6. Large N

Non-Abelian gauge theories are hard. We may have mentioned this previously. Indeed, it's not a bad summary of the lectures so far. The difficulty stems from the lack of a small, dimensionless parameter which we can use as the basis for a perturbative expansion.

Soon after the advent of QCD, 't Hooft pointed out that gauge theories based on the group $G = SU(N)$ simplify in the limit $N \to \infty$. This can then be used as a starting point for an expansion in $1/N$. Viewed in the right way, Yang-Mills does have a small parameter after all.

At first glance, it seems surprising that the theory simplifies in the large N limit. Naively, you might think that the theory only gets more complicated as the number of fields increase. However, this intuition breaks down when the fields are related by a symmetry, in which case the collective behaviour of the fields becomes stiffer as their number increases. This results in a novel, classical regime of the theory. The weakly coupled degrees of freedom typically look very different from the gluons that we start with in the original Lagrangian.

Large N limits are now commonplace in statistical and quantum physics. As a general rule of thumb, the large N limit renders a theory tractable when the number of degrees of freedom grows linearly with N. (We shall meet two examples in Section 7 when we discuss the $\mathbf{CP}^{N-1}$ model and the Gross-Neveu model.) In contrast, when the number of degrees of freedom grows as N^2, or faster, then the theory simplifies but, apart from a few special cases, cannot be solved. This is the case for Yang-Mills where the large N limit will not allow us to demonstrate, say, confinement. Nonetheless, it does provide an approach which allows us to compute certain properties. Moreover, it points to a deep connection between gauge theory and string theory, one which underlies many of the recent advances in both subjects.

You might reasonably wonder whether the large N expansion is likely to be relevant for QCD which has $N = 3$. We'll see as we go along how useful it is. A common rebuttal, originally due to Witten, is that in natural units the fine structure constant is

$$\alpha = \frac{e^2}{4\pi} \approx \frac{1}{137} \quad \Rightarrow \quad e \approx 0.30$$

This comparison is a little unfair. The true expansion parameter in QED is better phrased as $\alpha/4\pi \sim 10^{-3}$. In contrast, there are no factors of 4π that ride to the rescue

for Yang-Mills. The expansion parameter is $1/N$ or, in many situations, $1/N^2$. We might therefore hope that this approach will give us results that are quantitatively correct at the 10% level.

6.1 A Quantum Mechanics Warm-Up: The Hydrogen Atom

We start by providing a simple example where a large N limit offers a novel way to apply perturbation theory. The set-up is very familiar: the hydrogen atom.

In natural units, $\hbar = c = \epsilon_0 = 1$, the Hamiltonian of the hydrogen atom is

$$H = -\frac{1}{2m}\nabla^2 - \frac{\alpha}{r} \tag{6.1}$$

with α the fine structure constant. In our first course on Quantum Mechanics, we learn the exact solution for the bound states of this system. But suppose we didn't know this. Can we try to approximate the solutions using perturbation theory?

Since there's a small number, $\alpha \approx 1/137$, sitting in the potential term, you might think that you could expand in α. But this is misleading. In the context of atomic physics, the fine structure constant cannot be used as the basis for a perturbative expansion. This is because we can always reabsorb it by a change of scale. Define $r' = m\alpha r$. Then the Hamiltonian becomes,

$$H = m\alpha^2 \left[-\frac{1}{2}\nabla'^2 - \frac{1}{r'} \right]$$

We see that the fine structure constant simply sets the overall scale of the problem. This means that we expect the order of magnitude of bound state to be around

$$E_{\text{atomic}} = -m\alpha^2 \approx -27.2 \text{ eV}$$

In fact, the ground state energy is $E_{\text{atomic}}/2 \approx -13.6$ eV, the factor of $1/2$ coming from solving the Schrödinger equation.

For our purposes this means that the hydrogen atom is, like Yang-Mills, a theory with a scale but with no small, dimensionless parameter. How, then, to construct a perturbative solution? One possibility is to generalise the problem from three dimensions to N dimensions. The Hamiltonian remains (6.1), but now with ∇^2 denoting the Laplacian in $\mathbf{R}^N$ rather than $\mathbf{R}^3$. Clearly we have increased the number of degrees of freedom from 3 to N. We have also increased the symmetry group from $SO(3)$ to $SO(N)$.

We note in passing that we are not solving the higher dimensional version of the hydrogen atom, since in that case the Coulomb force would fall-off as $1/r^{N-2}$. Instead, we keep the Coulomb force fixed as $1/r$ and vary the dimension of space.

To see how this helps, we will focus on the s-wave sector. Here the Schrödinger equation becomes

$$H\psi = m\alpha^2 \left(-\frac{1}{2}\frac{d^2}{dr'^2} - \frac{(N-1)}{2r'}\frac{d}{dr'} - \frac{1}{r'} \right) \psi = E\psi$$

At leading order in $1/N$, we can replace the $(N-1)$ factor by N. We'll do this because the equations are a little simpler, although if we were serious about pursuing perturbation theory in $1/N$, we would have to be more careful. We can now remove the term that is first order in derivatives by redefining the wavefunction as $\psi(r') = \chi(r')/r'^{N/2}$, leaving us with the rescaled Schrödinger equation

$$H\chi = m\alpha^2 \left(-\frac{1}{2}\frac{d^2}{dr'^2} + \frac{N^2}{8r'^2} - \frac{1}{r'} \right) \chi = E\chi$$

We'll make one further rescaling, and define a new radial coordinate, $r' = N^2 R$. The Schrödinger equation now becomes

$$H\chi = \frac{m\alpha^2}{N^2} \left(-\frac{1}{2N^2}\frac{d^2}{dR^2} + V_{\text{eff}}(R) \right) \chi = E\chi \quad \text{with} \quad V_{\text{eff}}(R) = \frac{1}{8R^2} - \frac{1}{R}$$

This rescaling has removed all N dependence from the effective potential. Instead, we see that it appears in two places: the overall scale of the problem; and the effective (dimensionless) mass of the particle, which can be read off from the kinetic term and is $m_{\text{eff}} = N^2$.

We're left with a very heavy particle, moving in the one-dimensional effective potential $V_{\text{eff}}(R)$. In this limit, we can expand the potential in a Taylor series around the minimum $R_{\text{min}} = 1/4$. To leading order, we can then treat the problem as a harmonic oscillator, centred on R_{min}. Higher order terms in the Taylor series will affect the energy only at subleading order in $1/N$.

To leading order, the ground state energy is given by $V_{\text{eff}}(R_{\text{min}})$. (The zero point energy of the harmonic oscillator is suppressed by $1/m_{\text{eff}} \sim 1/N^2$.) This gives us our expression for the ground state of the harmonic oscillator,

$$E_{\text{ground}} = \frac{m\alpha^2}{N^2} \left(2 + \mathcal{O}\left(\frac{1}{N}\right) \right)$$

If we now revert to the real world with $N = 3$, we get $E_{\text{ground}} \approx 2m\alpha^2/9$. The true answer, as we mentioned above, is $E_{\text{ground}} = m\alpha^2/2$.

Of course, it's a little perverse to apply perturbation theory to a problem for which there is an exact solution. But the key idea remains: the extra degrees of freedom, together with the restriction of $O(N)$ symmetry, combine to render the problem weakly coupled in the limit $N \to \infty$. We will now see how a similar effect occurs for Yang-Mills theory.

6.2 Large N Yang-Mills

The action for $SU(N)$ Yang-Mills theory is

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

There is an immediate hurdle if we try to naively take the large N limit. As we saw in Section 2.4, confinement and the mass gap all occur at the strong coupling scale Λ_{QCD} which, at one-loop, is given by

$$\Lambda_{QCD} = \Lambda_{UV}\exp\left(-\frac{3}{22}\frac{(4\pi)^2}{g^2N}\right)$$

If we keep both the UV cut-off Λ_{UV} and the gauge coupling g^2 fixed, and send $N \to \infty$, then there is no parametric separation between the physical scale Λ_{QCD} and the cut-off. This is bad. To rectify this, we define the *'t Hooft coupling*,

$$\lambda = g^2N$$

We will consider the theory in the limit $N \to \infty$, with both Λ_{UV} and λ held fixed. This ensures that the physical scale Λ_{QCD} also remains fixed in this limit. Indeed, throughout this section we will discuss how masses, lifetimes and scattering amplitudes of various states scale with N. In all cases, it is Λ_{QCD} which fixes the dimensions of these properties.

With these new couplings, the Yang-Mills action is

$$S_{\rm YM} = -\frac{N}{2\lambda}\int d^4x \ {\rm tr}\, F^{\mu\nu}F_{\mu\nu} \tag{6.2}$$

This is the form we will work with.

6.2.1 The Topology of Feynman Diagrams

To proceed, we're going to look more closely at the Feynman diagrams that arise from the Yang-Mills action 6.2. We'll see that, in the 't Hooft limit $N \to \infty$, λ fixed, there is a rearrangement in the importance of various diagrams.

We will write down the Feynman rules for Yang-Mills. Each gluon field is an $N \times N$ matrix,

$$(A_\mu)^i{}_j \quad , \; i, j = 1, \ldots, N$$

The propagator has the index structure

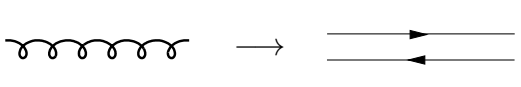
$$\langle A^i_{\mu\, j}(x)\, A^k_{\nu\, l}(y)\rangle = \Delta_{\mu\nu}(x-y)\, \left(\delta^i_l\, \delta^k_j - \frac{1}{N}\delta^i_j\, \delta^k_l\right)$$

where $\Delta_{\mu\nu}(x)$ is the usual photon propagator for a single gauge field. The $1/N$ term arises because we're working with traceless $SU(N)$ gauge fields, rather than $U(N)$ gauge fields. But clearly it is suppressed by $1/N$ and so, at leading order in $1/N$, we don't lose anything by dropping this term. We then have

$$\langle A^i_{\mu\, j}(x)\, A^k_{\nu\, l}(y)\rangle = \Delta_{\mu\nu}(x-y)\; \delta^i_l\, \delta^k_j$$

This means that we're really working with $U(N)$ gauge theory rather than $SU(N)$ gauge theory.

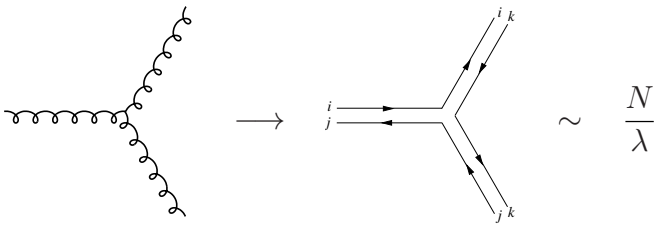
At this point, it is useful to introduce some new notation. The fact that the gauge field has two indices, i, j, suggests that we can represent it as two lines in a Feynman diagram rather than one. One of these lines represents the top index, which transforms in the $\bar{\mathbf{N}}$ representation; the other the bottom index which transforms in the $\mathbf{N}$ representation. Instead of the usual curly line notation for the gluon propagator, we have



$$\sim \quad \frac{\lambda}{N} \tag{6.3}$$

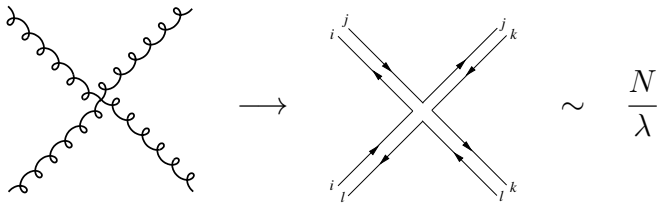
Note that each line comes with an arrow, and the arrows point in opposite ways. This reflects the fact that the upper and lower lines are associated to complex conjugate representations. The propagator scales as λ/N, as can be read off from the action (6.2).

Similarly, the cubic vertex that comes from expanding out the Yang-Mills action takes the form



$$\sim \quad \frac{N}{\lambda}$$

where we've now included the $i, j, k = 1, \ldots, N$ indices to show how these must match up as we follow the arrows. (There is also a second diagram from the cubic vertex in which the arrows are reversed.) Similarly, the quartic coupling vertex becomes



$$\sim \quad \frac{N}{\lambda}$$

Each vertex comes with a factor of N/λ. This also follows from the action 6.2. The fact that the vertex comes with an inverse power of the coupling might be unfamiliar, but it is because of the way we chose to scale our fields. It will all come out in the wash, with the propagators compensating so that increasingly complicated diagrams are suppressed by powers of λ as expected. We'll see examples shortly.

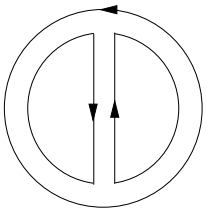
As we evaluate the various Feynman diagrams, we will now have a double expansion in both λ and in $1/N$. We'd like to understand how the diagrams arrange themselves. The general scaling will be

$$\text{diagram} \sim \left(\frac{\lambda}{N}\right)^{\#\text{propagators}} \left(\frac{N}{\lambda}\right)^{\#\text{vertices}} N^{\#\text{index contractions}} \tag{6.4}$$

where the index contractions come from the loops in the diagram. To see this more clearly, it's best to look at some examples.

Vacuum Bubbles

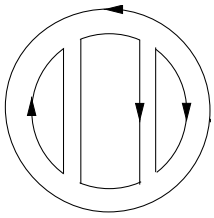
To understand the Feynman diagram expansion, let's start by considering the vacuum bubbles. The leading order contribution is a diagram which, in double line notation, looks like,



$$\sim \left(\frac{\lambda}{N}\right)^3 \left(\frac{N}{\lambda}\right)^2 N^3 \sim \lambda N^2 \tag{6.5}$$

Here the first two factors come from the 3 propagators and the 2 vertices in the diagram. The final factor is important: it comes from the fact that we have three contractions over the indices $i, j, k = 1, \ldots, N$. These are denoted by the three arrows in the diagram. Note that we get a contribution from the outside circle since we're dealing with vacuum bubbles.

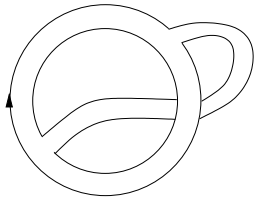
Similarly, at the next order in λ, we have the diagram



$$\sim \left(\frac{\lambda}{N}\right)^6 \left(\frac{N}{\lambda}\right)^4 N^4 \sim \lambda^2 N^2 \tag{6.6}$$

There are now four contractions over internal loops. This diagram has the same N^2 behaviour as our first one-loop diagram, but it is down in the expansion in 't Hooft coupling. It is easy to convince yourself that the two diagrams above give the leading contribution (in N) to the free energy, which scales as $\sim \mathcal{O}(N^2)$. This reflects the fact that Yang-Mills theory has N^2 degrees of freedom.

However, there is another diagram that we could have drawn. This has the same momentum structure as (6.5), but a different index structure. In double line notation it takes the form,



$$\sim \left(\frac{\lambda}{N}\right)^3 \left(\frac{N}{\lambda}\right)^2 N \sim \lambda \tag{6.7}$$

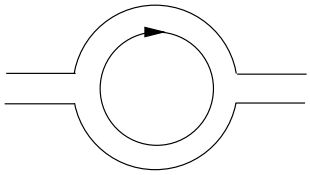
If you follow the loop around, you will find that there is now just a single contraction of the group indices. The result is a contribution to the vacuum energy which occurs at the same value of λ as (6.5), but is suppressed by $1/N^2$ relative to the first two diagrams. This means that in the limit $N \to \infty$, with λ fixed this diagram will be sub-dominant.

We see that, among all the possible Feynman diagrams, a subset dominate in the large N limit. The dominant diagrams are those which, like (6.5) and (6.6), can be drawn flat on a plane in the double line notation. These are referred to as *planar diagrams*. In contrast, diagrams like (6.7) need a third dimension to draw them. These non-planar diagrams are subleading.

The large N limit has seemed to simplify our task. We no longer need to sum over all Feynman diagrams; only the planar ones. This remains daunting. Nonetheless, as we will see below, this new structure does give us some insight into the strong coupling dynamics of non-Abelian gauge theory.

The Gluon Propagator

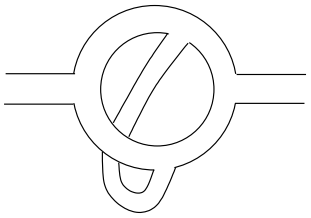
The ideas above don't just apply to the vacuum bubbles. A similar distinction holds for any Feynman diagram. We can, for example, consider the gluon propagator (6.3). A planar, one-loop correction is given by the diagram



$$\sim \left(\frac{\lambda}{N}\right)^4 \left(\frac{N}{\lambda}\right)^2 N \sim \frac{\lambda^2}{N}$$

Now we sum only over the indices on the internal loop, because we have fixed the external legs. We see that this again gives a contribution with the same $1/N$ scaling as the original propagator (6.3), but is down by a power of the 't Hooft coupling.

Meanwhile, the following two-loop, non-planar graph scales as



$$\sim \left(\frac{\lambda}{N}\right)^7 \left(\frac{N}{\lambda}\right)^4 \sim \frac{\lambda^3}{N^3}$$

and is suppressed by $1/N^2$ compared to the earlier contributions.

The Topology of Feynman Diagrams

Let's understand better how to order the different diagrams. We'll return to the vacuum diagrams. The key idea is that each of these can be inscribed on the surface of a two dimensional manifold of a given topology.

The planar diagrams can all be drawn on the surface of a sphere. This is because for any graph on a sphere, you can remove one of the faces and flatten out what's left to give the planar graph. The simplest example is the vacuum diagram (6.5) which sits nicely on the sphere as shown on the right.

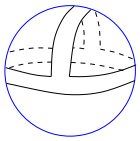
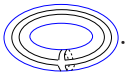


Figure 48:

In contrast, the non-planar diagrams must be drawn on higher genus surfaces. For example, the non-planar vacuum diagram (6.7) cannot be inscribed on a sphere, but requires a torus. It also requires more artistic skill than I can muster, but looks something like  .

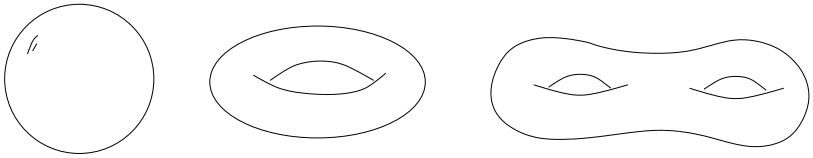


Figure 49: Examples of the simplest Riemann surfaces with $\chi = 2, 0$ and -2.

In general, the Feynman diagram tiles a two dimensional surface Σ. The map is

$$\begin{aligned} E &= \#\text{ of edges} = \#\text{ of propagators} \\ F &= \#\text{ of faces} = \#\text{ of index loops} \\ V &= \#\text{ of vertices} \end{aligned}$$

From (6.4), a given diagram then scales as

$$\text{diagram} \sim N^{F+V-E}\lambda^{E-V}$$

But there is a beautiful fact, due to Euler, which says that the following combination determines the topology of the Riemann surface

$$\chi(\Sigma) = F + V - E \tag{6.8}$$

The quantity $\chi(\Sigma)$ is called the *Euler character*. It is related to the number of handles H of the Riemann surface, also called the *genus*, by

$$\chi(\Sigma) = 2 - 2H \tag{6.9}$$

The simplest examples are shown in the figure. The sphere has $H = 0$ and $\chi = 2$; the torus has $H = 1$ and $\chi = 0$; the thing with two holes has $H = 2$ and $\chi = -2$. In this way, the large N expansion is a sum over Feynman diagrams, weighted by their topology

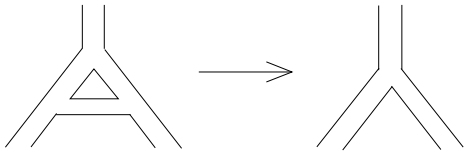
$$\text{diagram} \sim N^{\chi}\lambda^{E-V}$$

For each genus, the Riemann surface can be tiled in different ways by Feynman diagram webs, giving the expansion in the 't Hooft coupling. There is no topological interpretation of this exponent $V - E$. We'll shortly discuss the implication of this large N expansion.

The Euler Character

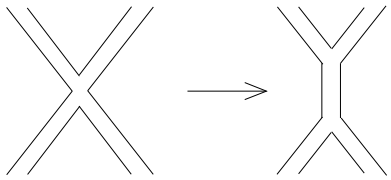
Before we proceed, it will be useful to get some intuition for why the Euler character (6.8) is a topological invariant, and why it is given by (6.9).

To see the former, it's best to play around a little bit by deforming various diagrams. The key manipulation is to take a face and shrink it to vanishing size. For example, we have



Under such a transformation, the number of faces shrinks by 1: $F \to F-1$. The number of vertices has also decreased, $V \to V-2$, as has the number of edges, $E \to E-3$. But the combination $\chi = F + V - E$ remains unchanged.

In all the examples above, we used only the cubic Yang-Mills vertex. Including the quartic vertex doesn't change the counting. This is because we can always split the quartic vertex into two cubic ones,



The left hand side has $V = 1$ and $E = 4$, which transforms into the right hand side with $V = 2$ and $E = 5$. We see neither χ, nor the power of λ depend on the kind of vertex that we use.

This should help explain why the Euler character does not vary under manipulations that make the diagram more and more complicated, but leave the underlying topology unchanged. For the sphere, the example we drew above shows that $\chi = 2$. For each extra handle, we can consider first cutting a hole in the surface. We do this by removing a face, leaving us with a boundary. To build a handle, we cut out two faces, each of which is an n-gon. This reduces the number of faces $F \to F - 2$. Now we glue the faces together by identifying the perimeters of the holes. This act reduces $E \to E - n$ and $V \to V - n$. But the net effect is that for each handle we add, $\chi \to \chi - 2$.

6.2.2 A Stringy Expansion of Yang-Mills

The large N limit of Yang-Mills has been repackaged as a sum over Riemann surfaces of different topologies. But this is the defining feature of weakly coupled string theory. This is discussed in much detail in the lectures on String Theory; here we'll just mention some pertinent facts.

In string theory, the sum over Riemann surfaces is weighted by the string coupling constant g_s. By analogy, we see that

$$g_s = \frac{1}{N}$$

But there are also differences. In string theory, the Riemann surfaces are smooth objects, which suffer quantum fluctuations governed by the inverse string tension α'. This is a quantity with dimension $[\alpha'] = -2$ and it is often written as $\alpha' = l_s^2$ with l_s the typical size of a string. The fluctuations of the Riemann surface are really governed by α'/L^2 where L is the spatial size of the background in which the string propagates.

In contrast, the Riemann surfaces that arise in the large N expansion are not smooth at all; they are tiled by Feynman diagrams and in the perturbative limit, $\lambda \ll 1$, the diagrams with the fewest vertices dominate. However, taken naively, it appears that in the opposite limit $\lambda \gg 1$, the diagrams with large numbers of vertices are important. With some imagination, these can be viewed as diagrams which finely cover the Riemann surface, so that it looks more and more like a classical geometry. This suggests that, in the 't Hooft limit, strongly coupled Yang-Mills may be a weakly coupled string theory in some background, with

$$\lambda^{-1} \sim \left(\frac{\alpha'}{L^2}\right)^{\#}$$

where I've admitted ignorance about the positive exponent #.

This is a bold idea. Weakly coupled string theory is a theory of quantum gravity, and gives rise to general relativity at long distances. If we can somehow make the idea above fly, then Yang-Mills theory would contain general relativity! But the strings and gravity would not live in the $d = 3 + 1$ dimensions of the Yang-Mills theory. Instead, we would find gravity in the "space in which the Feynman diagrams live", whatever that means.

So far, no one has made sense of these ideas for pure Yang-Mills. However, it is now understood how these ideas fit together in a very closely related theory called maximally supersymmetric (or $\mathcal{N} = 4$) Yang-Mills which is just $SU(N)$ Yang-Mills coupled to a bunch of adjoint scalars and fermions. In that case, the strongly coupled 't Hooft limit is indeed a theory of gravity in a $d = 9 + 1$ dimensional spacetime that has the form $AdS_5 \times \mathbf{S}^5$. The $d = 3 + 1$ dimensional world in which the Yang-Mills theory lives is the boundary of AdS_5. This remarkable connection goes by the name of the AdS/CFT correspondence or, more generally, gauge-gravity duality. It is a topic for another course.

It's an astonishing fact that, among the class of gauge theories in $d = 3{+}1$ dimensions, is a theory of quantum gravity in higher dimensional spacetime. It leaves us wondering just what else is hiding in the land of strongly coupled quantum field theories.

6.2.3 The Large N Limit is Classical

We can use the large N counting described above to understand the scaling of correlation functions.

In what follows, we consider gauge invariant operators which cannot be further decomposed into colour singlets. Since Yang-Mills has only adjoint fields, this means that we are interested in operators that have just a single trace. The simplest is

$$\mathcal{G}_{\mu\nu,\rho\sigma}(x) = \mathrm{tr}\, F_{\mu\nu}F_{\rho\sigma}(x)$$

There's a slew of further operators in which we add more powers of $F_{\mu\nu}$ inside the trace. However, it's important that the number of fields inside the trace is kept finite as we take $N \to \infty$, otherwise it will infect our N counting. This means, for example, that we can't discuss operators like $\det F_{\mu\nu}F^{\mu\nu}$. Of course, Yang-Mills also has non-local operators – the Wilson loops – and much of what we say will hold for them. But, for once, our main interest will be on the local, single trace operators.

We could also consider coupling our theory to adjoint matter, either scalars or fermions. Restricting to the adjoint representation means that these new fields are also $N \times N$ matrices, and the same $1/N$ counting that we developed above holds for their Feynman diagram expansion. This gives us the option to build more single trace operators, such as $\mathcal{G} = \mathrm{tr}(\phi^m)$ for a scalar ϕ, or combinations of scalars and field strengths. Once again, we insist only that the number of fields inside the trace does not scale with N.

We can compute correlation functions of any of these operators by adding sources in the usual way,

$$S_{YM} = N \int d^4x \ -\frac{1}{2\lambda}\mathrm{tr}\, F^{\mu\nu}F_{\mu\nu} + \ldots + J_a\mathcal{G}_a$$

where the $\ldots$ is any further adjoint matter that we've included, and where the operators $\mathcal{G}_a$ denote any single trace involving strings of the field strength, the other adjoint matter, or their derivatives. Note that we've scaled both fields and operators to keep an overall factor of N in front of the action. The connected correlation functions can be computed in the usual way by differentiating the partition function,

$$\langle \mathcal{G}_1 \ldots \mathcal{G}_p \rangle_c = \frac{1}{N^p}\frac{\delta}{\delta J_1}\ldots\frac{\delta}{\delta J_p}\ \log\, Z[J] \tag{6.10}$$

where the subscript c is there to remind us that we're dealing with connected correlators. Because the action, including the source terms, has the form $S = N \,\mathrm{tr}\,(\text{something})$, our previous large N counting goes over unchanged, and the free energy is dominated by planar graphs at order $\log Z \sim N^2$. (This conclusion would no longer hold if we included multi-trace operators as sources, or if there were some other powers of N that had somehow snuck unseen into the action). We learn that connected correlation functions of single trace operators have the leading scaling

$$\langle \mathcal{G}_1 \dots \mathcal{G}_p \rangle_c \sim N^{2-p} \tag{6.11}$$

where in this formula, and others below, we're ignoring the dependence on the 't Hooft coupling λ.

The simple formula (6.11) is telling us something interesting: the leading contribution to any correlation function comes from disconnected diagrams, rather than connected diagrams. For example, any two-point function has a connected piece $\langle \mathcal{G}\mathcal{G} \rangle \sim \langle \mathcal{G} \rangle \langle \mathcal{G} \rangle \sim N^2$. This should be contrasted with the connected piece which scales as $\langle \mathcal{G}\mathcal{G} \rangle_c \sim N^0$.

This means that the strict $N \to \infty$ limit of Yang-Mills is a free, classical theory. All correlation functions of single trace, gauge invariant operators factorise. Said slightly differently, quantum fluctuations are highly suppressed in the large N limit, with the variance of any gauge singlet operator $\mathcal{O}$ given by

$$(\Delta \mathcal{G})^2 = \langle \mathcal{G}\mathcal{G} \rangle - \langle \mathcal{G} \rangle \langle \mathcal{G} \rangle = \langle \mathcal{G}\mathcal{G} \rangle_c \sim N^0 \quad \Rightarrow \quad \frac{(\Delta \mathcal{G})^2}{\langle \mathcal{G} \rangle^2} \sim \frac{1}{N^2}$$

Usually when we hear the words "free, classical theory", we think "easy". That's not the case here. The large N limit is a theory of an infinite number of single trace operators $\mathcal{G}_a(x)$. If the theory is confining and has a mass gap, like Yang-Mills, each of these corresponds to a particle in the theory. (We will make this connection clearer below.) Or, to be more precise, each of the operators $\mathcal{G}(x)$ corresponds to some complicated linear combination of particles in the theory. After diagonalising the Hamiltonian, we will have a free theory of an infinite number of massive particles. Determining these masses is a difficult problem which remains unsolved.

The large N limit does not only hold for confining theories. For example, maximally supersymmetric Yang-Mills is a conformal field theory and does not confine. Now the goal in the large N limit is to diagonalise the dilatation operator to find the conformal dimensions of single trace operators. This is a difficult problem that is largely solved using techniques of integrability.

The fact that the large N limit is free leads to the concept of the *master field*. There should be a configuration of the gauge fields A_μ on which we can evaluate any correlation function to get the correct $N \to \infty$ answer. (If we add more adjoint matter fields, we would need to specify their value as well.) Once we have this master field, there is nothing left to do: no fluctuations, no integrations. We just evaluate. Furthermore, the master field should be translationally invariant so, at least in a suitable gauge, the A_μ are just constant. In other words, all of the information about Yang-Mils in the $N \to \infty$ limit is contained in four matrices, A_μ. The twist, of course, is that these are $\infty \times \infty$ matrices and, as a well known physicist is fond of saying, "you can hide a lot in a large N matrix". For pure Yang-Mills in $d = 3 + 1$ dimensions, no progress has been made in understanding the master field in decades. For maximally supersymmetric Yang-Mills, the master field should be equivalent to saying that the theory is really ten dimensional gravity in disguise.

6.2.4 Glueball Scattering and Decay

The strict $N \to \infty$ limit is free, with the degrees of freedom organised in single trace operators $\mathcal{G}(x)$. All of the difficulties of the strong coupling dynamics goes into diagonalising the Hamiltonian to determine masses (or scaling dimensions) of the corresponding states.

At large, but finite N, we introduce interactions between these degrees of freedom, which must scale as some power of $1/N$. Even though we can't solve the $N \to \infty$ limit, we can still get some useful intuition for the theory by looking at these interactions in a little more detail.

To see this, let's revert to pure Yang-Mills. We will assume that this theory confines in the large N limit. There is no reason to think this is not the case but it's important to stress that we can currently no more prove confinement in the large N limit than at finite N[11]. We consider the local glueball operators

$$\mathcal{G}(x) = \mathrm{tr}\, F^m(x) \tag{6.12}$$

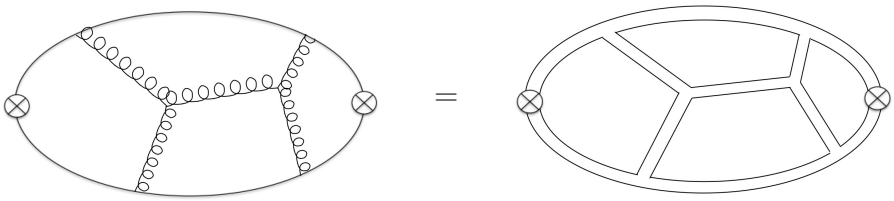
for some $m \geq 2$. We've ignored the Lorentz indices, which endow each operator with a certain spin. We could also include derivatives to increase the spin yet further.

[11]The Millennium Prize Problem requires that you prove confinement for *all* compact, simple non-Abelian gauge groups. This stipulation was put in place to avoid a scenario where confinement was proven only in the large N limit. Apparently, the authors of the problem originally meant to find a different phrasing, one that avoided the caveat of large N but would award a proof of confinement in, say, $SU(3)$ Yang-Mills. But they never got round to changing the wording. Like with all such prizes, if you're genuinely interested in the million dollars then you are probably in the wrong field.

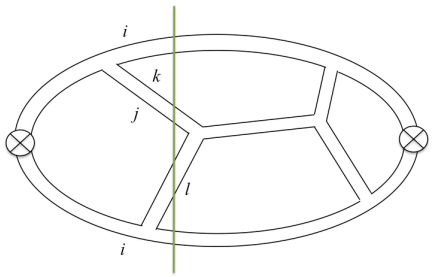
At large N, there is a connected component to the two point function which, with the normalisation (6.11), scales as

$$\langle \mathcal{G}(x)\,\mathcal{G}(0)\rangle_c \sim N^0$$

which means that $\mathcal{G}(x)$ creates a glueball state with amplitude of order 1. In terms of our original Feynman diagrams, this picks up contributions from very complicated processes, such as the one below



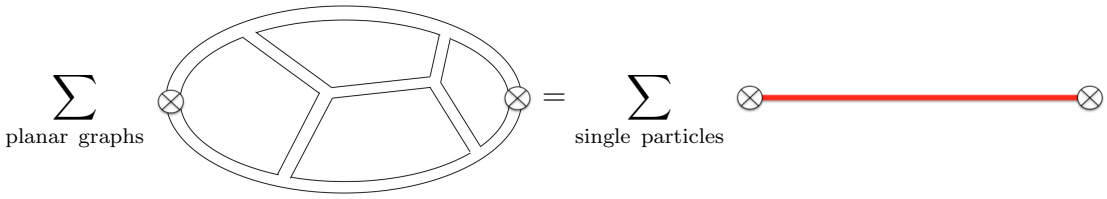
In the large N limit, this is converted into tree-level propagation of gauge singlet operators created by $\mathcal{G}(x)$. Importantly, the operator $\mathcal{G}(x)$ creates only single-particle states. To see this, we can cut the diagram to see the intermediate state, as shown below



We've now included $i, j = 1, \ldots, N$ indices to help keep track. To make something gauge invariant, we need to take the trace, which means combining each index with its partner. The only way to do this is to include all the internal legs together. This is the statement that the internal state corresponds to a single trace operator. In contrast, multi-particle states only propagate in non-planar diagrams where the internal lines can be combined into multi-trace colour singlets.

The fact that the single-trace operator $\mathcal{G}(x)$ creates single particle states also follows from the scaling of the correlation function (6.11). To see this, first suppose that the statement isn't true, and $\mathcal{G}$ creates a two particle state with amplitude order 1. Then one could construct a suitable correlation function which has the value $\langle \mathcal{G}\tilde{\mathcal{G}}\tilde{\mathcal{G}}\mathcal{G}\rangle \sim 1$, with the operators $\tilde{G}$ each interacting, with amplitude 1, with one of the the two intermediate particles. But we know from large N counting (6.11) that $\langle \mathcal{G}\tilde{\mathcal{G}}\tilde{\mathcal{G}}\mathcal{G}\rangle \sim 1/N^2$. (There is actually an implicit assumption here that there is no degeneracy of states at order N. But this is precisely the assumption of confinement.)

So we can think of any two-point function $\langle \mathcal{G}(x)\,\mathcal{G}(0)\rangle_c$ as the tree-level propagation of confined, single particle states. We are repackaging



In general, the only singularities in tree-level graphs are poles. (This is to be contrasted with one-loop diagrams where we can have two-particle cuts, and higher loop diagrams with multi-particles cuts.) This means that there should be some expansion of the two-point function in momentum space as

$$\langle \mathcal{G}(k)\,\mathcal{G}(-k)\rangle_c = \sum_n \frac{|a_n|^2}{k^2 - M_n^2} \tag{6.13}$$

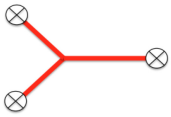
where $a_n = \langle 0|\mathcal{G}|n\rangle$, with $|n\rangle$ the single particle state with mass M_n. But now there's something of a puzzle. At large k, Yang-Mills theory is asymptotically free, and we can compute this correlation function to find that it scales as

$$\langle \mathcal{G}(k)\,\mathcal{G}(-k)\rangle_c \to k^2 \log\, k^2$$

Yet naively the propagator (6.13) would appear to scale as $1/k^2$ for large momentum. The only way we can reproduce the expected log behaviour is if there are an infinite number of stable intermediate states $|n\rangle$, with an infinite tower of masses m_n. This coincides with our earlier expectations: as $N \to \infty$ Yang-Mills is a theory of an infinite number of free particles.

At large but finite N, there can no longer be an infinite tower of stable, massive particles. The heavy ones surely decay to the light ones. But this process is captured by the correlation functions of the schematic form

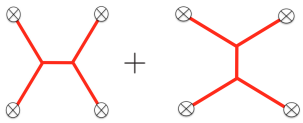
$$\langle \mathcal{G}\mathcal{G}\mathcal{G}\rangle \sim \sum$$



$$+ \ldots \sim \frac{1}{N}$$

which tells us that the amplitude for a glueball to decay to two glueballs scales as $1/N$, so their lifetime scales as N^2. Similarly, for scattering we can turn to the four-point function

$$\langle \mathcal{G}\mathcal{G}\mathcal{G}\mathcal{G}\rangle \sim \sum$$



$$+ \ldots \sim \frac{1}{N^2}$$

So the amplitude for gluon-gluon scattering scales as $1/N^2$.

6.2.5 Theta Dependence Revisited

We saw in Section 2.2 that Yang-Mills theory comes with an extra, topological parameter: the theta-term. How does this fare in the large N limit? The Lagrangian is

$$\begin{aligned}\mathcal{L}_{YM} &= -\frac{1}{2g^2}\text{tr}\, F_{\mu\nu}F^{\mu\nu} + \frac{\theta}{16\pi^2}\text{tr}\, F_{\mu\nu}{}^\star F^{\mu\nu} \\ &= N\left(-\frac{1}{2\lambda}\text{tr}\, F_{\mu\nu}F^{\mu\nu} + \frac{\theta}{16\pi^2 N}\text{tr}\, F_{\mu\nu}{}^\star F^{\mu\nu}\right)\end{aligned}$$

With the appropriate factor of N sitting outside the action, we see that we should keep θ/N fixed as we send $N \to \infty$. The first question that we should ask is: does the physics still depend on θ?

At first glance, it appears that the answer to this question should be no. The reasons for this are two-fold. At leading order in perturbation theory, none of the planar graphs appear to depend on θ. Moreover, the instanton effects which, at weak coupling, give us θ dependence now scale as $\sim e^{-8\pi^2/g^2} \sim e^{-8\pi^2 N/\lambda}$ and so are exponentially suppressed in the large N limit.

Although both of these arguments appear compelling, the conclusion is thought to be wrong. It is believed that, at leading order in the $1/N$ expansion, the physics continues to depend on θ (or, more precisely, on θ/N). Perhaps the simplest observable is the ground state energy, defined schematically in the Euclidean path integral as

$$e^{-V\,E(\theta)} = \int \mathcal{D}A\, \exp\left(-\int d^4x\, \mathcal{L}_{YM}\right) \tag{6.14}$$

where V is the spacetime volume. Recall that, in Euclidean space, the theta term weights the path integral as $e^{i\theta\nu}$ where ν is the topological winding of the configuration. The large N arguments that we've seen above tell us that $E \sim N^2$. It is believed that the θ dependence affects this quantity at leading order

$$E(\theta) = N^2 h\left(\frac{\theta}{N}\right) \tag{6.15}$$

for some function $h(x)$.

There are two main reasons for thinking that θ dependence survives in the large N limit. The first is that, in the presence of light quarks, the dependence can be seen in the chiral Lagrangian; we will describe this in Section 6.4. The second is that both the arguments we gave above also hold in toy models in two-dimensions (specifically the $\mathbf{CP}^N$ model that we will introduce in 7.3) where one can see that they lead to the wrong conclusion. The loophole lies in the first argument; at leading order in the $1/N$ expansion we must sum an infinite number of diagrams, and interesting things can happen for infinite series that don't arise for finite sums.

To make this more concrete, let's introduce the *topological susceptibility*,

$$\chi(k) = \int d^4x\, e^{ik\cdot x} \langle \mathrm{tr}\,(F_{\mu\nu}{}^\star F^{\mu\nu}(x))\,\mathrm{tr}\,(F_{\rho\sigma}{}^\star F^{\rho\sigma}(0))\rangle \tag{6.16}$$

(Not to be confused with the Euler character that we encountered earlier.) Roughly speaking, this tells us how the theory responds to changes in θ. In particular, the ground state energy $E(\theta)$ has the dependence

$$\frac{d^2E}{d\theta^2} = \left(\frac{1}{16\pi^2 N}\right)^2 \lim_{k\to 0}\chi(k) \tag{6.17}$$

We can compute contributions to $\chi(k)$ in perturbation theory. One finds that, at leading order in $1/N$, each individual diagram has $\chi(k) \to 0$ as $k \to 0$. Nonetheless, it is expected that the sum of all such diagrams does not vanish. No one has managed to perform this calculation explicitly in four-dimensional Yang-Mills theory. To see that such behaviour is indeed possible, you need only consider the series

$$f(k) = k^2 \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \log^n k^2 = k^2 \exp^{-\log k^2} = 1$$

The behaviour of the ground state energy (6.15) brings a new puzzle. The energy depends on θ/N, but must obey $E(\theta) = E(\theta + 2\pi)$. How can we reconcile these two properties? The accepted answer – and the one which is seen in the $\mathbf{CP}^N$ model – is that there is a level crossing in the ground state as θ is varied. This works as follows: at large N the theory is thought to have a large number of meta-stable, Lorentz-invariant states that differ in energy. There are order N such states and, in the k^{th}, the energy is given by

$$E_k(\theta) = N^2 h\left(\frac{\theta + 2\pi k}{N}\right)$$

The ground state energy is then

$$E(\theta) = \min_k E_k \tag{6.18}$$

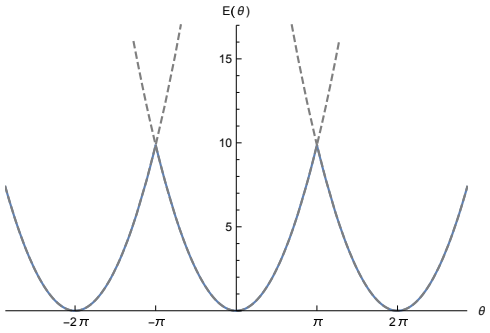


Figure 50: The vacuum energy as a function of θ.

We're left with a function which is periodic, but not smooth. In particular, when $\theta = \pi$ two levels cross.

What does the function $E(\theta)$ look like? First, we know that it has its minimum at $\theta = 0$. This is because the Euclidean path integral (6.14) is a sum over configurations weighted by $e^{i\theta}$. Only for $\theta = 0$ is this real and positive, hence maximising $e^{-VE(\theta)}$, and so minimising $E(\theta)$. Taylor expanding, we therefore expect that

$$E(\theta) = \min_k \frac{1}{2}C(\theta + 2\pi k)^2 + \mathcal{O}\left(\frac{1}{N}\right)$$

where $C = \chi(0)/(16\pi^2 N)^2$. This is shown in the figure.

A general value of θ explicitly breaks time-reversal or, equivalently, $\mathcal{CP}$. The two exceptions are $\theta = 0$ and $\theta = \pi$. (We explained why $\theta = \pi$ is time reversal invariant in Section 1.2.5.) But, at $\theta = \pi$, there are two degenerate ground states and time-reversal invariance maps one to the other. We learn that, at large N Yang-Mills, time-reversal invariance is spontaneously broken at $\theta = \pi$. This coincides with our conclusion from Section 3.6 using discrete anomalies.

6.3 Large N QCD

Our discussion in the previous section focussed purely on matrix valued fields. To get closer to QCD, we add quarks, as Dirac fermions in the fundamental representation.

We rescale the quark field $\psi \to \sqrt{N}\psi$, so that the action continues to have a factor of N sitting outside,

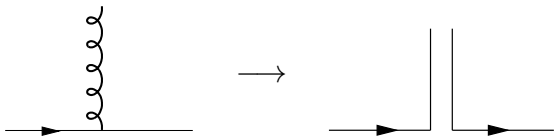
$$S_{QCD} = N \int d^4x \ -\frac{1}{2\lambda}\text{tr}\, F^{\mu\nu}F_{\mu\nu} + i\bar{\psi}\not{D}\psi$$

We'll stick with just a single quark field for now, but everything that we say will go over for N_f flavours of quarks provided that we keep N_f fixed as $N \to \infty$.

The quark field carries just a single gauge index, ψ^i with $i = 1, \ldots, N$. Correspondingly, it is represented by just a single line in a Feynman diagram,

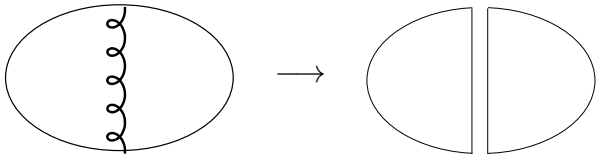
$$\sim \frac{\lambda}{N}$$

Meanwhile, the quark-gluon vertex is represented by



$$\sim N$$

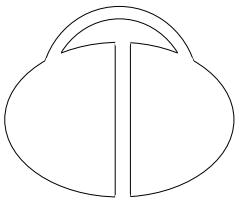
We can now repeat the large N counting that we saw previously. We can start by looking at contributions to the vacuum energy that include a quark loop. For example, we have



$$\sim \left(\frac{\lambda}{N}\right)^3 N^2\, N^2 \sim \lambda^3 N$$

where the first factor of N^2 comes from the two quark-gluon vertices, while the second factor comes from the index loops. We see that this is subleading compared to the pure glue vacuum diagrams which are $\sim N^2$. Including extra internal gluons, all planar diagrams with a single quark loop on the boundary will continue to scale as $\sim N$. This is the leading order contribution to the vacuum energy that includes quarks. This is simple to understand: the amplitude to create a quark is the same as the amplitude to create a gluon, but there are N^2 gluon degrees of freedom and only N quark degrees of freedom.

If the quark loop does not run around the boundary, the diagram is suppressed yet further. For example, consider the diagram



$$\sim \left(\frac{\lambda}{N}\right)^6 N^4\, N \sim \lambda^6 N^{-1}$$

Similarly, if we include internal quark lines in other Feynman diagrams, say the gluon propagator, we again get a suppression factor of $1/N$.

We can again interpret the large N Feynman diagrams in terms of 2d surfaces. However, now the surfaces are no longer closed. Instead, each quark loop should be thought of as the boundary of a hole on the Riemann surface. Each boundary increases the number of edges E by one, so a given Feynman diagram again scales as

$$\text{diagram} \sim N^{F+V-E}\,\lambda^{E-V} = N^{\chi}\lambda^{E-V}$$

which is the same result that we had before. But now the expression for the Euler character is

$$\chi = 2 - 2H - B$$

where B is the number of boundaries, or holes, in the surface.

In terms of string theory, the addition of quarks means that the large N limit includes open strings, with boundaries, as well as closed strings. This is closely related to the concept of D-branes in string theory.

6.3.1 Mesons

We can now rerun the arguments of Sections 6.2.3 and 6.2.4 for large N QCD. In addition to the glueball operators (6.12), we also have the meson operators

$$\mathcal{J}(x) = \sqrt{N}\,\bar{\psi}F^m\psi \tag{6.19}$$

where the F^m can denote any number of field strengths, derivatives and gamma matrices, so that $\mathcal{J}(x)$ is a local, gauge invariant operator that cannot be decomposed into smaller colour singlets.

Note that we've included an overall factor of $\sqrt{N}$ in (6.19). To see why this is, we compute correlation functions

$$\langle \mathcal{J}_1 \dots \mathcal{J}_p \rangle_c \sim N^{1-p/2} \tag{6.20}$$

The first factor of N comes from the planar diagrams with a quark loop running along the boundary. The normalisation factor of $\sqrt{N}$ in (6.19) means that correlation function scale as $N^{-p/2}$ rather than as N^{-p}. This normalises the two-point function as $\langle \mathcal{J}\mathcal{J} \rangle_c \sim N^0$, so $\mathcal{J}$ creates a meson state with amplitude 1.

The same arguments that we used for pure Yang-Mills still apply here. The strict $N \to \infty$ limit is again a free theory, now including infinite towers of both glueball and meson states. In momentum space, the analog of the propagator (6.13) is

$$\langle \mathcal{J}(k)\, \mathcal{J}(-k)\rangle_c = \sum_n \frac{|b_n|^2}{k^2 - m_n^2} \tag{6.21}$$

where $b_n = \langle 0|\mathcal{J}|n\rangle$, with $|n\rangle$ the single particle meson state with mass m_n. As for glueballs, this expression is only compatible with the log behaviour of asymptotic freedom if there is an infinite tower of massive meson states.

At large N, the three point function of meson fields

$$\langle \mathcal{J}\mathcal{J}\mathcal{J}\rangle \sim \frac{1}{\sqrt{N}}$$

tells us that the amplitude for a meson to decay into two lighter mesons scales as $1/\sqrt{N}$. The lifetime of a meson is then typically of order N. They are shorter lived than the glueballs. Similarly, the four point function of meson fields is

$$\langle \mathcal{J}\mathcal{J}\mathcal{J}\mathcal{J}\rangle \sim \frac{1}{N}$$

The amplitude for meson-meson scattering scales as $1/N$.

We can also compute correlation functions of both glueballs and mesons. At leading order, we have

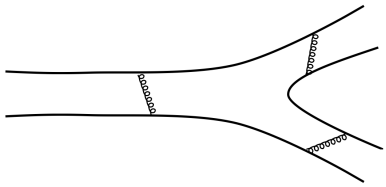
$$\langle \mathcal{J}_1 \ldots \mathcal{J}_p\, \mathcal{G}_1 \ldots \mathcal{G}_q\rangle \sim N\, N^{-p/2}\, N^{-q}$$

This means that the two-point function $\langle \mathcal{J}\mathcal{G}\rangle \sim 1/\sqrt{N}$, so mesons and glueballs don't mix at large N, even if they share the same quantum numbers. (We had assumed this when talking separately about meson and glueballs above, so it's good to know it's true.) We can also extract the amplitude for a gluon to decay into two mesons which is $\langle \mathcal{G}\mathcal{J}\mathcal{J}\rangle \sim 1/N$, which is the same order as the decay into two gluons. Meanwhile, the amplitude for a meson to decay into two gluons is $\langle \mathcal{J}\mathcal{G}\mathcal{G}\rangle \sim 1/N^{3/2}$. We see that a gluon doesn't much mind who it decays into, while a meson greatly prefers decaying into other mesons.

The OZI Rule

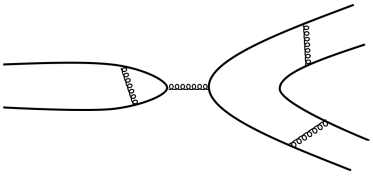
The large N approach helps explain a couple of phenomenological facts that had been previously observed to hold for QCD. In particular, note that the leading order meson

decays have the form



$$\sim \frac{1}{\sqrt{N}}$$

In such a process, one of the original quarks ends up in each of the final decay products. In contrast, a process in which the two original quarks decay into pure glue which subsequently produces two further mesons, is suppressed by an extra factor of $1/N$,



$$\sim \frac{1}{N^{3/2}}$$

This suppression was observed experimentally in the early days of meson physics and goes by the name of the OZI rule (for Okubo, Zweig and Iizuka; it is also sometimes called the Zweig rule).

The standard example is the ϕ vector meson, which has quark content $s\bar{s}$. On energy considerations alone, one would have thought this would decay to $\pi^+\pi^-\pi^0$, none of which contain a strange quark. In reality, this decay is suppressed by QCD dynamics, and the ϕ meson decays primarily to K^+K^-, where the positively charged kaon has quark content $u\bar{s}$. This fact is clearest in the $1/N$ expansion.

The large N expansion also makes it clear that we don't expect to see meson bound states or, more generally, $\bar{q}q\bar{q}q$ states with four quarks. Such states are referred to as *exotics*. The amplitude for meson interactions scales as $1/N$, so such exotics certainly don't form in the large N limit. The lack of exotics in particle data book suggests that this suppression extends down to $N = 3$.

6.3.2 Baryons

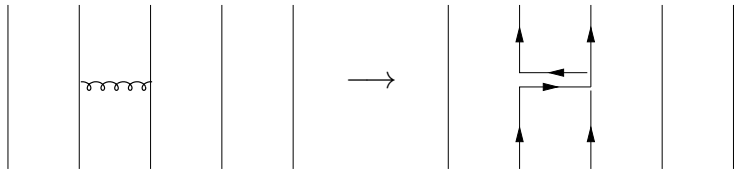
We now turn to baryons. These are a little more subtle because they contain N quarks, anti-symmetrised over the colour indices. Nonetheless, as first explained by Witten, they are naturally accommodated in the large N limit of QCD.

In what follows we will consider the large N limit with just a single flavour of quark, although it is not difficult to include $N_f > 1$ flavours. The baryon is then

$$B = \epsilon^{i_1 \ldots i_N} \psi_{i_1} \ldots \psi_{i_N} \tag{6.22}$$

This is the large N analog of, say, the Δ^{++} in QCD which contains three up quarks, or the Δ^- which contains three down quarks.

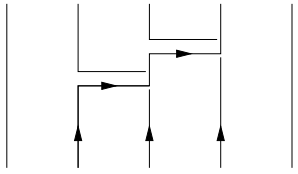
We can start by modelling these as N distinct quark lines. A gluon exchange between any pair of quarks is



$$\sim \frac{1}{N} \tag{6.23}$$

where we've been more careful in the second diagram in showing how the arrows flow. However, there $\frac{1}{2}N(N-1) \sim N^2$ different pairs of quarks, so the total amplitude for a gluon exchange within a baryon is order N.

There is a similar story for three body interactions. The gluon exchange is now



$$\sim \frac{1}{N^2} \tag{6.24}$$

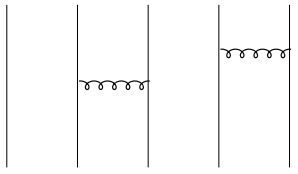
but there are order N^3 triplets of quarks, so again the total amplitude scales as N.

These simple arguments suggest that many-body interactions are all equally important, and contribute to the energy of the baryon at order N. It is therefore natural to guess that

$$M_{\rm baryon} \sim N \tag{6.25}$$

This is perhaps not a surprise since the baryon contains N quarks, and is certainly to be expected in the non-relativistic quark model.

There's a calculation which may give you pause. Consider the the gluon exchange between two different pairs of quarks,



$$\sim \frac{1}{N^2} \tag{6.26}$$

But now there are $\sim N^4$ ways of picking two pairs of quarks, so it looks as if this contributes to the energy at order $N^4 \times N^{-2} \sim N^2$. It seems like we get increasingly

divergent answers as we look at more and more disconnected pieces. In fact, this is the kind of behaviour that we would expect if the baryon mass scales as (6.25). The propagator for large times T then takes the form

$$e^{-iM_{\rm baryon}T} \approx 1 - iM_{\rm baryon}T - \frac{1}{2}M^2_{\rm baryon}T^2 + \ldots$$

For the diagram (6.26), each of the gluons can be exchanged at any time and so it corresponds to the second order term in the expansion above which, we see, should indeed scale as $M^2_{\rm baryon} \sim N^2$.

At this point, we could start to explore the interactions between baryons and mesons, and build towards a fuller phenomenology of QCD. However, we won't go in this direction. Instead, I will point out a nice connection between baryons in the large N expansion and another recurring topic from these lectures.

The Hartree Approximation

A particularly simple way to proceed is to assume that the quarks are non-relativistic. This is not particularly realistic for QCD, but it will provide a simple way to shine a light on the structure of the baryon. If each quark has mass m, we could try to model their physics inside a baryon by the following Hamiltonian

$$H = Nm + \frac{1}{2m}\sum_{i=1}^{N} p_i^2 + \frac{1}{2N}\sum_{i\neq j} V_2(x_{ij}) + \frac{1}{6N^2}\sum_{i\neq j\neq k} V_3(x_{ij}, x_{jk}) + \ldots$$

where $x_{ij} = x_i - x_j$ and the coefficients in front of the potentials are taken from (6.23) and (6.24). We should also include all multi-particle potentials. As we have seen, it is a mistake to think that these potentials are genuinely suppressed by the $1/N$ factors in the Hamiltonian: these are compensated by the sums over particles, so each term ends up of order N.

There is a straightforward variational approach to such many-body Hamiltonians called the Hartree approximation. It is the first port of call in atomic physics, when studying atoms with many electrons, and we met it in the lectures on Topics in Quantum Mechanics. The idea is to work with the ansatz for the ground state wavefunction given by

$$\psi(\mathbf{x}_1, \ldots, \mathbf{x}_N) = \prod_{i=1}^{N} \phi_0(\mathbf{x}_i)$$

Note that the quarks are fermions, but they have already been anti-symmetrised over the colour indices (6.22), so it is appropriate that the wavefunction for the remaining degrees of freedom is symmetric.

The Hartree ansatz neglects interactions between the quarks. Instead, it is a self-consistent approach in which each quark experiences a potential due to all the others. This approach becomes increasingly accurate as the number of particles becomes large, so it is particularly well suited to baryons in the large N limit.

Evaluating the Hamiltonian on the Hartree wavefunction gives

$$\langle\psi|H|\psi\rangle = N\ \left[m + \frac{1}{2m}\int d^3x\ |\phi(\mathbf{x})|^2 + \frac{1}{2}\int d^3x_1 d^3x_2\ V_2(x_{12})|\phi(\mathbf{x}_1)\phi(\mathbf{x}_2)|^2 + \frac{1}{6}\int d^3x_1 d^3x_2 d^3x_3\ V_3(x_{12},x_{23})|\phi(\mathbf{x}_1)\phi(\mathbf{x}_2)\phi(\mathbf{x}_3)|^2\right]$$

We then find the $\phi(\mathbf{x})$ which minimises this expression. This, obviously, is a hard problem. But fortunately it is not one we need to solve in order to extract the main lessons. These come simply from the fact that there is a factor of N outside the bracket, but nothing inside. This confirms our earlier conclusion (6.25) that the mass of the baryon indeed scales as $M_{\rm baryon} \sim N$. But we also learn something new, because whatever function $\phi(\mathbf{x})$ ends up being, it certainly does not depend on N. This means that the size of the baryon – its spatial profile in $\phi(\mathbf{x})$ – is order 1.

The mass and size of the baryon are rather suggestive. Recall that the large N limit is a theory of weakly coupled gauge singlets, interacting with coupling $1/N$. This means that the mass of the baryon scales as the inverse coupling, N, with the size independent of the coupling. But this is the typical behaviour of solitons. For example, the 't Hooft Polyakov monopole that we met in Section 2.8 has a mass which scales as $1/g^2$ and a size which is independent of g^2. This strongly suggests that the baryon should emerge as a soliton in large N QCD.

We have, of course, already seen a context in which baryons emerge as solitons: they are the Skyrmions in the chiral Lagrangian that we met in Section 5.3. To my knowledge, this connection has not been fully explained.

Before we move on, there is one further twist to the "baryons as solitons" story. The mass of the baryon, N, is not quite like the mass of the monopole: it is proportional to the inverse coupling, rather than the square of the inverse coupling. Returning to the language of string theory that we introduced in Section 6.2.2, the mass of the baryon scales as

$$M_{\rm baryon} \sim \frac{1}{g_s}$$

with $g_s = 1/N$ the string coupling constant. This suggests that baryons are a rather special kind of soliton: they are D-branes. These are objects in string theory on which

strings can end, and have a number of magical properties. (You can read more about D-branes in the lectures on String Theory.) With its N constituent quarks, the baryon is indeed a vertex on which N QCD flux tubes can end.

6.4 The Chiral Lagrangian Revisited

In this section, we will see what becomes of the chiral Lagrangian at large N. Let's first recall the usual story: Yang-Mills coupled to N_f massless fermions has a classical global symmetry

$$G = U(N_f)_L \times U(N_f)_R \tag{6.27}$$

However, the anomaly means that $U(1)_A$ does not survive the quantisation process, leaving us just with $U(1)_V \times SU(N_f)_L \times SU(N_f)_R$. This is subsequently broken to $U(1)_V \times SU(N_f)_V$, and the resulting Goldstone modes are described by the chiral Lagrangian.

How does this story change at large N. The key lies in the anomaly, which is given by

$$\partial_\mu J^\mu_A = \frac{g^2 N_f}{8\pi^2} \text{tr}\, F_{\mu\nu}{}^\star F^{\mu\nu} \tag{6.28}$$

In the large N limit, we send $g^2 \to 0$ keeping $\lambda = g^2 N$ fixed. This suggests that the anomaly is suppressed in the large N limit and the quantum theory enjoys the full chiral symmetry (6.27). This means that there is one further Goldstone mode that appears: the η' meson. In this section we will see how this plays out.

6.4.1 Including the η'

Our first steps are a straightforward generalisation of the chiral Lagangian derived in Section 5.2. The chiral condensate takes the form

$$\langle \bar\psi_{-\tilde i} \psi_{+\tilde j} \rangle = \sigma \Sigma_{\tilde i \tilde j}$$

but now with $\Sigma \in U(N_f)$ rather than $SU(N_f)$. (The ugly $\tilde i, \tilde j = 1, \ldots, N_f$ flavour indices are to ensure that we don't confuse them with the i, j colour indices we've used elsewhere in this Section.) As before, we promote the order parameter to a dynamical field, $\Sigma \to \Sigma(x)$, whose ripples describe our massless mesons, transforming under the chiral symmetry G as

$$\Sigma(x) \to L^\dagger \Sigma(x)\, R \tag{6.29}$$

with $L \times R \in G$. The overall phase of Σ is our new Goldstone boson, η',

$$\det \Sigma = e^{i\eta'/f_{\eta'}} \tag{6.30}$$

We would now like to construct the Lagrangian consistent with the chiral symmetry (6.29). Unlike in Section 5.2, we now have two different terms with two derivatives,

$$(\mathrm{tr}\, \Sigma^\dagger \partial_\mu \Sigma)^2 \quad \text{and} \quad \mathrm{tr}(\partial_\mu \Sigma^\dagger \partial^\mu \Sigma)^2 \tag{6.31}$$

The first term vanishes when $\Sigma \subset SU(N_f)$, but survives when $\Sigma \subset U(N_f)$. In other words, it provides a kinetic term only for η'. Meanwhile, the second term treats all Goldstone modes on the same footing.

Large N-ology tells us that all these mesons have the same properties and, in particular, to leading order in $1/N$ we have $f_{\eta'} = f_\pi$. This means that we need only the second kinetic term and the chiral Lagrangian takes the same form as (5.7),

$$\mathcal{L} = \frac{f_\pi^2}{4} \mathrm{tr}(\partial_\mu \Sigma^\dagger \partial^\mu \Sigma)^2$$

We can compute the expected scaling of f_π with N. Recall that the pion decay constant f_π is defined by (5.13)

$$\langle 0 | J^a_{L\,\mu}(x) | \pi^b(p) \rangle = -i \frac{f_\pi}{2} \delta^{ab}\, p_\mu e^{-ix\cdot p}$$

with J_L a generator of the $SU(N_f)$ flavour current. At this point we need to be a little careful about normalisations. The current J above is defined with the usual kinetic kinetic term $\mathcal{L} \sim i\bar{\psi} \not{D} \psi$. Meanwhile, our large N counting used a different normalisation in which there was an overall factor of N outside the action. Chasing this through, means that the current J_L is related to the appropriate normalised large N current (6.19) by

$$J_L = \sqrt{N} \mathcal{J}_L$$

We can then use the general result (6.20) to find

$$\langle J_L J_L \rangle = \sum_n \langle 0 | J_L | n \rangle \langle n | J_L | 0 \rangle \sim N \quad \Rightarrow \quad \langle 0 | J_L | n \rangle \sim \sqrt{N}$$

This means that the pion decay constant scales as

$$f_\pi \sim \sqrt{N}$$

6.4.2 Rediscovering the Anomaly

So far, things are rather easy. Now we would like to consider what happens at the next order in $1/N$. Obviously, we could add the other kinetic term in (6.31), splitting $f_{\eta'}$ and f_π. This doesn't greatly change the physics and we will ignore this possibility below. Instead, there is a much more dramatic effect that we must take into account, because the anomaly now gives η' a mass. How do we describe that?

We can isolate η' by taking the determinant (6.30), and therefore introduce a mass term by

$$\mathcal{L} = \frac{f_\pi^2}{4}\,\mathrm{tr}(\partial_\mu\Sigma^\dagger\partial^\mu\Sigma)^2 - \frac{1}{2}f_\pi^2\,m_{\eta'}^2\,(-i\log\det\Sigma)^2$$

Here $m_{\eta'}^2$ is the mass which must vanish as $N\to\infty$. We will see shortly that $m_{\eta'}^2\sim 1/N$.

It is unusual to include a log term in an effective action. However, as we will now see, it captures a number of aspects of the anomaly. To illustrate this, let's first add masses for the other quarks. As we saw in Section 5.2.3, this is achieved by including the term

$$\mathcal{L} = \int d^4x\;\frac{f_\pi^2}{4}\,\mathrm{tr}\,(\partial^\mu\Sigma^\dagger\,\partial_\mu\Sigma) - \frac{\sigma}{2}\mathrm{tr}\,\left(M\Sigma+\Sigma^\dagger M^\dagger\right) - \frac{1}{2}f_\pi^2\,m_{\eta'}^2\,(-i\log\det\Sigma)^2$$

with M a complex mass matrix. By a suitable $SU(N_f)\times SU(N_f)$ rotation, we can always choose

$$M = e^{i\theta/N}\mathcal{M}$$

where $\mathcal{M}$ is diagonal, real and positive. This final phase can be removed by a $U(1)_A$ rotation, $\Sigma\to e^{-i\theta/N}\Sigma$ to make the mass real. But this now shows up in the mass term for the η',

$$\mathcal{L} = \int d^4x\;\frac{f_\pi^2}{4}\,\mathrm{tr}\,(\partial^\mu\Sigma^\dagger\,\partial_\mu\Sigma) - \frac{\sigma}{2}\mathrm{tr}\,\left(\mathcal{M}\Sigma+\Sigma^\dagger\mathcal{M}^\dagger\right) - \frac{1}{2}f_\pi^2\,m_{\eta'}^2\,(-i\log\det\Sigma-\theta)^2$$

However, we've played these games before: in Section 3.3.3, we saw that rotating the phase of the mass matrix in equivalent to introducing a theta angle. We conclude that this is how the QCD theta angle appears in the chiral Lagrangian.

We can now minimise this potential to find the ground state. With $\mathcal{M}$ diagonal, the ground state always takes the form

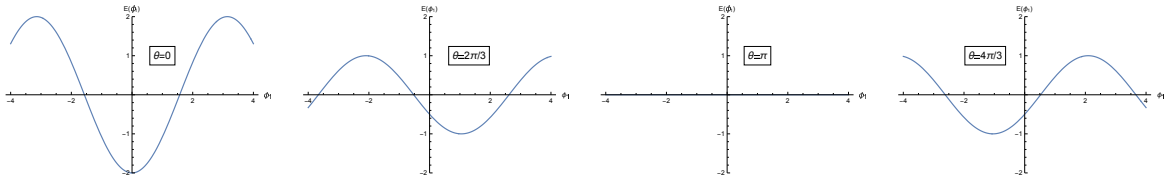
$$\Sigma = \mathrm{diag}\left(e^{i\phi_1},\ldots,e^{i\phi_{N_f}}\right)$$

The exact form depends in a fairly complicated manner on the choices of mass matrix $\mathcal{M}$ and theta angle. To proceed, we must make some assumptions. We will take m_η much bigger than all other masses, which means that we first impose the second term as a constraint

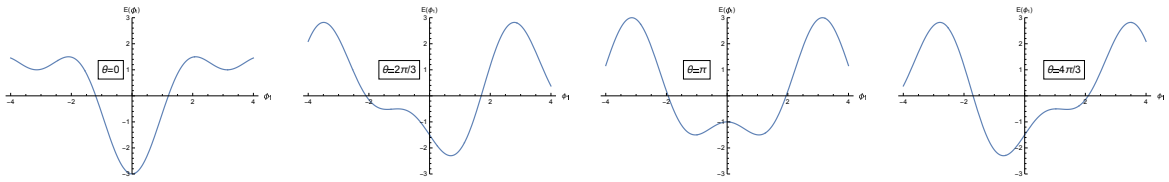
$$\sum_{i=1}^{N_f} \phi_i = \theta$$

We further look at the simplest case of a diagonal mass matrix: $\mathcal{M} = m\mathbf{1}_{N_f}$ with $m > 0$. We will then see how the ground states change as we vary θ.

For $\theta = 0$, the ground state sits at $\Sigma = 1$. Now we increase θ. What happens next differs slightly for $N_f = 2$ and $N_f > 2$. Let's start with $N_f = 2$. As we increase θ, the ground state moves to $\phi_1 > 0$ and the overall magnitude of the potential decreases. At $\theta \to \pi^-$, the ground state tends towards $\phi_1 = \pi/2$. At $\theta = \pi$ itself, the potential vanishes for all ϕ_1, which is symptomatic of a second order phase transition. If we now increase θ just a little more, the ground state jumps to $\phi_1 = -\pi/2$, before moving back towards $\phi_1 = 0$ as $\theta \to 2\pi$. The sequence is shown in the plots below for $\theta = 0, \frac{2\pi}{3}, \pi$ and $\frac{4\pi}{3}$



The fact that the potential vanishes when $\theta = \pi$ is special to $N_f = 2$. The story for $N_f \geq 3$ is similar, except that there are now just two degenerate vacua at $\theta = \pi$. This is characteristic of a first order phase transition. The potential for $N_f = 3$ for $\theta = 0, \frac{2\pi}{3}, \pi$ and $\frac{4\pi}{3}$ is shown below.



6.4.3 The Witten-Veneziano Formula

So far, we've happily incorporated the new η' Goldstone boson into our chiral Lagrangian. However, this brings something of a puzzle, which is to reconcile the following facts:

- The ground state energy is $E(\theta) \sim N^2$ and depends on θ.
- Quarks contribute to quantities such as $E(\theta)$ at order N.
- All θ dependence vanishes if we have a massless fermion.

These three facts seem incompatible. How can the $\sim N$ contribution from quarks cancel the $\sim N^2$ contribution from gluons to render $E(\theta)$ independent of θ?

To see how this might work, let's consider schematically the contribution to the susceptibility (6.16)

$$\chi(k) = \sum_{\text{glueballs}} \frac{N^2 a_n^2}{k^2 - M_n^2} + \sum_{\text{mesons}} \frac{N b_n^2}{k^2 - m_n^2}$$

where M_n are the masses of glueballs, m_n the masses of mesons, and a_n and b_n the amplitudes for $\text{tr}\, F_{\mu\nu}{}^\star F^{\mu\nu}$ to create these states from the vacuum,

$$\langle 0|\text{tr}\, F^\star F|n^{\text{th}}\text{ glueball}\rangle = N a_n \quad , \quad \langle 0|\text{tr}\, F^\star F|n^{\text{th}}\text{ meson}\rangle = \sqrt{N} b_n$$

We want the second term to cancel the first in the limit $k \to 0$. We can achieve this only if there is some meson whose mass scales as $m^2 \sim 1/N$. But this tallies with our discussion above; we expect that the η' becomes a genuine Goldstone boson in the large N limit. We're therefore led to the conclusion

$$\chi(0)\Big|_{\text{Yang-Mills}} = \frac{N b_{\eta'}^2}{m_{\eta'}^2} \tag{6.32}$$

But we can now use our anomaly equation (6.28) to write

$$\sqrt{N} b_{\eta'} = \langle 0|F^\star F|\eta'\rangle = \frac{8\pi^2 N}{\lambda N_f}\langle 0|\partial_\mu J_A^\mu|\eta'\rangle = \frac{8\pi^2 N}{\lambda N_f} p_\mu \langle 0|J_A^\mu|\eta'\rangle$$

But we know from our discussion of currents in the chiral Lagrangian (5.13) that $\langle 0|J_A^\mu|\eta'\rangle = -i\sqrt{N_f} f_\pi p^\mu$. (The factor of $\sqrt{N_f}$ here is a novel normalisation, but ensures that f_π is independent of N_f in the large N limit.) We therefore find that $\sqrt{N} b_{\eta'} = (8\pi^2 N/\sqrt{N_f}\lambda) f_\pi m_{\eta'}^2$. Inserting this into (6.32), and using (6.17), we have

$$m_{\eta'}^2 = \frac{4N_f}{f_\pi^2}\frac{d^2 E}{d\theta^2}\Big|_{\theta=0}$$

This is the *Witten-Veneziano formula*. Rather remarkably, it relates the mass of the η' meson to the vacuum energy $\chi(0)$ of large N, pure Yang-Mills theory without quarks.

It's worth pausing to see how the N scaling works in this formula. While $E(\theta) \sim N$, we expect that $d^2E/d\theta^2$ is of order 1. Meanwhile, $f_\pi \sim \sqrt{N}$. We then see that $m_{\eta'}^2 \sim 1/N$ as anticipated previously.

We don't know how to measure the topological susceptibility $\chi(0)$ experimentally. Nonetheless, we can use the Witten-Veneziano formula, with $m_{\eta'} \approx 950$ and $f_\pi \approx 93$ MeV and $N_f = 3$ to get $d^2E/d\theta^2 \approx (150\text{MeV})^4$.

6.5 Further Reading

The large N expansion in Yang-Mills was introduced by 't Hooft in 1974 [97]. ('t Hooft was astonishingly productive in those years!) Although we didn't cover it in these lectures, 't Hooft quickly showed how these methods could be used to solve QCD in two dimensions, a theory that is now referred to as the 't Hooft model [98].

The discussion of baryons in the $1/N$ expansion is due to Witten [221], as is the $1/D$ expansion in atomic physics [223]. Witten goes on to apply the $1/D$ expansion to helium. It's clever, but also shows why chemists tend not to adopt this approach.

The fact that, despite all appearances, dependence on the θ angle survives in the large N limit was first emphasised by Witten in [220]. The large N limit of the chiral Lagrangian was constructed in [224, 41], and the Witten-Veneziano formula was introduced in [222, 199]. The symmetry breaking pattern needed for the chiral Lagrangian can be proven in the large N limit: this result is due to Coleman and Witten [31]. The idea that QCD at $\theta = \pi$ spontaneously breaks time reversal was pointed out pre-QCD and pre-theta by Dashen [38] and is sometimes referred to as the Dashen phase.

The tantalising connection between string theory and the large N expansion can be made explicit in a number of low dimensional examples; the lectures by Ginsparg and Moore are a good place to start [75]. In $d = 3+1$ dimensions, this relationship underlies the AdS/CFT correspondence [129].

Coleman's lectures remain the go-to place for a gentle introduction to the $1/N$ expansion [32]. Manohar has written an excellent review of the phenomenology of large N QCD [132]. Any number of reviews on the gauge/gravity duality also contain a discussion of $1/N$ and its relationship to string theory: I particularly like the lectures by McGreevy [135].