On Confinement and Duality

Matthew J. Strassler*

University of Pennsylvania, Philadelphia, Pennsylvania, USA

Lectures given at the
Spring School on Superstrings and Related Matters
Trieste, 2 – 10 April 2001

LNS027003

*strassler@physics.upenn.edu

Abstract

Confinement in four-dimensional gauge theories is considered from several points of view. General features are discussed, and the mechanism of confinement is investigated. Dualities between field theories, and duality between field theory and string theory, are both put to use.

Contents

1 Introduction to Confinement

One of the most important discoveries of the twentieth century is that our world consists of atoms, of size 10^{-10} meters, made from electrons bound to positively charged nuclei. The size of the atom is set by the uncertainty principle; the electron is nonrelativistic, with a velocity of order α, so the size of the atom is of order $\delta x \sim 1/\delta p \sim (m_e \alpha)^{-1}$, where α is the QED coupling constant. While all experiments to date indicate that the electron itself has a size smaller than 10^{-18} meters, nuclei of atoms have a definite size, of order 10^{-15} meters. They consist of weakly-bound clumps of protons and neutrons. It was learned in the 1950s that the protons and neutrons have a size comparable to the nuclei which contain them. In the 1960s, evidence emerged that nucleons have pointlike constituents, weakly coupled in high-energy scattering processes, but highly relativistic, and therefore strongly bound, inside the proton. By the 1970s the theory of QCD emerged to explain how this strange effect was possible. The QCD interaction is weak in high-energy processes, and grows, through renormalization effects, to become strong in the low-energy processes that bind the quarks in the nucleons. The energy scale Λ_{QCD} at which it becomes strong is a few hundred MeV, corresponding to the size of the nucleon. The pointlike objects in the nucleons are the quarks suggested by Gell-Mann, interacting through the color charge suggested by Greenberg. These quarks are now themselves known to be smaller than 10^{-18} meters. They are also very light; most of the mass of the proton comes from their kinetic energy and from the powerful interactions binding the quarks together.

Yet no one has ever seen a quark, or its fractional electric charge, sitting by itself somewhere. So why should we believe this story? We all know the words: quarks are confined in hadrons — nucleons, pions, etc. — and never come out. But all too often we overlook the subtleties involved in this statement. What actually happens if we send an electron deep into a proton and try to kick a quark away from its two friends? A large amount of energy, in the form of chromoelectric field, appears in the region between the escaping quark and the remaining parts of the proton. Then what? We are familiar with the idea that large electric fields beyond a certain magnitude cannot survive; sufficiently strong fields, with energy densities bigger than $m_e^4 \sim 1$ MeV^4, are able to decay by producing pairs of electrons and positrons, the lightest electrically charged particles. The same holds for chromoelectric fields; when they become sufficiently strong, of order $\Lambda_{QCD}^4 \sim$

$(300\ \mathrm{MeV})^4$, they can pair-produce light quarks and antiquarks. How does this affect the departing quark? Well, as it moves away, the field between it and the other two quarks starts producing pairs. If for example a single pair is created, the new antiquark can end up bound to the escaping quark, and the new quark can end up bound to the other two quarks in the proton, making a new nucleon. Or perhaps multiple pairs will be created, and many quark-antiquark bound states will result. But in any case, the original quark succeeds in its escape. The force between it and the remaining quarks in the proton drops to zero as it moves away. Is this really "confinement"?

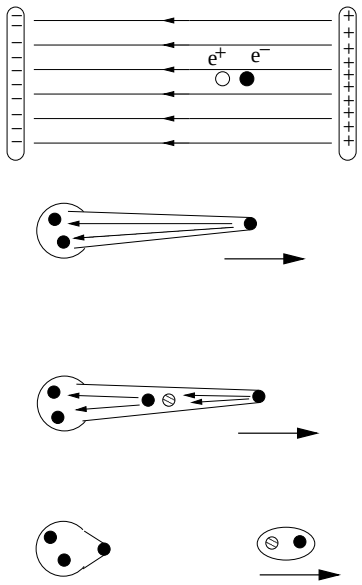

Figure 1: As with pair production of electrons in a strong electric field, pair production occurs as a quark tries to escape from a proton.

Let's contrast this with what might happen in an imaginary world in which all of the quarks had masses much larger than 100 MeV. In fact, let's take *all* of their masses to all be, oh, say, about 1000 GeV. Now the proton is a very heavy object, with mass of order 3000 GeV, and it is now quite a bit smaller than usual, about 10^{-17} meters in size (the factor of ten compared to $(1000\ \mathrm{GeV})^{-1}$ comes from the fact that the strong coupling constant is about 1/10 at these energies.) But let's imagine trying to kick a quark out of the proton now. As it rushes away, the chromoelectric field becomes very large, but the energy density, of order $\Lambda_{QCD}^4 \sim (300\ \mathrm{MeV})^4$, is far too low to produce pairs of 1000 GeV quarks. (Notice that for pair-production to be

impossible, it is essential that *all* flavors of quarks be heavy; if even *one* type of quark is light, the field will pair-produce it, independent of whether the quarks in the proton are themselves heavy.) So what happens now? Does the quark escape?

No; it cannot — or at least, it is extremely difficult. In this imaginary world, where all the quark masses m_q are very large compared to Λ_{QCD}, the quark is truly imprisoned. The force between the escaping quark and the remains of the proton goes to a constant; as we will discuss further, a "string" or "tube" of chromoelectric flux, of thickness $\Lambda_{QCD}^{-1} \sim 10^{-15}$ meters, and of tension (energy per unit length) Λ_{QCD}^2, connects the two colored objects to one another. Unless the tube becomes very long, of length m_q/Λ_{QCD}^2 (which in this case $\sim 10^{-12}$ meters, many times larger than the proton radius), there is insufficient energy in the chromoelectric field to pair-produce quarks. Even if the string does become this long, there is an exponentially low probability that all of its energy, spread out over 10^{-12} meters, will find itself localized in a region of radius $m_q^{-1} \sim 10^{-18}$ meters, as would be necessary to produce a pair of heavy quarks. So this tube of flux, stable if short, metastable if long but with a exponentially long lifetime, makes it essentially impossible for the kicked quark to escape. Eventually, the constant force from the flux tube will bring it to a stop, and pull it back into its protonic prison. This is true confinement, no doubt about it. The word really means something here.

Notice that it is not just the quarks which can be said to be "confined". *The chromoelectric field emitted by the quarks, rather than spreading out across space as in electromagnetism, is confined into "tubes," or "strings".* This is important, because even when we take the quarks away — say, by taking their masses to infinity — it might still be true that flux is confined, though there are no confined particles. In fact, we will soon see this is a more precise definition of confinement.

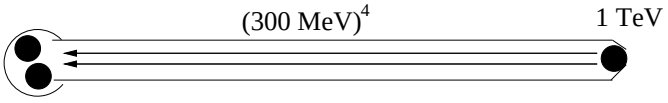

Figure 2: If all quarks were heavy, then flux tubes would break much less readily.

Strict confinement, of flux and of quarks, is thus a property of QCD only when all of its quarks are heavy. [More precisely, it is seen in the limit where the number of flavors N_f of light quarks is much smaller than the number

of colors N; the number of flavors need not be strictly zero, because the amplitude for the pair production process that splits flux tubes is of order N_f/N.] QCD with at least one light quark shows only a few remnants of these properties; there are hints of flux tubes between an escaping quark and the proton it leaves behind (though they break very quickly and never become long); and there are hints that some of the bound states of the theory behave as bits of spinning flux tube (though this is a very imprecise statement, and has as its strongest merit that it helped to motivate string theory).

So what is the right way to describe what happens in real-world QCD? We do not live in a truly confining world, and it might have been better for our own conceptual thinking if we had come up with another word for what QCD does to quarks. "Cloaking" or "maximal screening" might have been a better term. What QCD really does is ensure that a quark seeking to be free has a region in its vicinity, of size Λ_{QCD}^{-3} in volume, with chromoelectric energy density that is of order Λ_{QCD}^4. This by itself will cause an antiquark (and its partner quark) to pop out nearby, cloaking — that is, completely screening — the charge of the original quark. Compare this with electrons; in their vicinity there are regions with energy density of order m_e^4, but since the energy density is $(\alpha/r^2)^2$, the size of the region with this energy is too small to pair-produce electrons and positrons by a factor of $\alpha^{3/2}$. Thus to have this cloaking effect we need a strong coupling constant, but it hardly requires something as drastic as the flux tubes and the imprisonment found in worlds with only heavy quarks. (Indeed you might amuse yourselves by considering the possible physical properties of a hypothetical point particle of electric charge greater than $\sqrt{137}e$.)

In these lectures, we are going to explore truly confining gauge theories in some detail. Such theories may indeed exist in nature, but it is important to remember that real-world QCD is not among them.

1.1 Confinement in pure Yang-Mills

How do we even know that true confinement does in fact occur in some theories? This is a long story, and there are many ways to tell it. Let us begin in the middle, by assuming that confinement of flux occurs in pure Yang-Mills (YM) theory.

So instead of QCD, let us discard the quarks, leaving only a gauge boson A_μ in the adjoint of $SU(N)$. The group $SU(N)$ consists of $N \times N$ matrices

$U^{\alpha}_{\bar{\beta}}$ (with row indices α and column indices $\bar{\beta}$.) which are special ($\det U = 1$) and unitary ($U^\dagger = U^{-1}$). The "gluon" field A_μ takes values in the algebra of $SU(N)$,

$$(A_\mu)^{\alpha}_{\bar{\beta}} = A^a_\mu (T^a)^{\alpha}_{\bar{\beta}} \ .$$

Here T^a is a generator of the group $SU(N)$, also an $N \times N$ matrix (normalized to $\mathrm{tr}\ T^a T^b = \frac{1}{2}\delta^{ab}$,) and the group index a runs from 1 to the dimension of $SU(N)$, namely $N^2 - 1$. The theory has the simplest possible Lagrangian; defining $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu + i[A_\mu, A_\nu]$ (here F and A are matrices and the brackets indicate a matrix commutator), we write the Lagrangian as

$$\mathcal{L} = -\frac{1}{2g^2}\mathrm{tr}\ F_{\mu\nu}F^{\mu\nu} \ .$$

This normalization of the field A_μ differs from the one in standard textbooks on perturbation theory. There is good reason for this. We will not be doing perturbation theory. In perturbative calculations, it is more convenient to absorb the $1/g$ into A_μ; then $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu + ig[A_\mu, A_\nu]$. The quadratic terms in the Lagrangian are then the free Maxwell equations, and do not depend on g. We may then think of the theory as a set of free fields — simply $(N^2 - 1)$ independent photons — coupled together by interactions of order g. However, in these lectures we will not assume small g, and will rarely expand in powers of g. The normalization chosen here is more profound; it puts the coupling constant in its proper place, multiplying $\hbar$ and therefore determining the size of all quantum effects. Most nonperturbative properties of the theory will involve either $1/g^2$ or e^{-1/g^2}, as we will soon see.

Pure Yang-Mills theory is weakly coupled at high energy, like QCD, and becomes strongly coupled at a scale Λ. More accurately, we can show, using perturbation theory, that it *cannot* become strongly coupled at energies *above* a scale Λ; below this point we simply don't know what it does. The scale Λ can be estimated using one-loop graphs; at this order, the running of the gauge coupling is given by

$$\beta_g = \frac{\partial g}{\partial \ln \mu} = \frac{g^3}{16\pi^2}\left(-\frac{11}{3}N\right)$$

for $SU(N)$. The solution is

$$\frac{8\pi^2}{g^2(\mu)} = \frac{8\pi^2}{g^2(\mu_0)} + \frac{11}{3}N\ln(\mu/\mu_0)$$

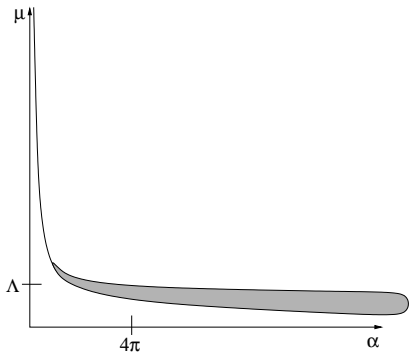

Figure 3: The coupling constant α versus energy scale μ; the one-loop calculation is valid at $\mu \gg \Lambda$ but becomes only approximate at low energy.

where μ_0 is an arbitrary starting point. Thus, the coupling is small above the energy scale $\mu \sim \Lambda$, where

$$\Lambda^{11N/3} \sim \mu_0^{11N/3} \ \exp\left[\frac{-8\pi^2}{g^2(\mu_0)}\right] \ . \tag{1}$$

This is reliable since higher loop graphs and nonperturbative effects are comparatively small above Λ.

As is standard in renormalization, the scale Λ is physical and thus independent of the arbitrary starting point μ_0. Near and below this energy regime, the coupling constant is strong; above it, perturbation theory in g^2 is possible. Also, notice that Λ involves e^{-1/g^2}. All of the really interesting physics in Yang-Mills theory is related to Λ; it is therefore nonperturbative in g^2, and cannot show up at any order in an ordinary Feynman graph expansion.

Now we must consider two more profound claims, which are fully nonperturbative, and are based on a combination of experiment, theoretical reasoning, and both analytic and especially numerical lattice gauge theory. First, the quantum Yang-Mills theory is known to develop a mass gap (that is, it has no massless fields in its spectrum, and instead has a discrete set of states with masses of order Λ) and second, it apparently becomes confining, in the true sense, at the scale Λ. Both of these effects are through strongly-coupled physics not visible semiclassically.

Both statements are strange. The gluons in the above Lagrangian are massless; how can there be no massless particles in the spectrum of the

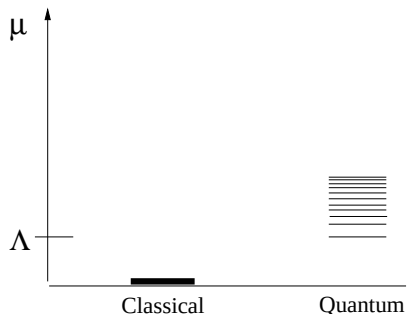

Figure 4: Classically the theory has many massless particles, but the quantum theory has a mass gap and a spectrum of gauge-neutral hadrons.

theory? Well, let us assume that, as in QCD, the effect of the strong interactions will be that we will observe only colorless bound states. What kinds of bound states can we make from gluons? We might say that we can make a bound state of two gluons, or three gluons, or four. But this is surely wrong. The interactions of the theory do not conserve the number of gluons even in perturbation theory; there are terms cubic and quartic in A_μ in the Lagrangian, so one gluon may become two or three, and vice versa. The situation will be worse once the interactions of Yang-Mills become strong. We clearly cannot use "gluon number" as a quantum number describing a state. In fact, the strong coupling dynamics makes it impossible to talk about gluons at low energies. Instead, we have only bound states, whose name "glueballs" is reasonably accurate, in that these gluey states do not really consist of a fixed number of gluons, but rather of a shifting mass of chromoelectric flux lines. There are a large number of these states. Below the scale Λ we might try to write an effective theory of these glueballs. Unlike the gluons, for which mass terms are forbidden (since they have only two polarization states and massive vectors require three), the glueballs include scalars (for which mass terms cannot be forbidden) and vectors with three polarizations (for which mass terms also cannot be forbidden) and similar higher spin particles. Their masses can't be much larger than Λ since that would contradict perturbation theory, but nothing stops them from having masses of order Λ. Essentially, there is a mass gap because there are no symmetries which forbid mass terms for any of the glueballs.

The statement about confinement is also, at first, strange. The theory has only gluons; are they confined? What happens when we try to pull a gluon out of a bound state? Does a flux tube form between it and the other

gluons? What does this mean, since the flux tube itself is made from gluons? How is it possible that pair-production of gluons is forbidden? In fact, it is not forbidden, but that is fortunately irrelevant. The statement about confinement has nothing to do with the gluons. The gluons are no more confined in Yang-Mills than light quarks are in real-world QCD; in fact they are even less so, since there is no parameter analogous to the quark mass which when large can make the gluons confined. *"Confinement" means that chromoelectric field is confined; it cannot spread out in space over regions larger than about* Λ^{-1} *in radius.*

One might ask if there is a connection between the mass gap and the confinement of flux. We will return to this issue later.

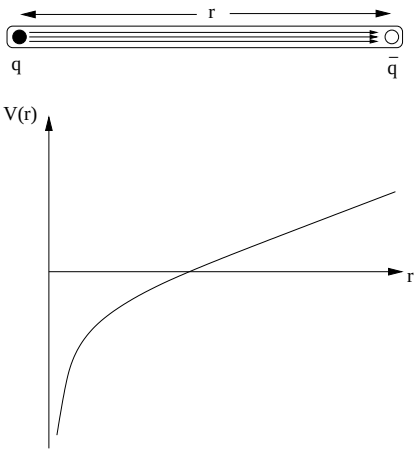

Figure 5: The confined field lines between a heavy quark and antiquark form a tube; the potential energy of the system goes as $1/r$ at distances short compared to Λ^{-1} but becomes linear at larger distances.

Now, how can we detect the presence of the strings which contain the chromoelectric flux? Ideally we would like to find a long and straight flux tube and find its tension (energy per unit length) but we might have trouble convincing one to stay straight long enough to do this measurement. So here we need a new idea. Recall how the heavy quarks of QCD-with-no-light-quarks were truly confined. This suggests that the way to detect confinement of flux in Yang-Mills theory is to put some extremely heavy quarks in it — so heavy that they can't affect the dynamics of the Yang-Mills theory — and see that these quarks are confined! That is, we can compute the quark-antiquark potential $V(r)$ and see that it grows without bound (indicating

confinement) and more specifically is linear in r (indicating confinement by flux tube.) Why is the linear potential characteristic of a flux tube? Well, consider Gauss's law. In an unconfined theory, the electric flux is uniformly distributed over a sphere surrounding a charge, and therefore falls off as $1/r^2$. In a confining theory with flux tubes, the flux tube has a fixed cross-sectional area of order Λ^{-2} no matter how long it is; and thus, for any sphere of radius $r \gg \Lambda^{-1}$ surrounding a charge, the flux on the sphere is zero everywhere except in a region of area Λ^{-2} where the flux tube passes through the sphere. From this we conclude that the electric field in that region has a magnitude which is r-independent! In turn, this implies the force that it generates on a test charge is also r-independent, and finally, that the potential between charges grows linearly with r.

So, let us add a charged fermion (or scalar) to the Yang-Mills theory, one whose mass M is so much larger than Λ that it cannot play a role in the strong-coupling physics. Adding a quark ψ we make the Lagrangian

$$\mathcal{L} = -\frac{1}{2g^2}\mathrm{tr}\ F_{\mu\nu}F^{\mu\nu} + i\bar{\psi}\not{D}\psi - M\bar{\psi}\psi\ .$$

The quark ψ is charged under $SU(N)$, but for the moment let us not specify the representation R of $SU(N)$ under which it transforms. Now let us consider the potential $V(r)$ between ψ, placed at one position, and $\bar{\psi}$, placed a distance r away. Since the quarks are very heavy, we can expect that they can be placed at rest and will move only very slowly, allowing us to do this computation. Confinement means that when r is large, a string — a tube of chromoelectric flux — stretches between ψ and $\bar{\psi}$, of constant tension T_R, such that the potential $V(r) = T_R r$ [1]. The force between two such fermions goes to a constant, and never drops off to zero. (That these facts are true in Yang-Mills theory does not follow from any direct theoretical calculation. Highly quantum mechanical in nature, they have only been checked using direct numerical simulation of Yang-Mills theory.)

In the limit where $M \to \infty$, the quarks become completely non-dynamical [1]; they are what we may call "chromoelectrostatic sources", probes which never appear in any loop diagram and thus are purely classical. What remains dynamical is the flux tube. Thus, we didn't really need the quarks as physical particles; using nondynamical chromoelectric sources, we could have detected the confinement of chromoelectric field, which is a property of the Yang-Mills theory without the added quarks. (An equivalent way to make this statement, without introducing the quarks, is to talk about Wilson loops in various representations R; in a confining theory the value of

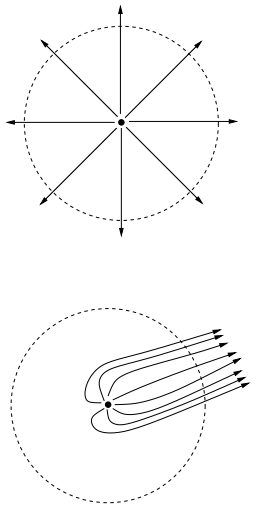

Figure 6: Gauss's law for unconfined and confined flux.

the Wilson loop is proportional to the exponential of minus its area, with proportionality constant T_R [1].)

In general, the string tension, and the corresponding force, between quark and antiquark can depend on the representation R. After all, why not? In particular, for R the adjoint, we already know $T_{adjoint} = 0$: any fermion in the adjoint can combine with a light gluon to make something gauge neutral, so two such fermions will each cloak themselves with a gluon and will feel no long-range force as we pull them apart. So clearly we need to think about how things depend on the representation R. Clearly the map from representations to flux tubes cannot be one-to-one (since both the trivial representation and the adjoint representation have $T_R = 0$.) Lie groups have an infinite number of representations, but the stable flux tubes number at most $\dim C_G$, the dimension of the center of the gauge group. Let us see why this is so.

What is the center of $SU(N)$? A matrix $U^{\alpha}_{\bar{\beta}} = e^{2\pi ik/N}\delta^{\alpha}_{\bar{\beta}}$, $k = 0, \ldots, N-1$, is an element of $SU(N)$. Being proportional to the identity, it obviously commutes with everything in $SU(N)$; in short, U is in the center $C_{SU(N)}$. The elements of the center are thus labelled by the integer k, which from the definition of U is only determined modulo N, so the labels form the group $\mathbf{Z}_N$, the additive integers mod N. Now consider any representation R. An element ρ of this representation is labelled by a certain number n

of unbarred (upper) and $\bar{n}$ of barred (lower) indices; that is, it takes the form $\rho^{\alpha_1 \alpha_2 \cdots \alpha_n}_{\bar{\beta}_1 \bar{\beta}_2 \cdots \bar{\beta}_{\bar{n}}}$. Under a group transformation, each unbarred index is rotated by the matrix U, while each barred index is rotated by $U^\dagger$. Consequently, the transformation of the representation R under the center C_G is by the phase $e^{2\pi i k(n-\bar{n})/N}$, where $n - \bar{n}$ is called the "N-ality" of the representation. The adjoint representation, with one upper and one lower index, is invariant under the center. The fundamental $\mathbf{N}$ representation (one unbarred index) rotates by $e^{2\pi i k/N}$; the antifundamental $\bar{\mathbf{N}}$ rotates by $e^{-2\pi i k/N}$. Both the antisymmetric-tensor and symmetric-tensor representations $\mathbf{N}(\mathbf{N} \pm \mathbf{1})/\mathbf{2}$, which have two unbarred indices, rotate by $e^{2\pi i (2k)/N}$. Indeed, all p-upper-index tensors carry charge p under $\mathbf{Z}_N$ — that is, they rotate by $e^{2\pi i p k/N}$ under the k^{th} element of $\mathbf{Z}_N$. In short, the representations R of $SU(N)$ break up into equivalence classes under the center, and can be labelled by their "N-ality" charge p [2, 3]. Note that the conjugate representation of R has "N-ality" $N - p$, since the number of barred and unbarred indices is exchanged.

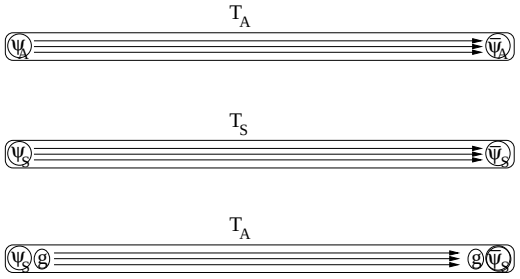

Figure 7: The flux between different quarks, or combinations of quarks and gluons, all with N-ality 2.

Why is this interesting? First consider, for example, adding a heavy quark ψ_A, in the antisymmetric representation, to Yang-Mills theory; the potential between $\bar{\psi}_A$ and ψ_A is $V(r) = T_A r$. Now consider instead adding a heavy quark ψ_S in the symmetric representation; the quark-antiquark potential between $\bar{\psi}_S$ and ψ_S is now $V(r) = T_S r$. Suppose that $T_S > T_A$ in Yang-Mills theory. (This is probably true, but what I'm about to say won't depend on the specific assumption.) Nothing prevents the theory from taking one of its light gluons (remember their number is not conserved so it need not be pair-produced) and putting it very near ψ_S. The combination of the gluon A_μ and the fermion ψ_S looks, from a distance, as though it were a single object. What is its charge? Well, we must consider the group theory

of $SU(N)$; what is (adjoint)$\otimes$(symmetric)? It is a direct sum of a number of representations, *all of which have the same "N-ality" as the symmetric representation, namely 2.* Said another way, the product of $(A_\mu)^\alpha_{\bar{\beta}}$ and $\psi_S^{\gamma\delta}$ can lead, no matter how the indices are contracted, only to representations with two more upper indices than lower indices. Among these representations is the antisymmetric representation. (In $SU(3)$, for instance, the symmetric tensor is **6**, the antisymmetric tensor is $\bar{\mathbf{3}}$, and $\mathbf{8}\otimes\mathbf{6} = \bar{\mathbf{3}}+\mathbf{6}+\overline{\mathbf{15}}+\mathbf{24}$.) But then, since we assumed $T_A < T_S$, there exists a dynamical process by which the theory may lower its energy! By popping a gluon out of the vacuum and putting it near ψ_S, the theory can make ψ_S look more like a fermion in the antisymmetric representation. The same goes for $\bar{\psi}_S$. Then, instead of a string of tension T_S, a string of tension T_A can link these two fermion-gluon combinations. The energy cost is that of making two extra gluons — at most of order Λ — while the energy gain is $(T_S - T_A)r$, which for r sufficiently large always wins. The reverse process will hold if $T_A > T_S$.

More generally, the fact that gluons are in the "N-ality"-zero adjoint representation implies that *the presence of nearby gluons can change one representation to another but only in a way that conserves N-ality.* Thus in Yang-Mills, the representation R of a chromoelectric source is not a conserved quantum number; only its "N-ality" is actually conserved. Consequently, we should expect that for the entire class of representations with the same N-ality charge, there will be only one stable configuration of strings (which might involve one or more tubes — for "N-ality"=2 there might be one tube with two units of flux or two tubes with one unit each.) *The tensions of the stable strings, or combinations of strings, are labelled not by R but by the N-ality p of R.* Charge conjugation symmetry also ensures that $T_p = T_{N-p}$; thus we have of order $N/2$ stable flux tube configurations in $SU(N)$ Yang-Mills theory.

Can we see this in $SU(3)$ Yang-Mills? Yes and no. There is N-ality 0,1, and 2; but $T_0 = 0$ while $T_2 = T_1$, so only one confining string is predicted. The nontrivial statements are then only that, for example, the symmetric **6** representation of $SU(3)$ is confined by the same string tension as the antisymmetric tensor, the $\bar{\mathbf{3}}$; this in turn has the same tension as the fundamental **3**. To have a nontrivial set of strings we must go to $SU(4)$; here the antisymmetric tensor **6** should have a tension T_6 different from that of the **4** and $\bar{\mathbf{4}}$, T_4. There is still a question as to whether $T_6 < 2T_4$; if not, the flux between two **6** fields may be carried by two strings of N-ality 1 rather than a single string of N-ality 2. Theoretical arguments [4] and lattice calculations

[5, 6, 7] support the view that $T_6 < 2T_4$ (and similarly in other theories) so that there really are two independent stable flux tubes, of N-ality 1 and 2 (and again $T_3 = T_1$.)

To summarize, we expect that Yang-Mills theories have stable flux tubes labelled by a charge in the center of the group [2]; for $SU(N)$ this is its N-ality, a charge under the $C_{SU(N)} = \mathbf{Z}_N$ group action. While the gluons are not confined by these strings, any heavy quark with nonzero N-ality will experience a linear potential energy and a constant force which will confine it to an antiquark, or more generally, to some combination of quarks and antiquarks which have the opposite N-ality. (For example, it could combine with $N-1$ other quarks to form a baryon. As another example, a **6** of $SU(4)$ could combine with two $\bar{\mathbf{4}}$ quarks to form an exotic object not found in real-world $SU(3)$ QCD.)

1.2 Confinement in $\mathcal{N} = 1$ Super-Yang-Mills

Let us now consider $\mathcal{N} = 1$ supersymmetric Yang-Mills theory (SYM.) This theory is very interesting in that (1) many of its properties can be exactly or approximately determined, (2) it resembles Yang-Mills theory, in that it has confinement and flux tubes, has a mass gap, and lacks light particles similar to pions, yet (3) it resembles QCD in that it has chiral symmetry breaking and an anomaly which makes a would-be Nambu-Goldstone boson, the η', massive, while (4) it differs from both in that it has multiple isolated, degenerate vacua.

The $SU(N)$ SYM theory is nothing more than $SU(N)$ gauge theory with a vector boson (gluon) A_μ and a massless Majorana spinor (gluino) λ_α, both in the adjoint representation of the gauge group. The Lagrangian is simply

$$\mathcal{L} = \frac{1}{2g^2}\left[\mathrm{tr}\ F_{\mu\nu}F^{\mu\nu} + i\bar{\lambda}\not{D}\lambda\right] \tag{2}$$

Pure $\mathcal{N} = 1$ SYM, like pure non-supersymmetric YM, is a confining theory. (Convincing arguments confirming earlier expectations are given in [8, 9].) It will have stable flux tubes, just like YM, despite the presence of the gluinos. The gluino carries the same gauge charge as the gluon, and is neutral under $C_{SU(N)} = \mathbf{Z}_N$. Therefore, like the gluon, it does not break flux tubes carrying $\mathbf{Z}_N$; no flux tube which carries such a charge can end on a $\mathbf{Z}_N$-neutral gluino. (This is in contrast to $SU(3)$ QCD, where the quarks, which carry charge under the $\mathbf{Z}_3$ center, do indeed break the flux tubes.) Thus SYM is a good place to study confining strings.

The theory also has an anomalous $U(1)$ global symmetry, just like QCD. We won't need this, but it is useful for you to know a bit about it. How does this work? Classically, the Lagrangian of the theory has a global symmetry $\lambda \to \lambda e^{i\alpha}$, where α is any real number. However, the path integral of SYM does *not* have this symmetry. There isn't time in these lectures to study anomalies in detail, so let me just quote the classic result: under this rotation, the path integal itself is not invariant unless $2N\alpha$ is a multiple of 2π. (A similar statement applies in QCD with N_f massless flavors of quarks ψ_i and $\tilde{\psi}_{\tilde{j}}$; under the global rotation $\psi_i \to \psi_i e^{i\alpha}$, $\tilde{\psi}_{\tilde{i}} \to \tilde{\psi}_{\tilde{i}} e^{i\alpha}$, the path integral is not invariant unless $2N_f\alpha$ is a multiple of 2π.) Thus the $U(1)$ is a fake; only a discrete $\mathbf{Z}_{2N}$ subgroup of this $U(1)$ is actually a symmetry.

In QCD, with Lagrangian[1]

$$\mathcal{L} = \frac{1}{2g^2}\mathrm{tr}\ F_{\mu\nu}F^{\mu\nu} + \sum_{i=1}^{N_f} i\bar{\psi}_i \not{D}\, \psi_i + \sum_{\tilde{j}=1}^{N_f} i\overline{\tilde{\psi}}_{\tilde{j}} \not{D}\, \tilde{\psi}_{\tilde{j}} - \sum_{i,\tilde{j}} m^{i\tilde{j}}\tilde{\psi}_{\tilde{j}}\psi_i \ ,$$

there is an entire $SU(N_f)$ symmetry for the quarks ψ_i, another $SU(N_f)$ for the antiquarks $\tilde{\psi}_{\tilde{j}}$, a $U(1)$ "baryonic" symmetry under which the ψ_i and $\tilde{\psi}_{\tilde{j}}$ have opposite charge, and finally the fake "axial" $U(1)$ mentioned above of which only a $\mathbf{Z}_{2N_f}$ is a true symmetry. These symmetries do not all appear at low energy, however. First, the nonzero quark masses $m^{i\tilde{j}}\tilde{\psi}_{\tilde{j}}\psi_i$ break most of the two $SU(N_f)$ symmetries; but the masses are relatively small for the up, down, and strange quarks, so let us imagine for a moment that they are zero, and, forgetting the heavier quarks (which are dynamically less important,) take $N_f = 3$. But even then, for $m^{i\tilde{j}} = 0$, the vacuum does not show all of the symmetries of the theory. For reasons not entirely understood, a quark-antiquark bilinear operator $\tilde{\psi}_{\tilde{j}}\psi_i$ develops a nonzero expectation value[2] proportional to $\delta_{i\tilde{j}}$, with a magnitude of order $(\Lambda_{QCD})^3$. This quark-antiquark condensate is not invariant under the $SU(N_f)$ symmetries mentioned above;

[1] Note the fermion fields $\psi, \tilde{\psi}$ written here are not each others' complex conjugates! They are left-handed quarks and left-handed antiquarks; they form two separate sets of two-component Weyl fermions, transforming in the $\mathbf{N}$ and $\bar{\mathbf{N}}$ representations. Mass terms $m^{i\tilde{j}}\tilde{\psi}_{\tilde{j}}\psi_i$ make them into massive four-component Dirac fermions, but without the masses they are independent fields, with independent generation indices $i = 1, \ldots, N_f$ and $\tilde{j} = 1, \ldots, N_f$.

[2] All of the following statements about chiral symmetry breaking apply at least for small N_f; they are certainly not true for $N_f > (11/2)N$, at which point $SU(N)$ QCD has a positive one-loop beta function and can't possibly be strongly-coupled in the infrared. At what value of N_f they stop being true is not known, although most guesses these days for $N = 3$ range from 5 to 12.

it is only invariant under *simultaneous* rotations of the quarks ψ_i by a matrix U in *their* $SU(N_f)$ and of the antiquarks $\tilde{\psi}_{\tilde{j}}$ by the conjugate matrix $U^\dagger$ in the *other* $SU(N_f)$. These "diagonal" rotations define a group $SU(N_f)_D$, which remains a symmetry of the vacuum. All other $SU(N_f) \times SU(N_f)$ rotations change the vacuum, and thus are not symmetries of it. This is known as "spontaneous chiral symmetry breaking"; the equations of the theory still have an $SU(N_f) \times SU(N_f)$ symmetry, but the vacuum itself, a particular solution of those equations, is invariant only under its $SU(N_f)_D$ subgroup. As both Nambu and Goldstone taught us years ago, this implies, as an automatic consequence, that there are massless particles corresponding to the broken rotations. These are the pions. They tell us that QCD has not one vacuum, but in fact a continuous set of degenerate vacua (if the quarks are strictly massless!) The pions are massive in nature only because the quark masses are in fact not zero, and the $SU(N_f) \times SU(N_f)$ flavor symmetry is only approximate. Note that the baryonic $U(1)$ is unbroken. If the axial $U(1)$ had been a true symmetry, it would have been broken, and we would have expected a Goldstone boson for it, the η', which corresponds to shifts of the phase of the condensate $\langle \tilde{\psi}_{\tilde{j}} \psi_i \delta^{i\tilde{j}} \rangle$. However, the $U(1)$ is a fake; and although the $\mathbf{Z}_{2N_f}$ axial symmetry mentioned above is also spontaneously broken by the condensate to a $\mathbf{Z}_2$ subgroup, only continuous symmetries give continuous sets of degenerate vacua and corresponding massless particles. The η' in fact has a periodic potential, with N_f minima rotated by the $\mathbf{Z}_{2N_f}$ symmetry. In each of these minima the potential has some upward curvature, so the η' has a mass. Note however, that these minima are not actually isolated since they are connected via $SU(N_f) \times SU(N_f)$ rotations.

What happens in SYM? In this case the operator $\lambda\lambda$ develops an expectation value (this is a largely rigorous statement, for which there are many fairly strong proofs; see for example [10].) The $\mathbf{Z}_{2N}$ axial symmetry is broken to $\mathbf{Z}_2$. Because there are no continuous global symmetries, we have no continuous space of vacua. Instead we have N isolated, degenerate vacua, in which

$$\langle \lambda\lambda \rangle \propto \Lambda^3 e^{2\pi i r/N} \ , \quad r = 0, 1, 2, \ldots, N-1 \ .$$

In this theory, the beta function has coefficient $3N$, so the strong-coupling scale satisfies $\Lambda^{3N} = \mu_0^{3N} e^{-8\pi^2/g^2(\mu_0)}$. Notice that the $\mathbf{Z}_{2N}$ symmetry rotates one vacuum into the next, so the N vacua, though distinct from one another, are isomorphic. This guarantees they are degenerate with one another. Again, in each vacuum the $\mathbf{Z}_{2N}$ is broken, but the *space* of N vacua is $\mathbf{Z}_{2N}$ symmetric, and the symmetry rotates one vacuum into the next. The

η' particle in this theory is the phase of $\langle\lambda\lambda\rangle$, and it has a periodic potential, with N degenerate minima. Thus, like QCD, SYM has a fermion bilinear condensate which breaks global symmetries, and it has an η' with a periodic potential; but unlike QCD, and similar to YM, it has no massless or very light particles.

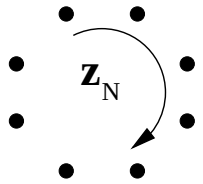

Figure 8: The vacua of $\mathcal{N} = 1$ $SU(N)$ SYM (shown in the complex $\langle\lambda\lambda\rangle$ plane) are rotated by a $\mathbf{Z}_N$ global symmetry.

2 Confinement of Magnetic Flux

Now let us try to understand why and how confinement occurs. In Yang-Mills theory it occurs through a process requiring strong coupling; detailed investigations have revealed no small parameter in which we can do perturbation theory, and no simple calculation that we can perform. From where can we gain some insight? We might ask: where we have seen tubes of confined flux before?

2.1 Superconductors and the Abelian Higgs Model

In Type I superconductors, magnetic flux is excluded from the material. This occurs through the appearance of surface currents, which can exist without energy cost due to the absence of any resistance in the material. These currents generate an exactly-compensating magnetic field which cancels any external magnetic field trying to penetrate the material, and instead produces some additional magnetic field outside. This makes it appear that all external magnetic fields are "expelled" from the superconductor. This famous piece of physics is called the "Meissner effect."

In Type II superconductors, however, the situation is a bit more complicated. Flux can indeed penetrate the superconductor in this case, although only in a very specific way. The material becomes nonsuperconducting in a narrow tube running from one side of the material to another, and the magnetic flux threads that tube. The magnetic field, which would have been

free to roam in a normal material, is trapped inside "Abrikosov vortices" [11] traversing the superconductor. These vortices carry one or more quanta of flux; in short, they carry an integer charge, $q \in \mathbf{Z}$. *Superconductors confine magnetic flux into quantized vortices.*

Indeed this looks familiar. We have learned that $SU(N)$ YM and $\mathcal{N} = 1$ SYM both confine *electric* flux into tubes which carry a discrete charge in $\mathbf{Z}_N$. This looks similar enough to set off alarm bells. We had better look at this more closely.

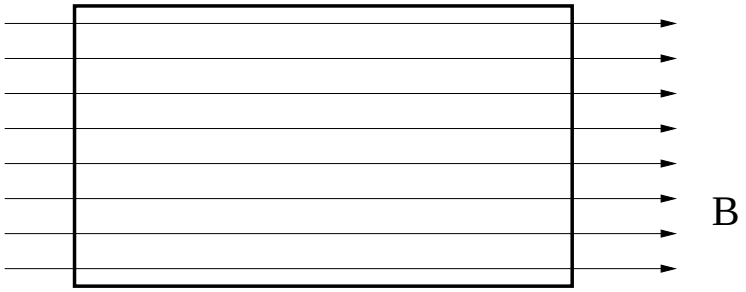

Figure 9: Normal materials can sustain magnetic fields.

How does a superconductor accomplish this? The superconductivity occurs because electrons form Cooper pairs, which are bosons. Let us call the density of these pairs ϕ. Since the pairs carry electric charge 2, ϕ must be complex, and couples to the photon. More specifically, the photon must couple to a conserved current, namely

$$J^\mu = \phi^\dagger \partial^\mu \phi - (\partial^\mu \phi^\dagger)\phi \tag{3}$$

Now suppose that there were a magnetic field attempting to pass through the material. Since the Cooper pairs can flow without resistance, they can respond by creating a compensating current. For instance, suppose we have a long cylinder of material of radius R; let us use cylindrical coordinates r, θ, z. Suppose we attempt to apply a uniform magnetic field $B_z > 0$ along the axis of the cylinder. The Cooper pairs can respond by generating a current J^θ, which can propagate without resistance, at the surface of the cylinder $r = R$. This completely cancels the applied field, reducing the energy density inside the superconductor. It also generates a dipole field outside the cylinder. The field appears to have been "expelled" from the material.

However, the material could also respond in an additional way, and does so in the type II case. In addition to generating a current at $r = R$, it could

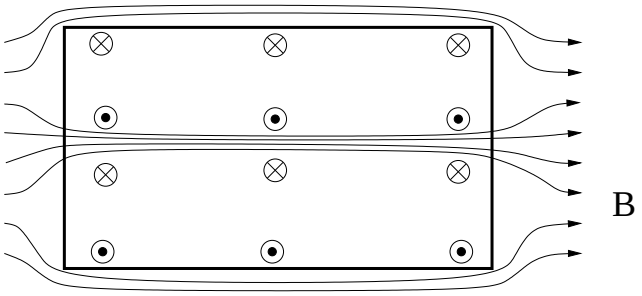

Figure 10: In superconductors, Cooper pair currents (shown into and out of the plane of the paper) are induced, causing the magnetic flux to be expelled or trapped in vortices.

also generate a current in the opposite direction at $r = r_0 \ll R$, deep within the material. This current, like the current in a solenoid, generates a field in the positive B_z direction, all confined within the region $r < r_0$. This is a magnetic flux tube.

What does ϕ do near this flux tube? Consider a circle of radius $r_1 > r_0$. The integral of the magnetic flux inside this circle, $\int_{r<r_1} B_z \; r \; dr \; d\theta$, should be independent of r_1 if flux is indeed confined. On the other hand, it is also equal to $\oint_{r=r_1} d\theta \; A_\theta$. By cylindrical symmetry, A_θ can be only a function of r. From this we learn that A_θ is a constant for large r. But this poses problems. The kinetic terms for ϕ itself surely include $\vec{\nabla}\phi \cdot \vec{\nabla}\phi$, where $\nabla_i = \partial_i + iA_i$, and thus $A_\theta^2|\phi^2|/r^2$. If ϕ is a constant v at infinity, then the integral of such a term in the Hamiltonian density is divergent! So this cannot give a finite energy solution. The only way out is to have $\partial_\theta \phi = -iA_\theta \phi$, which can be accomplished if $\phi(r) = ve^{is\theta}$ at large r, where s a real constant. Furthermore, we can avoid a divergent potential energy only if v is at a potential minimum; and at the minimum $v \neq 0$ (or we would not have superconductivity!) But then single-valuedness of ϕ requires that s is an integer. Therefore this approach only works if $A_\theta = s \in \mathbf{Z}$, and thus if $\int B_z \; r dr d\theta$ is an integral multiple of a fundamental flux quantum.

From Eq. (3), we see that J^θ is now nonzero; as advertised, the flux is of necessity enclosed by a current of Cooper pairs. Furthermore, because the phase of ϕ is winding as we go once around in θ, the radial derivatives of ϕ will be ill-defined at $r = 0$ unless ϕ has a zero there. Thus we have $\phi = ve^{is\theta} f(r)$, where $f(0) = 0$ and $f(r \to \infty) \to 1$, and s an integer. The material becomes nonsuperconducting at the vortex core, paving the way for the magnetic field to pass through unobstructed.

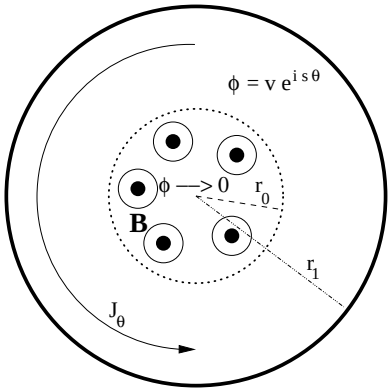

Figure 11: A flux tube of radius r_0; the phase of ϕ winds as one circles the core, in which the magnetic flux is trapped and $|\phi| < v$.

This configuration, with quantized magnetic flux and a zero for ϕ at its center, and a winding of A_θ and a corresponding winding of the phase of ϕ outside its core, is the Abrikosov vortex. Let us consder the topology associated with this vortex. We have a $U(1)$ gauge group, under which ϕ is charged. When the vacuum expectation value of ϕ is nonzero, the $U(1)$ group is broken spontaneously; gauge transformations will rotate the phase of $\langle\phi\rangle$. [However, remember that gauge transformations are not real symmetries! Therefore, unlike the case of sponaneously broken global symmetries, we do not have a continuous set of physically distinct vacua and associated Nambu-Goldstone bosons. Instead we will get a massive photon!] To make a magnetic flux tube, it must be that as we traverse a circle around the flux tube in space, the phase of the field ϕ makes a closed loop inside the $U(1)$ group. We may think of this as a map from a circle in space to a closed loop in the broken gauge group. Such a map may wind s times around the $U(1)$ as we make a single circle in space. In short, the topology of such maps, given by the first homotopy group of $U(1)$, is the group $\pi_1[U(1)] = \mathbf{Z}$. Every element in the group is labelled by an integer, the winding number s.

To round out the story, it is a bit more convenient to look at a slightly different system. Instead of studying superconductors — three-dimensional nonrelativistic systems — I will take us on a quick tour of the relativistic version, the "abelian Higgs model". This model has Nielsen-Olesen vortices [12], magnetic flux tubes very similar to those of Abrikosov.

Let us take a photon — a $U(1)$ gauge field — and a charged scalar field ϕ. The action for ϕ must be invariant under local $U(1)$ rotations $\phi \to \phi e^{i\alpha(x)}$,

which can only happen if all derivatives of ϕ are covariant, that is, of the form $D_\mu\phi \equiv (\partial_\mu + iA_\mu)\phi$, where A_μ is the photon vector potential. In particular, the kinetic term for ϕ must be of the form

$$(D_\mu\phi)^\dagger D^\mu\phi \ .$$

There can also be a potential for ϕ, but gauge invariance requires it be a function only of $\phi^\dagger\phi$. In addition we should add the action for the photon. The action is thus of the form

$$-\frac{1}{4g^2}F_{\mu\nu}F^{\mu\nu} + (D_\mu\phi)^\dagger D^\mu\phi - V(\phi^\dagger\phi) \ .$$

The potential V may have its minimum at $\phi^\dagger\phi = 0$. In this case the vacuum of the theory is much like the one we live in; the photon is massless, propagates at maximum speed, and generates a long-range force. Magnetic and electric fields are related by a symmetry; both fall off as $1/r^2$ from magnetic and electric point charges.

However, the potential might instead have its minimum at $\phi^\dagger\phi = |v|^2 \neq 0$. Now the physics is very different. First, the photon is now massive. To see this, consider small fluctuations of electric fields A_μ for fixed $\phi = ve^{i\sigma}$. The Lagrangian for these modes is

$$-\frac{1}{4g^2}F_{\mu\nu}F^{\mu\nu} - |v|^2(A_\mu A^\mu)$$

A massive photon, which can be brought to rest, must have three polarization states ($J_3 = 1, 0, -1$) unlike a photon which has only two, $J_3 = \pm 1$. Where does this extra state come from? It comes from σ, the phase of ϕ! Let us see this; if we write $\phi = ve^{i\sigma(x)}$ the Langrangian density now becomes

$$-\frac{1}{4g^2}F_{\mu\nu}F^{\mu\nu} - |v|^2(\partial_\mu\sigma + A_\mu)^\dagger(\partial^\mu\sigma + A_\mu)$$

from which we see that σ and A_μ mix. We cannot think of them any longer as separate fields, and thus σ and A_μ together form a massive, three-polarization-state spin-one particle. (If we like, we can use a gauge transformation to set $\sigma = 0$ and absorb it into A_μ, but this merely puts the degree of freedom of σ into A_μ. It will not always be useful to do this.) This is the Higgs mechanism, discovered by Anderson (always remember that condensed matter physicists have much to teach us) and then rediscovered by many others independently, including Higgs.

Finally, we still have the magnitude of ϕ. Writing $\phi = v + \delta\phi$, we can quickly see from the Lagrangian that $\delta\phi$ acts as a neutral, massive field. I will leave this as an exercise. This means *the theory has a mass gap!* There are no massless modes and no long-range forces.

Now, what happens to electric fields in this context? Suppose I put an electric charge at the origin. The equation of free electrostatics

$$\nabla^2 A^0 = g^2 \delta(x)$$

whose solution is the usual $1/r$ electrostatic potential, is now modified. The new equation is

$$[\nabla^2 + (gv)^2] A^0 = g^2 \delta(x)$$

The solution to this equation is the Yukawa potential for a massive field with mass $m_\gamma = gv$, $V(r) \propto e^{-m_\gamma r}/r$. The electrostatic field falls off exponentially rapidly at distances larger than the inverse of m_γ. *Electric fields are screened!*.

What about magnetic fields? We cannot expel magnetic fields from an infinite system, but we can make currents, just as in superconductors, from the charged scalar ϕ, and use them to confine magnetic flux. Since the photon is massive, it is energetically preferable for the magnetic field to be localized in tubes where ϕ shrinks to zero and the photon is lighter than m_γ. On the other hand, the presence of the magnetic field in a confined region requires, as we saw, that the phase of ϕ wind an integer number of times around the center of the vortex. Classical solutions to the above equations satisfying these conditions can be found; they are called Nielsen-Olesen vortices. Their tensions can be calculated, and are proportional to $1/g^2$. Thus, *magnetic flux is confined!* The topological analysis that we did for the Abrikosov vortex — that the charges of these vortices is given by the first homotopy group of $U(1)$, the group $\pi_1[U(1)] = \mathbf{Z}$ — goes through here as well, without alteration.

Magnetic flux tubes can arise in other gauge groups as well when they are broken via the Higgs mechanism. If we have a gauge group G broken down to a smaller gauge group H (which might be the identity, as in the example above) we will get magnetic flux tubes if $\pi_1(G/H)$ is not trivial. For example, if we have the group $SU(N)$, and it is broken down to nothing, then there are no flux tubes; $SU(N)$ is simply connected, so all closed curves on it can be shrunk down to nothing, and all of its homotopy groups are trivial. However, if we break $SU(N)$ down to its center $\mathbf{Z}_N$, then since

$\pi_1(G/H) = \pi_0(H)$ if G is simply connected, and since $\pi_0(H)$ is the number of distinct components of H, we have simply $\pi_1(SU(N)/\mathbf{Z}_N) = \mathbf{Z}_N$. Magnetic flux tubes are generated, and they carry a charge in $\mathbf{Z}_N$, the integers modulo N [2]. [As an example, consider $SU(2)$. The matrices $\mathrm{diag}(e^{i\alpha}, e^{-i\alpha})$ are in $SU(2)$; for $\alpha = 0$ and π they are in the center. The path from $\alpha = 0$ to $\alpha = 2\pi$ is a closed path in $SU(2)$, but the path from $\alpha = 0$ to $\alpha = \pi$ is not closed. However, in $SU(2)/\mathbf{Z}_2$, the matrices with $\alpha = 0$ and $\alpha = \pi$ are identified, so the second path is also closed and forms the nontrivial element of $\pi_1(SU(2)/\mathbf{Z}_2) = \mathbf{Z}_2$.]

2.2 Electric Sources and Fluxes

Let us review what we learned in the first lecture, but a bit more formally. Consider a pure gauge theory with gauge group G. Suppose we have a source — an infinitely massive, static, electrically charged particle — in a representation R of G. If we surround the source with a large sphere, what characterizes the flux passing through the sphere? If G is $U(1)$, the flux measures the electric charge directly. However, in non-abelian gauge theories the gauge bosons carry charge. Since there may be a number (varying over time) of gauge bosons inside the sphere, the representation under which the charged objects in the sphere transform is not an invariant. But, by definition, the gauge bosons are neutral under the discrete group C_G, the center of G. It follows that the charge of R under the center *is* a conserved quantity, and that the total flux exiting the sphere carries a conserved quantum number under C_G.

Electric sources and fluxes in pure gauge theories carry a conserved C_G quantum number. If the gauge group confines, then the confining electric flux tubes will also carry this quantum number.

If the theory also contains light matter charged under C_G but neutral under a subgroup C_m of C_G, then the above statements are still true with C_G replaced with C_m. For example, if we take $SU(N)$ with light fields in the $\mathbf{N}$ representation, then C_m is just the identity, reflecting the fact that all sources can be screened and all flux tubes break. If we take $SO(10)$ with fields in the $\mathbf{10}$, then the center $\mathbf{Z}_4$ is replaced with spinor-number $\mathbf{Z}_2$. Sources in the $\mathbf{10}$ will be screened and have no flux tube between them, while sources in the $\mathbf{16}$ or $\overline{\mathbf{16}}$ will be confined by a single type of flux tube.

2.3 Magnetic Sources and Fluxes

Before discussing the magnetic case, I review some basic topology. [The presentation which follows is overly naive, though it serves for present purposes. A more rigorous story requires a study of the relevant fiber bundles.] The p-th homotopy group of a manifold $\mathcal{M}$, $\pi_p(\mathcal{M})$, is the group of maps from the p-sphere into $\mathcal{M}$, where we identify maps as equivalent if they are homotopic (can be continuously deformed into one another) in $\mathcal{M}$. All we will need for present purposes are the following examples. Suppose a Lie group G has rank r, so that its maximal abelian subgroup is $U(1)^r$; then

$$\pi_2[G] = \mathbf{1} \quad \Rightarrow \quad \pi_2\left[G/U(1)^r\right] = \pi_1[U(1)^r] = \mathbf{Z} \times \mathbf{Z} \times \cdots \times \mathbf{Z} \equiv [\mathbf{Z}]^r \ . \quad (4)$$

Similarly,

$$\pi_1[G] = \mathbf{1} \quad \Rightarrow \quad \pi_1\left[G/C_G\right] = \pi_0[C_G] = C_G \ . \quad (5)$$

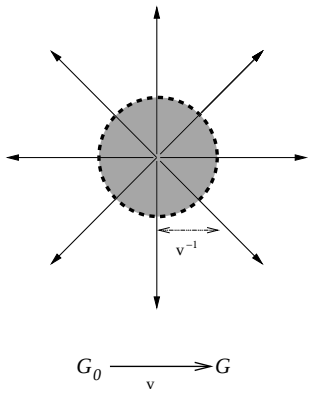

Figure 12: A magnetic monopole soliton of size v^{-1}.

We will need to investigate both monopole solitons and string solitons below. The classic monopole soliton is that of 't Hooft and of Polyakov, which arises in $SU(2)$ broken to $U(1)$; in this case the important topological relation is $\pi_2[SU(2)/U(1)] = \pi_1[U(1)] = \mathbf{Z}$. This leads to a set of monopole solutions carrying integer charge. Note that the stability of, for example, a single monopole which has charge two against decay to two monopoles, each of charge one, is determined not by topology but by dynamics. The situation is similar for the Nielsen-Olesen magnetic flux tube of the abelian Higgs model; here the relevant topological relation is $\pi_1[U(1)] = \mathbf{Z}$. This again leads to solutions with an integer charge, whose stability against decay to minimally charged vortices is determined dynamically.

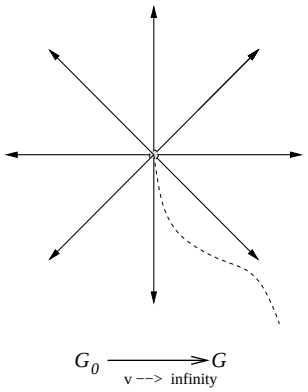

Figure 13: A pointlike Dirac monopole, with its unphysical Dirac string.

More generally, if we have a *simply connected* gauge group G_0 which breaks to a group G at a scale v, there will be solutions to the classical equations in the form of magnetic monopoles carrying a quantum number in $\pi_2[G_0/G]$ (see, for example, [13].) These will have mass [radius] proportional to v $[1/v]$. Now imagine that we take $v \to \infty$. In this limit the gauge group G_0 disappears from the system. The monopoles become pointlike and infinitely massive; their only non-pointlike feature is their (nonphysical) Dirac string, which stems from our having discarded G_0, and which carries a quantum number in $\pi_1[G]$. In short, the solitonic monopoles become fundamental Dirac monopoles in this limit. Note that since $\pi_2[G_0/G] = \pi_1[G]$, the charges carried by the solitonic monopoles and their Dirac monopole remnants are the same. At this point, we can forget about G_0, which is only relevant at infinitely high energies. Since the Dirac monopoles are heavy, we may use them as magnetic sources in a theory with gauge group G.

Let's further suppose that the gauge group G is broken completely at some scale v'. In this case no Dirac strings can exist in the low-energy theory, and so the monopoles allowed previously have seemingly vanished. However, solitonic magnetic flux tubes, carrying charges under $\pi_1[G]$, will be generated; they will have tension [radius] of order v'^2 $[1/v']$. Their $\pi_1[G]$ quantum numbers are precisely the ones they need to confine the $\pi_1[G]$-charged Dirac monopole sources of the high-energy theory. Thus, when G is completely broken, the Dirac monopoles disappear because they are confined by flux tubes.

Magnetic sources and fluxes in pure gauge theories carry a conserved

$\pi_1[G]$ quantum number. If the gauge group is completely broken, then the confining magnetic flux tubes will also carry this quantum number.

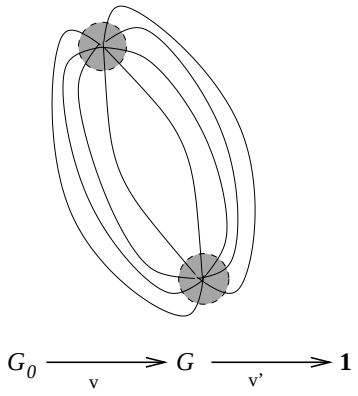

Figure 14: Confined monopole solitons in a theory with flux tubes.

3 Electric-Magnetic Duality?

So let us observe something about $SU(N)$. The electric fluxes of $SU(N)$ are in $C_{SU(N)} = \mathbf{Z}_N$, while its magnetic fluxes are in $\pi_1[SU(N)] = \mathbf{1}$. The electric fluxes of $SU(N)/\mathbf{Z}_N$ are in $C_{SU(N)/\mathbf{Z}_N} = \mathbf{1}$, while its magnetic fluxes are in $\pi_1[SU(N)/\mathbf{Z}_N] = \mathbf{Z}_N$. In fact, more generally, for k a divisor of N, a theory with $SU(N)/\mathbf{Z}_k$ has electric fluxes in $\mathbf{Z}_{(N/k)}$ and magnetic fluxes in $\mathbf{Z}_k$. This electric-magnetic symmetry appears very interesting. What does it mean?

3.1 Duality in Maxwell's theory

The symmetry between electric and magnetic fields in the case of classical electromagnetism is well known. If there are no electric charges present, the Maxwell equations have a symmetry $E \to B$, $B \to -E$. This is physically meaningful, since E and B are both gauge invariant. Without charges, there is no way to say which type of field is which.

Let us be more explicit. Under this transformation, the Bianchi identities $\nabla \times E + \dot{B} = 0, \nabla \cdot B = 0$ are exchanged with the equations of motion $\nabla \times B - \dot{E} = 0, \nabla \cdot E = 0$. Said more covariantly,

$$F_{\mu\nu} \to \tilde{F}_{\mu\nu} = \epsilon_{\mu\nu\rho\sigma} F^{\rho\sigma}$$

and the Bianchi identity $\epsilon^{\mu\nu\rho\sigma}\partial_\rho F_{\mu\nu} = 0$ goes to the equation of motion $\partial^\mu F_{\mu\nu} = 0$.

None of this is particularly obvious if one uses the formalism of potentials, and with good reason. Because of the Bianchi identities, we are free to write $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$, which defines A_μ. The symmetry of the equations under $A_\mu(x) \to A_\mu(x) + \partial_\mu \chi(x)$ is the $U(1)$ gauge symmetry — let us call it the "electric" gauge symmetry. Notice it is not a symmetry of anything physical! It is a symmetry of the variables A_μ! The physical quantities — E and B — are gauge invariant, and are trivial under this "symmetry." This is a good thing, because under exchange of E and B, we cannot exchange A_μ with anything. We must introduce a new, and entirely different, vector potential C_μ, with $\tilde{F}_{\mu\nu} = \partial_\mu C_\nu - \partial_\nu C_\mu$. *No local expression will convert A_μ to C_μ.* Furthermore, C_μ has its own $U(1)$ symmetry — let us call it "magnetic" — $C_\mu(x) \to C_\mu(x) + \partial_\mu \rho(x)$. This is just as unphysical as the first $U(1)$. Even if we were to find a transformation from A_μ to C_μ, nonlocal as it would be, *we are free to redefine C_μ through a magnetic $U(1)$ transformation separately from any redefinition of A_μ through an electric $U(1)$ transformation!* There are two $U(1)$ groups here; they are two entirely distinct symmetries of two entirely distinct sets of variables, and both are unphysical. When we say that the Maxwell equations are the equations of a $U(1)$ gauge theory, we are being extremely careless with the truth.

Let's see this as a path integral statement. (I learned the following from Seiberg and Witten's first paper [9].) They start with the free Maxwell theory

$$\int \mathcal{D}A \; e^{-i\int \frac{F^2}{4g^2}} \;\; \delta(\partial \cdot A)$$

(I will suppress indices except where clearly needed.) Notice the gauge fixing term. This expression is just a number. Let us instead write something more useful. Let's introduce a source $J^{\mu\nu}$ for $F_{\mu\nu}$, and write

$$Z[J] = \int \mathcal{D}A \; e^{-i\int \frac{F^2}{4g^2} + \int JF} \; \delta(\partial \cdot A) \; .$$

Functional derivatives of $\ln Z$ with respect to J now give the correlation functions of F.

Let us now change variables in this path integral. Up to an overall constant. the path integral can be rewritten as an integral, not over A, but over F. We have to be careful, though, because F is subject to the Bianchi

identities, which are exact operator identities. Consider the expression

$$Z[J] = \int \mathcal{D}F \; e^{-i\int \frac{F^2}{4g^2} + \int JF} \; \delta(\epsilon^{\mu\nu\rho\sigma}\partial_\rho F_{\mu\nu}) \; .$$

There's no gauge fixing needed now, but the Bianchi identities must be implemented through a delta function. Let us rewrite this Bianchi identity using a Lagrange multiplier which we will for some unknown reason call C_μ,

$$Z[J] = \int \mathcal{D}F\mathcal{D}C \; e^{-i\int \frac{F^2}{4g^2} + \int JF + \frac{i}{4\pi}\int \epsilon^{\mu\nu\rho\sigma} C_\sigma \partial_\rho F_{\mu\nu}} \; \delta(\partial \cdot C) \; .$$

Notice that the integral over C enforces the Bianchi identity, but since $\epsilon^{\mu\nu\rho\sigma}\partial_\rho\partial_\sigma F_{\mu\nu} = 0$, the Langrange multiplier field C itself has a gauge invariance, which must be fixed by the new delta function. Now let us integrate by parts

$$\epsilon^{\mu\nu\rho\sigma} C_\sigma \partial_\rho F_{\mu\nu} = -\epsilon^{\mu\nu\rho\sigma}\partial_\rho C_\sigma F_{\mu\nu} + \partial^\rho(\cdots) \equiv -\frac{1}{2}\tilde{F}_C F + \partial^\rho(\cdots)$$

where F_C is the field strength of C and $\sim$ represents contraction with an ϵ tensor. We next carry out the integral over F, obtaining

$$Z[J] = e^{-ig^2\int J^2} \int \mathcal{D}C \; e^{-i\int \frac{g^2}{64\pi^2}F_C^2 - \frac{g^2}{4\pi}\int J\tilde{F}_C} \; \delta(\partial \cdot C) \; ,$$

where I have used $\tilde{F}_C^2 = F_C^2$. Thus we recover a free Maxwell theory for C! It looks identical to the original one, except (1) g has been replaced with $\tilde{g} = 4\pi/g$ — weak coupling and strong coupling have been exchanged — (2) the source J, which coupled to F, now couples to $\frac{g^2}{4\pi}\tilde{F}_C$, so the electric field F/g of A_μ is proportional to the magnetic field $\tilde{F}_C/\tilde{g}$ for C_μ, and (3) there is a contact term proportional to J^2 (a typical quantum subtlety which does not affect Green's functions of fields at different points — you may want to experiment with Fourier transforms of Gaussian integrals to see why it is there.) Thus we have found that we can express a single quantum theory (in the form of a generating function for gauge-invariant correlation functions) using two, entirely distinct, integral representations, both of which are nice-looking and well-behaved. One theory, two descriptions, each with its own $U(1)$ gauge (non-)symmetry. This is *duality*.

3.2 The addition of charged fields

Only when we add charges to the theory do we start to learn the distinction between electric and magnetic fields. We know that in nature we have only

electric charges, and all symmetry between electric and magnetic charges is lost. And yet — what if there are magnetic monopoles? Could the symmetry be restored?

Yes, and no. If we put both electric and magnetic currents in the classical Maxwell equations, they look beautifully symmetric: the Bianchi identity $\epsilon^{\mu\nu\rho\sigma}\partial_\rho F_{\mu\nu} = J_m^\sigma$ is exchanged with the equation of motion $\partial_\mu F^{\mu\nu} = J_e^\nu$. All seems well. Electric charges have charge e, while magnetic charges are proportional to $1/e$; thus if electrons are weakly coupled, monopoles are strongly interacting with the photon, and vice versa.

So let us return to the question of confinement. We have seen that we can use condensing electric charges to cause electric charge to be screened, and make magnetic flux confined through the Meissner effect. Clearly, there should be a "dual" Meissner effect; *if we have condensing magnetic charges, then magnetic charge will be screened and electric flux will be confined.* In both cases there will be a mass gap in the theory. Thus we now have a guess as to how confinement will occur: if there are some magnetically charged objects around — perhaps composite ones not visible even semiclassically — then their condensation would cause electric flux to be confined via the dual Meissner effect. All we have to do now is write down the equations governing this process, and see that in such a world, electrons are confined by flux tubes...

But there's a problem. The Bianchi identities are now $\nabla \times E + \dot{B} = J_{mag}$, $\nabla \cdot B = q_{mag}$. This means we *cannot* introduce A_μ anymore; the very introduction of the vector potential imposes the Bianchi identities with zero for the right-hand sides. If we want to introduce a *magnetically* charged field, we will have to use C_μ. In this case, the equations for C_μ will look exactly the same as they did before for A_μ, simply relabelled. And that's not good, if we want to see that *electrically* charged particles are confined. Fields for *electrically* charged particles must have kinetic terms defined using covariant derivatives which contain A_μ! We cannot write a local expression for an electron's kinetic terms if we only have C_μ. Even worse, the presence of the electron field ruins the Bianchi identity for C_μ, so we can't really introduce C_μ either. There isn't going to be a local Lagrangian, and there isn't going to be an ordinary, classical analysis. All we have is a mess.

And that's before quantum mechanics. These complications prevent us from repeating the argument for duality using the path integral. Once there are charged fields, we do not know how (as of yet — though see [14]) to write a path integral which converts an electric description of a theory to an mag-

netic one. (Furthermore, in contrast to $U(1)$ without charged matter, there is in fact little reason to expect that a $U(1)$ theory with charged matter is actually quantum-mechanically dual to an identical theory; it could easily be dual to a nonabelian gauge theory, and/or have multiple dual representations [15].)

We could, of course, forgo the electrically charged particles. Then we would just have a photon coupled to magnetically charged particles; but this would look exactly the same as the superconductor we just considered. That won't help us with Yang-Mills theory, or any other theory with electric confinement that we would like to understand. In such theories, the gluons themselves are chromoelectrically charged, and we can't simply choose to discard all possible chromoelectrically charged objects.

3.3 Duality in pure Yang-Mills?

Can we find a similar duality for the pure Yang-Mills theory? We know that Yang-Mills has the property that it generates electric flux tubes with $\mathbf{Z}_N$ quantum numbers. We might hope that Yang-Mills has an obvious duality to some theory with a gauge group H which has $\pi_1(H) = \mathbf{Z}_N$, so that when H is broken by a condensing field, it generates magnetic flux tubes with $\mathbf{Z}_N$ charges. A natural guess for H would be $SU(N)/\mathbf{Z}_N$. Of course we will need some additional matter — at least a couple of scalar fields — if we are to break this gauge group completely, so the dual description of this theory *can't itself be pure Yang-Mills.* Is there any hope that there exists a dual $SU(N)/\mathbf{Z}_N$ gauge theory of some type, which gives a weakly-coupled (and therefore calculable) dual description analogous to the Meissner effect of confinement in Yang-Mills?

This type of idea, popular briefly in the 1970s, has a few serious problems. First, unlike the case of $U(1)$ gauge theories, the electric and magnetic fields of $SU(N)$ are in the adjoint representation of the gauge group and are not themselves gauge-invariant. This makes the Bianchi identities $\epsilon^{\mu\nu\rho\sigma} D_\rho F^{\mu\nu} = 0$ nonlinear. Secondly, their equations of motion $D_\mu F^{\mu\nu} = 0$ are nonlinear. In both expressions, covariant derivatives appear, which means we always have to write expressions using the vector potential A_μ. This means we cannot simply exchange electric and magnetic fields as we did in the classsical Maxwell equations; the potential appears in the classical equations. At the quantum level, this is equally problematic; the path-integral trick used above for $U(1)$ is useless here, since it required we write the path integral only in

terms of F. (Note Halpern [16] showed in the 1970s that it is consistent to write A nonlocally in terms of F inside a path integral, but no one has figured out how to make use of this fact.)

Another complication is that Yang-Mills theory has a running coupling constant. At high energies it is weak. (Any magnetic description therefore will be strongly coupled at these high energies, but we don't mind that, since the original description is weakly coupled, and extremely useful, in this regime.) At low energies, below Λ, it is strong — but how strong? Is it infinite, or merely order 1? This is important, because we are interested in trying to find a dual description of confinement which presumably inverts the coupling constant $g \to 1/g$. Unless the gauge coupling is much larger than one, our dual description will itself have a large coupling (of order one) and we won't be able to use it for a semiclassical description of the physics. In this case the dual magnetic description will be as hard to use as our original, electric one.

Unfortunately, all indications are that the coupling in the region near Λ is closer to $\sqrt{4\pi}$ than to infinity. There is no evidence that the theory at low energies has a weakly-coupled magnetic description, and the dynamics of the theory does not seem to have any small parameters, or large separations of scales, which could make it easier to analyze. The nonperturbative physics of Yang-Mills may just be a hard problem.

We might be stuck. But here's an idea. What happens if we make the gauge coupling g artificially large? Maybe in that limit a dual description can be found, and its description of confinement will be easier to study and to use. And maybe from there we can get back to the Yang-Mills theory that we want to understand.

How could we do this? Well, let's review... why does the coupling become small in the ultraviolet? It does so because the theory is asymptotically free; its beta function is negative, so the coupling becomes smaller and smaller as we go to high energy. We can't avoid this region of small coupling unless we do something drastic...

Well, one drastic thing we can do is put the theory on a lattice. This means there is a shortest distance below which there can be no vibrations; the theory only looks like pure Yang-Mills at much longer scales. The ultraviolet modes are simply removed, so we won't have to worry about the theory becoming weakly coupled at high energy. In fact, we are free to choose the coupling constant $g(a^{-1})$ at the energy a^{-1} corresponding to the lattice spacing a. Instead of choosing it small and allowing the theory to run to

strong coupling at low energy, let's just choose $g(a^{-1})$ very large. What happens?

In this case we can do a "strong-coupling expansion". I won't review this here, but the expansion on the lattice in powers of $1/g^2$ can in fact be performed [1], and one sees the existence of confining strings right away. There, we're done. Yang-Mills confines chromoelectric field, and Strassler's lectures are over.

Or does it? The problem is that the theory on the lattice has very different dynamics from that of pure Yang-Mills. If $g(a^{-1})$ is very large, then the confinement scale Λ will be at the same order as $1/a$. This can be seen from Eq. (1) with $\mu_0 = a^{-1}$, using the fact that $e^{-8\pi^2/g^2} \sim 1$ if $g^2 \gg 1$. There will be no separation between the scale of the lattice and the scale of confinement. The mass gap will be at this scale also, so there will be no long-distance physics at all. All of the glueball spectrum will be sensitive to the lattice. Thus the theory is very different from Yang-Mills, in fact. If we change the lattice from a square lattice to a triangular one, we will change the glueball spectrum significantly. So why should the fact that the lattice theory confines convince us that when we take the limit

$$a \to 0,\ g^2 \to 0,\ \Lambda^{11N/3} = a^{-11N/3} e^{-8\pi^2/g^2(a)} \text{ fixed},$$

thereby recovering the pure Yang-Mills theory, that the confinement, the flux tubes, and the mass gap will actually survive? Couldn't there easily be a phase transition at some value of g which would change the physics completely?

It's a serious objection. Indeed, we see here a general approach at work, and its basic advantages and disadvantages. Let's review them. We can't study Yang-Mills directly; it is too hard. But let's *change the theory* in a way that allows us to artificially make a parameter small (in this case $1/g^2$.) By doing so, we permit a new expansion in powers of the small parameter. This gives us a calculational technique in which it may be possible to show that confinement and other nonperturbative properties do actually occur, and explain how and why they arise. That's a great idea; and it works, too! But we changed the theory; it is related continuously to Yang-Mills, but that's all. Let's now try to go back to Yang-Mills itself. The problem is that our small parameter will become large again as we do so, and we have no guarantee that confinement, etc., and especially the explanation for confinement, will survive as we make our way back to our starting point. This is especially true since the dynamics of the theory with the small parameter *depends in*

detail on how we changed the theory.

In fact, experience shows that in considering a variety of weakly-coupled variations on pure Yang-Mills, one finds (1) all of the reasonable variations confine, tending to confirm that Yang-Mills confines, and (2) each variation has its own, separate explanation as to how confinement happens. The various explanations have a few things in common but their details are different. Is this progress? I leave this as a question for you to decide.

Similar issues arise for $\mathcal{N} = 1$ SYM. There are many ways to distort the theory (see for example [17, 3, 9]) so that it becomes easier to study; each shows that the theory confines, although each gives a somewhat different explanation. In the remaining part of these lectures, we will be choosing a couple of these variations, and studying how confinement occurs in these cases. We will embed the $\mathcal{N} = 1$ SYM theory into $\mathcal{N} = 4$ SYM — the most symmetric of all gauge theories — and use the dualities of $\mathcal{N} = 4$ to study the confinement in $\mathcal{N} = 1$ (and possibly, if the mathematics is kind, of pure Yang-Mills itself.)

3.4 $\mathcal{N} = 4$ Supersymmetric Gauge Theory

We now need to review some properties of $\mathcal{N} = 4$ supersymmetric gauge theory. We will take the gauge group to be $SU(N)$ unless otherwise noted. The theory consists of one gauge field, four Majorana fermions, and six real scalars, all in the adjoint representation. It is useful to combine these using the language of $\mathcal{N} = 1$ supersymmetry, in which case we have one vector multiplet (the gauge boson A_μ and one Majorana fermion λ) and three chiral multiplets (each with a fermion ψ^s and a complex scalar Φ^s, $s = 1, 2, 3$.)

These fields have the usual gauged kinetic terms, along with additional interactions between the scalars and fermions. I won't write them all here (you can find them in many books and review articles on supersymmetry) and will instead focus on the potential energy for the scalars.

$$V(\Phi^s) = \sum_{a=1}^{dim\ G} |D_a|^2 + \sum_{s=1}^{3} |F_s|^2 \tag{6}$$

where

$$D_a = \left(\sum_{s=1}^{3} [\Phi^{s\dagger}, \Phi^s] \right)_a \tag{7}$$

(here a is an index in the adjoint of G) and

$$F_s = \epsilon_{stu} [\Phi^t, \Phi^u] \ . \tag{8}$$

Supersymmetry requires that $\langle V(\Phi^s)\rangle = 0$, and so all D_a and F_s must vanish separately. The solution to these requirements is that the matrices are all diagonal, namely

$$\langle \Phi^s \rangle = \mathrm{diag}(v_1^s, v_2^s, \cdots, v_N^s) \ . \tag{9}$$

If the v_i^s, thought of as N vectors $\vec{v}_i$, $i = 1, \ldots N$, in a three-dimensional complex space, are all distinct, this breaks G to $U(1)^r$. Since $\pi_2[G/U(1)^r] = [\mathbf{Z}]^r$ [see Eq. (4)] the theory has monopoles carrying r integer charges under $U(1)^r$. (Quantum mechanically, the theory also has dyons, carrying r electric and r magnetic charges (n_e, n_m) [18].)

The space of vacua written in Eq. (9) is not altered by quantum mechanics. In the generic $U(1)^r$ vacuum, each $U(1)$ has no charged matter, and consequently has the usual electric-magnetic duality of the Maxwell equations.

When all v_i^s are zero, the gauge group is unbroken. The theory is conformally invariant. All reasonable Green's functions are power laws. All reasonable operators have a definite, fixed, dimension. The gauge coupling g has an exactly-zero beta function, and does not run. Thus, in contrast to QCD, YM, and $\mathcal{N} = 1$ SYM, the $\mathcal{N} = 4$ SYM theory has a truly dimensionless coupling constant; there is no strong-coupling scale Λ, no dimensional transmutation. We can dial this truly dimensionless g to be whatever we like — it can be small, or it can be large — and it will stay that way at all energy scales. And this nonabelian gauge theory, with lots of charged matter, has a generalization of electric-magnetic duality, suggested first by Montonen and Olive in 1977 [19], in which this coupling constant is inverted.

3.5 Montonen-Olive Duality

Like the pure Maxwell theory, the $\mathcal{N} = 4$ theory has more than one description. There's lots of evidence for this, although it has not been proven directly. Consider this an open challenge.

There is actually an infinite set of alternate descriptions (one has to talk about the θ angle of the theory to obtain them, and I will not have time to cover this here) but the most important one, for our purposes, exchanges electric and magnetic charges. It is generated by a change of variables **S** analogous to the one we discussed above for electromagnetism, but whose explicit form remains a mystery. It has the effect

$$\mathbf{S} \ : g \to \frac{4\pi}{g}; q_e \leftrightarrow q_m \ ; \ G \to \tilde{G} \ . \tag{10}$$

S exchanges electric and magnetic charge, inverts the gauge coupling [19], and changes the gauge group [20, 21] from G to its dual group $\tilde{G}$, as defined below.

The group G has a root lattice Γ_G which lies in an $r = \mathrm{rank}(G)$ dimensional vector space. This lattice has a corresponding dual lattice $(\Gamma_G)^*$. It is a theorem that there exists a Lie group whose root lattice $\Gamma_{\tilde{G}}$ equals $(\Gamma_G)^*$ [20]. Here are some examples:

$$\begin{array}{ll} SU(N) \leftrightarrow SU(N)/\mathbf{Z}_N \ ; & SO(2N+1) \leftrightarrow USp(2N) \ ; \\ SO(2N) \leftrightarrow SO(2N) \ ; & Spin(2N) \leftrightarrow SO(2N)/\mathbf{Z}_2 \ . \end{array} \tag{11}$$

Notice that this set of relationships depends on the global structure of the group, not just its Lie algebra; $SO(3)$ (which does not have spin-1/2 representations) is dual to $USp(2) \approx SU(2)$ (which does have spin-1/2 representations.) These details are essential in that they affect the topology of the group, on which Montonen-Olive duality depends.

In particular, there are two topological relations which are of great importance to Montonen-Olive duality. The first is relevant in the generic vacuum, in which G is broken to $U(1)^r$. The electric charges under $U(1)^r$ of the massive electrically charged particles (spin $0, \frac{1}{2}, 1$) lie on the lattice Γ_G. The massive magnetic monopoles (*also* of spin $0, \frac{1}{2}, 1$) have magnetic charges under $U(1)^r$ which lie on the dual lattice $(\Gamma_G)^*$ [20, 21]. Clearly, for the **S** transformation, which exchanges the electrically and magnetically charged fields and the groups G and $\tilde{G}$, to be consistent, it is essential that $\Gamma_{\tilde{G}} = (\Gamma_G)^*$ — which, fortunately, is true.

The second topological relation is the one we will use below. We have seen that the allowed electric and magnetic sources for a gauge theory with adjoint matter (such as $\mathcal{N} = 4$) are characterized by quantum numbers in C_G and $\pi_1(G)$ respectively. Consistency of the **S** transformation would not be possible were these two groups not exchanged under its action. Fortunately, it is a theorem of group theory that [20]

$$\pi_1(G) = C_{\tilde{G}} \ ; \ \pi_1(\tilde{G}) = C_G \ . \tag{12}$$

For example, $\pi_1[SU(N)] = C_{SU(N)/\mathbf{Z}_N} = \mathbf{1}$ while $C_{SU(N)} = \pi_1[SU(N)/\mathbf{Z}_N] = \mathbf{Z}_N$.

Thus, as a consequence of Eq. (12) and the results discussed in our earlier discussions of electric and magnetic fluxes and sources, the allowed magnetic sources of G are the same as the allowed electric sources for $\tilde{G}$, and vice

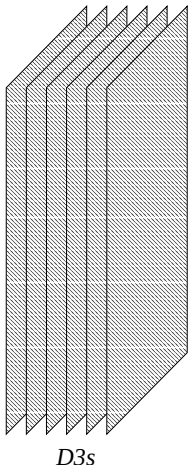

Figure 15: N D3 branes have a $U(N)$ $\mathcal{N} = 4$ SYM on their world volume.

versa. This is a significant piece of evidence in favor of S-duality, and will be essential later on.

Now, this is not the only way to approach $\mathcal{N} = 4$ SYM, as you have already heard in Prof. Maldacena's lectures. As he showed you, the world-volume theory on a stack of N D3 branes of Type IIB string theory has a complicated action, but at low energy it reduces to $\mathcal{N} = 4$ $U(N)$ SYM theory. The extra $U(1)$ decouples, and all of the interesting physics is in the $SU(N)$ part of the theory.

Do we see signs of S-duality in this string construction of $\mathcal{N} = 4$ SYM? We certainly do! Type IIB string theory itself has an S-duality — for which, again, there is tremendous evidence but no proof (see for example [22] and [23].) The duality inverts the string coupling: $g_s \to 1/g_s$. It also changes various extended objects into one another. The theory has (among other things) fundamental strings, Neveu-Schwarz 5-branes, and D1, D3 and D5 branes. (It also has D(-1) and D7 branes but we won't discuss them.) Now, under S-duality, the D1 and F1 (fundamental) strings are exchanged, as are the D5 and NS5 branes. The D3 branes, however, are unchanged. The $\mathcal{N} = 4$ $SU(N)$ SYM theory goes back to itself, except that its coupling constant $g^2_{YM} = g_s/4\pi$ is inverted — just as we expected! Furthermore, a fundamental string ending on a D3 brane looks like a point electric charge from the perspective of an observer stuck on the D3 brane. A D1 brane ending on a D3 brane looks like a point magnetic charge. Thus S-duality in

Type IIB string theory correctly inverts the $\mathcal{N} = 4$ SYM coupling constant, exchanges its electric and magnetic charges, and exchanges the gauge groups of the electric and magnetic descriptions.[3]

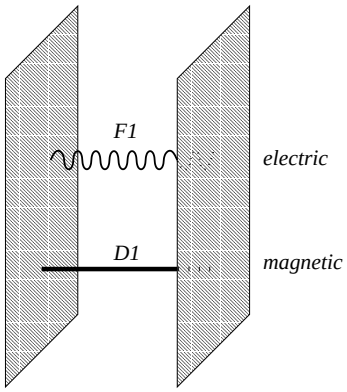

Figure 16: F1 (D1) strings appear as electrically (magnetically) charged particles.

A word of warning about this beautiful structure. Most examples of duality are much more complicated than this! The identification of the dual group is vastly more difficult, and the relations which we have used in arguing that it is $\tilde{G}$ do not work. So don't be fooled into thinking that most of the other known dualities are this elegant. They are both less straightforward and much richer in content. A good example for you to look at is the Seiberg duality of $\mathcal{N} = 1$ supersymmetric gauge theories [17, 24, 25], which could actually be relevant in nature. But the example of $\mathcal{N} = 4$ duality proves to be a good one for examining confinement in $\mathcal{N} = 1$ SYM and pure YM, so we'll stick with it.

4 Breaking $\mathcal{N} = 4$ to $\mathcal{N} = 1$

It's time to return to our goal of discussing confinement in $\mathcal{N} = 1$ SYM theory. Let's try to apply the trick we discussed earlier in the context of the strong-coupling expansion on the lattice. Is there, perhaps, a way to take $\mathcal{N} = 1$ SYM, make its coupling artificially large, and do a strong-coupling expansion? The lattice badly breaks supersymmetry, so it won't help us very

[3]Well, almost. Actually, the D3-branes give $U(N)$, whose dual is $U(N)$ again. To remove the $U(1)$ factors, and see the $\mathbf{Z}_N$, is subtle. It is much easier to see that $SO(2N+1)$ is exchanged with $USp(2N)$, so you might try that instead.

much (although it might be worth revisiting this point after recent advances in lattice theory [26].) A different approach would be to put $\mathcal{N} = 1$ Yang-Mills theory inside of $\mathcal{N} = 4$ Yang-Mills. How might we do this?

We could add to the $\mathcal{N} = 1$ SYM theory three chiral multiplets (that's three Majorana fermions and six real scalars) in the adjoint representation of the group, all with a common mass m. We'll also add some additional interactions, so that when m goes to zero the theory has $\mathcal{N} = 4$ supersymmetry. We take all scalars to have expectation values less than or of order m (an assumption which will be justified later.)

At energies well above m, the theory is approximately $\mathcal{N} = 4$ SYM. Since the masses m are comparatively tiny at these energy scales, the theory will be approximately conformally invariant. The gauge coupling will run very little for energies bigger than m, and for very high energy it goes to a constant g_0. But at energies well below m, the classically massless particles will be those of $\mathcal{N} = 1$ SYM. Quantum mechanically, the gauge coupling will run below the scale m, and confinement will presumably occur at some scale $\Lambda < m$.

Thus this $\mathcal{N} = 1$ supersymmetric theory — which we will call "$\mathcal{N} = 1^*$", for short — interpolates between $\mathcal{N} = 4$ SYM and $\mathcal{N} = 1$ SYM. As required for our trick, we have kept the basic $\mathcal{N} = 1$ SYM infrared dynamics but have changed the ultraviolet behavior of the theory in such a way that we can, if we wish, ensure the coupling constant is always large! In particular, we can simply choose the ultraviolet value of the coupling g_0 much larger than one. Since $g(\mu) \approx g_0$ for $\mu > m$, the coupling constant at $\mu = m$ will also be large — and thus, just below the scale m, we obtain a theory with the matter content of $\mathcal{N} = 1$ SYM, but with an artificially large coupling constant. All we have to do now is expand in $1/g_0$. But that's exactly what Montonen-Olive duality allows us to do! The magnetic dual description of this physics will be weakly coupled, with coupling constant $1/g_0 \ll 1$.

But how close will the $\mathcal{N} = 1^*$ theory be to $\mathcal{N} = 1$ SYM? What properties will they share? It is worth examining the strong coupling scale of the $\mathcal{N} = 1^*$ theory. Below the scale m, the coupling constant $g(\mu)$ will run as it does in pure $\mathcal{N} = 1$ SYM theory, so the one-loop relation between $g(\mu)$ at the scale $\mu = m$ and the scale Λ reads

$$\Lambda^{3N} = m^{3N} e^{-8\pi^2/g^2(m)} \approx m^{3N} e^{-8\pi^2/g_0^2}$$

Notice that if g_0 is small, $\Lambda \ll m$, but if g_0 is large, as we will want for our strong-coupling expansion, $\Lambda \sim m$. Thus, just as in the lattice strong-

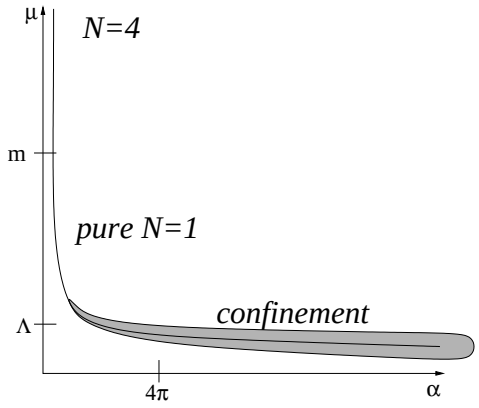

Figure 17: $\mathcal{N} = 1^*$ for small g_0.

coupling expansion, there will not be a separation of scales between the new physics (in this case the three massive adjoint multiplets) and the scale of confinement, glueball masses, etc. We will not be doing much better than the lattice case. Our strong-coupling expansion will depend on the details of our the mass scale m. For example, if we give the extra chiral multiplets different masses instead of a common mass m, the glueball spectrum will reflect this change, although there would be no such change at small g_0 where $m \gg \Lambda$. This is the standard limitation; we accept it and move on.

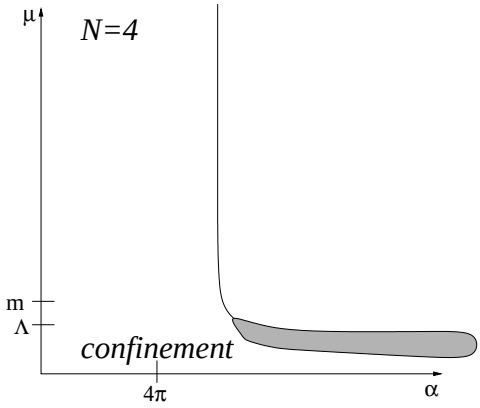

Figure 18: $\mathcal{N} = 1^*$ for large g_0.

You might wonder if there is some danger that the massive chiral multiplets will ruin the confinement we want to study. In fact, there is not

much to worry about. As we noted earlier, $\mathcal{N} = 1$ SYM has confining strings because neither gluons nor gluinos can break these flux tubes; fields in the adjoint representation are neutral under the center of the gauge group C_G. The addition of massive matter in the adjoint representation does not change this; heavy particles would only obstruct confinement by breaking flux tubes, which adjoint matter cannot do. We therefore can expect that $\mathcal{N} = 1^*$ should share some qualitative features with pure $\mathcal{N} = 1$ SYM: both should have mass gaps and confine flux into tubes carrying a C_G quantum number.

Now let's examine things more closely. Let's first take g_0 very small so we can do a semiclassical analysis. When we break the $\mathcal{N} = 4$ supersymmetry by adding masses m for the fields Φ^s, the F_s functions of (8) become

$$F_s = \epsilon_{stu}[\Phi^t, \Phi^u] + m\Phi^s \ , \tag{13}$$

so that $F_s = 0$ implies $\epsilon_{stu}[\Phi^t, \Phi^u] = -m\Phi^s$ [8]. Up to normalization, these are the commutation relations for an $SU(2)$ algebra; thus solutions will take the form

$$\Phi^1 = -imJ_x \ ; \Phi^2 = -imJ_y \ ; \Phi^3 = -imJ_z \ , \tag{14}$$

where J_x, J_y, J_z are $N \times N$ matrices satisfying $[J_x, J_y] = iJ_z$, *etc.*, a representation of $SU(2)$. Each possible gauge-inequivalent choice for the J's gives a separate, isolated vacuum of the classical $\mathcal{N} = 1^*$ theory [8].

How does this work, explicitly, in $SU(N)$? We can write the Φ^s as $N \times N$ traceless matrices, so the J_s should be an N-dimensional (generally reducible and possibly trivial) representation of $SU(2)$ [8, 3]. The trivial choice corresponds to $J_i = 0$; clearly if $\Phi^s = 0$ the JJ commutation relations are satisfied. We will call the corresponding vacuum the "unbroken" vacuum, since the $SU(N)$ gauge group is preserved. Another natural choice is to take the J_s in the irreducible spin-$\frac{N-1}{2}$ representation of the $SU(2)$. In this case $SU(N)$ is completely broken (this is left as an exercise); we will call this the "Higgs vacuum". We may also choose the J_s in a reducible representation

$$J_s = \left[\begin{array}{c|c|c} \sigma_s & & 0 \\ -- & -|- & -- \\ 0 & & 0 \end{array}\right] \ ; \tag{15}$$

here the σ_s are the Pauli matrices. In this case $SU(N)$ is partly broken. There are many vacua like this last one, but they will play no role in today's story; we will only need the unbroken vacuum and the Higgs vacuum.

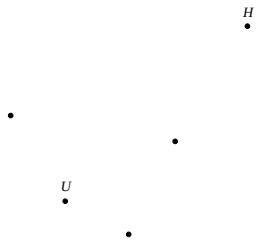

Figure 19: A few of the classical vacua of $\mathcal{N} = 1^*$, including the unbroken (U) and completely Higgesd (H) vacua.

In all of these vacua, the scalar fields are massive, as are most of the fermions. However, in any vacuum with unbroken gauge symmetry, there are both massless gauge bosons and their massless fermionic superpartners. Thus, the Higgs vacuum has a mass gap — there are no massless fields — while the unbroken vacuum has the massless gauge bosons and fermions of an $SU(N)$ $\mathcal{N} = 1$ SYM theory.

As an example, let's take the case of an $SU(2)$ gauge group [8]. This is a rather degenerate one, but it has all the essential features. In this case we need two-by-two matrices which satisfy the above commutation relations; the only solutions are $J_s = 0$ and $J_s = im\sigma_i$. We thus have two classical vacua, one with unbroken $SU(2)$ gauge symmetry, and one in which the $SU(2)$ is completely broken by the Higgs mechanism. (The expectation value for Φ^3 breaks $SU(2)$ to $U(1)$, while the expectation values for Φ^1 and Φ^2 break the remaining $U(1)$.)

In summary, the classical analysis of the $SU(N)$ $\mathcal{N} = 1^*$ theory shows that it has isolated supersymmetric vacua scattered about, with the unbroken (U) vacuum at the origin of field space and the Higgs vacuum (H) at large Φ^s expectation values (of order m) [8, 3]. The Higgs vacuum has a mass gap, while the unbroken vacuum has the matter content of an $SU(N)$ $\mathcal{N} = 1$ SYM theory.

4.1 OM Duality and the Yang-Mills String

The above picture is modified by quantum mechanics. The U vacuum has the matter content of $SU(N)$ $\mathcal{N} = 1$ SYM theory. Remember we are still working at small g_0. We know this theory is asymptotically free, so at an energy scale exponentially small compared to m — more precisely, at an energy $\Lambda \sim me^{-8\pi^2/3Ng_0^2} \ll m$ — the gauge coupling will become strong.

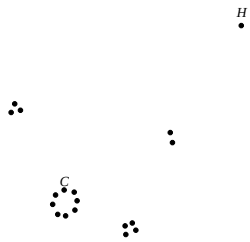

Figure 20: Quantum mechanically, the vacua with unbroken gauge groups split; the U vacuum splits into N, one of which (C) has confinement via magnetic monopole condensation.

Since this scale is so small, the physics at energies of order m cannot affect it. We already know, then, what this theory will do; it will confine, generate a mass gap of order Λ, and it will have not one but N vacua due to the breaking of the $\mathbf{Z}_{2N}$ axial symmetry down to $\mathbf{Z}_2$. As we noted earlier, these vacua are related by this $\mathbf{Z}_{2N}$ symmetry, so we can focus on just one of them.[4] Let's call it the confining (C) vacuum.

By contrast, in the H vacuum the gauge group is completely broken at the scale $m \gg \Lambda$, and there is a mass gap of order m, so there is no way for non-trivial low-energy dynamics to take place. Consequently, the H vacuum remains a single vacuum.

Now let's compare the Higgs vacuum and the confining vacuum. Recall that we took the gauge group to be $SU(N)$. The confining vacuum has a strongly-coupled process of confinement and generation of a mass gap of order $\Lambda \ll m$. We expect the confining electric flux tubes to have tension of order Λ, and for them to carry a $\mathbf{Z}_N$ charge. In the Higgs vacuum, on the other hand, there is a weakly-coupled breaking of the gauge group. We can see classically that a mass gap is generated. But actually the gauge group is not completely broken. The adjoint scalar fields carry no charge under the center of the group, so $SU(N)$ is in fact broken down to its center $\mathbf{Z}_N$! Now, we have already learned that there will be solitonic magnetic flux tubes in any breaking of a gauge theory $G \to H$ if $\pi_1(G/H)$ is nontrivial, and these strings will carry charges in $\pi_1(G/H)$. Here we have $SU(N) \to \mathbf{Z}_N$, and $\pi_1[SU(N)/\mathbf{Z}_N]$ is $\mathbf{Z}_N$. So the Higgs vacuum has *confining magnetic flux tubes*, carrying charge $\mathbf{Z}_N$, as a result of condensation of the electrically

[4] A caution: this symmetry is actually only exact when $m \to \infty$, and is approximate if $m \gg \Lambda$. However, for reasons explained below, the number of vacua cannot change when m varies, so our counting of vacua is correct for any m. The vacua are actually related by shifting the θ angle by $2\pi k$, $k \in \mathbf{Z}$.

charged fields Φ^s. The scale of these flux tubes and of the mass gap is $\sim m$.

This is extremely suggestive. Let us attempt to rewrite this physics using the magnetic description of the theory. Montonen-Olive duality converts g_0 to $1/g_0 \gg 1$... oops. The physics we were just discussing for small g_0 will now be converted to a very strongly coupled description. In such a highly-fluctuating set of variables, we won't know how to calculate anything. Bad move.

So instead, let's first *continously* vary g_0 from small to large, as we had discussed doing earler. Now Λ and m will gradually become of the same order. The classical analysis we performed of the Higgs vacuum will become invalid, as will our semiclassical analysis of the unbroken vacuum. However, we may now appeal to a special property of supersymmetric field theories. Even an $\mathcal{N} = 1$ supersymmetric theory has the property that the energy of any field configuration is positive. All supersymmetric vacua have exactly zero energy, and are global minima of the potential. Furthermore, the potential energy is proportional to the square of a complex function, whose zeroes are controlled by complex analysis. These zeroes cannot simply disappear. Even if we change g_0 (which, when combined with the θ angle of the gauge theory, is actually complex) the number of zeroes cannot suddenly change. (This hand-waving argument is vastly improved by consideration of Witten's index [27], discovered around 1980.) This gives us great confidence that even at large g_0, the H vacuum will still exist, with a mass gap and confining magnetic flux tubes, and so will the C vacuum, with its own mass gap and confining electric flux tubes. This is not quite a proof, but the evidence is very strong. (The mathematics of [3] elevates the argument to a near-proof.)

Now, having moved to a theory with $g \gg 1$ which still has the flux tubes of interest, let's apply a strong-coupling expansion by switching over to the magnetic description of the theory, using $SU(N)/\mathbf{Z}_N$ variables whose gauge coupling is $\tilde{g} = 4\pi/g_0$. What happens in the magnetic description? Not only does Montonen-Olive duality invert the gauge coupling, exchange electric and magnetic charge, and switch $SU(N)$ with $SU(N)/\mathbf{Z}_N$, giving a new description in terms of new adjoint gauge, spinor, and scalar fields $\hat{\Phi}^s$, magnetically charged, *it also exchanges the H vacuum with the C vacuum* [8, 3]!

It's important not to get confused, so let's review. In the electric theory, there is an H vacuum, described at small g_0 by simple breaking of a gauge group by condensation of the Φ^s fields. We don't have a good electric de-

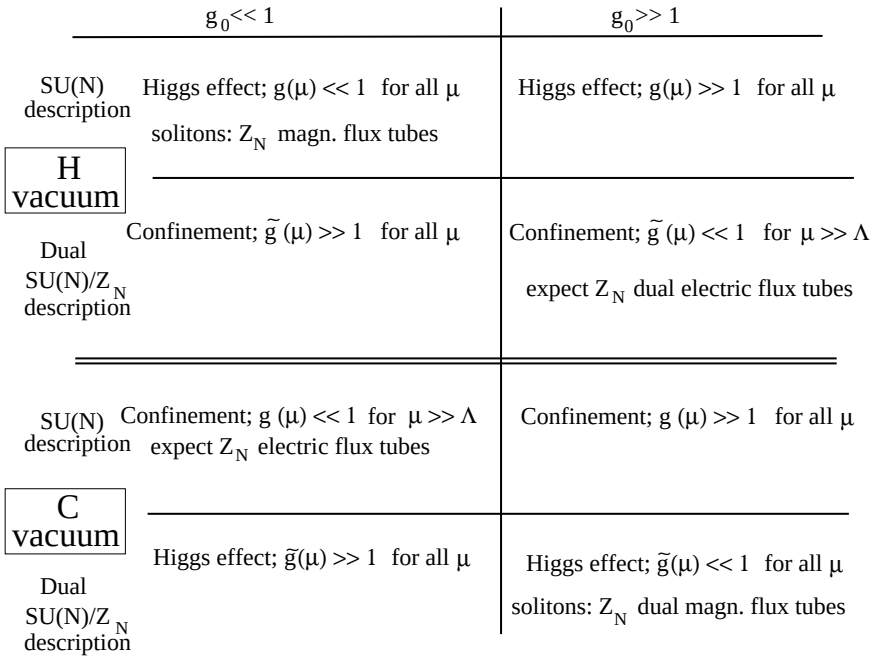

Figure 21: The Higgs and Coulomb vacua, in the regions of large and small g_0, as described by the two different sets of variables.

scription of it at large g_0, but we know it still exists. We also know there is a C vacuum, and we don't have a good electric description of it even at small g_0, much less at large g_0. Each of these two vacua may also be described using the magnetic variables of the $\mathcal{N} = 4$ theory. In these variables, we do not have any good descriptions when g_0 is small, since $1/g_0$ is big. However, when g_0 is large, and $1/g_0$ is small, we have a good description of the C vacuum (!) which is exactly *isomorphic* to the small-g_0 electric description of the H vacuum at small g_0. And that's what we want: a magnetic description of the C vacuum, valid at $g_0 \gg 1$, which makes it easy to see the confining electric flux tubes of the C vacuum. In this magnetic description of the C vacuum, the electric flux tubes are simply the semiclassical (remember $\tilde{g}_0 \ll 1$) solitonic strings which emerge from the condensation of the scalars $\hat{\Phi}^s$, which are *magnetically* charged and break the *magnetic* gauge group from $SU(N)/\mathbf{Z}_N$ to nothing. These solitons carry $\mathbf{Z}_N = \pi_1[SU(N)/\mathbf{Z}_N]$ charge — which is exactly what we need! Furthermore, we can easily see how the mass gap is generated in this context, just as it is generated classically at small g_0 in the H vacuum.

So we have found our strong-coupling description of confinement, and it is precisely as we originally suggested: it is a non-Abelian generalization of the dual Meissner effect, in which condensation of magnetically charged scalar fields generates a mass gap and confines electric flux. The picture even gives us flux tubes with the correct charges!

Can we go back to $\mathcal{N} = 1$ SYM? No; that would require varying $m \to \infty, g_0 \to 0$, which would make the magnetic description of the C vacuum strongly coupled and unreliable. But by supersymmetry, the physics should not change too much as we vary g_0. We may therefore consider this a near-proof that $\mathcal{N} = 1$ SYM does indeed have a mass gap and confinement. It is a strong argument that the corresponding flux tubes carry $\mathbf{Z}_N$ charges for the flux tubes. However, it is no proof at all that confinement occurs via a simple picture of condensing, weakly-coupled magnetically-charged objects. In fact, it firmly suggests that the magnetic condensation process is *strongly coupled.* This means, for example, that any calculation of the string tension, or even of ratios of tensions of different flux tubes, will be suspect. Qualitatively things look great; but a quantitative tool this is not.

Should we expect this picture to survive to the non-supersymmetric case? Take the theory with $\mathcal{N} = 4$ supersymmetry broken to $\mathcal{N} = 1$, and futher break $\mathcal{N} = 1$ supersymmetry by adding an $SU(N)$ gluino mass $m_\lambda \ll m$. Duality is in fact enough to tell us how to implement this breaking at leading order in m_λ/m. However we don't need to think very hard. We know that the theory has a mass gap, so small supersymmetry-breaking can only change some properties of the massive fields, *without altering the fact that* $SU(N)/\mathbf{Z}_N$ *is completely broken.* The strings, whose existence depends only on this breaking, thus survive for small m_λ. To reach pure YM, however, requires taking m, m_λ all to infinity together as $g_0 \to 0$. It seems probable, given what we know of YM physics, that the strings undergo no transition as these masses are varied. In particular, we may hope that there is no phase transition for the strings between pure $\mathcal{N} = 1$ SYM and pure YM. Note that this conjecture can, and should, be tested numerically on the lattice.

If in fact the strings of $\mathcal{N} = 1$ SYM and of YM are continously related, without a transition as a function of the gluino mass, then the arguments given above for $\mathcal{N} = 1$ SYM extend to YM, establishing a direct link between Montonen-Olive duality of $\mathcal{N} = 4$ gauge theory and the confining $\mathbf{Z}_N$-strings of pure YM theory.

4.2 A gravitational description of confinement

We have used up most of these lectures, and yet still not reached the latest developments. I will give an overview of some recent work with Polchinski [28] which gives a new and remarkable picture of confinement. A somewhat different picture emerged earlier in this context [29], and other pictures were discovered later [30, 31, 32]. The reason for the existence of all of these different pictures is the same as before: each of them represents a distinct modification of the confining theory of interest into a regime where there is a new small parameter, and each therefore agrees that confinement occurs but disagrees on the precise mechanism.

Let me comment on these disagreements. We should abstract a lesson from all this, namely that confinement is a generic property of gauge theories for which there can be many causes. The various causes we are learning about need not be directly relevant for pure YM, or $\mathcal{N} = 1$ SYM, which is too bad, since it means that we are not yet learning any quantitative method for computing in such a theory. But it may be that neither of these theories has enough small parameters to permit simple computation. We are not guaranteed that a given physical phenomenon has a perturbative expansion in some parameter, any more than we are guaranteed a similar property for a generic function. It may be that the only way to understand Yang-Mills theory is either to simulate it or solve it exactly. The latter goal is far beyond any mathematical problem ever solved. Simulation may be the end of the line. [Fortunately, in real-world QCD, there are large global symmetries among the *quarks* which are only weakly broken. Expansions around an exactly-globally-symmetric theory in the small symmetry-breaking parameters has allowed many *relations between quantities* in nonperturbative QCD to be predicted. This was essential in the development of the theory of the strong interactions.]

But even if our new descriptions of confinement are less relevant for YM and $\mathcal{N} = 1$ SYM (and we already know they are even less relevant for QCD,) they still provide new phenomena for us to think about, ones which could be relevant in yet other contexts. The goal of these lectures is not merely to explore confinement in YM and SYM. It is to show you the variety of phenomena in gauge theories, and encourage you to consider the possibility that confinement occurs elsewhere in nature, perhaps in unexpected ways and in unexpected places.

In particular, the most strange and wonderful of all of the developments

of the 1990s has been the discovery that string theory and field theory are not even distinct mathematical entities. In the Maldacena [33] conjecture, sharpened further by Witten [34] and by Gubser, Klebanov and Polyakov [35], there is strong evidence for a new form of duality. We saw earlier that we may take a generating functional and give it multiple integral representations, each of them with a four-dimensional local Lagrangian in its integrand, giving us a local quantum field theory. But it turns out that we may also rewrite this functional as a well-known string theory in 9+1 dimensions, with five of the dimensions compact. Even though Polyakov [36] has argued for years that we should seek a five-dimensional string to describe gauge theories in four-dimensions, it is astonishing that the needed string is one that we already know. (Of course the string theory has its own dualities, so we mustn't limit ourselves to a single set of variables for it either.)

There are many technical problems with this duality. First, we don't know how to write a path integral for string fields. (The usual two-dimensional world-sheet path integral is analogous to a one-dimensional particle world-line path integral, not to the path integral of a four-dimensional field theory. The first is "first-quantization", the second is "second-quantization".) We therefore have no explicit way to write the equating of the field theory and the string theory. Second, the string theory is particularly nasty. The presence of large curvatures and large Ramond-Ramond fields makes the usual techniques of classical string theory invalid. But fortunately there is a limit in which these issues are unimportant, and it is in that limit that we may hope to study new properties of field theory. This is the limit in which the quantum string theory reduces simply to classical supergravity. (Actually this is too restrictive as has been shown very recently [37, 38].) In the remaining time, we will seek to study the $\mathcal{N} = 1^*$ theory in a regime where it is simply described by semiclassical supergravity coupled to strings and to branes.

Both pure YM and the $\mathcal{N} = 1$ SYM theory have two parameters, the QCD scale Λ and the number of colors N. (The coupling $g(\mu)$ runs with scale and is a determined function of μ and Λ; thus it is not an independent parameter.) However, Λ is simply the only scale in the problem, so it is not a dimensionless quantity that it is meaningful to vary. The only other parameter available is N, and it has long been suggested that as $N \to \infty$ gauge theory might simplify, and might even be soluble. The solution to large N gauge theory has remained elusive, however.

By contrast, the $\mathcal{N} = 4$ theory has *two* dimensionless parameters: N

and the high-energy coupling g_0. As Maldacena has shown you, the two parameters play an essential role in the string theoretic description of the $\mathcal{N} = 4$ theory. The coupling $g_0^2/4\pi$ is the string coupling g_s, which when small makes the string theory classical. However, this is not enough, since even the classical theory in background Ramond-Ramond fields is too complicated. When $g_0^2 N/4\pi \equiv \lambda$, the 't Hooft coupling, is large, then the space on which the classical string theory is defined becomes very large, with very low curvature; then the string theory reduces to its low-energy limit, namely type IIB supergravity.

Here we see that the hope of the previous paragraphs, that the large N limit of gauge theory might simplify, appears to be partially realized. At large N we do indeed find a new description, a classical string theory. But only if we simultaneously take λ large do we obtain a well-understood theory, one in which anything can be calculated. At small λ the theory is very complicated. This is unfortunate, because the YM and SYM theories we might want to study do not have a dimensionless parameter corresponding to λ. The gauge coupling runs from small to large, so we are guaranteed that at high energy $\mu \gg \Lambda$ the running $\lambda(\mu)$ will be small (which is not a problem, because we can use field theory perturbation theory in that regime) and that $\lambda(\mu)$ becomes potentially large only near to the energy scale Λ. Unfortunately, there is no evidence that $\lambda \gg 1$ at $\mu \sim \Lambda$. More likely, it is only of order 2π, which (when you check the factors of 2π) is not sufficiently large for gravity to work. In particular, in $\mathcal{N} = 1$ SYM, the scale of confinement and the mass gap is

$$\Lambda \sim \mu e^{-2\pi/3\lambda(\mu)}$$

(in pure YM, replace 3 with 11/3) so the energy scale μ is of the same order as the confining scale when λ is of order 2π. Thus, even if gravity were to actually describe confinement in YM or $\mathcal{N} = 1$ SYM, it could only do so at energy scales extremely close to Λ, corresponding to a ten-dimensional space whose curvature would be large everywhere except (at best) in a small region.

If this is true, then gravity cannot provide a nice description of confining YM or $\mathcal{N} = 1$ SYM. The confinement occuring in these theories can only be studied using the classical but extremely complicated theory of strings in Ramond-Ramond fields and on a highly curved space. This duality is not much better than the electric-magnetic duality we had before. But we can consider our by-now familiar trick; can we find a way to distort YM or

$\mathcal{N} = 1$ SYM in such a way that we can take λ artificially large?

Yes; just as before, let us consider $\mathcal{N} = 1^*$. The $\mathcal{N} = 1^*$ theory has *three* parameters: N, g_0 and m. The first two are those of $\mathcal{N} = 4$ and are the important ones in the ultraviolet. In the infrared, g_0 and m are combined into Λ, leaving N as the only dimensionless constant. As in our earlier discussion, we may take g_0 small but $g_0^2 N$ large; then the ultraviolet theory will be approximately $\mathcal{N} = 4$ SYM *in the supergravity regime!* We can then consider the effect of $m \neq 0$ in the context of supergravity, and see if we can obtain a picture of how confinement occurs. As always, the corresponding picture will be special to this particular deformation of $\mathcal{N} = 1$ SYM — note that $\Lambda \sim m e^{-8\pi^2/3Ng_0^2}$ and m will be of the same order, so as usual our confining scale will not be well-separated from the physics of the massive adjoint chiral multiplets — but we'll accept this limitation and move forward.

4.3 Confinement in the supergravity regime of $\mathcal{N} = 1^*$

This is a long story, and I can't describe it all here. One needs a nice discussion of branes, fluxes, and all the rest. So let me be schematic, and give you a brief but telling overview of what happens in this theory. Needless to say, a significantly more rigorous discussion appears in our paper [28].

The key idea was provided by Rob Myers, in a slightly different context [39]. What he showed was this. Suppose you take a collection of flat Dp branes, forming $p+1$ dimensional Minkowski space $\mathcal{M}^{p+1}$ embedded in $9+1$ dimensional flat space. Now subject them to a certain electric field, not an ordinary $F_{\mu\nu} = \partial_{[\mu} A_{\nu]}$ but rather a derivative of an antisymmetric-tensor potential with $p+3$ indices — in short, an electric field with $p+4$ indices. In this background field, the Dp branes link together and expand into a D$(p+2)$ brane, with a $p+3$-dimensional worldvolume in the form of a two-sphere [40] times $\mathcal{M}^{p+1}$.

Myers called this "dielectric branes", and with good reason. Take an atom; it is electrically neutral, but carries a global charge, its atomic number. Now subject it to an electric field. It will polarize, as in a dielectric. It is still electrically neutral, but it locally has electric charge. Also, it still has its atomic number charge, which is unaffected. Here, our N Dp branes carry a charge, the total number N. After they expand into a D$(p+2)$ brane, what do they have? First, the number of Dp branes hasn't changed; that charge remains. Second, the total D$(p+2)$ brane charge hasn't changed; a brane

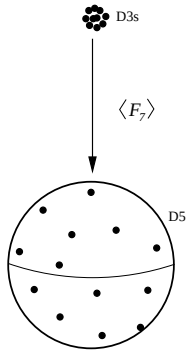

Figure 22: The Myers effect for D3-branes.

in the form of a two-sphere can collapse and disappear, so our D$(p+2)$-brane will vanish if we turn off the electric flux, and there is no net charge associated with it. Still, locally on the two-sphere, there *is* D$(p+2)$ brane charge. Go near to the two-sphere and you can feel it; the other side of the sphere, with cancelling charge, is far away. Thus the Dp branes have expanded into a D$(p+2)$-brane *dipole*! Particles form dipoles by moving apart a certain distance; strings and other branes form dipoles by forming closed surfaces; but the idea is the same.

What's the connection? Take the $\mathcal{N}=4$ theory, described as type IIB string theory on $AdS_5 \times S^5$. Now modify the gauge theory by adding mass terms as in $\mathcal{N}=1^*$. It turns out that the modification of the Lagrangian by the mass operators corresponds, in supergravity, to turning on a background electric field, a tensor with 7 indices. The D3-branes, whose near-horizon geometry formed the $AdS_5 \times S^5$ spacetime, expand, as Myers suggested, into a 5-brane. However, they have two choices (actually many more, but we'll only consider these two for now.) They can expand into a D5-brane. But by S-duality, under which D3-branes are invariant and D5-branes are exchanged with NS5-branes, it must also be possible for the D3-branes to expand into an NS5-brane. Solving the equations, one finds that both of these possibilities are realized. The first corresponds to the Higgs vacuum of $\mathcal{N}=1^*$, the second to the confining vacuum!

What does this do to the supergravity? The full supergravity solution has still not been found. However, we were able to show that there exists a good perturbative expansion in this theory which allows us to demonstrate solutions of the following form: at large AdS radius r, near the boundary,

we have $AdS_5 \times S^5$ modified slightly by corrections of order $1/r$ to a power. At a radius of order $m\alpha' N$ these corrections become large. A singularity is avoided, however, by the presence of a D5-brane (or NS5-brane) carrying N units of D3-brane charge. The brane has world-volume S^2 (placed on an equator of the S^5) times $\mathcal{M}^4$ (parallel to the boundary of AdS_5.) Specifically, this prevents the 7-form electric flux from diverging and causing the metric to do the same. Instead, there is a smooth solution (except at the position of the 5-brane, where there is a standard and understood singularity) which rounds off nicely at $r = 0$, without a horizon or singularity at that point. In fact, for $r \ll m\alpha' N$, the spacetime is approximately flat ten-dimensional space.

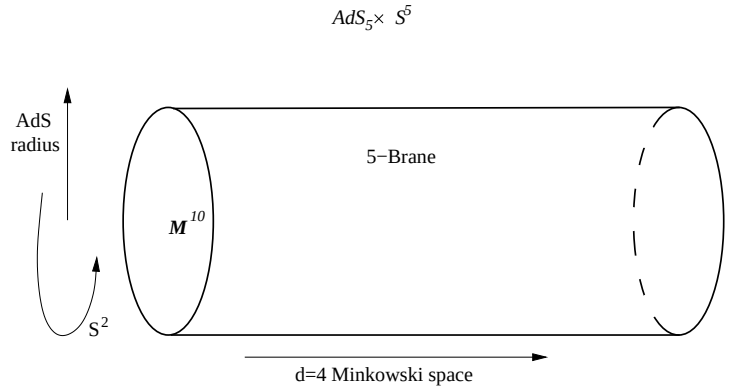

Figure 23: A useful geometrically-reduced representation of a 5-brane of the sort found in the $\mathcal{N} = 1^*$ solution.

What about confinement? Can we see that magnetic flux is confined in the H vacuum and that electric flux is confined in the C vacuum? Indeed we can. D-branes, by definition, are places where strings can end. In particular, F1-strings can end on D3- and D5-branes. But then, by S-duality, D1-branes can end on D3- and NS5-branes. On the other hand, F1-strings cannot end on NS5 branes, nor D1-branes on D5-branes. Another important feature is that D1 branes, and F1-strings, if placed parallel to D3-branes, can dissolve in them. But D1-branes cannot dissolve into D5-branes, nor can F1-strings dissolve into NS5-branes.

All of these facts have physical implications for the $\mathcal{N} = 4$ and $\mathcal{N} = 1^*$ field theories. F1-strings ending on D3-branes look like electrically charged particles; D1-strings look like magnetic monopoles. We can create a pair of oppositely-oriented F1-strings, for example, and move them apart without

large energy cost; thus the electric charges are unconfined, as expected in $\mathcal{N} = 4$ SYM. An F1-string placed parallel to and inside a stack of D3-branes corresponds to putting a line of electric flux into the $\mathcal{N} = 4$ theory. The dissolving of this line indicates that electric flux prefers to minimize its energy by expanding to infinity. Thus electric flux is, as expected, unconfined. The same holds for magnetic flux, a dissolving D1-brane.

However, in the $\mathcal{N} = 1^*$ theory the vacua of the theory correspond to 5-branes with D3-brane charge. Now, in the H vacuum, we have a spherical D5-brane, on which D1-branes cannot end! Magnetic charges can no longer appear with finite energy. And suppose we put a D1-brane parallel to and near a D5-brane which also carries D3-brane charge. Here a remarkable thing happens; the D1-brane can only *partially* dissolve. The D3-branes try to make the D1-brane expand, but the D5-brane charge prevents its complete dissolution. We are left with a diffuse, but nonetheless finite-thickness, D1-brane–D5/D3-brane bound state. The magnetic flux corresponding to the D1-brane expands, but only to a tube of fixed size; it is *confined* in this tube. Furthermore, if we attempt to produce a pair of magnetic monopoles in the form of D1-branes ending on this D5/D3-brane composite, we will find instead that they are connected by this diffuse flux tube. The charges, kinematics and dynamics of D-branes tell us that magnetic charge is confined in the Higgs vacuum of $\mathcal{N} = 1^*$!

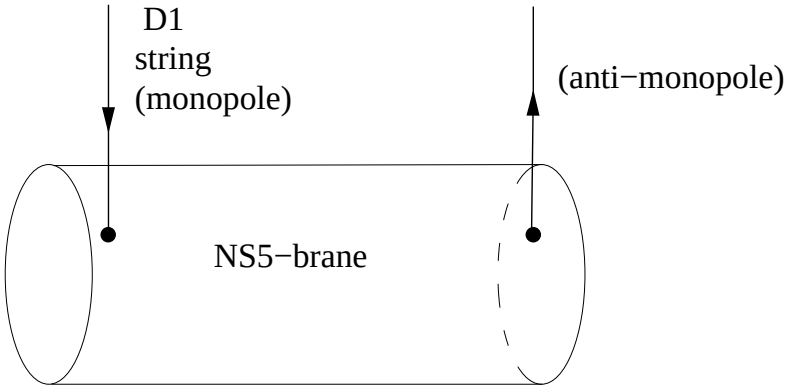

Figure 24: Monopoles (D1 strings) can end on NS5-branes; they are not confined.

The S-dual story holds in the confining vacuum. We can repeat the entire previous paragraph, exchanging D1 with F1, D5 with NS5, and magnetic with electric. The conclusion is also exchanged: the charges, kinematics and

dynamics of NS5-branes, D3-branes and fundamental strings tell us that electric charge is confined in the appropriate vacuum of $\mathcal{N} = 1^*$. We have found a new picture for confinement. It occurs through the appearance of an NS5-brane dipole in the 9+1-dimensional spacetime. The dipole prevents flux tubes, in the form of fundamental strings, from dissolving into the D3-branes contained within the dipole, and instead makes them into flux tubes which are fundamental strings bound to the NS5-brane![5]

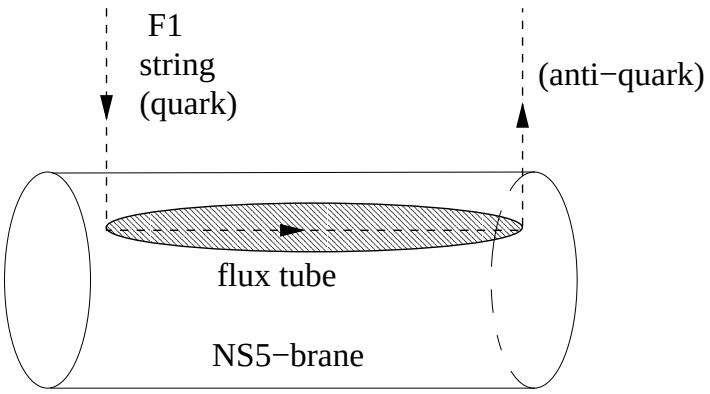

Figure 25: Heavy quarks (fundamental strings) cannot end on NS5-branes; however, there is an NS5-F1 bound state that serves as a flux tube connecting the quark and antiquark.

Of course this is not the end of the story. One should (and can) check that there is a mass gap, that strings carry $\mathbf{Z}_N$ charges, that various expectation values come out correctly, etc. For a few quantities, there are exact results from field theory that are complicated functions of λ and N; comparison with our gravity solution shows precise agreement, even for the numerical coefficients. There is also an exciting new form of duality, which is beyond the scope of these lectures, which takes not $g_0 \to 1/g_0$ but $\lambda \to 1/\lambda$! This is still largely unexplored territory, although it has been discussed further in [41].

It is important to remember that we have not been constructing an analogy. We have not found a new "model" for confinement in field theory. This *is* confinement in field theory. The string theory is just a convenient

[5] In principle it is also possible to break supersymmetry. If the supersymmetry breaking is small the story does not change much. For large supersymmetry breaking, of order m, the technical challenges become greater. It is not known whether reliable computations can be done in that regime, although there are no known obstructions.

description of it; but we are not dealing with a different theory, just an alternative description of the *same* theory. This mechanism for confinement is a new behavior of ordinary, four-dimensional continuum field theory which was not previously known. It is one of several which have been uncovered in the regime of large 't Hooft coupling.

However, as always, this is not confinement in pure $\mathcal{N} = 1$ SYM. To reach that theory, we would have to take the 't Hooft coupling λ small. In that limit, the NS5-brane dipole would shrink in size, its radius becoming of order the string scale. All calculational control would be lost. That's the price we paid for our new picture. Like Moses, we can see the promised land but never quite manage to reach it.

5 Wrap-up

In these lectures I have given you an overview of some of the key ideas underlying confinement as a property of field theory, and now, of string theory as well. This is a tiny fraction of what field theory (and now string theory) is capable of, and we are still uncovering new features on a monthly basis. In fact, most field theories do not have confinement, for reasons entirely different from those of QCD. Many become nontrivial conformal field theories at low energy. Others become composite, weakly-coupled gauge theories (the so-called "free-magnetic phase" [17].) Dualities of many stripes are found everywhere. Ordinary dimensional analysis in string theory is totally wrong in the regime where it looks like weakly-coupled field theory, and ordinary dimensional analysis in field theory is totally wrong in the regime where it looks like weakly-coupled supergravity. There's much more. You are encouraged to stride into the midst of these developments, to search with us for new features of both field theory and string theory (or, better said, of the single theory of which both are a part,) and most importantly, and most difficult, to explain to us what all these dualities really mean, and where they come from. Good hunting.

References

[1] K. G. Wilson, Phys. Rev. D **10**, 2445 (1974).

[2] G. 't Hooft, Nucl. Phys. B **190**, 455 (1981), Nucl. Phys. B **153**, 141 (1979), Nucl. Phys. B **138**, 1 (1978).

[3] R. Donagi and E. Witten, Nucl. Phys. B **460**, 299 (1996) [hep-th/9510101].

[4] M. J. Strassler, Nucl. Phys. Proc. Suppl. **73**, 120 (1999) [hep-lat/9810059].

[5] M. Wingate and S. Ohta, Phys. Rev. D **63**, 094502 (2001) [hep-lat/0006016].

[6] B. Lucini and M. Teper, hep-lat/0103027.

[7] B. Lucini and M. Teper, Phys. Lett. B **501**, 128 (2001) [hep-lat/0012025].

[8] C. Vafa and E. Witten, Nucl. Phys. B **431**, 3 (1994) [hep-th/9408074].

[9] N. Seiberg and E. Witten, Nucl. Phys. B **426**, 19 (1994) [Erratum-ibid. B **430**, 485 (1994)] [hep-th/9407087].

[10] D. Finnell and P. Pouliot, Nucl. Phys. B **453**, 225 (1995) [hep-th/9503115].

[11] A. A. Abrikosov, Sov. Phys. JETP **5**, 1174 (1957) [Zh. Eksp. Teor. Fiz. **32**, 1442 (1957)].

[12] H. B. Nielsen and P. Olesen, Nucl. Phys. B **61**, 45 (1973).

[13] S. Coleman *Aspects of Symmetry, Selected Erice lectures* (Cambridge University Press, Cambridge, 1985)

[14] A. Kapustin and M. J. Strassler, JHEP **9904**, 021 (1999) [hep-th/9902033].

[15] N. Seiberg and E. Witten, Nucl. Phys. B **431**, 484 (1994) [hep-th/9408099].

[16] M. B. Halpern, Phys. Rev. D **19**, 517 (1979).

[17] N. Seiberg, Nucl. Phys. B **435**, 129 (1995) [hep-th/9411149].

[18] E. Witten, Phys. Lett. B **86**, 283 (1979).

[19] C. Montonen and D. Olive, Phys. Lett. B **72**, 117 (1977).

[20] P. Goddard, J. Nuyts and D. Olive, Nucl. Phys. B **125**, 1 (1977).

[21] H. Osborn, Phys. Lett. B **83**, 321 (1979).

[22] J. Polchinski *String Theory, Volume II* (Cambridge University Press, Cambridge, 1998)

[23] J. H. Schwarz, Nucl. Phys. Proc. Suppl. **55B**, 1 (1997) [hep-th/9607201].

[24] K. Intriligator and N. Seiberg, Nucl. Phys. Proc. Suppl. **45BC**, 1 (1996) [hep-th/9509066].

[25] M. J. Strassler, Prog. Theor. Phys. Suppl. **131**, 439 (1998) [hep-lat/9803009].

[26] F. Niedermayer, Nucl. Phys. Proc. Suppl. **73**, 105 (1999) [hep-lat/9810026].

[27] E. Witten, Nucl. Phys. B **188**, 513 (1981).

[28] J. Polchinski and M. J. Strassler, hep-th/0003136.

[29] E. Witten, Adv. Theor. Math. Phys. **2**, 505 (1998) [hep-th/9803131].

[30] I. R. Klebanov and M. J. Strassler, JHEP **0008**, 052 (2000) [hep-th/0007191].

[31] J. M. Maldacena and C. Nunez, Phys. Rev. Lett. **86**, 588 (2001) [hep-th/0008001].

[32] C. Vafa, hep-th/0008142.

[33] J. Maldacena, Adv. Theor. Math. Phys. **2**, 231 (1998) [Int. J. Theor. Phys. **38**, 1113 (1998)] [hep-th/9711200].

[34] E. Witten, Adv. Theor. Math. Phys. **2**, 253 (1998) [hep-th/9802150].

[35] S. S. Gubser, I. R. Klebanov and A. M. Polyakov, Phys. Lett. B **428**, 105 (1998) [hep-th/9802109].

[36] A. M. Polyakov, hep-th/9304146.

[37] J. Polchinski and M. J. Strassler, Phys. Rev. Lett. **88**, 031601 (2002) [arXiv:hep-th/0109174].

[38] D. Berenstein, J. Maldacena and H. Nastase, JHEP **0204**, 013 (2002) [arXiv:hep-th/0202021].

[39] R. C. Myers, JHEP **9912**, 022 (1999) [hep-th/9910053].

[40] D. Kabat and W. I. Taylor, Adv. Theor. Math. Phys. **2**, 181 (1998) [hep-th/9711078].

[41] O. Aharony, N. Dorey and S. P. Kumar, JHEP **0006**, 026 (2000) [hep-th/0006008].