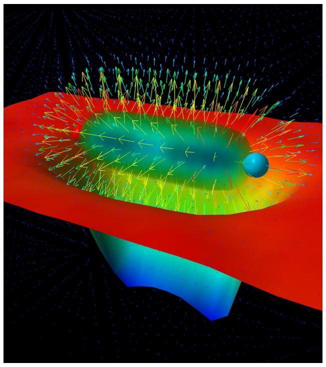


Figure 17: A simulation of a separated quark anti-quark pair in QCD.

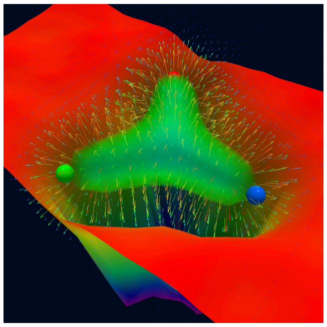


Figure 18: A simulation a separated baryon state in QCD.

described above As we have seen, happily the magnetic flux carried by a single vortex coincides with the magnetic flux emitted by a single Dirac monopole. The energy cost in separating the monopole and anti-monopole by a distance r is now

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

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

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

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

We might also wish to take away another lesson from the superconducting story. In the Abelian Higgs model, the electrically charged field ϕ condenses, resulting in the confinement of monopoles. Duality then suggests that to confine electrically charged

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

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

Regge Trajectories

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

$$E = p + \sigma r$$

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

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

For a fixed J, this is minimized at $r = \sqrt{J/\sigma}$, which gives us the relationship between the energy and angular momentum of the states,

$$E^2 \sim \sigma J$$

We can now compare this to the data for hadrons. A plot of the mass^2 vs spin is known as a Chew-Frautschi plot. It is shown on the right for light vector mesons[6]. We see that families of meson and their resonances do indeed sit on nice straight lines, referred to as *Regge trajectories*. The slope of the lines is determined by the QCD string tension, which turns out to be around $\sigma \sim 1.2\ GeV^2$. Perhaps more surprisingly, the data also reveals nice straight Regge trajectories in the baryon sector.

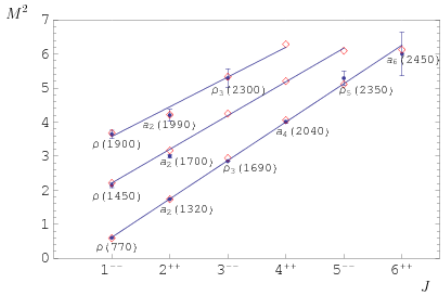


Figure 19:

2.5.3 Wilson Loops Revisited

Above we identified two different possible phases of Yang-Mills theory: the Coulomb phase and the confining phase. The difference between them lies in the forces experienced by two well-separated probe particles.

[6] This plot was taken from the paper by D. Ebert, R. Faustov and V. Galkin, arXiv:0903.5183.

- Coulomb: $V(r) \sim 1/r$
- Confining: $V(r) \sim r$

To this, we could add a third possibility that occurs when the gauge field is Higgsed, so that electric charges are completely screened. In this case we have

- Higgs: $V(r) \sim$ constant

We'll discuss this phase more in Section 2.7.3.

Usually in a quantum field theory (or in a statistical field theory) we identify the phase by computing the expectation value of some order parameter. The question that we would like to ask here is: what is the order parameter for confinement?

To answer this, we can rephrase our earlier discussion in terms of the path integral. To orient ourselves, let's first return to Maxwell theory. If we want to compute the path integral in the presence of an electrically charged probe particle, we simply introduce the particle by its associated current J^μ, which now acts as a source. We then add to the action the term $A_\mu J^\mu$. Moreover, for a probe particle which moves along a worldline C, the current J is a delta-function localised on C. We then compute the partition function with the insertion $e^{i\oint_C A}$,

$$\left\langle \exp\left(i\oint_C A\right)\right\rangle = \int \mathcal{D}A\ \exp\left(i\oint_C A\right)\, e^{iS_{\text{Maxwell}}} \tag{2.73}$$

where we're being a little sloppy on the right-hand-side, omitting both gauge fixing terms and the normalisation factor coming from the denominator.

In Yang-Mills, there is a similar story. The only difference is that we can't just stipulate a fixed current J^μ because the term $A_\mu J^\mu$ is not gauge invariant. Instead, we must introduce some internal colour degrees of freedom for the quark, as we described previously in Section 2.1.3. As we saw, integrating over these colour degrees of freedom leaves us with the Wilson loop $W[C]$, which we take in the fundamental representation

$$W[C] = \operatorname{tr}\mathcal{P}\exp\left(i\oint A\right)$$

Performing the further path integral over the gauge fields A leaves us with the expectation value of this Wilson loop

$$\Big\langle W[C]\Big\rangle = \int \mathcal{D}A\ \operatorname{tr}\mathcal{P}\exp\left(i\oint_C A\right) e^{iS_{YM}} \tag{2.74}$$

Now consider the specific closed loop C shown in the figure. We again take this to sit in the fundamental representation. It has the interpretation that we create a quark anti-quark pair, separated by a distance r, at some time in the past. These then propagate forward for time T, before they annihilate back to the vacuum.

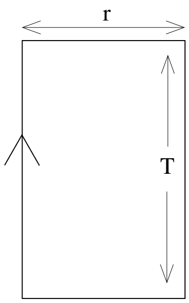


Figure 20:

What behaviour would we expect from the expectation value $\langle W[C]\rangle$? We'll work in Euclidean space. Recall from our earlier lectures on quantum field theory that, for long times, the path integral projects the system onto the lowest energy state. Before the quarks appear, and after they've gone, this is the ground state of the system which we can take to have energy zero. (Actually, you can take it to have any energy you like; its contribution will disappear from our analysis when we divide by the normalisation factor that missing on the right-hand-side of (2.73) and (2.74).) However, in the presence of the sources, the ground state of the system has energy $V(r)$. This means that we expect the Euclidean path integral to give

$$\lim_{r,T\to\infty}\Big\langle W[C]\Big\rangle \sim e^{-V(r)T}$$

This now gives us a way to test for the existence of the confining the phase directly in Yang-Mills theory. If the theory lies in the confining phase, we should find

$$\lim_{r,T\to\infty}\Big\langle W[C]\Big\rangle \sim e^{-\sigma A[C]} \tag{2.75}$$

where $A[C]$ is the area of the the loop C. This is known as the *area law criterion* for confinement. We won't be able to prove that Wilson loops in Yang-Mills exhibit an area law, although we'll offer an attempt in Section 4.2 when we discuss the strong coupling expansion of lattice gauge theory. We will have more success in Section 7 and 8 when we demonstrate confinement in lower dimensional gauge theories.

If a theory does not lie in the confining phase, we get different behaviour for the Wilson loop. For example, we could add scalar fields which condense and completely break the gauge symmetry. This is the *Higgs phase*, and we will discuss it in more detail in Section 2.7 where we first introduce dynamical matter fields. In the Higgs phase, we have

$$\lim_{r,T\to\infty}\Big\langle W[C]\Big\rangle \sim e^{-\mu L}$$

where $L = 2(r+T)$ is the perimeter of the loop and μ is some mass scale associated to the energy in the fields that screen the particle. This kind of perimeter law is characteristic of the screening phase of a theory.

Wilson Loops as Operators

There is a slightly different perspective on Wilson loops that will also prove useful: we can view them as operators on the Hilbert space of states. Since we are now dealing with Hilbert spaces and states, it's important that we are back in Lorentzian signature.

In quantum field theory, states are defined as living on a spacelike slice of the system. For this reason, we should first rotate our Wilson loop so that C is a spacelike, closed curve, sitting at a fixed point in time. The interpretation of the operator $W[C]$ is that it adds to the state a loop of electric flux along C. To see this, we can again revert to the canonical formalism that we introduced in Section 2.2. The electric field is $E^i = -i\delta/\delta A_i(x)$, so we have

$$E^i W[C] = \operatorname{tr}\mathcal{P}\left(\left[\frac{\delta}{\delta A_i(x)}\oint_C A\right] W[C]\right)$$

which indeed has support only on C.

The expectation value $\langle W[C]\rangle$ is now interpreted as the amplitude for a loop of electric flux $W[C]|0\rangle$ to annihilate to the vacuum $\langle 0|$. In the confining phase, this is unlikely because the flux tube is locally stable. The flux tube can, of course, shrink over time and disappear, but that's not what $\langle W[C]\rangle$ is measuring. Instead, it's looking for the amplitude that the flux tube instantaneously disappears. This can happen only through a tunnelling effect which, in Euclidean space, involves a string stretched across the flux tube acting. This Euclidean action of this string is proportional to its area, again giving $\langle W[C]\rangle \sim e^{-\sigma A}$ with $A[C]$ the minimal area bounding the curve.

In contrast, in the Higgs phase the string is locally unstable. Each part of the string can split into pieces and dissolve away. This is still unlikely: after all, it has to happen at all parts of the string simultaneously. Nonetheless, it is more likely than the corresponding process in the confining phase, and this is reflected in the perimeter law $\langle W[C]\rangle \sim e^{-\mu L}$.

2.6 Magnetic Probes

Much of our modern understanding of gauge theories comes from the interplay between electric and magnetic degrees of freedom. In the previous section we explored how Yang-Mills fields respond to electric probes. In this section, we will ask how they respond to magnetic probes.

A warning: the material in this section is a little more advanced than what we covered until now and won't be required for much of what follows. (An exception is Section 3.6 which discusses discrete anomalies and builds on the machinery we develop here.) In particular, sections 2.7 and 2.8 can both be read without reference to this section.

2.6.1 't Hooft Lines

Our first task is to understand how to construct an operator that corresponds to the insertion of a magnetic monopole. These are referred to as *'t Hooft lines*. For electric probes, we could build the corresponding Wilson line out of local fields A_μ. But there are no such fields that couple to magnetic charges. This means that we need to find a different way to describe the magnetic probes.

We will achieve this by insisting that the fields of the theory have a prescribed singular behaviour on a given locus which, in our case, will be a line C in spacetime. Because such operators disrupt the other fields in the theory, they are sometimes referred to as *disorder operators*.

't Hooft Lines in Electromagnetism

To illustrate this idea, we first describe 't Hooft lines in $U(1)$ electromagnetism. We have already encountered magnetic monopoles in Section 1.1. Suppose that a monopole of charge m traces out a worldline C in $\mathbf{R}^{3,1}$. (We referred to magnetic charge as g in Section 1.1, but this is now reserved for the Yang-Mills coupling so we have to change notation.) For any $\mathbf{S}^2$ that surrounds C, we then have

$$\int_{\mathbf{S}^2} \mathbf{B}\cdot d\mathbf{S} = m \tag{2.76}$$

We normalise the $U(1)$ gauge field to have integer electric charges. As explained in Section 1.1, the requirement that the monopole is compatible with these charges gives the Dirac quantisation condition (1.3), which now reads

$$e^{im} = 1 \quad \Rightarrow \quad m \in 2\pi\mathbf{Z} \tag{2.77}$$

For the magnetic field to carry flux (2.76), we must impose singular boundary conditions on the gauge field. As an example, suppose that we take the line C to sit at the spatial origin $\mathbf{x}=0$ and extend in the temporal direction t. Then, as explained in Section 1.1 we can cover the $\mathbf{S}^2$ by two charts. Working in polar coordinates with $A_r = 0$ gauge, in the northern hemisphere, we take the gauge field to have the singular behaviour

$$A_\phi \to \frac{m(1-\cos\theta)}{2r\,\sin\theta} \quad \text{as} \quad r\to 0$$

There is a similar condition (1.7) in the southern hemisphere, related by a gauge transformation.

We now define the 't Hooft line $T[C]$ by requiring that we take the path integral only over fields subject to the requirement that they satisfy (2.76) on C. This is a rather unusual definition of an "operator" in quantum field theory. Nonetheless, despite its unfamiliarity, , we can – at least in principle – use to compute correlation functions of $T[C]$ with other, more traditional operators.

't Hooft Lines in Yang-Mills

What's the analogous object in Yang-Mills theory with gauge group G. To explain the generalisation of Dirac quantisation to an arbitrary, semi-simple Lie group we need to invoke a little bit of Lie algebra-ology that was covered in the *Symmetries and Particles* course.

We work with a Lie algebra $\mathfrak{g}$. We denote the Cartan sub-algebra as $\mathbf{H} \subset \mathfrak{g}$. Recall that this is a set of r mutually commuting generators, where r is the rank of the Lie algebra. Throughout the rest of this section, bold (and not silly gothic) font will denote an r-dimensional vector.

We again define a 't Hooft line for a timelike curve C sitting at the origin. We will require that the magnetic field B^i, $i = 1, 2, 3$, takes the form

$$B^i \to \frac{x^i}{4\pi r^3} Q(x) \quad \text{as } r \to 0$$

where $Q(x)$ is a Lie algebra valued object which specifies the magnetic charge of the 't Hooft line. Spherical symmetry requires that $Q(x)$ be covariantly constant. We can again cover the $\mathbf{S}^2$ with two charts, and in each pick $Q(x)$ to be a constant which, by a suitable gauge transformation, we take to sit in the Cartan subalgebra. We write

$$Q = \mathbf{m} \cdot \mathbf{H}$$

for some r-dimensional vector $\mathbf{m}$ which determines the magnetic charge. We can think of this as r Dirac monopoles, embedded in the Cartan subalgebra.

The requirement that the 't Hooft lines are consistent in the presence of Wilson lines gives the generalised Dirac quantisation condition,

$$\exp\left(i\mathbf{m} \cdot \mathbf{H}\right) = 1 \tag{2.78}$$

The twist is that this must hold for all representations of the Lie algebra. To see why this requirement affects the allowed magnetic charges, consider the case of $G = SU(2)$.

We can pick a $U(1) \subset SU(2)$ in which we embed a Dirac monopole of charge m. The W-bosons have electric charge $q = \pm 1$ and are consistent with a 't Hooft line of charge $m = 2\pi$. However, our 't Hooft line should also be consistent with the insertion of a Wilson line in the fundamental representation, and this carries charge $q = \pm 1/2$. This means that, for $G = SU(2)$, the 't Hooft line must carry $m = 2$, twice the charge of the simplest Dirac monopole.

To extend this to a general group and representation, we need the concept of *weights*. Given a d dimensional representation, $|\mu_a\rangle$ with $a = 1, \ldots, d$ of $\mathfrak{g}$, we may introduce a set of weights, which are the eigenvalues

$$\mathbf{H}|\mu_a\rangle = \boldsymbol{\mu}_a|\mu_a\rangle \tag{2.79}$$

All such weights span the *weight lattice* $\Lambda_w(\mathfrak{g})$.

The weights of the adjoint representation are special and are referred to as *roots*. Recall that these roots $\boldsymbol{\alpha}$ can be used to label the other generators of the Lie algebra, which are denoted as $E_{\boldsymbol{\alpha}}$. In the adjoint representation, the eigenvalue condition (2.79) becomes the commutation relation $[\mathbf{H}, E_{\boldsymbol{\alpha}}] = \boldsymbol{\alpha} E_{\boldsymbol{\alpha}}$. Importantly, the roots also span a lattice

$$\Lambda_{\text{root}}(\mathfrak{g}) \subset \Lambda_w(\mathfrak{g})$$

The weights and roots have the property that

$$\frac{\boldsymbol{\alpha} \cdot \boldsymbol{\mu}}{\boldsymbol{\alpha}^2} \in \frac{1}{2}\mathbf{Z}$$

for all $\boldsymbol{\mu} \in \Lambda_w(\mathfrak{g})$ and $\boldsymbol{\alpha} \in \Lambda_{\text{root}}(\mathfrak{g})$. This is exactly what we need to solve the Dirac quantisation condition (2.78), which becomes $\mathbf{m} \cdot \boldsymbol{\mu} \in 2\pi\mathbf{Z}$ for all $\boldsymbol{\mu} \in \Lambda_w(\mathfrak{g})$. We define the *co-root*

$$\boldsymbol{\alpha}^\vee = \frac{2\boldsymbol{\alpha}}{\boldsymbol{\alpha}^2}$$

These co-roots also span a lattice, which we call $\Lambda_{\text{co-root}}(\mathfrak{g})$. Clearly, we have $\boldsymbol{\alpha}^\vee \cdot \boldsymbol{\mu} \in \mathbf{Z}$ for all $\boldsymbol{\alpha}^\vee \in \Lambda_{\text{co-root}}(\mathfrak{g})$ and $\boldsymbol{\mu} \in \Lambda_w(\mathfrak{g})$. If the magnetic charge vector sits in the co-root lattice, then the Dirac quantisation condition is obeyed. More generally, it turns out that for simply connected groups we have

$$\mathbf{m} \in 2\pi\, \Lambda_{\text{co-root}}(\mathfrak{g}) \tag{2.80}$$

This is sometimes referred to as the Goddard-Nuyts-Olive (or GNO) quantisation condition. We will look at the possible magnetic charges for non-simply connected groups shortly.

There is one last part of this story. The co-root lattice can be viewed as the root lattice for a Lie algebra $\mathfrak{g}^\vee$, so that $\Lambda_{\rm co-root}(\mathfrak{g}) = \Lambda_{\rm root}(\mathfrak{g}^\vee)$ For simply laced algebras (these are the ADE series, and so includes $su(N)$), all roots have the same length and are normalised to $\boldsymbol{\alpha}^2 = 2$. In this case, the roots and co-roots are the same and $\mathfrak{g}^\vee = \mathfrak{g}$. For non-simply laced groups, the long and short roots get exchanged. This means that, for example, $so(2N+1)^\vee = sp(N)$ and $sp(N)^\vee = so(2n+1)$.

2.6.2 $SU(N)$ vs $SU(N)/\mathbf{Z}_N$

There seems to be something of an imbalance between the Wilson line operators and the 't Hooft line operators. Of course, these electric and magnetic probes are defined in rather different ways, but that's not our concern. Instead, it's slightly disconcerting that there are more Wilson line operators than 't Hooft line operators. This is because Wilson line operators are labelled by representations R which, in turn, are associated to elements of the weight lattice $\Lambda_w(\mathfrak{g})$. In contrast, 't Hooft lines are labelled by elements of $\Lambda_{\rm root}(\mathfrak{g}^\vee)$ which is a subset of $\Lambda_w(\mathfrak{g}^\vee)$. Roughly speaking, this means that Wilson lines can sit in any representation, including the fundamental, while 't Hooft lines can only sit in representations that arise from tensor products of the adjoint. Why?

To better understand the allowed magnetic probes, we need to look more closely at the global topology of the gauge group. We will focus on pure Yang-Mills with $G = SU(N)$. Because the gauge bosons live in the adjoint representation, they are blind to any transformation which sits in the centre $\mathbf{Z}_N \subset SU(N)$,

$$\mathbf{Z}_N = \Big\{ e^{2\pi i k N},\ k = 0, 1, \ldots, N-1 \Big\}$$

The gauge bosons do not transform under this centre $\mathbf{Z}_N$ subgroup. In the older literature, it is sometimes claimed that the correct gauge group of Yang-Mills is actually $SU(N)/\mathbf{Z}_N$. But this is a bit too fast. In fact, the right way to proceed is to understand that there are two different Yang-Mills theories, defined by the choice of gauge group

$$G = SU(N) \quad \text{or} \quad G = SU(N)/\mathbf{Z}_N$$

Indeed, more generally we have a different theory with gauge group $G = SU(N)/\mathbf{Z}_p$ for any $\mathbf{Z}_p$ subgroup of $\mathbf{Z}_N$. The difference between these theories is rather subtle. We can't distinguish them by looking at the action, since this depends only on the shared $su(N)$ Lie algebra. Moreover, this means that the correlation functions of all local operators are the same in the two theories so you don't get to tell the difference by doing any local experiments. Nonetheless, different they are. The first place this shows up is in the kinds of operators that we can use to probe the theory.

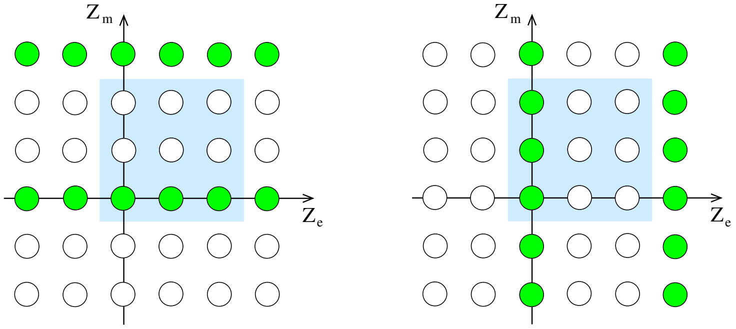


Figure 21: The figure on the left shows that allowed Wilson and 't Hooft lines (in green) for the gauge group $SU(3)$. The figure on the right shows the allowed lines for gauge group $SU(3)/\mathbf{Z}_3$.

Let's start with the Wilson lines. As we saw in Section 2.5, these are labelled by a representation of the group. The representations of $G = SU(N)/\mathbf{Z}_N$ are a subset of those of $G = SU(N)$; any representation that transforms non-trivially under $\mathbf{Z}_N$ is prohibited. This limits the allowed Wilson lines. In particular, the theory with $G = SU(N)/\mathbf{Z}_N$ does not admit the Wilson line in the fundamental representation, but Wilson lines in the adjoint representation are allowed. Similarly, the theory with gauge group $G = SU(N)/\mathbf{Z}_N$ cannot be coupled to fundamental matter; it can be coupled to adjoint matter.

This has a nice description in terms of the lattices that we introduced. For $G = SU(N)$, the representations are labelled by the weight lattice $\Lambda_w(\mathfrak{g})$. (The precise statement is that there is a one-to-one correspondence between representations and $\Lambda_w(\mathfrak{g})/W$ where W is the Weyl group.) However, for $G = SU(N)/\mathbf{Z}_N$, the representations are labelled by the root lattice $\Lambda_{\rm root}(\mathfrak{g})$. Indeed, the difference between the weight and root lattice for $\mathfrak{g} = su(N)$ is precisely the centre,

$$\Lambda_w(\mathfrak{g})/\Lambda_{\rm root}(\mathfrak{g}) = \mathbf{Z}_N$$

Now we come to the 't Hooft lines. When we introduced 't Hooft lines in the previous section, we were implicitly working with the universal cover of the gauge group, so that all possible Wilson lines were allowed. The requirement that magnetic charges are compatible with all representations and, in particular, the fundamental representation, resulted in the GNO condition (2.80) in which 't Hooft lines are labelled by $\Lambda_{\rm root}(\mathfrak{g})$. But what if we work with $G = SU(N)/\mathbf{Z}_N$? Now we have fewer Wilson lines, and so the demands of Dirac quantisation are less onerous. Correspondingly, in this theory the 't Hooft lines are labelled by $\Lambda_w(\mathfrak{g})$.

We can summarise the situation by labelling any line operator by a pair of integers

$$(z^e, z^m) \in \mathbf{Z}_N^e \times \mathbf{Z}_N^m \tag{2.81}$$

These describe how a given line operator transforms under the electric and magnetic centres of the group. If we have two line operators, labelled by (z^e, z^m) and (z'^e, z'^m) then Dirac quantisation requires $z^e z'^m - z^m z'^e = 0 \bmod N$. Note the similarity with the quantisation condition on dyons (1.4) that we met earlier.

For gauge group $G = SU(N)$, the line operators are labelled by $(z^e, 0)$ with $z^e = 0, \ldots, N-1$. Note that this doesn't mean that there are no magnetically charged 't Hooft lines: just that these lines sit in the root lattice and so have $z^m = 0 \bmod N$.

In contrast, for $G = SU(N)/\mathbf{Z}_N$ the line operators are labelled by $(0, z^m)$ with $z^m = 0, \ldots, N-1$. This time the Wilson lines must transform trivially under the centre of the group, so $z^e = 0 \bmod N$. The resulting line operators for $G = SU(3)$ and $G = SU(3)/\mathbf{Z}_3$ are shown in Figure 21. Yang-Mills with $G = SU(N)$ has more Wilson lines; Yang-Mills with $G = SU(N)/\mathbf{Z}_N$ has more 't Hooft lines.

There is a slightly more sophisticated way of describing these different line operators using the idea of generalised symmetries. We postpone this discussion until Section 3.6 where we will find an application in discrete anomalies.

The Theta Angle and the Witten Effect

The Witten effect gives rise to an interesting interplay between 't Hooft lines and the theta angle of Yang-Mills. Recall from Section 1.2.3, that a Dirac monopole of charge m in Maxwell theory picks up an electric charge proportional to the θ angle, given by

$$q = \frac{\theta m}{2\pi}$$

This analysis carries over to 't Hooft lines in both Maxwell and Yang-Mills theories. In the latter case, a shift of $\theta \to \theta + 2\pi$ changes the electric charge carried by a line operator,

$$\theta \to \theta + 2\pi \quad \Rightarrow \quad (z^e, z^m) \to (z^e + z^m, z^m)$$

For $G = SU(N)$, this maps the spectrum of line operators back to itself. However, for $G = SU(N)/\mathbf{Z}_N$ there is something of a surprise, because after a shift by 2π, the spectrum of line operators changes. This is shown in Figure 22 for $G = SU(3)/\mathbf{Z}_N$. We learn that the theory is *not* invariant under a shift of $\theta \to \theta + 2\pi$. Instead, to return to

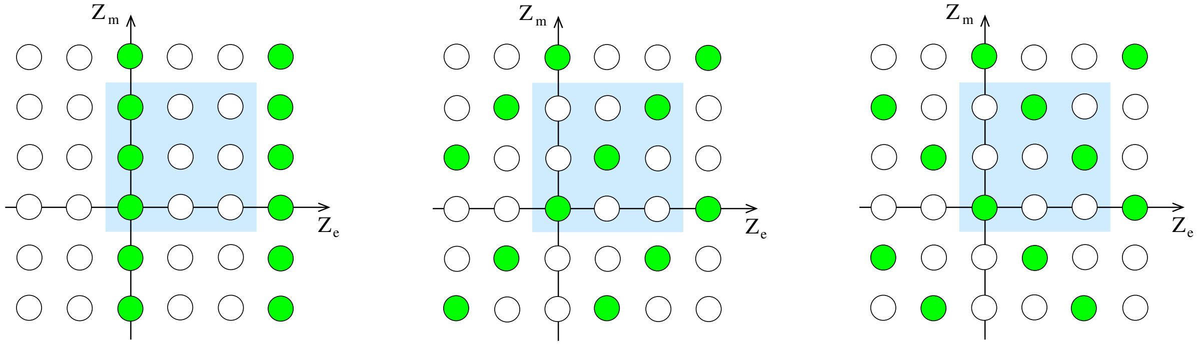


Figure 22: The spectrum of dyonic line operators in gauge group $SU(3)/\mathbf{Z}_3$, shown for $\theta = 0$ (on the left), $\theta = 2\pi$ (in the middle) and $\theta = 4\pi$ (on the right).

our original theory, with the same line operators, we must send $\theta \rightarrow \theta + 2\pi N$. In other words,

$$G = SU(N) \text{ has } \theta \in [0, 2\pi) \quad , \quad G = SU(N)/\mathbf{Z}_N \text{ has } \theta \in [0, 2\pi N)$$

We'll explore some consequences of this in Section 3.6 when we discuss anomalies in discrete symmetries.

One of the arguments we gave in Section 2.2 for the periodicity $\theta \in [0, 2\pi)$ was the appropriate quantisation of the topological charge $\int d^4x \; \mathrm{tr} \, {}^\star F^{\mu\nu} F_{\mu\nu}$. Instantons provide solutions to the equations of motion with non-vanishing topological charge. For Yang-Mills with $G = SU(N)/\mathbf{Z}_N$, the enlarged range of θ suggests that there might be "fractional instantons", configurations that carry $1/N^{\text{th}}$ the charge of an instanton. In fact, there are no such non-singular configurations on $\mathbf{R}^4$. But these fractional instantons do arise on manifolds with non-trivial topology. For example, if we take Euclidean spacetime to be $\mathbf{T}^4$, we can impose twisted boundary conditions in which, upon going around any circle, gauge fields come back to themselves up to a gauge transformation which lies in the centre $\mathbf{Z}_N$. Such boundary conditions are allowed for gauge group $G = SU(N)/\mathbf{Z}_N$, but not for $G = SU(N)$. One can show that these classes of configurations carry the requisite fractional topological charge.

't Hooft Lines as Order Parameters

One of the primary motivations for introducing line operators is to find order parameters that will distinguish between different phases of the theory. When $G = SU(N)$ we have the full compliment of Wilson lines. As we saw in Section 2.5, an area law for the fundamental Wilson loop signals that the theory lies in the confining phase, which is the expected behaviour for pure Yang-Mills. If we also add scalar fields to the theory, these could condense so that we sit in the Higgs phase; in this case the Wilson loop exhibits a perimeter law.

If the gauge group is $G = SU(N)/\mathbf{Z}_N$, we no longer have the fundamental Wilson line at our disposal. Instead, we have the fundamental 't Hooft line with $z^m = 1$, and this now acts as our order parameter. Since the local dynamics is independent of the global topology of the gauge group, pure Yang-Mills theory is again expected to confine. But, as in our discussion of superconductors in Section 2.5.2, the confinement of electric charge is equivalent to the screening of magnetic charge. This means that the signature of electric confinement is now a perimeter law for the 't Hooft line.

We can also consider $G = SU(N)/\mathbf{Z}_N$ Yang-Mills in the Higgs phase. The theory does not admit scalar fields in the fundamental representation, so we introduce adjoint scalars which subsequently condense. A single adjoint scalar will break the gauge group to its maximal torus, $U(1)^{N-1}$, but with two misaligned adjoint Higgs fields we can break the gauge symmetry completely. This is the Higgs phase. As described in Section 2.5.2, the Higgs phase can be thought of as confinement of magnetic charges. Correspondingly, the 't Hooft line now exhibits an area law.

That's All Well and Good, but...

The difference between Yang-Mills with $G = SU(N)$ and $G = SU(N)/\mathbf{Z}_N$ seems rather formal. As we mentioned above, all correlation functions of local operators in the two theories coincide, which means that any local experiment that we can perform will agree. The theories only differ in the kinds of non-local probes that we can introduce. You might wonder whether this is some pointless intellectual exercise.

If we consider Yang-Mills on flat $\mathbf{R}^{3,1}$, then there is some justification in ignoring these subtleties: the physics of the two theories is the same, and we're just changing the way we choose to describe it. However, even in this case these subtleties will help us say something non-trivial about the dynamics as we will see in Section 3.6 when we discuss discrete anomalies.

The real differences between the two theories arise when we study them on background manifolds with non-trivial topology. Here the two theories can have genuinely different dynamics. Perhaps the most straightforward case arises for Yang-Mills coupled to a single, massless adjoint Weyl fermion. This theory turns out to have supersymmetry and goes by the name of $\mathcal{N} = 1$ super Yang-Mills. Although supersymmetry is beyond the scope of these lectures, it turns out that it provides enough of a handle for us to make quantitative statements about their dynamics. If we consider these theories on spacetime $\mathbf{R}^{2,1} \times \mathbf{S}^1$, the low energy dynamics, specifically the number of ground states, does depend on the global topology of the gauge group.

2.6.3 What is the Gauge Group of the Standard Model?

We all know the answer to the question in the heading. The gauge group of the Standard Model is

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

Or is it?

The fermions in a single generation sit in the following representations of G,

$$\begin{array}{rrlcl} \text{Leptons:} & l_L: & (\mathbf{2},\mathbf{1})_{-3} & \Rightarrow & (z_2^e, z_3^e)_Y = (1,0)_{-3} \\ & e_R: & (\mathbf{1},\mathbf{1})_{-6} & \Rightarrow & (z_2^e, z_3^e)_Y = (0,0)_{-6} \\ \text{Quarks:} & q_L: & (\mathbf{2},\mathbf{3})_{+1} & \Rightarrow & (z_2^e, z_3^e)_Y = (1,1)_{+1} \\ & u_R: & (\mathbf{1},\mathbf{3})_{+4} & \Rightarrow & (z_2^e, z_3^e)_Y = (0,1)_{+4} \\ & d_R: & (\mathbf{1},\mathbf{3})_{-2} & \Rightarrow & (z_2^e, z_3^e)_Y = (0,1)_{-2} \end{array}$$

where the subscript denotes $U(1)_Y$ hypercharge Y, normalised so that $Y \in \mathbf{Z}$. We could add to this the right-handed neutrino ν_R which is a gauge singlet. In the table above, we have also written the charges z_2^e and z_3^e under the $\mathbf{Z}_2 \times \mathbf{Z}_3$ centre of $SU(2) \times SU(3)$. Finally, the Higgs boson sits in the representation $(\mathbf{2},\mathbf{1})_3 \;\Rightarrow\; (z_2^e, z_3^e)_Y = (1,0)_3$.

Each of these representations has the property that

$$Y = 3z_2^e - 2z_3^e \mod 6$$

This means that there is a $\mathbf{Z}_6$ subgroup of $G = U(1)_Y \times SU(2) \times SU(3)$ under which all the fields are invariant: we must simultaneously act with the $\mathbf{Z}_6 = \mathbf{Z}_2 \times \mathbf{Z}_3$ centre of $SU(2) \times SU(3)$, together with a $\mathbf{Z}_6 \subset U(1)_Y$. Because nothing transforms under this $\mathbf{Z}_6$ subgroup, you can sometimes read in the literature that the true gauge group of the Standard Model is

$$G = \frac{U(1)_Y \times SU(2) \times SU(3)}{\Gamma} \tag{2.82}$$

where $\Gamma = \mathbf{Z}_6$. But this is also too fast. The correct statement is that there is a fourfold ambiguity in the gauge group of the Standard Model: it takes the form (2.82), where Γ is a subgroup of $\mathbf{Z}_6$, i.e.

$$\Gamma = 1,\ \mathbf{Z}_2,\ \mathbf{Z}_3,\ \text{or } \mathbf{Z}_6$$

We note in passing that we can embed the Standard Model in a grand unified group, such as $SU(5)$ or $Spin(10)$, only if $\Gamma = \mathbf{Z}_6$.

As we mentioned above, the choice of Γ does not affect any local correlations functions and, in particular, does not affect physics at the LHC. Nonetheless, each choice of Γ defines a different theory and, in principle, the distinction could have observable consequences. One place that the difference in Γ shows up is in the magnetic sector. Previously we discussed the allowed 't Hooft lines. However, there is a folk theorem that when a quantum field theory is coupled to gravity then any allowed electric or magnetic charge has a realisation as a physical state. In other words, particles (or groups of particles) should exist with each of the allowed electric and magnetic charges.We'll see in Section 2.8 how magnetic monopoles can arise as dynamical particles in a non-Abelian gauge theory.

The arguments for this are far from rigorous and, for magnetic charges, boil down to the fact that an attempt to define an infinitely thin 't Hooft line in a theory coupled to gravity will result in a black hole. If we now let this black hole evaporate, and insist that there are no remnants, then it should spit out a particle with the desired magnetic charge.

So what magnetic monopoles are allowed for each choice of Γ? First, let's recall how electromagnetism arises from the Standard Model. The electromagnetic charge q of any particle is related to the hypercharge Y and the $SU(2)$ charge τ^3 by

$$q = -\frac{Y}{6} + \tau^3$$

This gives us the familiar electric charges: for the electron $q = -1$; for the up quark $q = +2/3$; and for the down quark $q = -1/3$.

We denote the magnetic charge under $U(1)_Y$ as m_Y. As we explained in Section 2.5.2, when a Higgs field condenses, many of the magnetically charged states are confined. In the Standard Model, those that survive must have

$$\frac{6m_Y}{2\pi} = z_2^m \mod 2$$

The magnetic charge under $U(1)_Y$ and $SU(2)$ then conspires so that these states are blind to the Higgs field. For such states, the resulting magnetic charge under electromagnetism is

$$m = 6m_Y$$

Now we're in a position to see the how the global structure of the gauge group affects the allowed monopole charge. Suppose that we take $\Gamma = 1$. Here, the monopoles must obey the Dirac quantisation condition with respect to each gauge group individually. This means that $m_Y \in 2\pi\mathbf{Z}$, and so the magnetic charge of any particle is quantised as $m \in 12\pi\mathbf{Z}$. This is six times greater than the magnetic charge envisaged by Dirac. Of course, Dirac only knew about the existence of the electron with charge $q = 1$. The quarks, together with the structure of the electroweak force, impose a more stringent constraint.

In contrast, if $\Gamma = \mathbf{Z}_6$, more magnetic charges are allowed. This is entirely analogous to the situation that we saw in the previous section. The Dirac quantisation condition now imposes a single constraint on the combined gauge charges from each factor of the gauge group,

$$3z_2^e z_2^m + 2z_3^e z_3^m - \frac{6Y m_Y}{2\pi} \in 6\mathbf{Z}$$

But this gives us more flexibility. Now we are allowed a magnetic monopole with $m_Y = \frac{1}{6} \times 2\pi$ provided that it also carries a magnetic charge under the other groups, $z_2^m = 1$ and $z_3^m = 1$. In other words, the Standard Model with $\Gamma = \mathbf{Z}_6$ admits the kind of magnetic monopole that Dirac would have expected, with $m = 2\pi$. Of course, this obeys Dirac quantisation with respect to the electron. But it also obeys Dirac quantisation with respect to the fractionally charged quarks because it carries a compensating non-Abelian magnetic charge.

2.7 Dynamical Matter

Until now, we have (mostly) focussed on pure Yang-Mills, without any additional, dynamical matter fields. It's time to remedy this. We will consider coupling either scalar fields, ϕ, or Dirac spinors ψ to Yang-Mills.

Each matter field must transform in a representation R of the gauge group G. In the Lagrangian, the information about our chosen representation is often buried in the covariant derivative, which reads

$$\mathcal{D}_\mu = \partial_\mu - iA_\mu^a T^a(R)$$

where $T^a(R)$ are the generators of the Lie algebra in the representation R. For scalar fields, the action is

$$S_{\text{scalar}} = \int d^4x\; \mathcal{D}_\mu\phi^\dagger \mathcal{D}^\mu\phi - V(\phi)$$

where $V(\phi)$ can include both mass terms and ϕ^4 interactions. For spinors, the action is

$$S_{\text{fermion}} = \int d^4x\; i\bar{\psi}\not{\mathcal{D}}\psi - m\bar{\psi}\psi$$

If we have both scalars and fermions then we can also include Yukawa interactions between them.

Our ultimate goal is to understand the physics described by non-Abelian gauge theories coupled to matter. What is the spectrum of excitations of these theories? How do these excitations interact with other? How does the system respond to various probes and sources? In this section, we will start to explore this physics.

2.7.1 The Beta Function Revisited

The first question we will ask is: how does the presence of these matter degrees of freedom affect the running of the gauge coupling $g^2(\mu)$? This is simplest to answer for massless scalars and fermions. Suppose that we have N_s scalars in a representation R_s and N_f Dirac fermions in a representation R_f. The 1-loop running of the gauge coupling is

$$\frac{1}{g^2(\mu)} = \frac{1}{g_0^2} - \frac{1}{(4\pi)^2}\left[\frac{11}{3}I(\text{adj}) - \frac{1}{3}N_s I(R_s) - \frac{4}{3}N_f I(R_f)\right]\log\left(\frac{\Lambda_{UV}^2}{\mu^2}\right) \qquad (2.83)$$

This generalises the Yang-Mills beta function (2.56). Recall that the Dynkin indices $I(R)$ are group theoretic factors defined by the trace normalisations,

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

and we are working in the convention in which $I(F) = \frac{1}{2}$ for the fundamental (or minimal) representation of any group.

When a field has mass m, it contributes the running of the coupling only at scales $\mu > m$, and decouples when $\mu < m$. There is a smooth crossover from one behaviour to the other at scales $\mu \sim m$, but the details of this will not be needed in these lectures.

Here we will briefly sketch the derivation of the running of the coupling, following Section 2.4.2. We will then look at some of the consequences of this result.

The Beta Function for Scalars

If we integrate out a massless, complex scalar field, we get a contribution to the effective action for the gauge field given by

$$S_{\text{eff}}[A] = \frac{1}{2g^2}\int d^4x\ \text{tr}F_{\mu\nu}F^{\mu\nu} + \text{Tr}\,\log(-\mathcal{D}^2)$$

But this is something we've computed before, since it is the same as the ghost contribution to the effective action. The only differences are that we get a plus sign instead of a minus sign, because our scalars are the sensible kind that obey spin statistics, and that we pick up the relevant trace coefficient $I(R)$, as opposed to $I(\text{adj})$ for the ghosts. We can then immediately import our results from Section 2.4.2 to get the scalar contribution in (2.83)

The Beta Function for Fermions

If we integrate out a massless Dirac fermion, we get a contribution to the effective action for the gauge field given by

$$S_{\text{eff}}[A] = \frac{1}{2g^2}\int d^4x\ \text{tr} F_{\mu\nu}F^{\mu\nu} - \log\det(i\not{\mathcal{D}})$$

To compute the determinant, it's useful to expand as

$$\begin{aligned}\det(i\not{\mathcal{D}}) &= \det{}^{1/2}(-\gamma^\mu\gamma^\nu\mathcal{D}_\mu\mathcal{D}_\nu)\\ &= \det{}^{1/2}\Big(\frac{1}{2}\{\gamma^\mu,\gamma^\nu\}\mathcal{D}_\mu\mathcal{D}_\nu - \frac{1}{2}[\gamma^\mu,\gamma^\nu]\mathcal{D}_\mu\mathcal{D}_\nu\Big)\\ &= \det{}^{1/2}\Big(-\mathcal{D}^2 + \frac{i}{4}[\gamma^\mu,\gamma^\nu]F_{\mu\nu}\Big)\end{aligned}$$

where, to go to the final line, we have used both the Clifford algebra $\{\gamma^\mu,\gamma^\nu\} = 2\delta^{\mu\nu}$, as well as the fact that $[\mathcal{D}_\mu,\mathcal{D}_\nu] = -iF_{\mu\nu}$. The contribution to the effective action is then

$$\begin{aligned}-\log\det(i\not{\mathcal{D}}) &= -\frac{1}{2}\text{Tr}\ \log\Big(-\mathcal{D}^2\mathbf{1}_4 + \frac{i}{4}[\gamma^\mu,\gamma^\nu]F_{\mu\nu}\Big)\\ &= -2\text{Tr}\log(-\mathcal{D}^2) + [\gamma^\mu,\gamma^\nu]F_{\mu\nu}\ \text{terms}\end{aligned}$$

Here the $\frac{1}{2}$ has changed into a 2 after tracing over the spinor indices. We're left having to compute the contribution from the $[\gamma^\mu,\gamma^\nu]F_{\mu\nu}$ terms. This is very similar in spirit to the extra term that we had to compute for the gauge fluctuations in Section 2.4.2. However, the difference in spin structure means that it differs from the gauge contribution by a factor of 1/2. The upshot is that we have

$$-\log\det(i\not{\mathcal{D}}) = -\frac{1}{2}\left[\frac{4}{3}-4\right]\frac{T(R)}{(4\pi)^2}\int\frac{d^4k}{(2\pi)^4}\ \text{tr}\left[\bar{A}_\mu(k)\bar{A}_\nu(-k)\right](k^\mu k^\nu - k^2\delta^{\mu\nu})\log\left(\frac{\Lambda_{UV}^2}{k^2}\right)$$

which gives the fermionic contribution to the running of the gauge coupling in (2.83). Note that, once again, contributions from the extra spin term (the -4) overwhelm the contribution from the kinetic term (the $+4/3$). But, because we are dealing with fermions, there is an overall minus sign. This means that fermions, like scalars, give a positive contribution to the beta function.

2.7.2 The Infra-Red Phases of QCD-like Theories

We will start by ignoring the scalars and considering non-Abelian gauge theories coupled to fermions. In many ways, this is the most subtle and interesting class of quantum field theories and we will devote Sections 3 and 5 to elucidating some of their properties. Here we start by giving a brief tour of what is expected from these theories.

Obviously, there are many gauge groups and representations that we could pick. We will restrict ourselves to gauge group $SU(N_c)$, where N_c is referred to as the number of *colours*. We will couple to this gauge field N_f Dirac fermions, each transforming in the fundamental representation of the gauge group. Here N_f is referred to as the number of *flavours*. We will further take the fermions to be massless, although we will comment briefly on what happens as they are given masses. This class of theories will be sufficient to exhibit many of the interesting phenomena that we care about. Moreover, this class of theories boasts QCD as one of its members (admittedly you should relax the massless nature of the quarks just a little bit.)

At one-loop, the running of the gauge coupling can be read off from (2.83)

$$\frac{1}{g^2(\mu)} = \frac{1}{g_0^2} - \frac{1}{(4\pi)^2}\left[\frac{11N_c}{3} - \frac{2N_f}{3}\right]\log\left(\frac{\Lambda_{UV}^2}{\mu^2}\right) \tag{2.84}$$

These theories exhibit different dynamics depending on the ratio N_f/N_c.

The Infra-Red Free Phase

Life is simplest when $N_f > 11N_c/2$. In this case, the contribution to the beta function from the matter overwhelms the contribution from the gauge bosons, and the coupling g^2 becomes weaker as we flow towards the infra-red. Such theories are said to be *infra-red free*. This means that, for once, we can trust the classical description at low energies, where we have weakly coupled massless gauge bosons and fermions.

The force between external, probe electric charges takes the form

$$V_{\text{electric}}(r) \sim \frac{1}{r\,\log(r\Lambda_{UV})}$$

which is Coulombesque, but dressed with the extra log term which comes from the running of the gauge coupling. This is the same kind of behaviour that we would get in (massless) QED. Meanwhile, the potential between two external magnetic charges takes the form

$$V_{\text{magnetic}} \sim \frac{\log(r\Lambda_{UV})}{r}$$

The log in the numerator reflects the fact that magnetic charges experience a force proportional to $1/g^2$ rather than g^2.

When $N_f = 11N_c/2$, the one-loop beta function vanishes. To see the fate of the theory, we must turn to the two-loop beta function which we discuss below. It will turn out that the theory is again infra-red free.

These theories are ill-defined in the UV, where there is a Landau pole. However, it's quite possible that theories of these types arise as the low-energy limit of other theories.

The Conformal Window

Next, consider N_f just below $11N_c/2$. To understand the behaviour of the theory, we can look at the two-loop contribution to the beta function,

$$\beta(g) = \mu\frac{dg}{d\mu} = \beta_0 g^3 + \beta_1 g^5 + \ldots$$

with the one-loop beta function extracted from (2.84)

$$\beta_0 = \frac{1}{(4\pi)^2}\left(-\frac{11N_c}{3} + \frac{2N_f}{3}\right)$$

We won't compute the two-loop beta function here, but just state the result:

$$\beta_1 = \frac{1}{(16\pi^2)^2}\left(-\frac{34N_c^2}{3} + \frac{N_f(N_c^2-1)}{N_c} + \frac{10N_fN_c}{3}\right)$$

Note that $\beta_1 > 0$ as long as the number of flavours sits in the range $N_f < 34N_c^3/(13N_c^2 - 3)$. But $\beta_0 < 0$ provided $N_f < 11N_c/2$ and so we can play the one-loop beta function against the two-loop beta function, to find a non-trivial fixed point of the RG flow, at which $\beta(g_\star) = 0$. This is given by

$$g_\star^2 = -\frac{\beta_0}{\beta_1}$$

Importantly, for $N_f/N_c = 11/2 - \epsilon$, with ϵ small, we have $g_\star^2 \ll 1$ and the analysis above can be trusted. We learn that the low-energy physics is described by a weakly coupled field theory which, as a fixed point of RG, is invariant under scale transformations. This is known as the *Banks-Zaks fixed point.* There is a general expectation (although not yet a complete proof) that relativistic theories in $d = 3+1$ which are scale invariant are also invariant under a larger conformal symmetry.

At any such fixed point, the scale invariance is enough to ensure that both external magnetic and electric probes experience a Coulomb force

$$V(r) \sim \frac{1}{r}$$

Such a phase could be described as a non-Abelian Coulomb phase, comprised of massless gluons and fermions.

What happens if we now lower N_f with fixed N_c? The formal result above says that the fixed point remains (at least until $N_f \approx 34N_c^3/(13N_c^2 - 3)$ but the value of the coupling $g_\star^2$ gets larger so that we can no longer trust the analysis. In general, we expect there to be a conformal fixed point for

$$N_\star < N_f < \frac{11N_c}{2} \tag{2.85}$$

for some critical value $N_\star$. This range of N_f is referred to as the *conformal window*. The obvious question is: what is the value of $N_\star$?

We don't currently know the answer to this question. At the lower end of the conformal window, the theory is necessarily strongly coupled which makes it difficult to get a handle on the physics. There is evidence from numerical work that when $N_c = 3$ (which is the case for QCD) then the lower end of the conformal window sits somewhere in the window $N_\star \in [8, 12]$, and probably closer to the middle than the edges. One would also expect the conformal to scale with N_c, so one could guess that $N_\star \approx 3N_c$ to $4N_c$. There are various arguments that give values of $N_\star$ in this range, but none of them are particularly trustworthy.

We've seen that there are a set of conformal fixed point, labelled by N_c and N_f in the range (2.85). We met such fixed points before in the course on Statistical Field Theory. In that context, we came across the powerful idea of *universality*: many different ultra-violet theories all flow to the same fixed point. This is responsible for the observation that all gases, regardless of their microscopic make-up, have exactly the same divergence in the heat capacity at their critical point. We could ask: is there a form of universality in gauge theories? In other words, can we write down two gauge theories which look very different in the ultra-violet, but nonetheless flow to the same infra-red fixed point?

We don't yet know of any examples of such universality in the QCD-like gauge theories that we discuss in these lectures, although this is most likely due to our ignorance. However, such examples are known in supersymmetric theories, which consist of gauge fields, scalars and fermions interacting with specific couplings. In that context, it is known that supersymmetric $SU(N_c)$ gauge theories coupled to N_f fundamental flavours flows to the same fixed point as $SU(N_f - N_c)$ gauge theory coupled to N_f flavours. (The latter flavours should also be a coupled to a bunch of gauge neutral fields.) Furthermore, the two descriptions can be identified as electric and magnetic variables for the system. This phenomenon is known as *Seiberg duality*. However, it is a topic for a different course.

Confinement and Chiral Symmetry Breaking

What happens when $N_f \leq N_\star$ and we are no longer in the conformal window? The expectation is that for $N_f < N_\star$ the coupling is once again strong enough to lead to confinement, in the sense that all finite energy excitations are gauge singlets.

Most of the degrees of freedom will become gapped, with a mass that is set parametrically by $\Lambda_{QCD} = \mu e^{1/2\beta_0 g^2(\mu)}$. However, there do remain some massless modes. These occur because of the formation of a vacuum condensate

$$\langle \bar{\psi}_i \psi_j \rangle \sim \delta_{ij} \qquad i, j = 1, \ldots, N_f$$

This spontaneously breaks the global symmetry of the model, known as the chiral symmetry. The result is once again a gapless phase, but now with the massless fields arising as Goldstone bosons. We will have a lot to say about this phase. We will say it in Section 5.

For pure Yang-Mills, we saw in Section 2.5 that a Wilson line, $W[C] = \operatorname{tr} \mathcal{P} \exp \left(i \oint A \right)$ in the fundamental representation provides an order parameter for the confining phase, with the area law, $\langle W[C] \rangle \sim e^{-\sigma A}$, the signature of confinement. However, in the presence of dynamical, charged fundamental matter – whether fermions or scalars – this criterion is no longer useful. The problem is that, for a sufficiently long flux tube, it is energetically preferable to break the string by producing a particle-anti-particle pair from the vacuum. If the flux tube has tension σ and the particles have mass m, this will occur when the length exceeds $L > 2m/\sigma$. For large loops, we therefore expect $\langle W[C] \rangle \sim e^{-\mu L}$. This is the same behaviour that we previously argued for in the Higgs phase. To see how they are related, we next turn to theories with scalars.

2.7.3 The Higgs vs Confining Phase

We now consider scalars. These can do something novel: they can condense and spontaneously break the gauge symmetry. This is the *Higgs phase*.

Consider an $SU(N_c)$ gauge theory with N_s scalar fields transforming in the fundamental representation. If the scalars are massless, then the gauge coupling runs as

$$\frac{1}{g^2(\mu)} = \frac{1}{g_0^2} - \frac{1}{(4\pi)^2}\left[\frac{11N_c}{3} - \frac{N_s}{6}\right]\log\left(\frac{\Lambda_{UV}^2}{\mu^2}\right)$$

and, correspondingly, the coefficient of the one-loop beta function is

$$\beta_0 = \frac{1}{(4\pi)^2}\left(-\frac{11N_c}{3} + \frac{N_s}{6}\right)$$

For $N_s < 22N_c$, the coupling becomes strong at an infra-red scale, $\Lambda_{QCD} = \Lambda_{UV}e^{1/2\beta_0 g_0^2}$. It is thought that the theory confines and develops a gap at this scale. We expect no massless excitations to survive.

What now happens if we give a mass m^2 to the scalars? For $m^2 > 0$, we expect these to shift the spectrum of the theory, but not qualitatively change the physics. Indeed, for $m^2 \gg \Lambda_{QCD}^2$, we can essentially ignore the scalars at low-energies and where we revert to pure Yang-Mills. The real interest comes when we have $m^2 < 0$ so that the scalar condense. What happens then?

Suppose that we take $m^2 \ll -\Lambda_{QCD}^2$. This means that the scalars condense at a scale where the theory is still weakly coupled, $g^2(|m|) \ll 1$, and we can trust our semi-classical analysis. If we have enough scalars to fully Higgs the gauge symmetry ($N_s \geq N_c - 1$ will do the trick), then all the gauge bosons and scalars again become massive.

It would seem that the Higgs mechanism and confinement are two rather different ways to give a mass to the gauge bosons. In particular, the Higgs mechanism is something that we can understand in a straightforward way at weak coupling while confinement is shrouded in strongly coupled mystery. Intuitively, we may feel that the Higgs phase is not the same as the confining phase. But are they really different?

The sharp way to ask this question is: does the theory undergo a phase transition as we vary m^2 from positive to negative? We usually argue for the existence of a phase transition by exhibiting an order parameter which has different behaviour in the two phases. For pure Yang-Mills, the signature for confinement is the area law for the Wilson loop. But, as we argued above, in the presence of dynamical fundamental matter the confining string can break, and the area law goes over to a perimeter law. But this is the expected behaviour in the Higgs phase. In the absence of an order parameter to distinguish between the confining and Higgs phases, it seems plausible that they are actually the same, and one can vary smoothly from one phase to another. To illustrate this, we turn to an example.

An Example: $SU(2)$ with Fundamental Matter

Consider $SU(2)$ gauge theory with a single scalar ϕ in the fundamental representation. For good measure, we'll also throw in a single fermion ψ, also in the fundamental representation. We take the action to be

$$S = \int d^4x \; -\frac{1}{2g^2}\mathrm{tr}\, F^{\mu\nu}F_{\mu\nu} + |\mathcal{D}_\mu\phi|^2 - \frac{\lambda}{4}(\phi^\dagger\phi - v^2)^2 + i\bar{\psi}\,\not{\mathcal{D}}\psi + m\bar{\psi}\psi$$

Note that it's not possible to build a gauge invariant Yukawa interaction with the matter content available. We will look at how the spectrum changes as we vary from v^2 from positive to negative.

Higgs Phase, $v^2 > 0$: When $v^2 \gg \Lambda_{QCD}$ we can treat the action semi-classically. To read off the spectrum in the Higgs phase, it is simplest to work in unitary gauge in which the vacuum expectation value takes the form $\langle\phi\rangle = (v, 0)$. We can further use the gauge symmetry to focus on fluctuations of the form $\phi = (v + \tilde{\phi}, 0)$ with $\tilde{\phi} \in \mathbf{R}$. You can think of the other components of ϕ as being eaten by the Higgs mechanism to give mass to the gauge bosons. The upshot is that we have particles of spin 0,1/2 and 1, given by

- A single, massive, real scalar $\tilde{\phi}$.
- Two Dirac fermions $\psi_i = (\psi_1, \psi_2)$. Since the $SU(2)$ gauge symmetry is broken, these no longer should be thought of as living in a doublet. As we vary the mass $m \in \mathbf{R}$, there is a point at which the fermions become massless. (Classically, this happen at $m = 0$ of course.)
- Three massive spin 1 W-bosons A^a_μ, with $a = 1, 2, 3$ labelling the generators of $su(2)$.

Confining Phase, $v^2 < 0$: When $v^2 < 0$, the scalar has mass $m^2 > 0$ and does not condense. Now we expect to be in the confining phase, in the sense that only gauge singlets have finite energy. We can list the simplest such states: we will see that they are in one-to-one correspondence with the spectrum in the Higgs phase

- A single, real scalar $\phi^\dagger\phi$. This is expected to be a massive excitation. If we were to evaluate this in the Higgs phase then, in unitary gauge, we have $\phi^\dagger\phi = v^2 + v\tilde{\phi} + \ldots$ and so the quadratic operator corresponds to the single particle excitation $\tilde{\phi}$, plus corrections.

  There are further scalar operators that we can construct, including $\mathrm{tr}\, F_{\mu\nu}F^{\mu\nu}$ and $\bar{\psi}\psi$. These have the same quantum numbers as $\phi^\dagger\phi$ and are expected to mix with it. In the confining phase, the lightest spin 0 excitation is presumably created by some combination of these.

- Two Dirac fermions. The first is $\Psi_1 = \phi^\dagger\psi$. The second comes from using the ϵ^{ij} invariant tensor of $SU(2)$, which allows us to build $\Psi_2 = \epsilon^{ij}\phi_i\psi_j$. If we expand these operators in unitary gauge in the Higgs phase, we have $\Psi_1 = v\psi_1 + \ldots$ and $\Psi_2 = v\psi_2 + \ldots$.

  It's now less obvious that each of these fermions becomes massless for some value of $m \in \mathbf{R}$, but it remains plausible. Indeed, one can show that this does occur. (A modern perspective is that the fermionic excitation is in a different topological phase for $m \gg 0$ and $m \ll 0$, ensuring a gapless mode as we vary the mass between the two.)

- Finally, we come to the spectrum of spin 1 excitations. Since we want these to be associated to gauge fields, we might be tempted to consider gauge invariant operators such as $\text{tr}\, F^{\mu\nu}F_{\mu\nu}$, but this corresponds to a scalar glueball. Instead, we can construct three gauge invariant, spin 1 operators. We have the real operator $i\phi^\dagger\mathcal{D}_\mu\phi$, and the complex operator $\epsilon^{ij}\phi_i(\mathcal{D}_\mu\phi_j)$. In unitary gauge, these become $v^2A^3_\mu$ and $v^2(A^1_\mu + iA^2_\mu)$ respectively.

This is a strongly coupled theory, so there may well be a slew of further bound states and these presumably differ between the Higgs and confining phases. Nonetheless, the matching of the spectrum suggests that we can smoothy continue from one phase to the other without any discontinuity. We conclude that, for this example, the Higgs and confining phases are actually the same phase.

Another Example: $SU(2)$ with an Adjoint Scalar

It's worth comparing what happened above with a slightly different theory in which we can distinguish between the two phases. We'll again take $SU(2)$, but this time with an *adjoint* scalar field ϕ. We'll also throw in a fermion ψ, but we'll keep this in the fundamental representation. The action is now

$$S = \int d^4x \; -\frac{1}{2g^2}\text{tr}\left(F^{\mu\nu}F_{\mu\nu} + (\mathcal{D}_\mu\phi)^2\right) - \frac{\lambda}{4}\left(\text{tr}\,\phi^2 - \frac{v^2}{2}\right)^2 + i\bar{\psi}\,\not{D}\psi + \lambda'\bar{\psi}\phi\psi + m\bar{\psi}\psi$$

where we've now also included a Yukawa coupling between the scalar and fermion.

Once again, we can look at whether there is a phase transition as we vary v^2. For $v^2 < 0$, the scalar field is massive and we expect the theory to be gapped and confine. Importantly, in this phase the spectrum contains only bosonic excitations. There are no fermions because it's not possible to construct a gauge invariant fermionic operator.

In contrast, when $v^2 > 0$ the scalar field will get an expectation value, breaking the gauge group $SU(2) \to U(1)$, resulting in a gapless photon. There are also now two fermionic excitations which carry charge $\pm\frac{1}{2}$. The spectrum now looks very different from the confining phase.

Clearly in this case the Higgs and confining phases are different. Yet, because we have fermions in the fundamental representation, we will still have dynamical breaking of the flux tube and so fundamental Wilson loop $W[C]$ does not provide an order parameter for confinement. Nonetheless, the existence of finite energy states which transform under the $\mathbf{Z}_2$ centre of $SU(2)$ – which here coincides with $(-1)^F$, with F the fermion number – provides a diagnostic for the phase.

2.8 't Hooft-Polyakov Monopoles

Coupling dynamical, electrically charged particles to Yang-Mills theory is straightforward, although understanding their dynamics may not be. But what about dynamical magnetically charged particles?

For Abelian gauge theories, this isn't possible: if you want to include Dirac monopoles in your theory then you have to put them in by hand. But for non-Abelian gauge theories, it is a wonderful and remarkable fact that, with the right matter content, magnetic monopoles come along for free: they are solitons in the theory.

Magnetic monopoles appear whenever we have a non-Abelian gauge theory, broken to its Cartan subalgebra by an adjoint Higgs field. The simplest example is $SU(2)$ gauge theory coupled to a single adjoint scalar ϕ. As explained previously, we use the convention in which ϕ sits in the Lie algebra, so $\phi = \phi^a T^a$. For $G = SU(2)$ the generators are $T^a = \sigma^a/2$, with σ^a the Pauli matrices. We take the action to be

$$S = \int d^4x \; -\frac{1}{2g^2}\mathrm{tr}F^{\mu\nu}F_{\mu\nu} + \frac{1}{g^2}\mathrm{tr}(\mathcal{D}_\mu\phi)^2 \; - \frac{\lambda}{4}\left(\mathrm{tr}\,\phi^2 - \frac{v^2}{2}\right)^2 \tag{2.86}$$

Note that we've rescaled the scalar ϕ so that it too has a $1/g^2$ sitting in front of it.

The potential is positive definite. The vacuum of the theory has constant expectation value $\langle\phi\rangle$. Up to a gauge transformation, we can take

$$\langle\phi\rangle = \frac{1}{2}\begin{pmatrix} v & 0 \\ 0 & -v \end{pmatrix} \tag{2.87}$$

This breaks the gauge group $SU(2) \to U(1)$. The spectrum consists of a massless photon – which, in this gauge, sits in the T^3 part of the gauge group — together with massive W-bosons and a massive scalar.

There are, however, more interesting possibilities for the expectation value. Any finite energy excitation must approach a configuration with vanishing potential at spatial infinity. Such configurations obey $\mathrm{tr}\,\phi^2 \to v^2$ as $|\mathbf{x}| \to \infty$. Decomposing the Higgs field into the generators of the Lie algebra, $\phi = \phi^a T^a$, $a = 1, 2, 3$, the requirement that the potential vanishes defines a sphere in field space,

$$\mathbf{S}^2 := \left\{ \phi : \phi^a \phi^a = v^2 \right\} \tag{2.88}$$

We see that for any finite energy configuration, we must specify a map which tells us the behaviour of the Higgs field asymptotically,

$$\phi : \mathbf{S}^2_\infty \mapsto \mathbf{S}^2$$

The fact that these maps fall into disjoint classes should no longer be a surprise: it's the same idea that we met in Sections 2.2 and 2.3 when discussing theta vacua and instantons, and again in Section 2.5.2 when discussing vortices. This time the relevant homotopy group is

$$\Pi_2(\mathbf{S}^2) = \mathbf{Z}$$

Given a configuration ϕ, the winding number is computed by

$$\nu = \frac{1}{8\pi v^3} \int_{\mathbf{S}^2_\infty} d^2 S_i \; \epsilon^{ijk} \epsilon_{abc} \phi^a \partial_j \phi^b \partial_k \phi^c \in \mathbf{Z} \tag{2.89}$$

In a sector with $\nu \neq 0$, the gauge symmetry breaking remains $SU(2) \to U(1)$. The difference is that now the unbroken $U(1) \subset SU(2)$ changes as we move around the asymptotic $\mathbf{S}^2_\infty$.

The next step is to notice that if the Higgs field has winding $\nu \neq 0$, then we must also turn on a compensating gauge field. The argument is the same as the one we saw for vortex strings. Suppose that we try to set $A_i = 0$. Then, the covariant derivatives are simply ordinary derivatives and, asymptotically, we have $(\mathcal{D}_i \phi)^2 = (\partial_i \phi)^2 \sim (\partial_\theta \phi)^2 / r^2$, with ∂_θ denoting the (necessarily non-vanishing) variation as we move around the angular directions of the asymptotic $\mathbf{S}^2_\infty$. The energy of the configuration will then include the term

$$E = \frac{1}{g^2} \int d^3x \; \mathrm{tr}\,(\partial_i \phi)^2 \sim \frac{1}{g^2} \int_{\mathbf{S}^2_\infty} d^2\Omega \int dr \; r^2 \, \mathrm{tr} \frac{(\partial_\theta \phi)^2}{r^2}$$

This integral diverges linearly. We learn that if we genuinely want a finite energy excitation in which the Higgs field winds asymptotically then we must also turn on the gauge fields A_i to cancel the $1/r$ asymptotic fall-off of the angular gradient terms, and ensure that $\mathcal{D}_\theta \phi \to 0$ as $r \to \infty$. We want to solve

$$\mathcal{D}_i \phi = \partial_i \phi - i[A_i, \phi] \to 0 \quad \Rightarrow \quad A_i \to \frac{i}{v^2}[\phi, \partial_i \phi] + \frac{a_i}{v}\phi$$

Here the first term works to cancel the fall-off from $\partial_i \phi$. To see this, you will need to use the fact that $\operatorname{tr} \phi^2 \to v^2$, and so $\operatorname{tr}(\phi \partial_i \phi) \to 0$, as well as the $su(2)$ commutation relations. The second term in A_i does not contribute to the covariant derivative $\mathcal{D}_i \phi$. The function a_i is the surviving, massless $U(1)$ photon which can be written in a gauge invariant way as

$$a_\mu = \frac{1}{v} \operatorname{tr}(\phi A_\mu) \tag{2.90}$$

We can also compute the asymptotic form of the field strength. The same kinds of manipulations above show that this lies in the same direction in the Lie algebra as ϕ,

$$F_{ij} = \frac{1}{v} \mathcal{F}_{ij}\, \phi$$

with

$$\mathcal{F}_{ij} = f_{ij} + \frac{i}{v^3} \operatorname{tr}\left(\phi\, [\partial_i \phi, \partial_j \phi]\right)$$

Here $f_{ij} = \partial_i a_j - \partial_j a_i$ is the Abelian field strength that we may have naively expected. But we see that there is an extra term, and this brings a happy surprise, since it contributes to the magnetic charge m of the $U(1)$ field strength. This is given by

$$m = -\int d^2 S_i\ \frac{1}{2} \epsilon^{ijk} \mathcal{F}_{jk} = \frac{1}{2v^3} \int d^2 S_i\ \epsilon^{ijk} \epsilon^{abc} \phi^a \partial_j \phi^b \partial_k \phi^c = 4\pi \nu \tag{2.91}$$

with ν the winding number defined in (2.89). We learn that any finite energy configuration in which the Higgs field winds asymptotically necessarily carries a magnetic charge under the unbroken $U(1) \subset SU(2)$. This object is a soliton and goes by the name of the *'t Hooft-Polyakov monopole*.

The topological considerations above have led us to a quantised magnetic charge. However, at first glance, the single 't Hooft-Polyakov monopole with $\nu = 1$ seems to have twice the charge required by Dirac quantisation (1.3), since the W-bosons have electric charge $q = 1$. But there is nothing to stop us including matter in the fundamental representation of $SU(2)$ with $q = \pm\frac{1}{2}$, with respect to which the 't Hooft-Polyakov monopole has the minimum allowed charge.

2.8.1 Monopole Solutions

We have not yet solved the Yang-Mills-Higgs equations of motion with a given magnetic charge. In general, no static solutions are expected to exist with winding $\nu > 1$, because magnetically charged objects typically repel each other. For this reason, we restrict attention to the configurations with winding $\nu = \pm 1$.

We can write an ansatz for a scalar field with winding $n = 1$,

$$\phi^a = \frac{x^a}{r^2} h(r) \quad \text{with } h(r) \to \begin{cases} 0 & r \to 0 \\ vr & r \to \infty \end{cases}$$

This is the so-called "hedgehog" ansatz, since the direction of the scalar field $\phi = \phi^a T^a$ is correlated with the direction x^a in space. Just like a hedgehog. In particular, this means that the $SU(2)$ gauge action on ϕ^a and the $SO(3)$ rotational symmetry on x^a are locked, so that only the diagonal combination are preserved by such configurations. We can make a corresponding ansatz for the gauge field which preserves the same diagonal $SO(3)$,

$$A_i^a = -\epsilon_{aij} \frac{x^j}{r^2} \left[1 - k(r)\right] \quad \text{with } k(r) \to \begin{cases} 1 & r \to 0 \\ 0 & r \to \infty \end{cases}$$

We can now insert this ansatz into the equations of motion

$$\mathcal{D}^\mu F_{\mu\nu} - i[\phi, \mathcal{D}_\nu \phi] = 0 \quad \text{and} \quad \mathcal{D}^2 \phi = 2g^2 \lambda (\mathrm{tr}\, \phi^2 - v^2)\phi \tag{2.92}$$

This results in coupled, ordinary differential equations for $h(r)$ and $k(r)$. In general, they cannot be solved analytically, but it is not difficult to find numerical solutions for the minimal 't Hooft-Polyakov monopole.

BPS Monopoles

Something special happens when we set $\lambda = 0$ in (2.86). Here the scalar potential vanishes which means that, at least classically, we can pick any expectation value v for the scalar. The choice of v should be thought of as extra information needed to define the vacuum of the theory. (In the quantum theory, one typically expects to generate a potential for ϕ. The exception to this is in supersymmetric theories, where cancellations ensure that the quantum potential also vanishes. Indeed, the monopole that we describe below have a nice interplay with supersymmetry, although this is beyond the scope of these lectures.)

When the potential vanishes, it is possible to use the Bogomolnyi trick to rewrite the energy functional. In terms of the non-Abelian magnetic field $B_i = -\frac{1}{2}\epsilon_{ijk}F_{jk}$, the energy of a static configuration with vanishing electric field is

$$\begin{aligned} E &= \frac{1}{g^2}\int d^3x \;\mathrm{tr}\left(B_i^2 + (\mathcal{D}_i\phi)^2\right) \\ &= \frac{1}{g^2}\int d^3x \;\mathrm{tr}\,(B_i \mp \mathcal{D}_i\phi)^2 \pm 2\,\mathrm{tr}\,B_i\mathcal{D}_i\phi \\ &\geq \pm\frac{2}{g^2}\int d^3x\; \partial_i\,\mathrm{tr}\,B_i\phi \end{aligned}$$

where, to get to the last line, we have discarded the positive definite term and integrated by parts, invoking the Bianchi identity $\mathcal{D}_iB_i = 0$. We recognise the final expression as the magnetic charge. We find that the energy of a configuration is bounded by the magnetic charge

$$E \geq \frac{2v|m|}{g^2} \tag{2.93}$$

A configuration which saturates this bound is guaranteed to solve the full equations of motion. This is achieved if we solve the first order Bogomolnyi equations

$$B_i = \pm\mathcal{D}_i\phi \tag{2.94}$$

with the ± sign corresponding to monopoles (with $m > 0$) and anti-monopoles (with $m < 0$) respectively. It can be checked that solutions to (2.94) do indeed solve the full equations of motion (2.92) when $\lambda = 0$.

Solutions to (2.94) have a number of interesting properties. First, it turns out that the equations of motion for a single monopole have a simple analytic solution,

$$h(r) = vr\coth(vr) - 1 \quad \text{and} \quad k(r) = \frac{vr}{\sinh vr}$$

This was first discovered by Prasad and Sommerfield. In general, solutions to (2.94) are referred to as BPS monopoles, with Bogomolnyi's name added as well.

A warning on terminology: these BPS monopoles have rather special properties in the context of supersymmetric theories where they live in short multiplets of the supersymmetry algebra. The term "BPS" has since been co-opted and these days is much more likely to refer to some kind of protected object in supersymmetry, often one that has nothing to do with the monopole.

The Bogomolnyi equations (2.94) also have solutions corresponding to monopoles with higher magnetic charges. These solutions include configurations that look like far separated single charge monopoles. This is mildly surprising. Our earlier intuition told us that such solutions should not exist because the repulsive force between magnetically charged particles would ensure that the energy could be lowered by moving them further apart. That intuition breaks down in the Bogomolnyi limit because we have a new massless particle – the scalar ϕ – and this gives rise to a compensating attractive force between monopoles, one which precisely cancels the magnetic repulsion. You can learn much more about the properties of these solutions, and the role they play in supersymmetric theories, in the lectures on Solitons.

Monopoles in Other Gauge Groups

It is fairly straightforward to extend the discussion above the other gauge groups G. We again couple a scalar field ϕ in the adjoint representation and give it an expectation value that breaks $G \to H$ where $H = U(1)^r$, with r is the rank of the gauge group.

Given an expectation value for ϕ, we can always rotate it by acting with G. However, by definition, H leaves the scalar untouched which means that in configurations are now classified by maps from $\mathbf{S}^2_\infty$ into the space G/H. (In our previous discussion we had $G/H = SU(2)/U(1) = \mathbf{S}^2$ which coincides with what we found in (2.88).) A result in homotopy theory tells us that, for simply connected G,

$$\Pi_2(G/H) = \Pi_1(H) = \mathbf{Z}^r$$

We learn that the 't Hooft-Polyakov monopoles are labelled by an r-dimensional magnetic charge vector $\mathbf{m}$. This agrees with our analysis of 't Hooft lines in Section 2.6. A closer look reveals that the 't Hooft-Polyakov monopoles have magnetic charge $\mathbf{m} \in 2\pi\, \Lambda_{\text{co}-\text{root}}(\mathfrak{g})$, as required by the Goddard-Nuyts-Olive quantisation (2.80).

2.8.2 The Witten Effect Again

We saw in Section 1.2.3 that, in the presence of a θ term, a Dirac monopole picks up an electric charge. As we now show this phenomenon, known as the Witten effect, also occurs for the 't Hooft-Polyakov monopole.

To see this, we simply need to be careful in identifying the electric charge operator in the presence of a monopole. We saw in (2.90) that the unbroken $U(1) \subset SU(2)$ is determined by the ϕ. The corresponding global gauge transformation is

$$\delta A_\mu = \frac{1}{v}\mathcal{D}_\mu \phi$$

But we already did the hard work and computed the Noether charge Q associated to such a gauge transformation in (2.30), where we saw that it picks up a contribution from the θ term (2.22); we have

$$Q = \frac{1}{g^2}\int d^3x \ \text{tr}\left(E_i + \frac{\theta g^2}{8\pi^2}B_i\right)\frac{1}{v}\mathcal{D}_i\phi$$

In our earlier discussion, around equation (2.30), we were working inthe vacuum and could discard the contribution from θ. However, in the presence of a monopole both terms contribute. The total electric charge Q is now

$$Q = q + \frac{\theta g^2 m}{8\pi^2} \tag{2.95}$$

with the naive electric charge q defined as

$$q = \frac{1}{v}\int d^3x \ \text{tr}\,\mathcal{D}_i\phi\, E_i$$

and the magnetic charge m defined, as in (2.91), by

$$m = \frac{1}{v}\int d^3x \ \text{tr}\,\mathcal{D}_i\phi\, B_i$$

We see that the theta term does indeed turn the monopole into a dyon. This agrees with our previous discussion of the Witten effect (1.19), with the seemingly different factor of 2 arising because, as explained above, q is quantised in units of 1/2 in the non-Abelian gauge theory.

2.9 Further Reading

Trinity College, Cambridge boasts many great scientific achievements. The discovery of Yang-Mills theory is not among the most celebrated. Nonetheless, in January 1954 a graduate student at Trinity named Ronald Shaw wrote down what we now refer to as the Yang-Mills equations. Aware that the theory describes massless particles, which appear to have no place in Nature, Shaw was convinced by his supervisor, Abdus Salam, that the result was not worth publishing. It appears only as a chapter of his thesis [181].

Across the Atlantic, in Brookhaven national laboratory, two office mates did not make the same mistake. C. N. Yang and Robert Mills constructed the equations which now bear their name [232]. It seems likely that that they got the result slightly before Shaw, although the paper only appeared afterwards. Their original motivation now seems somewhat misguided: their paper suggests that global symmetries of quantum field theory – specifically $SU(2)$ isospin – are not consistent with locality. They write

> "It seems that this [global symmetry] is not consistent with the localized field concept that underlies the usual physical theories"

From this slightly shaky start, one of the great discoveries of 20th century physics emerged,

In those early days, the role played by Yang-Mills theory was, to say the least, confusing. Yang gave a famous seminar in Princeton in which Pauli complained so vociferously about the existence of massless particles that Yang refused to go on with the talk and had to be coaxed back to the blackboard by Oppenheimer. (Pauli had a headstart here: in 1953 he did a Kaluza-Klein reduction on $\mathbf{S}^2$, realising an $SU(2)$ gauge theory but discarding it because of the massless particle [151]. A similar result had been obtained earlier by Klein [122].)

It took a decade to realise that the gauge bosons could get a mass from the Higgs mechanism, and a further decade to realise that the massless particles were never really there anyway: they are an artefact of the classical theory and gain a mass automatically when $\hbar \neq 0$. Below is a broad brush description of this history. A collection of reminiscences, "50 Years of Yang-Mills" [108], contains articles by a number of the major characters in this story.

Asymptotic Freedom

As the 1970s began, quantum field theory was not in fashion. Fundamental laws of physics, written in the language of field theory, languished in the literature, unloved and uncited [77, 205]. The cool kids were playing with bootstraps.

The discovery of asymptotic freedom was one of the first results that brought field theory firmly into the mainstream. The discovery has its origins in the deep inelastic scattering experiments performed in SLAC in the late 1960s. Bjorken [19] and subsequently Feynman [56] realised that the experiments could be interpreted in terms of the momentum distribution of constituents of the proton. But this interpretation held only if the interactions between these constituents became increasingly weak at high energies. Feynman referred to the constituents as "partons" rather than "quarks" [57]. It is unclear whether this was because he wanted to allow for the possibility of other constituents, say gluons, or simply because he wanted to antagonise Gell-Mann.

In Princeton, David Gross set out to show that no field theory could exhibit asymptotic freedom [86]. Having ruled out field theories based on scalars and fermions, all that was left was Yang-Mills. He attacked this problem with his new graduate student

Frank Wilczek. The minus signs took some getting right, but by April 1973 they realised that they had an asymptotically free theory on their hands [83] and were keenly aware of its importance.

Meanwhile, in Harvard, Sidney Coleman was interested in the same problem. He asked his graduate student Erick Weinberg to do the calculation but, content that he had enough for his thesis, Erick passed it on to another graduate student, David Politzer. Politzer finished his calculation at the same time as the Princeton team [156]. In 2004, Gross, Politzer and Wilczek were awarded the Nobel prize. Politzer's Nobel lecture contains an interesting, and very human, account of the discovery [157].

In fact, both American teams had been scooped. In June 1972, at a conference in Marseilles, a Dutch graduate student named Gerard 't Hooft sat in a talk by Symanzik on the SLAC experiments and their relation to asymptotic freedom. After the talk, 't Hooft announced that Yang-Mills theory is asymptotically free. Symanzik encouraged him to publish this immediately but, like Shaw 20 years earlier, 't Hooft decided against it. His concern was that Yang-Mills theory could not be relevant for the strong force because it had no mechanism for the confinement of quarks [107].

The failure to publish did not hurt 't Hooft's career. By that stage he had already shown that Yang-Mills was renormalisable, a fact which played a large role in bringing the theory out of obscurity [93, 94, 95]. This was enough for him to be awarded his PhD [96]. It was also enough for him to be awarded the 1999 Nobel prize, together with his advisor Veltman. We will be seeing much more of the work of 't Hooft later in these lectures.

The analogy between asymptotic freedom and paramagnetism was made by N. K. Nielsen [148], although the author gives private credit to 't Hooft. In these lectures, we computed the one-loop beta function using the background field method. This method was apparently introduced by (of course) 't Hooft in lectures which I haven't managed to get hold of. It first appears in published form in a paper by Larry Abbott [1] (now a prominent theoretical neuroscientist) and is covered in the textbook by Peskin and Schroeder [154].

Confinement and the Mass Gap

Asymptotic freedom gave a dynamical reason to believe that Yang-Mills was likely responsible for the strong force. Earlier arguments that quarks should have three colour degrees of freedom meant that attention quickly focussed on the gauge group $SU(3)$ [84, 65]. But the infra-red puzzles still remained. Why are the massless particles predicted by Yang-Mills not seen? Why are individual quarks not seen?

Here things were murky. Was the $SU(3)$ gauge group broken by a scalar field? Or was it broken by some internal dynamics? Or perhaps the gauge group was actually unbroken but the flow to strong coupling does something strange. This latter possibility was mooted in a number of papers [84, 207, 208, 65]. This from Gross and Wilczek in 1973,

> "Another possibility is that the gauge symmetry is exact. At first sight this would appear ridiculous since it would imply the existence of massless, strongly coupled vector mesons. However, in asymptotically free theories these naive expectations might be wrong. There may be little connection between the "free" Lagrangian and the spectrum of states."

This idea was slowly adopted over the subsequent year. The idea of dimensional transmutation, in which dimensionless constants combine with the cut-off to give the a physical scale, was known from the 1973 work of Coleman and E. Weinberg [27]. Although they didn't work with Yang-Mills, their general mechanism removed the most obvious hurdle for a scale-invariant theory to develop a gap. A number of dynamical explanations were mooted for confinement, but the clearest came only in 1974 with Wilson's development of lattice gauge theory [214]. This paper also introduced what we now call the Wilson line. We will discuss the lattice approach to confinement in some detail in Section 4.

The flurry of excitement surrounding these developments also serves to highlight the underlying confusion, as some of the great scientists of the 20th century clamoured to disown their best work. For example, in an immediate response to the discovery of asymptotic freedom, and six years after his construction of the electroweak theory [205], Steven Weinberg writes [208]

> "Of course, these very general results will become really interesting only when we have some specific gauge model of the weak and electromagnetic interactions which can be taken seriously as a possible description of the real world. This we do not yet have."

Not to be outdone, in the same year Gell-Mann offers [65]

> "We do not accept theories in which quarks are real, observable particles."

It's not easy doing physics.

Semi-Classical Yang-Mills

In these lectures, we first described the classical and semi-classical structure of Yang-Mills theory, and only then turned to the quantum behaviour. This is the logical way through the subject. It is not the historical way.

Our understanding of the classical vacuum structure of Yang-Mills theory started in 1975, when Belavin, Polyakov, Schwartz and Tyupkin discovered the Yang-Mills instanton [14]. Back then, Physical Review refused to entertain the name "instanton", so they were referred to in print as "pseudoparticles".

't Hooft was the first to perform detailed instanton calculations [101, 102], including the measure $K(\rho)$ that we swept under the carpet in Section 2.3.3. Among other things, his work clearly showed that physical observables depend on the theta angle. Motivated by this result, Jackiw and Rebbi [113], and independently Callan, Dashen and Gross [23], understood the semi-classical vacuum structure of Yang-Mills that we saw in Section 2.2.

Jackiw's lectures [115] give a very clear discussion of the theta angle and were the basis for the discussion here. Reviews covering a number of different properties of instantons can be found in [182, 191, 197].

Magnetic Yang-Mills

The magnetic sector of Yang-Mills theory was part of the story almost from the beginning. Monopoles in $SU(2)$ gauge theories were independently discovered by 't Hooft [99] and Polyakov [158] in 1974. The extension to general gauge groups was given in 1977 by Goddard, Nuyts and Olive [80]. This paper includes the GNO quantisation condition that we met in our discussion of 't Hooft line, and offers some prescient suggestions on the role of duality in exchanging gauge groups. (These same ideas rear their heads in mathematics in the Langlands program.)

Bogomolnyi's Bogomolnyi trick was introduced in [20]. Prasad and Sommerfeld then solved the resulting equations of motion for the monopole [162], and the initials BPS are now engraved on all manner of supersymmetric objects which have nothing to do with monopoles. (A more appropriate name for BPS states would be Witten-Olive states [217].) Finally, Witten's Witten effect was introduced in [216]. Excellent reviews of 't Hooft-Polyakov monopoles, both with focus on the richer BPS sector, can be found in Harvey's lecture notes [89] and in Manton and Sutcliffe's book [133]. There are also some TASI lectures [191].

The Nielsen-Olesen vortex was introduced in 1973 [145]. Their motivation came from string theory, rather than field theory. The fact that such strings would confine magnetic monopoles was pointed out by Nambu [142] and the idea that this is a useful analogy for quark confinement, viewed in dual variables, was made some years later by Mandelstam [130] and 't Hooft [100].

The 't Hooft line as a magnetic probe of gauge theories was introduced in [103]. This paper also emphasises the importance of the global structure of the gauge group. A more modern perspective on line operators was given by Kapustin [120]. A very clear discussion of the electric and magnetic line operators allowed in different gauge groups, and the way this ties in with the theta angle, can be found in [4].

Towards the end of the 1970s, attention began to focus on more general questions of the phases of non-Abelian gauge theories [103, 104]. The distinction, or lack thereof, between Higgs and confining phases when matter transforms in the fundamental of the gauge group was discussed by Fradkin and Shenker [63] and by Banks and Rabinovici [9]; both rely heavily on the lattice. The Banks-Zaks fixed point, and its implications for the conformal window, was pointed out somewhat later in 1982 [10].