_attice Gauge Theories: an Introduction, which explains most of the concepts and technology. There are also shorter papers by a number of authors. The purpose of these lectures is to explain to a field theorist, what the main ideas of lattice gauge theory are, and hopefully to see that it teaches us something about the nature of field theory. With this in mind, I will not emphasize algorithmic methods whatsoever. If these lectures are successful, the reader will be able to attend a lattice gauge theory talk and at least understand the basic ideas underpinning the techniques being presented; and they may gain a better mental picture of what a number of field theory concepts "really mean"–particularly, nonabelian gauge invariance, Wilson lines and Wilson loops, and the problems with the regularization of chiral fermions.

The original format of these notes was as a series of hour and a half lectures (which each ran over), one per chapter. There are fairly strong logical divisions between chapters, but each builds on the previous one.

1 Uses of Euclidean Correlation Functions

As we will see, lattice gauge theories are only capable of determining, directly, Euclidean correlation functions of a quantum field theory. At first sight, this makes them useless, because we live in Minkowski space; this means that lattice gauge theory will fail to compute any of the correlation functions which we actually want, in order to solve QCD. This is an exaggeration. Indeed, there are many correlation functions which we wish lattice QCD could compute, but which it will not be able to. However, Euclidean correlation functions do have a role, and a number of interesting questions can be answered with them. The purpose of this chapter is to explain some of the most prominent examples.

1.1 Equal time correlation functions

All correlation functions (all physics) can be found from analytic continuation of Euclidean correlation functions. Some things can be found directly in Euclidean space.

Equal time vacuum correlation functions

$$\langle 0|\, \mathcal{O}_1(x,t)\mathcal{O}_2(y,t)\mathcal{O}_3(z,t)\,|0\rangle \tag{1}$$

can be found as follows (assuming the theory has a mass gap, which is good enough for us since it is true in QCD):

Observe that

$$e^{-H\tau}\,|E_n\rangle = e^{-E_n\tau}\,|E_n\rangle \qquad \text{and in particular} \qquad e^{-H\tau}\,|0\rangle = |0\rangle\,, \tag{2}$$

which means that, for large τ, $e^{-H\tau}$ acts as a projection operator onto the vacuum,

$$e^{-H\tau}\left[\sum_n P_n\,|E_n\rangle\right] = \sum_n P_n e^{-E_n\tau}\,|E_n\rangle \to_{\tau>>E_1} P_0\,|0\rangle\,. \tag{3}$$

Now, sum over all states but with such a projection operator,

$$\begin{aligned} & \langle 0|\, \mathcal{O}_1(x,t)\mathcal{O}_2(y,t)\mathcal{O}_3(z,t)\,|0\rangle & (4)\\ = & \sum_n \langle E_n|\, e^{-H\tau}\mathcal{O}_1(x,t)\mathcal{O}_2(y,t)\mathcal{O}_3(z,t)e^{-H\tau}\,|E_n\rangle & (5)\\ = & \mathrm{Tr}\,\left[\mathcal{O}_1(x)\mathcal{O}_2(y)\mathcal{O}_3(z)e^{-2H\tau}\right] & (6)\end{aligned}$$

which can be rewritten as a path integral by the standard trick (which I assume you know),

$$\frac{\int \mathcal{D}[\text{fields}]\;\mathcal{O}_1(x)\mathcal{O}_2(y)\mathcal{O}_3(z)\;\exp(-\int_0^{2\tau} d\tau' \int d^3x \mathcal{L}_{\mathrm{E}}[\text{fields}])}{\int \mathcal{D}[\text{fields}]\;\;\exp(-\int_0^{2\tau} d\tau' \int d^3x \mathcal{L}_{\mathrm{E}}[\text{fields}])}\,, \tag{7}$$

where now the field operators appearing in $\mathcal{O}_1$ etc. are replaced by their values. The region of the path integral is $\Re^3$ (or space) times S^1 the circle of radius 2τ (the summation on $|E_n\rangle$ gives periodic boundary conditions, hence the space is the circle).

Therefore, we see that all vacuum, equal time correlation functions can be determined directly from a Euclidean path integral. Note that these expressions are only valid in the large τ limit; outside this limit, we are actually computing the equal time, finite temperature correlation functions (if fermions are given anti-periodic boundary conditions).

1.2 Spectrum of the theory

What do you get when you act with "some" operator $\mathcal{O}$ with a particular set of angular momentum, parity, charge conjugation, and other conserved numbers, summarized by an index α? (One usually calls this combination of conserved quantities the "channel" in which the operator acts.)

First define

$$\mathcal{O}_\alpha(\mathbf{p}) \equiv \int_x e^{i\mathbf{p}\cdot\mathbf{x}} \mathcal{O}_\alpha(\mathbf{x}) \,, \tag{8}$$

the operator which creates a state in a definite linear momentum state. Generically,

$$\mathcal{O}_\alpha(\mathbf{p}) \left|0\right\rangle = \sum_\beta c_\beta \left|\mathbf{p}, \alpha, \beta\right\rangle \tag{9}$$

where β is an index over all states in the Hilbert space with momentum $\mathbf{p}$ and conserved numbers α. There may be both single particle states and multiparticle (continuum) states included in such a sum. c_β are the coefficients giving how much of each state the operator $\mathcal{O}_\alpha$ turns out to produce. Typically there are many choices of operator within a symmetry channel, and these coefficients will be different for each operator.

It may be possible to carefully construct an operator for which c_1, the coefficient for the lowest lying state, is 1 and all others are 0. Such an operator would fulfil our intuitive idea of a creation operator for the particle. However, a generic operator will create some rich combination of the lightest state and excited and multi-particle states, and there is no good first-principles way of guessing what operator creates just the lowest lying state.

However, it is still possible to determine the particle spectrum, or at least the lowest lying states in each symmetry channel. To do so, we use the fact that the operator $e^{-H\tau}$ suppresses high energy states. Consider

$$\left\langle 0\right| \mathcal{O}_\alpha^\dagger(-\mathbf{p}) e^{-H\tau} \mathcal{O}_\alpha(\mathbf{p}) \left|0\right\rangle . \tag{10}$$

This equals

$$\sum_{\beta\delta} c_\beta^* c_\delta \left\langle \mathbf{p}, \alpha, \beta\right| e^{-E_\delta \tau} \left|\mathbf{p}, \alpha, \delta\right\rangle = \sum_\beta \left|c_\beta\right|^2 e^{-E_\beta \tau} \,, \tag{11}$$

and for suitably large τ, only the lowest energy state contributes. Therefore, at large τ we find the τ dependence is an exponential tail. The amplitude of that exponential tail gives us the coefficient $|c_1|^2$, which is some information about the normalization of our operator as an interpolating function. The exponential decay rate gives the energy of the lightest state in this symmetry channel, unless by accident c_1 happens to be zero.

One may improve on this method by using a collection of operators $\mathcal{O}_1$, $\mathcal{O}_2, \ldots$ all with the same symmetry properties. One then considers the matrix

$$\begin{bmatrix} \left\langle 0\right| \mathcal{O}_1^\dagger e^{-H\tau} \mathcal{O}_1 \left|0\right\rangle & \left\langle 0\right| \mathcal{O}_2^\dagger e^{-H\tau} \mathcal{O}_1 \left|0\right\rangle & \ldots \\ \left\langle 0\right| \mathcal{O}_1^\dagger e^{-H\tau} \mathcal{O}_2 \left|0\right\rangle & \left\langle 0\right| \mathcal{O}_2^\dagger e^{-H\tau} \mathcal{O}_2 \left|0\right\rangle & \ldots \\ \ldots & \ldots & \ldots \end{bmatrix} \tag{12}$$

at a few (suitably large) values of τ. The lowest lying eigenvalue's exponential falloff gives the lowest state energy; the next lowest eigenvalue gives the next state's energy. The lowest eigenvector gives a good interpolating function for the lowest energy state, the best interpolating function one can get as a linear combination of the $\mathcal{O}_i$. The second eigenvector gives an interpolating function which does not produce the lowest lying state.

The above discussion becomes more complicated when multi-particle states can carry the quantum number in question. Then, rather than a sum of exponentials, the large τ falloff will contain an integral over the spectral weight of the cut associated with the multi-particle states. This becomes especially problematic when we consider an unstable particle, such as the ρ meson, where the particle's mass is higher than the energy where the two-particle cut begins. If this is the case, then the large τ behavior will not be a pure exponential. One can learn something about resonances by Laplace transforming the τ dependence of the operator correlation function to get the spectral weight as a function of energy. However, such Laplace transforms from numerical data are more problematic, numerically, than the extraction of a few exponential coefficients. Note also that, depending on the operator used, there may be more overlap with the bound states or more overlap with the multiparticle states. For instance, for the ρ meson, an operator which is well localized and corresponds with the naive quark model content of a ρ meson is likely to have more overlap with the ρ, while an operator which is very extended and "looks like" two pions is liable to have more overlap with continuum states.

The point is that, again, the correlation functions we want can be written entirely in terms of Euclidean path integrals, this time with the operators at different points in the Euclidean time direction:

$$\langle 0|\, \mathcal{O}_1^\dagger(\mathbf{x}) e^{-H\tau} \mathcal{O}_2(\mathbf{y})\, |0\rangle = \frac{\int \mathcal{D}[\text{fields}] \mathcal{O}_1^\dagger(\mathbf{x},\tau)\mathcal{O}_2(\mathbf{y},0) \exp(-\int \mathcal{L}_\text{E}[\text{fields}])}{\text{same, no operators}} \tag{13}$$

Again the fields within the operators $\mathcal{O}$ are replaced by their values (the integration variables). The above expression is an "ordinary" integral, made up of c-numbers (albeit with an infinite number of integration variables, which is the next problem to address).

1.3 Matrix elements

It is also possible to find weak matrix elements from purely Euclidean field theory. The idea here is that we want the value of perfectly local operators, integrated over space, sandwiched between an initial and a final particle state. For instance, to study the exclusive decay of a B^+ meson into a particular hadronic final state and a lepton pair, we need the matrix element

$$\langle \mu^+ \nu_\mu D|\, \mathcal{O}_\text{weak}\, |B^+\rangle \tag{14}$$

with $\mathcal{O}_\text{weak}$ a 4-fermi operator containing a b annihilation and d creation operator and two fermionic operators, which can be determined by integrating out the W boson and then running the resulting operator down in scale to the renormalization point implicitly hiding in the above expression. The leptonic part of this calculation can be done analytically; the hard part is the QCD part. However, we have already seen how to construct the states $|B^+\rangle$ etc. using interpolating operators and $e^{-H\tau}$ to project to the lowest energy state which we want. Now, we will need to know the normalization $|c|^2$ of the interpolating operator from the previous discussion. This can be determined from the amplitude of operator-operator correlation functions, as discussed already. There can also be difficulties in controlling the

purity and normalization of the weak operator (in practice), but leave that aside for the current discussion. The point is that the quantity of interest can be determined from correlation functions of three operators, namely interpolating functions for initial and final states and $\mathcal{O}_{\rm weak}$, separated by Euclidean time.

2 Lattice theories: the general idea

Now we are convinced that Euclidean correlation functions are interesting. How are we going to determine them nonperturbatively?

2.1 Why we want to do lattice calculations

The idea of a lattice calculation is to look at some field theory problem which can be formulated in terms of a Euclidean path integral, say,

$$\frac{1}{Z}\int \mathcal{D}\Phi\ \mathcal{O}_1(\Phi)\mathcal{O}_2(\Phi)\ \exp\left(-\int d^4x \mathcal{L}_{\rm E}(\Phi)\right)\ , \tag{15}$$

and then actually "do the integral," for instance by some numerical Monte-Carlo procedure. This is an interesting thing to do when the action is so non-Gaussian that analytic techniques, such as perturbation theory, are not reliable. It allows us to compute from first principles nonperturbative phenomena. For the current discussion we will talk about scalar Φ^4 theory, to avoid issues about gauge invariance and special complications associated with fermions. We will return to these special complications, since, in fact, the only reason that lattice techniques are interesting is that QCD is asymptotically free, and this actually demands that we discuss gauge invariance.

The problem is that $\mathcal{D}\Phi$ really means an integral over the value of Φ, for every point in space. This is an infinite number of integrals. There is also a serious issue of renormalization which is being swept under the rug in the above discussion, which has to be dealt with.

The first step is to reduce the amount of work potentially needed, by making the spacetime considered finite. Last section explains why replacing $\int_{-\infty}^{\infty} d\tau$ with $\int_0^{N/m} d\tau$, with $N \gg 1$ and m the mass gap, is good enough. The same is true of the spatial extent; compactifying space with $L \gg 1/m$ leaves only exponentially small sensitivity to L, provided that

1. we choose a space with no boundaries (otherwise the regions near boundaries are messed up by boundary effects),

2. we choose a space with zero (metric) curvature (otherwise there is something local which tells you the space is finite).

a torus (or twisted torus) is an ideal choice. We make it now.

Next is the radical step; we replace continuous space with a lattice, so there are only a finite number of points, and therefore a finite number of integrations to be performed. So

the integral we want turns into

$$\int \mathcal{D}\Phi \Rightarrow \int \prod_{x_1=[a,\ldots,L];\, x_2=\ldots} d\Phi(x_1, x_2, x_3, \tau)\,, \tag{16}$$

which is a finite number $(L/a)^3(\tau/a)$ of integrations.[1] That means it is at least conceptually possible to do the integrals by some Monte-Carlo procedure. Doing the integral allows the computation of all the Euclidean correlation functions. In practice one gets them with error bars, because the integration has to be done by some approximate numerical technique. These notes will not discuss such approximate techniques, the reader should simply accept that such techniques exist.

At this point we should comment why lattice methods are only applicable to Euclidean correlation functions. The reason is that the integrand in Eq. (15) is positive definite. The integration is done by some sort of sampling procedure; it is essential that all the samples are adding to each other, and that the integral is absolutely convergent. If the action exponential were $\exp(i\sum_x \mathcal{L}_{\rm M}(\Phi(x)))$, the exponential of i times the Minkowski action, then each possible field configuration would be contributing equally, but with a different phase. The point is that numerical methods are far better suited to treat "steepest descent," positive definite integrands than they are to treat "stationary phase," non-absolutely-integrable integrands. Euclidean path integrals are generally purely real; Minkowski ones are purely complex. Note however that even Euclidean path integrals sometimes have problems with nontrivial phases, generally associated with the breaking of C symmetry, either explicitly (as in the case of finite $\Theta_{\rm QCD}$), or by a C non-symmetric density matrix (as in the case of finite chemical potential). We will not discuss these quite difficult problems further in these notes.

2.2 Lattice actions, derivatives

But what is the action? And do I actually get the theory I want? The potential term is easy to convert from continuous to lattice variables:

$$\int d^4x V(\Phi(x)) \Rightarrow a^4 \sum_x V_{\rm bare}(\Phi(x))\,. \tag{17}$$

We had better think about the coefficients of the potential as bare parameters. Roughly, they are at the renormalization point $\mu = 1/a$; eventually we will say something more precise here.

The choice of lattice derivative is not unique. It is constrained by the requirements,

1. Locality: the lattice version of $(\nabla\Phi)^2$ can contain Φ at different lattice points, say, $\Phi(x)K(x-y)\Phi(y)$, but K must have compact support. Presumably we will want an $a \to 0$ limit and in this limit the support of K will go towards zero.

[1] We have made a definite choice here, to take a *hypercubic* lattice. This is not the only lattice possible in 4 dimensions; for instance, the D_4 lattice (the hypercubic lattice plus a point at the center of each lattice cell, a generalization of the face centered cubic lattice) is in some respects a more elegant choice. A relatively small literature exists on these other choices, which are generally ignored because there are very practical reasons to prefer the hypercubic lattice; it is easier to think about, and it is easier to implement in code.

2. Right continuum limit: the lattice derivative must equal $(\nabla\Phi)^2$ for smoothly varying Φ, up to higher order in a terms, and, potentially, multiplicative renormalizations.

The simplest solution is

$$\frac{1}{2}(\nabla\Phi)^2(x) \;\Rightarrow\; \frac{1}{2}\sum_\mu \left[\frac{\Phi(x+a\hat{\mu})-\Phi(x)}{a}\right]^2 , \tag{18}$$

$$\int_x \;\Rightarrow\; a^2\sum_{x,\mu}\left[\Phi^2(x)-\Phi(x+a\hat{\mu})\Phi(x)\right] . \tag{19}$$

In the second expression we have re-arranged some terms, which is equivalent in the continuum to integrating by parts. The term with $\Phi(x+a\hat{\mu})\Phi(x)$ is often called a "hopping term."

2.3 IR effective theory, accidental $O(4)$ invariance

So, write down this Lagrangian. Does this theory actually correspond to a continuum Φ^4 theory? We can argue that it does if

1. We choose the mass to be small, $|ma| \ll 1$, so that correlation lengths much longer than the lattice spacing can exist.

2. We make the coupling weak at the lattice spacing scale, or more precisely we choose λ in a range which is not separated from $\lambda = 0$ by a phase transition (such things can happen).

These conditions are sufficient to ensure that interesting IR ($p \ll 1/a$) correlations exist. We should be able to describe what happens in the IR using an infrared effective theory. Since we took weak coupling, the field content of the effective theory is Φ the field content of the lattice theory.

The right IR effective theory is the most general theory we can write down, with the field content and symmetries of the lattice theory. The parameters of that effective theory must then be determined by a matching calculation. The symmetries are

- Discrete $\Phi \to -\Phi$ symmetry; Φ always appears in even powers in any operator in the Lagrangian.

- Discrete translation symmetry

- Hypercubic point symmetry

Discrete translation symmetry is what we have instead of continuous translation symmetry. Rather than ensuring momentum conservation, it ensures momentum conservation up to $2\pi/a$ (Umklapp). This is good enough since the IR effective theory only deals with small momenta $p \ll 1/a$; in practice discrete translation symmetry is just as good as full translation

symmetry. However, we will have to remember this Umklapp possibility when we think about fermions on the lattice.

Hypercubic symmetry is that subgroup of $O(4)$ which takes the hypercubic lattice (points with all integer coordinates) to itself. That is the same as permutations of the (x, y, z, τ) axes, with possible sign flips on each axis. The group therefore has $4!\, 2^4 = 384$ (in general, $D!\, 2^D$) elements. It is NOT $O(4)$ (Euclidean rotation) invariance. Since it is a smaller invariance it less severely constrains what operators can appear.

Is it good enough?

To answer this, let's list some operators and see if they are allowed under hypercubic symmetry.

$$(\partial_1 \Phi)^2 \qquad \text{no} \qquad \text{breaks } (x_1 \leftrightarrow x_2) \tag{20}$$

$$\sum_\mu (\partial_\mu \Phi)^2 \qquad \text{yes} \tag{21}$$

$$\sum_\mu (\partial_\mu \Phi) \qquad \text{no} \qquad \text{breaks } (x_1 \leftrightarrow -x_1) \tag{22}$$

$$\sum_\mu (\partial_\mu \Phi)^4 \qquad \text{yes} \tag{23}$$

The last of these is something impossible under $O(4)$ symmetry.

Generally,

- $O(4)$: no "hanging" Lorentz indices, Lorentz indices summed in pairs;
- Hypercubic: no "hanging" Lorentz indices, Lorentz indices summed in even numbers (that is, μ can appear 2 times, 4 times, etc.)

There are operators allowed by hypercubic but not $O(4)$ invariance, but they must have at least 4 derivatives; the first such operator is

$$\sum_\mu \Phi \partial_\mu^4 \Phi \tag{24}$$

(and others related to it by integration by parts). This is dimension 6, and so nonrenormalizable. (It is actually present at tree level in the derivative operator we constructed; even if it weren't it would get generated radiatively.)

However, it just happens that, when we list all *renormalizable* and hypercubic symmetric operators, that the set of operators we can write also display full $O(4)$ invariance. Therefore, $O(4)$ invariance is recovered in the infrared as an accidental symmetry (rather as parity is recovered as an accidental symmetry of QCD, even though it is absent in the Standard Model). This means that the lattice treatment IS "good enough;" differences between the lattice and continuum values of correlation functions will vanish as the second power of (ap), so the theories are the same in the infrared (which is the most we could have asked for).

I emphasize that renormalization will occur between lattice and continuum effective theories; in particular the mass term additively renormalizes, $m^2_{\rm eff} \sim \lambda/a^2 + m^2_{\rm bare}$. This makes

taking the continuum limit at fixed physical m^2 quite challenging. However, such mass generation is protected by symmetries from occurring additively, in QCD, so treating QCD on the lattice is arguably actually *easier* than treating scalar field theories on the lattice.

3 Gauge fields on the lattice

3.1 Problems with naive gauge field implementations

The most obvious and naive implementation of gauge fields on the lattice is that there should be 4 fields, labelled A_μ, living at each lattice point. Covariant derivatives are then filled in with the most obvious rule available:

$$\Phi^\dagger D_\mu^2 \Phi \;\rightarrow\; \sum_{x,\mu} \Phi^\dagger(x) \left\{ \frac{\Phi(x+\hat{\mu}) - 2\Phi(x) + \Phi(x-\hat{\mu})}{a^2} - iA_a^\mu T^a \frac{\Phi(x+\hat{\mu}) - \Phi(x-\hat{\mu})}{2a} + A_a^\mu A_b^\mu T^a T^b \Phi(x) \right\}, \tag{25}$$

where from here on, when I do not write a next to $\hat{\mu}$, it is assumed.

The problem with any such naive implementation is gauge invariance. Lattice implementations of gauge fields are a little tricky because a gauge symmetry is a relation between a field A^μ and the derivatives of other fields; and derivatives become finite differences on the lattice which do not necessarily obey all properties of derivatives. For instance, in the above implementation, on gauge transformation at the site $(x+\hat{\mu})$, what should change? Because the term $\Phi(x+\hat{\mu})$ appears in the above, to make the whole invariant, the field $A^\mu(x)$ must be modified. But then, the $A^\mu\Phi(x-\hat{\mu})$ and $A_a^\mu A_b^\mu \Phi(x)$ terms will change, and these don't contain $\Phi(x+\hat{\mu})$ so there is no compensating change. Therefore expressions like the one we wrote down are generally not gauge invariant.

Looking back to the last chapter, we see why we so badly want exact gauge invariance. Even if we make an action which is approximately gauge invariant for suitably slowly varying fields (the above expression can be made so up to a^2 suppressed corrections), all possible gauge non-invariant operators will in general be radiatively induced in the IR effective theory. These include such operators as

$$A_a^\mu A_a^\mu\,, \qquad A_a^\mu A_a^\mu A_b^\mu A_b^\mu\,. \tag{26}$$

The former is dimension 2, which is a disaster: the latter is $O(4)$ non-invariant, so we don't recover rotational symmetry as an "accidental" symmetry any more.

It is conceivable that one can fix this problem; write down a lattice Lagrangian with at least one coefficient per allowed dimension 2 or 4 operator in the IR effective theory, and tune the coefficients to make all undesirable operators vanish. Such tuning must in general be done in a nonperturbative way. Such things have been done (for chiral symmetry) but they are hard. Basically, for every gauge non-invariant, renormalizable operator, you must find one independent measurable whose expectation value vanishes when gauge invariance

is satisfied and does not vanish when it is not; then you have to tune coefficients to get the expectation values all to vanish. This is a formidable task.

This program is difficult, and is made even more so because one must generally break gauge invariance (fix the gauge) anyway to prevent runaway values for A^μ: the gauge kinetic term (inverse propagator) $k^2 g_{\mu\nu} - k_\mu k_\nu$ is not invertible–it has a zero eigenvalue–which means the corresponding A field eigenvector can grow unboundedly. (This is just the freedom to make A arbitrarily big by an arbitrarily strong gauge transformation.) Gauge fixing means we would have to deal with Fadeev-Popov ghosts on the lattice, and to look for BRST, not gauge, invariance.

This is a nightmare. It is much smarter to build in *exact* gauge invariance at the level of the lattice theory. That is enough to ensure that only desired operators appear; for instance the only gauge invariant dimension 4 operators of a pure glue QCD theory are F^2 and $F\tilde{F}$, and the latter can be forbidden by discrete symmetries.

3.2 Gauge field as a connection

Ken Wilson showed in 1974 how to put exact gauge invariance on the lattice. The key is to think about what gauge invariance and gauge theories mean geometrically–to understand the gauge field as a connection on the principle bundle.

Consider a theory with one scalar Φ in some irreducible representation, say the fundamental representation. That is, for an SU(N_c) theory, Φ will be a complex N_c of scalar fields. At each point, there is a space which is the space of allowed Φ values at that point. (See Fig. 1.) We will call that space (just $\Re^{2N_c}$) the *fiber* at the point x, and the manifold, consisting of the fiber over each point, as the fiber bundle over $\Re^4$–the bundle with fiber $\Re^{2N_c}$ and "base space" $\Re^4$. Also for the following discussion it is convenient to pretend that Φ is constrained to be of some length, though this is not essential.

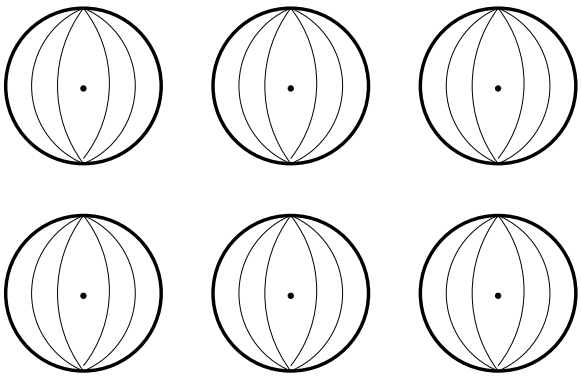


Figure 1: At each point in space (the dots) there is a "fiber" space (represented as the balls) of allowed Φ values at that point. A particular function $\Phi(x)$ means a choice of one point on each ball.

You already know about one fiber bundle, from general relativity; the tangent space at a point is a fiber, the fiber bundle made up of tying them all together is called the tangent

bundle. The gauge fields will play the same role in this fiber bundle we will study, as the Christoffel symbols play on the tangent bundle in general relativity.

How do you compare Φ on one fiber (say, at x) with Φ at another? In a theory with global symmetry, each ball pairs off with each other ball in the obvious way. If the space has "no" structure, then each ball is really its own space and there is no *a priori* way of relating one fiber to another. Even when there is global symmetry, we are free to perversely choose coordinates for one fiber which are different than others, that is, to apply a rotation of one fiber without rotating others, see Fig. 2. When we compare between different fibers (which a derivative operator has to do) we are obliged to undo this damage.

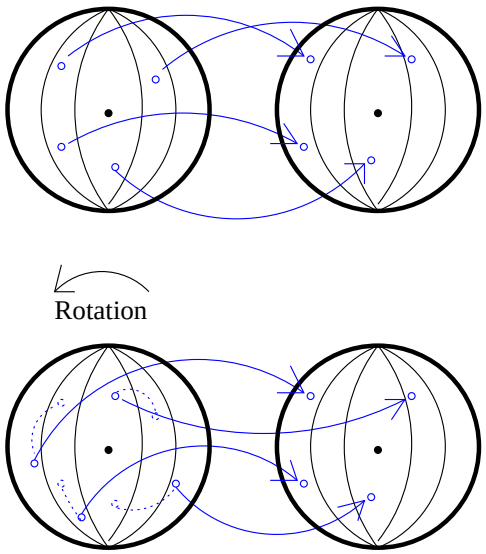


Figure 2: The connection tells how points on one fiber (values of Φ at one point) are mapped into points on another fiber. If we rotate one fiber (or simply change the coordinate basis we are using to describe it), the rotated points should still map to the same points on the neighboring fiber, which demands a change in the connection (or the description of the connection in the coordinates we have chosen).

A gauge theory has an intermediate amount of structure. The fiber bundle has a connection, which means a rule for relating fibers on infinitesimally close points of space. There is a rule to take a point on one fiber (the value of Φ at one point in space) and parallel transport it to an infinitesimally neighboring point. If some continuity conditions are satisfied by the connection and not violated by my coordinate choice, the parallel transport is generated infinitesimally. Since the fibers are vectors transforming under some representation of $\mathrm{SU}(N_\mathrm{c})$ the infinitesimal generators are the Lie algebra generators in that representation. That is, to take $\Phi(x+\delta x)$ to point x, we must multiply it by $(1+iA^a_\mu \delta x^\mu T^a)$. The gauge field's role is that it is the parallel transporter relating different fibers. It has a vector index μ which gets contracted with δx^μ, and two fiber space indices (the matrix indices (i,j) on T^a_{ij}), just as the Christoffel symbol $\Gamma_\mu{}^{\alpha\beta}$ has one vector index and two tangent space indices.

To perform longer distance comparison, we specify a path and perform a series of infinitesimal comparisons along that path. If the points $x_0, x_1, x_2, x_3, \ldots x_n$ form a series of infinitesimally separated points along a path, parallel transport of $\Phi(x_n)$ to the beginning of the path is conducted by

$$U(x_0,x_n)\Phi(x_n) = \Big(1 + iA^a_\mu(x_0)\,(x_1-x_0)^\mu T^a\Big)\Big(1 + iA^a_\mu(x_1)\,(x_2-x_1)^\mu T^a\Big)\ldots \\ \times\Big(1 + iA^a_\mu(x_{n-1})\,(x_n-x_{n-1})^\mu T^a\Big)\Phi(x_n)\,. \tag{27}$$

Here $U(x_0,x_n)$ means the operator which does the parallel transport. This product of a series of infinitesimal parallel transports, with x_0 (the path's beginning) on the left and x_n (the path's end) on the right, is the definition of the path ordered exponential,

$$U(x_0,x_n,C) = \mathrm{Pexp}\left(\int_C iA^a_\mu dx^\mu T^a\right)\,. \tag{28}$$

$U(x_0,x_n,C)$ is called the *Wilson line* from x_n to x_0 along path C. When the path C is a closed loop, we call U a *Wilson loop*.

When going around a closed path doesn't bring us back to where we started,

$$\mathrm{Pexp}\oint_C iA^a_\mu dx^\mu T^a \neq \mathbf{1}\,, \tag{29}$$

we say the connection has *curvature*. The local measure of curvature is the field strength $F^a_{\mu\nu}$. The limit of a small loop, stepping δx in the μ direction, δx in the ν direction, $-\delta x$ in the μ direction, and $-\delta x$ in the ν direction again, has a path ordered exponential of $\simeq 1 + iT^aF^a_{\mu\nu}(\delta x)^2$. This is the analog of $R_{\mu\nu}{}^{\alpha\beta}$ of general relativity, which also has two spacetime indices (μ,ν) and two fiber space (in the GR case, tangent space) indices (α,β).

The gauge connection is the thing which tells how to tie together the different fibers into a fiber bundle. The gauge *field* A^a_μ is the thing which does this when we make a particular choice for the coordinates of the fibers. A *gauge transformation* $\Lambda(x)$ is a change in our choice for coordinates on the fibers, where we rotate the (coordinates of the) fiber at x with the $\mathrm{SU}(N_\mathrm{c})$ matrix $\Lambda(x)$ (or, if we are not in the fundamental representation, with the appropriate representation matrix for $\Lambda(x) \in \mathrm{SU}(N_\mathrm{c})$). This change of coordinates induces a compensating change in A^a_μ because the fibers are still tied together in the same way even though we changed coordinates.

If you didn't get this, go back to the beginning of the subsection and read it again slowly, in the morning and with a cup of coffee. The geometrical notions here are key, not only to understanding how lattice gauge theory is going to work, but to understanding what a nonabelian gauge theory actually *is*.

3.3 Lattice implementation

The lattice implementation should now be clear. We should represent the gauge fields on the lattice with what they "really are," which is instructions for parallel transportation.

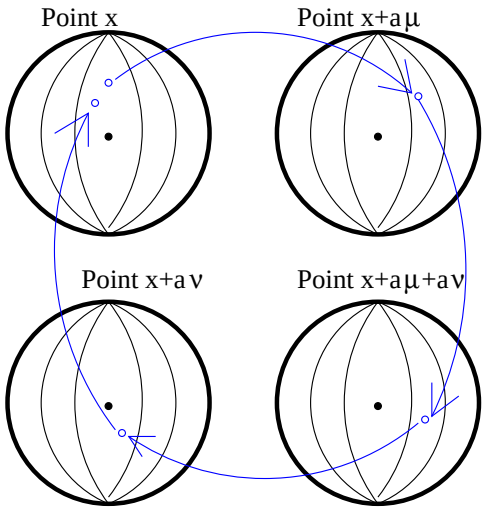


Figure 3: Curvature: when parallel transporting around a closed cycle doesn't bring you back to the same point.

Since fields Φ are only defined on the lattice points, parallel transformation properties are only required between lattice points. We make the further restriction that admissible paths are composed of "elementary links," which are the lines between neighboring lattice points. Then a choice for the connection, for a fixed coordinate choice of the Φ field space at each point, is an $\mathrm{SU}(N_\mathrm{c})$ matrix associated with each elementary link. That is, the lattice gauge field is

$$U_\mu(x) \equiv \mathrm{Pexp} \int_x^{x+a\hat{\mu}} A^a_\mu(x+dl) T^a dl = \exp(iaT^a A^a_\mu(x)) \,. \tag{30}$$

Our instinct is to think of $A_\mu(x)$ as the fundamental quantity but we should not; it is more convenient to work and think in terms of the link matrices $U_\mu(x) \in \mathrm{SU}(N_\mathrm{c})$ (or a representation of $\mathrm{SU}(N_\mathrm{c})$). The field content of lattice gauge theory with matter is illustrated in Fig. 4. Note that $U_\mu(x)$ "really" lives at $(x+\hat{\mu}a/2)$, not at x; putting it at the lattice point is a convenient but unfortunate convention.

$U_\mu(x)$ is a matrix, living on the link between x and $x+a\hat{\mu}$. It has two indices,

$$U_\mu(x)_{\alpha\beta} : \quad \begin{cases} \alpha \text{ fundamental rep. at point } x \,, \\ \beta \text{ antifundamental rep. at point } x+\hat{\mu} \end{cases} . \tag{31}$$

The parallel transport of $\Phi(x+\hat{\mu})$ to x is

$$U(x, x+\hat{\mu})\Phi(x+\hat{\mu}) = U_\mu(x)\Phi(x+\hat{\mu}) \,. \tag{32}$$

The parallel transport of $\Phi(x)$ to point $x+\hat{\mu}$ is

$$U(x+\hat{\mu}, x)\Phi(x) = U^\dagger_\mu(x)\Phi(x) \,. \tag{33}$$

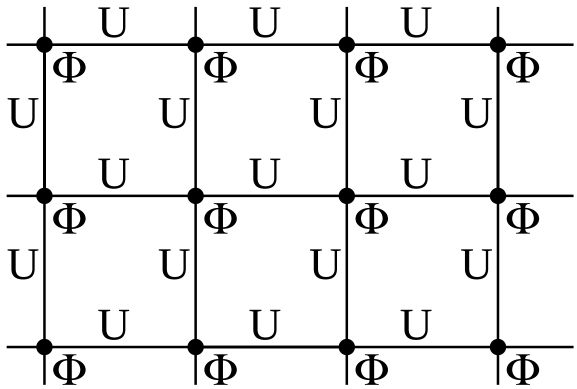


Figure 4: Lattice degrees of freedom; a Φ field at each lattice point, and a U field on each link.

The parallel transport of $\Phi(x{+}2\hat{\mu}{+}\hat{\nu})$ to x is path dependent; along the path through $x{+}\hat{\mu}{+}\hat{\nu}$ and $x{+}\hat{\mu}$, it is

$$U_C \Phi(x{+}2\hat{\mu}{+}\hat{\nu}) = U_\mu(x) U_\nu(x{+}\hat{\mu}) U_\mu(x{+}\hat{\mu}{+}\hat{\nu}) \Phi(x{+}2\hat{\mu}{+}\hat{\nu})\,. \tag{34}$$

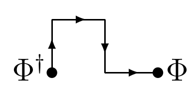
Gauge transformation is rotation of Φ at each point x by a different matrix $\Lambda(x)$. All fundamental representation objects at x are rotated by $\Lambda(x)$, all antifundamental by Λ^{-1}. The transformation rule for $U_\mu(x)$ is

$$U_\mu(x) \to \Lambda(x) U_\mu(x) \Lambda^{-1}(x{+}\hat{\mu})\,. \tag{35}$$

This makes it clear that $U_\mu(x)\Phi(x{+}\hat{\mu})$ transforms as an object in the fundamental representation at point x, as it should. Also, for small, smooth Λ, the transformation rule reduces to the expected $-\partial_\mu \Lambda + [\Lambda, A]$ rule we expect.

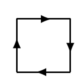
A gauge invariant object is something with no “hanging indices”; a fundamental index at point x may only be “tied off” against an antifundamental index at x, not at any other point. If fields are at different points, U matrices must be put between them to build gauge invariant objects. Examples of gauge invariant objects are

$$\Phi^\dagger(x) U_\mu(x) U_\nu(x{+}\hat{\mu}) U_\mu^{-1}(x{+}\hat{\nu}) U_\nu(x{+}\hat{\nu}) \Phi(x{+}2\hat{\nu}) \tag{36}$$



or

$$\mathrm{Tr}\, U_\mu(x) U_\nu(x{+}\hat{\mu}) U_\mu^{-1}(x{+}\hat{\nu}) U_\nu^{-1}(x)\,. \tag{37}$$



In this last example, the U form the smallest possible closed loop, which is called a plaquette; the trace means that the indices at the ends, which both live at point x, are contracted against each other. Of course, the Lagrangian and all measurables must be gauge invariant objects. The last example here is the simplest nontrivial gauge invariant object composed entirely of the U matrices and no other fields.

3.4 Integration measure

Before we construct a lattice action and interesting measurables, we must determine what the integration measure in the path integral is. Obviously we should make the replacement,

$$\mathcal{D}A_\mu \;\Rightarrow\; \prod_{x,\mu} dU_\mu(x)\,. \tag{38}$$

What does $dU_\mu(x)$ mean, though? If we think of $U_\mu(x)$ as generated by its Lie algebra elements,

$$U_\mu(x) = \exp(iaT^a A^a_\mu)\;: \qquad dU_\mu(x) \stackrel{?}{=} \prod_a dA^a_\mu(x)\,? \tag{39}$$

This is wrong; this measure is not gauge invariant. A gauge transformation, say, at x, must leave the measure $dU_\mu(x)$ (and the measure $dU_\mu(x-\hat{\mu})$) invariant. What that means is that we need a measure for the group manifold $\mathrm{SU}(N_\mathrm{c})$ which is preserved under "rotation" of the group manifold by right or left group multiplication.

It turns out that such a measure generally exists for compact, simple gauge groups, and it is unique up to a dilatation (which corresponds to an uninteresting overall multiplicative factor in the partition function). The measure is called the Haar measure. It can be generated as follows. Choose an orthonormal basis for the Lie algebra T^a. The infinitesimal volume element at the origin is $T^1 \wedge T^2 \wedge T^3 \cdots$. This means that we assign $\delta^{N_\mathrm{c}^2-1}$ measure to the box containing all points of form $g \in \{1 + i\delta \sum_a T^a e^a, ; e^a \in [0,1]\}$. Then, the measure near the point g' is that we assign the same volume to the set of points g for which $g(g')^{-1}$ lies in that box.

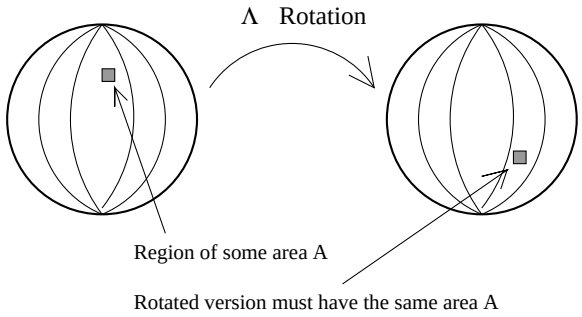


Figure 5: Requirement of the Haar measure; a patch on the gauge group, of measure (area) A, must go under rotation (induced by left multiplication by an element Λ of the group) into a patch of the same measure.

For an example, consider SU(2). For any $\vec{\theta}$ of length 2π, $\exp(i\vec{\theta}\cdot\vec{T})$ is the $-\mathbf{1}$ element of the group. (Recall that $T^a = \tau^a/2$ with τ^a the Pauli matrices; $\exp i\pi\tau^a = -\mathbf{1}$ for each a.) Obviously, then, the measure for such large $\vec{\theta}$ should approach zero. The correct measure turns out to be

$$dU_\mu(x) = \prod_{a=1,2,3} J(A^2) dA^a_\mu(x)\,, \qquad J(A^2) = \frac{1-\cos A^2/4}{A^2/4}\,, \tag{40}$$

up to uninteresting multiplicative rescaling.

The integration $\prod dU_\mu(x)$ obviously overcounts physical configurations; in fact it integrates uniformly over all possible gauge orbits. There is no problem with doing this, though; because the $U_\mu(x)$ and the $\Lambda(x)$ reside on compact spaces, this amounts to a finite overall multiplicative factor in the partition function, which is of no physical consequence. This is the lattice's answer to how to fix the gauge: Don't fix the gauge. Integrate over all possible gauge field configurations, including redundantly integrating over a configuration and its gauge copies. With the lattice implementation this space of redundant integrations, $\Pi_x \Lambda(x)$, is compact and of finite volume, so no harm is done.

Building a lattice gauge theory with these components and this integration measure, and using a gauge invariant action, automatically ensures *exact* gauge invariance of the IR effective theory, which is strong enough to ensure that only desired dimension 4 operators can appear.

4 Lattice action and perturbation theory

4.1 The Wilson action

The simplest possible gauge invariant object made out of U matrices is the product of four U matrices around a square:

$$\mathrm{Tr}\, \Box_{\mu\nu}(x) \equiv \mathrm{Tr}\, U_\nu(x) U_\mu(x+\hat{\nu}) U_\nu^\dagger(x+\mu) U_\mu^\dagger(x)\,, \tag{41}$$

which is usually called the "elementary plaquette" or "plaquette" in the μ, ν direction at point x (again, it really lives at $x+(\hat{\mu}+\hat{\nu})/2$).

How should we think about this object? From the last chapter, you should expect that the product of link matrices is proportional to the field strength:

$$\Box_{\mu\nu}(x) \simeq 1 - ia^2 T^a F^a_{\mu\nu}(x)\,. \tag{42}$$

We can see that this is correct by expanding order by order in weak and slowly varying fields. At leading order each U matrix is just 1, giving the 1 term. At next order, $U \simeq 1 + iaA$ and $U^\dagger = 1 - iaA$, so

$$\begin{aligned} \Box_{\mu\nu} &\simeq 1 + ia\Big(A_\nu(x+\hat{\nu}/2) + A_\mu(x+\hat{\nu}+\hat{\mu}/2) - A_\nu(x+\hat{\nu}/2+\mu) - A_\mu(x+\hat{\mu}/2)\Big) \\ &\simeq 1 + ia^2(\partial_\nu A_\mu - \partial_\mu A_\nu)\,. \end{aligned} \tag{43}$$

If we label A fields as living at the middles of links and $\Box_{\mu\nu}$ as living at the center of the plaquette, the first corrections to this expression are order a^4, that is, are cubic in derivatives.

Since the last term cancelled at $O(a)$, we have to go one order higher, where there are 6 contributions from pairs of A fields and also contributions from A^2 arising from a single

link;

$$\begin{aligned}\Box_{\mu\nu} &\simeq (\text{above}) - a^2\left(A_\mu^2 + A_\nu^2 + A_\nu A_\mu - A_\nu A_\nu - A_\nu A_\mu - A_\mu A_\nu - A_\mu A_\mu + A_\nu A_\mu)\right) \\ &= (\text{above}) - a^2\Big[A_\nu\,,\,A_\mu\Big]\,, \end{aligned} \tag{44}$$

which fills out the field strength. Therefore,

$$\Box_{\mu\nu} \simeq -ia^2 F^a_{\mu\nu} T^a\,. \tag{45}$$

We also know that $\Box_{\mu\nu}$ is unitary; therefore the $-ia^2F$ term ensures that there is also a $-a^4F^2$ term,

$$\Box_{\mu\nu} \simeq 1 - ia^2 F^a_{\mu\nu} T^a - \frac{a^4}{2} F^a_{\mu\nu} F^b_{\mu\nu} T^a T^b + \dots\,, \tag{46}$$

and so its trace contains $-a^4F^2/4$. We can get the standard gauge action by summing over the traces of plaquettes,

$$\sum_{x,\mu>\nu} \left[N_{\rm c} - \mathrm{Tr}\,\Box_{\mu\nu}(x)\right] \;\simeq\; \int d^4x \frac{1}{8} F^a_{\mu\nu} F^a_{\mu\nu}\,. \tag{47}$$

Lattice people always write the action with prefactor $\beta/N_{\rm c}$, β plays the part of the inverse gauge coupling. We see that at tree level,

$$S = \frac{\beta}{N_{\rm c}} \sum_{x,\mu>\nu} \left(N_{\rm c} - \mathrm{Tr}\,\Box_{\mu\nu}(x)\right) \quad \Rightarrow \quad \beta = \frac{2N_{\rm c}}{g_0^2}\,. \tag{48}$$

This action is called the Wilson action and is the simplest available. It gets corrections to F^2 at dimension 6, that is, $F_{\mu\nu}D_\mu^2 F_{\mu\nu}$; this correction is also not $O(4)$ invariant. One can make more sophisticated actions, such as one involving 1×2 boxes rather than just squares;

$$S_{\rm Symanzik}: \;\; \sum_{\mu>\nu} \mathrm{Tr}\,\Box_{\mu\nu} \to \frac{5}{3}\sum_{\mu>\nu} \mathrm{Tr}\,\Box_{\mu\nu} - \frac{1}{12}\sum_{\mu>\nu} (\mathrm{Tr}\;\square_{\mu\nu} + \mathrm{Tr}\;\square_{\nu\mu})\;. \tag{49}$$

This action has no tree level dimension 6 operators and gives improved rotational invariance convergence.

4.2 Perturbation theory

The lattice is a nonperturbative, well defined regulation. (In fact, it is the ONLY nonperturbatively defined regulation scheme I know.) We can study perturbation theory in this regulation scheme, both for its own sake and as a way to do matching calculations between the lattice and the continuum. This is especially important, for instance, when one wants to consider many or complicated operators such as electroweak operators needed in some weak matrix element problems.

What are the fields and Feynman rules? I will just outline the procedure for getting them, not give great detail. H. Rothe's book contains quite a complete and readable account.

First we must choose a definition for the gauge field, based on the link. We want $U \simeq 1 + iaT^a A^a_\mu$, but this is not a unitary matrix; how do we deal with the higher order in A piece? Two natural choices would be,

$$\begin{aligned} U_\mu &= \exp(iaT^a A^a_\mu)\,, \\ \text{or} \qquad U_\mu - U^*_\mu &= 2iaT^a A^a_\mu\,. \end{aligned} \tag{50}$$

The latter is more convenient for going from lattice U matrices to A fields. The former is more convenient for perturbation theory and people usually use it.

As usual you have to fix the gauge. This will introduce ghosts, which are not TOO different from normal, and I will not discuss them further. The interesting things are that

1. inverse propagators and vertices have trigonometric momentum dependence, and
2. there are extra, unexpected vertices.

These two are related to each other and are inevitable if one does gauge theory on a lattice.

Consider the scalar action

$$\frac{1}{2}(D_\mu\Phi)^2 \to \frac{a^2}{2}\sum_{x,\mu}\left(2\Phi^\dagger(x)\Phi(x) - \Phi^\dagger(x)U_\mu(x)\Phi(x+\hat\mu) - \Phi^\dagger(x+\hat\mu)U^\dagger_\mu(x)\Phi(x)\right)\,. \tag{51}$$

First expand it to zero order in A to find the inverse propagator. Go to momentum space by replacing

$$\Phi(x) = \int_{\rm BZ}\frac{d^3p}{(2\pi)^3}e^{ip\cdot x}\Phi(p)\,, \tag{52}$$

where $\int_{\rm BZ}$ means the integral only extends over the Brillouin zone, $p_1 \in (-\pi/a, \pi/a]$ and similarly for the other 3 components. Note that the zone is a hypercube, another reflection of the breaking of $O(4)$ invariance. Substituting and doing the sum gives an inverse propagator of

$$\Delta^{-1}(p) = \sum_\mu \frac{2}{a^2}(1-\cos ap_\mu) = \sum_\mu \frac{4}{a^2}\sin^2\frac{ap_\mu}{2} \equiv \tilde p^2\,. \tag{53}$$

The low ap limit of the propagator is p^2 but there are a^2p^4 corrections; for hard momenta the propagator is significantly different, as it must be because it is smooth and periodic. Non-smooth behavior can only be achieved by nonlocal actions, and would carry other costs.

The linear in A term, that is, the three point vertex, is determined by expanding U to linear order in A. Write A as living at the middle of the link, as is natural; the term in the action is

$$\frac{a^2}{2}\sum_{x,\mu} iaA^a_\mu(x+\hat\mu/2)\left(-\Phi^\dagger(x)T^a\Phi(x+\hat\mu) + \Phi^\dagger(x+\hat\mu)T^a\Phi(x)\right) \tag{54}$$

which, denoting the Φ momentum as k and the $\Phi^\dagger$ momentum as p (both incoming to the vertex), gives a Feynman rule

$$T^a\frac{2}{a}\sin\frac{(k-p)_\mu a}{2}\,. \tag{55}$$

This differs from the continuum vertex by the substitution of $(2/a)\sin[(k-p)_\mu a/2]$ for $(k-p)_\mu$. These are equivalent at small momentum and differ at $O(a^2)$. The quadratic in a term is also peculiar; since $\Phi^\dagger$ and Φ are evaluated one site off from each other, it behaves as

$$\{T^a\,,\,T^b\}g_{\mu\nu}\cos\frac{(k-p)_\mu a}{2} \tag{56}$$

whereas usually the cosine term would be 1.

One may continue with the expansion; at A^3 order there is a vertex with no continuum analog, which however has an explicit a^2 suppression. It is only needed in relatively high loop calculations; in fact there are infinitely many vertices but as they have ever more lines, they are only needed at ever higher orders in the loop expansion. The forms of all these terms can actually be determined completely by looking at the inverse propagator and insisting that the derivatives there be covariant derivatives.

The gauge field action can be expanded similarly, though it is more complicated. The inverse propagator is

$$G^{-1}_{\mu\nu} = g_{\mu\nu}\tilde{k}^2 - \tilde{k}_\mu\tilde{k}_\nu\,, \qquad \tilde{k}_\mu \equiv \frac{2}{a}\sin\frac{k_\mu a}{2}\,. \tag{57}$$

The cubic interaction term looks like its normal form but with the substitutions

$$(k_1{-}k_2)_\mu\delta_{\nu\lambda} \to (\widetilde{k_1{-}k_2})_\mu\cos\frac{(k_3)\nu a}{2}\delta_{\nu\lambda} \quad \text{etc.} \tag{58}$$

The 4 gluon piece looks horrible. Besides the type of complication already encountered, multiplying the expected piece $\Gamma^{ABCD} \propto f^{ABE}f^{CDE}(\ldots)+$ permutations, it has a completely unexpected *symmetric* in external indices piece, of form

$$\begin{aligned}\Gamma^{ABCD}_{\mu\nu\lambda\rho}(p,q,r,s) \;=\; & (\text{Antisymmetric piece}) + \\ & \frac{g^2}{12}\left\{\frac{2}{3}\delta_{AB}\delta_{CD} + d_{ABE}d_{CDE} + (AC,BD) + (AD,BC)\right\} \\ & \times(\tilde{p}\,\tilde{q}\,\tilde{r}\,\tilde{s}) \text{ with various Lorentz structure}\,.\end{aligned} \tag{59}$$

This extra piece comes about because $\Box \simeq \exp(-ia^2F_{\mu\nu})$; at a^8 level there is an F^4 term with the T indices totally symmetric as in the above, which looks like four A fields and four derivatives. This term is totally irrelevant as an IR effective operator, but in the ultraviolet it is very large.

4.3 Tadpole improvement

Such large but very UV interactions lead to significant "unexpected" renormalizations. This means that, for instance, the matching between the lattice coupling and the MOM scheme coupling doesn't give $\mu_{\rm MOM} \simeq 1/a$ as expected, but

$$g_0^2 = g^2_{\rm MOM}(\mu = 83.5/a) \text{ at one loop}\,. \tag{60}$$

In other words, one must put in a much larger β value than naively expected (make the gauge action "stiffer" than expected) in order to get a given physical coupling constant. This is dominantly due to large renormalizations from "tadpoles" (simple closed loops originating from a single vertex), mostly from the above, symmetrical piece of the vertex (and its higher point brothers).

These large terms are common (approximately proportional) in essentially all operators, and can be easily, approximately, cancelled by "tadpole improvement" as advocated by Lepage and Mackenzie. The idea is that the corrections are due to very UV corrections which make each link U look "shorter" than it is; averaging over UV fields on a compact manifold like $\mathrm{SU}(N_\mathrm{c})$ effectively averages with some weight over a patch of the manifold around the "infrared" value of the link. Think of $\mathrm{SU}(N_\mathrm{c})$ as an n-sphere (SU(2) actually is the 3-sphere); such averaging gives you a point on the interior of the sphere, that is, it effectively makes U "shorter". To compensate, measure the mean value of $\mathrm{Tr}\,\Box$, and multiply each link by $(\mathrm{Tr}\,\mathbf{1}/\mathrm{Tr}\,\Box)^{1/4}$ (or just multiply terms in the action by this quantity to the power of the number of links involved). This crude prescription comes startlingly close to undoing the large renormalizations and is now widely used under the rubric of "tadpole improvement."

5 Confinement

A characteristic and puzzling feature of QCD is that it is a confining theory, meaning, loosely, that quarks and gluons do not appear as physical particles but are always bound into colorless, integer charged hadrons. Since lattice QCD can compute the spectrum of the theory, we should be able to look for an explanation of this phenomenon on the lattice.

What does "confinement" mean? Naively, it means that we do not see free quarks and gluons as particles in QCD, but rather see composite objects with color combinations which add up to be singlets. Exactly how to say this more formally, and how we explore the issue in lattice QCD, is the purpose of this section. We will address some interesting issues, but arguably we won't give a very satisfying answer to the question.

5.1 Vanishing of $\langle\mathcal{O}\rangle$ if $\mathcal{O}$ is gauge variant

A "naive" definition of confinement is, that the interpolating fields for all particles are gauge invariant operators. Therefore, the particles are all colorless. Alternately, we might think that confinement means that all expectation values of gauge dependent operators vanish.

I will argue that this is not a good definition of confinement, since it

1. is automatically true, and

2. is also true of QED, which we agree is *not* a confining theory.

To see this, let me first be clear what I intend to show. Suppose an operator or product of operators, $\mathcal{O}$, transforms in some non-singlet, irreducible representation under gauge

transformations at the point x. That is,

$$\mathcal{O}(\Lambda U) = \mathcal{U}(\Lambda(x))\mathcal{O}(U) \neq \mathcal{O}(U)\,. \tag{61}$$

Some operators may transform reducibly; in this case any singlet part of them is gauge invariant, and the non-singlet part is gauge variant. Here $\mathcal{U}(\Lambda)$ is the action in the relevant representation of the gauge rotation Λ.

In this case, the claim is that the integral over all possible gauge connections weighted by the action, that is, $\langle\mathcal{O}\rangle$, automatically vanishes. This is just because the action is invariant under the substitution

$$U_\mu(x) \to \Lambda U_\mu(x)\,, \quad U_\mu(x-\hat{\mu}) \to U_\mu(x-\hat{\mu})\Lambda^{-1}\,, \tag{62}$$

the action is unchanged, and so is the integration measure,

$$S(U) = S(\Lambda U) \quad \text{and} \quad \prod_{x,\mu} dU_\mu(x) = \prod_{x,\mu} d(\Lambda U_\mu(x))\,. \tag{63}$$

Therefore, nothing stops us from changing coordinates to isolate $d\Lambda(x)$ as a separate integration in the measure;

$$\prod_{x,\mu} dU_\mu(x) e^{-S(U)}\mathcal{O}(U) = \frac{1}{\int d\Lambda(x)} \int d\Lambda(x) \prod_{x,\mu} d(\Lambda U_\mu(x)) e^{-S(\Lambda U)}\mathcal{O}(\Lambda U)\,; \tag{64}$$

the action and the measure are independent of Λ, so its integration can be done first;

$$\int d\Lambda(x) \prod_{x,\mu} d(\Lambda U_\mu(x)) e^{-S(\Lambda U)}\mathcal{O}(\Lambda U) = \prod_{x,\mu} dU_\mu(x) e^{-S(U)} \left(\int d\Lambda(x)\mathcal{O}(\Lambda U) \right)\,. \tag{65}$$

Now we need a little fact about compact, simple Lie groups; the average over the gauge group of any quantity transforming in a non-singlet irreducible representation is zero. (This is just orthogonality of the characters of the group.) In other words, for $\mathcal{U}(\Lambda)$ any non-singlet irreducible representation of Λ,

$$\int d\Lambda \mathcal{U}(\Lambda) = 0\,. \tag{66}$$

Since $\mathcal{O} \to \mathcal{U}(\Lambda)\mathcal{O}$ under gauge transformation, the average of this gauge transformed quantity vanishes.

Therefore the expectation value of any gauge variant operator, simply vanishes. The same is true for any combination of gauge variant (covariant rather than invariant) operators unless their indices are tied off to make a gauge invariant product. Therefore, "almost" the only questions which can be asked are questions about gauge invariant operators. The "almost" is, that one may also ask about gauge -fixed- operators, which I will discuss briefly at the end of this section.

The point is that the supposed criterion of confinement–that all interpolating operators and therefore all states of the theory are gauge invariant–is trivially true. It is also true of

electromagnetism, where it means that an electron creation operator is not a valid operator unless it is attached to a Wilson line which either goes to infinity (in infinite noncompact spaces) or to a positron creation operator (in compact spaces without boundary). (The problem of maintaining gauge invariance in a theory with boundary is one I don't want to address.) This corresponds with the fact that the electrical flux from an electron has to end somewhere–either at infinity or at a positron. However, we still say that QED is not confining, because the energy cost of putting the positron arbitrarily far away, becomes independent of separation at large values of the separation.

5.2 Area law behavior of Wilson loops

This suggests a second definition of confinement; that the energy cost of separating a quark from an antiquark (or other neutralizing combination, such as two more quarks) rises without limit as the separation is made large, rather than saturating as it does for the Coulomb potential of QED.

This is a nontrivial criterion. How do we probe whether it is satisfied? We want to know the potential as a function of separation, $V(r)$, for a fundamental and anti-fundamental test charge. The energy cost is converted into an action cost by integrating it over Euclidean time;

$$\int H d\tau = S_{\text{Euclid}} \,. \tag{67}$$

Therefore we can find the energy cost by

1. Creating a quark and anti-quark test particle at some point x,
2. Separating them a distance r,
3. Holding them separate for a Euclidean time τ (or propagating them forward by Euclidean time τ), where τ is taken large compared to r,
4. Measuring the action due to the interaction with the field, S, and differentiating it, $dS(r,\tau)/d\tau = V(r)$.

The way the fields interact with the test charges are through the Wilson lines parallel transporting their charges; the action S is the log of the expectation value of the trace of the Wilson loop the test quark and anti-quark form, see Fig. 6.

The potential may have an ill-defined zero point, but the derivative with respect to separation,

$$\frac{dV}{dr} \equiv \frac{d}{dr}\left(\frac{d}{d\tau}\log\langle \text{Tr}\, U_{C(r,\tau)}\rangle\right) \,, \tag{68}$$

is well defined. For the Coulomb potential, it goes as $e^2/(4\pi r^2)$. A linearly confining potential gives a constant value, that is,

$$\text{Linear confinement}: \; S \propto \tau r + O(r^0) + O(\tau^0) \,. \tag{69}$$

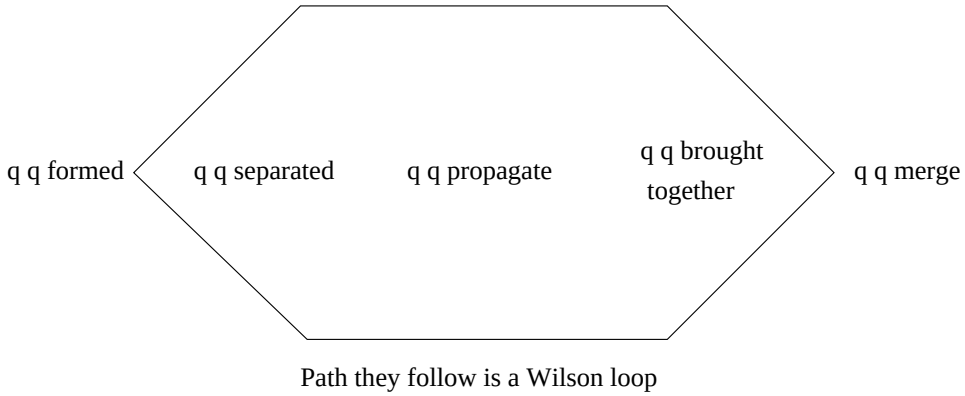


Figure 6: Creating, moving, and annihilating a $q\bar{q}$ pair; the motion of their charges along the path they follow, is a Wilson loop.

That is, linear confinement means the log of the trace of a Wilson loop scales with the area of the loop. There is a

> **Theorem** (Yaffe and Simon): The slowest possible growth for $\ln\langle \mathrm{Tr}\, U_C\rangle$ as C is uniformly expanded is with the perimeter of C; the fastest possible growth is with the area of C.

Therefore the strongest confining potential possible is a linear potential.

5.3 Wilson loop in QED

To see the relation between the trace of a Wilson loop and the potential between charges, consider QED below the electron mass. In this case the Wilson loop is

$$U_C = \exp\ i\int_C A\cdot dl\,. \tag{70}$$

Further, the theory is Gaussian; $\mathrm{Tr}\, U_C$ is determined by the two point function,

$$\langle \mathrm{Tr}\, U_C\rangle = \exp\left(-\frac{1}{2}\int_C dl_1^\mu \int_C dl_2^\nu \langle A_\mu(l_1)A_\nu(l_2)\right)\,. \tag{71}$$

The only contributions which grow with τ are from connecting the long straight segments of C:

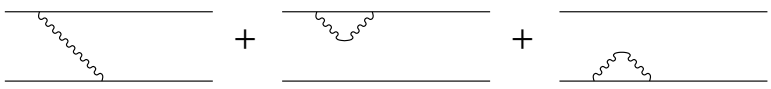


Figure 7: Relevant propagators for the Wilson loop

Only the propagator which connects the different sides of the Wilson loop (the first in the figure) depends on r, the other two contribute to the self-energy. Evaluating this diagram gives

$$(V(r) - \text{constant})\tau = \int dl_1 dl_2 G_{\tau\tau}(l_1 - l_2, \tau)\,, \tag{72}$$

with $G_{\tau\tau}$ the gauge field propagator in real space. Feynman gauge is convenient, for a massless field the propagator is

$$G_{\mu\nu}(x-y) = \frac{e^2 g_{\mu\nu}}{4\pi^2 |x-y|^2}\,. \tag{73}$$

This gives

$$(V(r) - \text{constant})\tau = \frac{e^2\tau}{4\pi^2}\int d\Delta l \frac{1}{(\Delta l)^2 + r^2} = \frac{e^2}{4\pi r}\,, \tag{74}$$

which is nothing but the Coulomb potential.

The same-line contributions give

$$\text{constant}\ \tau = \frac{e^2\tau}{4\pi^2}\int d\Delta l \frac{1}{(\Delta l)^2}\,, \tag{75}$$

which is linearly UV divergent. This is the self-energy of having a perfectly localized charge in QED. On the lattice it is of course cut off, the integral is best done in momentum space and gives

$$a\sum_{\Delta l}\int_{\text{BZ}} \frac{d^4k}{(2\pi)^4}\frac{e^{ik\cdot\Delta l}}{\tilde{k}^2}\,. \tag{76}$$

The sum gives $2\pi\delta(k_\tau)$, leaving

$$\text{constant} = e^2\int_{\text{BZ}}\frac{d^3k}{\tilde{k}^2} = \frac{e^2\Sigma}{4\pi a}\,, \qquad \Sigma_{\text{Wilson}} = 3.17591135\ldots\,. \tag{77}$$

Here I show the value of the lattice integral for the Wilson choice of the lattice action.

The point of the above is that the trace of a Wilson loop does indeed give the Coulombic attraction between the opposite sides, but there is a self-interaction which goes as the perimeter of the Wilson loop and is linearly UV divergent, cut off by the lattice spacing if evaluated on the lattice. (This corresponds to a severe renormalization sensitivity of the Wilson line as an object in field theory.)

5.4 Wilson loops in lattice QCD

Lattice practitioners have successfully observed area law growth of $\ln \text{Tr}\, U_C$ for large Wilson loops in pure glue QCD, showing that, in the absence of fermions in the theory, external test quarks attract each other with a linear potential. The value of the potential is of order 800MeV/fermi.

The calculations are challenging because of the perimeter contributions mentioned above. For any given gauge field configuration, the trace of the Wilson loop is order 1 (of either sign); the expectation value being exponentially small arises from cancellations between configurations which are severe because of the large perimeter contribution. Elaborate techniques have been developed to "build in" as much of this cancellation as possible into the measurement. This is done either by performing the averaging in the path integral independently over disconnected regions, or by averaging over a large set of Wilson loops which are "almost the same" as the basic Wilson loop of interest (or equivalently, using "fattened links"). These are basically numerical games so I will not address them here.

We don't expect area law to remain valid when dynamical quarks exist in the theory. Suppose we separate two test charges by several fermis; then plenty of energy is available to create a quark-antiquark pair. The quark pairs off with the test antiquark and the antiquark with the test quark. There are then two color-neutral mesons, which should have no long range attractive force. The potential should then "turn over" at some value and become flat, see Fig. 8.

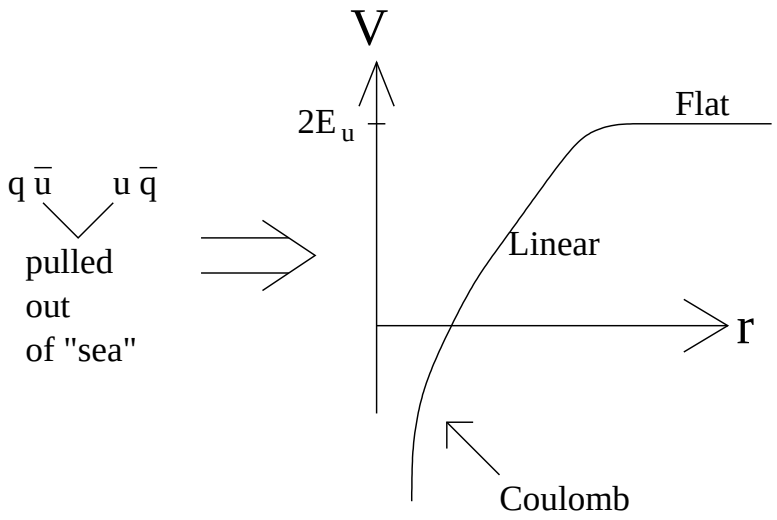


Figure 8: String breaking.

This behavior, dubbed "string breaking," has recently been observed, though mainly through measurables other than the Wilson loop; it is close to the forefront of current research in numerical lattice gauge theory.

This leaves an unsettling feeling. Area law behavior in Wilson loops, or a linearly rising potential, must *not* be a good definition of confinement. After all, they are true in pure glue QCD, which is undoubtedly confining, but they are not true in QCD with fermions and are not expected to be. Quarks form into colorless mesons because the potential is linearly rising until it is so high that a pair gets created to neutralize the color charges; but technically, the potential between two test charges then flattens out. "Full" QCD Wilson loops show perimeter behavior and a potential which becomes flat at large distances–in fact, it is flatter than in QED.

Perhaps the big difference with respect to QED is that QED has a massless degree of freedom, the photon, while the spectrum of QCD displays a mass gap. It is the absence of a mass gap which lets us form an interesting extended object like an electron with widely separated positron in QED. Therefore many people have settled on the existence of a mass gap as the definition of confinement in QCD. The mass gap is certainly observed on the lattice.

5.5 Gauge fixing and gauge fixed operators

Finally I return to what I meant by gauge invariant operators being "almost" the only ones allowed.

What is a gauge fixing prescription? Naively, it is a selection of a subset of the possible gauge fields to integrate over, so only physically distinct gauge field configurations are considered. However it is more useful to think of it as the following. A gauge fixing prescription is a way, given any gauge field U, to assign it to a "gauge fixed copy" $U_{\rm gf}(U)$, such that

1. There exists a gauge transformation $\Lambda(x)$ which carries U to $U_{\rm gf}(U)$, $U = \Lambda U_{\rm gf}(U)\Lambda^{-1}$, and

2. Any gauge copy of U is assigned to the same gauge fixed copy; that is, if

$$\text{if } U \overset{\text{gauge fixing}}{\longrightarrow} U_{\rm gf}(U) \text{ then } \Lambda U\Lambda^{-1} \overset{\text{gaugefixing}}{\longrightarrow} U_{\rm gf}(U) \text{ as well}. \tag{78}$$

Here $\Lambda U\Lambda^{-1}$ means the gauge transformation of U, namely, $U_\mu(x) \to \Lambda(x)U_\mu(x)\Lambda^{-1}(x+\hat{\mu})$.

In other words, gauge fixing is a rule for projecting the space of all gauge fields onto a subset which has one element for each physically distinct configuration, with the projection rule associating a gauge field with the physically equivalent choice.

If we can come up with a unique and well defined gauge fixing procedure, then we can consider gauge fixed operators. A gauge fixed operator is an operator on the *full* set of gauge field configurations, which takes on the value that a gauge variant operator takes on at the gauge fixed copy configuration; that is,

$$\mathcal{O}_{\text{gauge fixed}}(U) = \mathcal{O}_{\text{gauge dependent}}(U_{\rm gf}(U)). \tag{79}$$

Rather than integrating over only the gauge fixed space, we can integrate over all gauge fields, but evaluate the operator on the gauge fixed copy rather than the actual gauge field U. This is equivalent to the usual gauge fixing procedure, but it is more convenient since it is totally obvious that, if we choose to consider gauge invariant observables instead, we have not damaged the theory. Alternately we may view gauge fixed observables as just "very strange" gauge invariant observables. We now study just how strange they are.

How hard is it to specify a unique and consistent gauge fixing procedure? Perturbatively it is easy, but on nonperturbatively large fields it is hard. Consider for instance Landau

gauge. (Coulomb gauge turns out to be similar but even worse. Axial gauges deserve a special discussion, which I will not give.) The usual Landau gauge prescription is

$$\partial_\mu A^\mu = 0 \qquad \text{(Landau gauge)}\,. \tag{80}$$

This is equivalent to

$$\int d^4x \frac{1}{2} A^a_\mu A^\mu_a \quad \text{extremized with respect to all } \Lambda\,. \tag{81}$$

That is, Landau gauge is the choice among all gauge copies of $A_\mu(x)$, of the gauge copy which extremizes the integral $\int A^2$. To see this, vary with respect to infinitesimal $\Lambda(x)$;

$$\delta_\Lambda \int \frac{1}{2} A^a_\mu A^\mu_a = \int A^a_\mu \delta_\Lambda A^\mu_a = \int A^a_\mu \left(\partial^\mu \Lambda_a + i f_{abc} \Lambda^b A^c\right)\,. \tag{82}$$

The f_{abc} term vanishes because $A^a_\mu A^\mu_c$ is symmetric. The other term should be integrated by parts to give (ignoring subtleties of boundary terms, which are absent in compact Euclidean space and are related in noncompact Minkowski space to harmonic gauge freedoms)

$$\delta_\Lambda \int \frac{1}{2} A^a_\mu A^\mu_a = \int \Lambda^a \partial_\mu A^\mu_a\,. \tag{83}$$

Demanding an extremum requires $\partial_\mu A^\mu_a = 0$.

Perturbatively, an extremum is unique. Nonperturbatively, Gribov showed that this is not in general so. Stick to minima. The point is that the functional

$$\int d^4x \frac{1}{2} A^a_\mu A^\mu_a \quad \text{(or, on the lattice, } \sum_{x,\mu} -\mathrm{Tr}\, U_\mu(x) \text{)}\,, \tag{84}$$

can have multiple local minima. Landau gauge is not unique. In fact, on the lattice one generically finds that this functional has many local minima.

To fix the gauge completely (up to a global gauge freedom which can be fixed in some other way), one must choose, for instance, the prescription,

> Pick the *global* minimum over Λ of $\int A^2_\Lambda$.

This prescription for defining Landau gauge is completely unique (off a set which is measure zero and codimension 1). We see that it is also equivalent to, "make the gauge field as small as possible," or "choose the gauge which makes the link variables collectively as close to the identity as is possible." This gauge choice has however the following problems:

1. The gauge choice is not in general singularity free in the continuum. On the lattice, although as an average statement the links are close to the identity, there will be locations where a few links are very far from the identity. As the lattice is refined, these locations shrink in size but the corresponding continuum field strength diverges. For instance, in the continuum, in the field of an isolated instanton, there is a gauge field singularity at the center of the instanton.

2. The gauge fixing condition is not local. Changing one link $U_\mu(y)$ at a point y, changes the value of $\int A^2$ and requires a change of the gauge at all points x.

3. The gauge fixing procedure is not smooth on the space $\prod dU_\mu(x)$ of gauge field configurations. That is, a tiny change to the gauge fields (links) can make one local minimum become deeper than another, making a finite change to the gauge fixed copy. In other words, on the space $\prod dU_\mu(x)$ of gauge fields, the gauge fixing function has (codimension 1) surfaces of discontinuity.

4. It is difficult algorithmically to find the global minimum. Local minima are easy. (Much work has gone into evading this problem; most of the progress amounts to choosing a slightly different criterion than the global minimum of $\int A^2$. I believe these alternatives still suffer from the other problems listed–though for some cases the non-smoothness may only be strictly true in the thermodynamic limit.)

The most fatal problems are the nonlocality and non-smoothness, which are related. It is difficult to interpret gauge fixed correlation functions because the operators in question are not local. In particular it is not obvious that cluster decomposition or reflection positivity will hold. The discontinuous nature means it is not obvious that gauge fixed correlation functions are analytic; if they are not, then continuation to Minkowski time is impossible.

Gauge fixing directly in Minkowski time is hopeless because the functional $\int A^a_\mu A^\mu_a = \int (A^a_i A^i_a - A^0_a A^0_a)$ is not bounded above or below in Minkowski space; all extrema are saddlepoints and there is no unique prescription for isolating one. One may evade this problem only at the expense of choosing a Lorentz non-invariant gauge fixing procedure, which is also not very palatable.

I will not discuss the problems of axial gauges. Suffice it to say that they are even worse–auxiliary gauge fixing conditions are required, translation invariance is broken, and the axial condition must be violated on some surface if the space is compact in the direction in which axial gauge is enforced. If that dimension is noncompact, the gauge becomes arbitrarily ill behaved as one moves away from the surface where auxiliary gauge fixing is performed. One may also define "flow" gauges which are a hybrid of axial and Landau; but these break T invariance, which may sacrifice reflection positivity. I will not attempt a fair exposition of these topics.

6 Fermions I

6.1 Fermions in continuum field theory

In QCD, the fermionic part of the path integral is supposed to be

$$\begin{aligned} Z_{\text{fermi}} &= \int \mathcal{D}\bar{\phi}_m \mathcal{D}\phi_m \exp(-S(\bar{\phi},\phi,A)) \,, \\ S(\bar{\phi},\phi,A) &= M_m \bar{\phi}_m \phi_m + \bar{\phi}_m D_\mu \gamma^\mu \phi_m \,, \end{aligned} \tag{85}$$

where the index runs over the flavors of quarks appearing in QCD. Here the fields $\bar{\phi}$, ϕ are Grassmanian 4-component spinors, that is, at each point there are four Grassman variables associated with $\bar{\phi}_m$ and four associated with ϕ_m. Different quarks have different masses; realistically we can probably do without inclusion of the c, b, and t quarks (integrate them out). Generally, on the lattice people are happy enough with the approximation $M_u = M_d$, since error bars are always larger than any effect of this splitting (in a similar spirit one ignores electromagnetic effects).

If we were talking about the Standard Model (electroweak interactions) the difference would be that the integration is taken over Weyl (2-component) fermions and the mass term is absent (the γ matrices are replaced with σ^μ, $\sigma^0 = \mathbf{1}$); alternately the integration is over Majorana fermions, that is over Dirac fermions but with a delta function enforcing the Majorana condition.

There is a potential problem we see right away, which is that there is a dimension 3 operator $\bar{\psi}\psi$. When the regularization (lattice) is at some very UV point, what prevents a large $\propto \Lambda \sim 1/a$ coefficient for this operator from being induced? This isn't a problem, because the mass term breaks a symmetry of the Lagrangian, namely chiral symmetry:

$$\begin{aligned} \psi_m &\Rightarrow (L_{mn}P_L + R_{mn}P_R)\,\psi_n\,, \\ \bar{\psi}_m &\Rightarrow \bar{\psi}_n \left(L^\dagger_{nm}P_R + R^\dagger_{nm}P_L\right)\psi_n\,, \end{aligned} \tag{86}$$

with L, R independent $\mathrm{U}(N_\mathrm{f})$ matrices, $N_\mathrm{f}$the number of fermionic flavors, an $P_{L,R}$ the usual left and right handed projection operators. Because $\gamma^\mu P_{L,R} = P_{R,L}\gamma^\mu$, the kinetic term is invariant under the transformation;

$$\begin{aligned} \bar{\psi}\gamma^\mu\psi &\Rightarrow \bar{\psi}\left(L^\dagger P_R + R^\dagger P_L\right)\gamma^\mu\left(LP_L + RP_R\right)\psi \\ &= \bar{\psi}\,\gamma^\mu\left(L^\dagger P_L + R^\dagger P_R\right)\left(LP_L + RP_R\right)\psi \\ &= \bar{\psi}\gamma^\mu\psi\,. \end{aligned} \tag{87}$$

However the mass term is not invariant,

$$\begin{aligned} \bar{\psi}\psi &\Rightarrow \bar{\psi}\left(L^\dagger P_R + R^\dagger P_L\right)\left(LP_L + RP_R\right)\psi \\ &= \bar{\psi}\left(L^\dagger RP_R + R^\dagger LP_L\right)\psi\,. \end{aligned} \tag{88}$$

If the masses are unequal, it is not even invariant under the vector ($L = R$) subgroup.

Therefore, provided that *only* the mass term violates this symmetry, it is protected from being radiatively induced; if one puts a mass term into the theory, then while it gets multiplicatively renormalized (that is, $M_\mathrm{bare} = ZM_\mathrm{renorm}$), it is not additively renormalized.

6.2 Naive lattice fermions

What is a lattice implementation of $\bar{\psi}\gamma^\mu D_\mu\psi$? Obviously the difference between D_μ and ∂_μ is that we should parallel transport group indices using the gauge links whenever we compare things at different points. What should ∂_μ mean?

The *most* naive implementation would be

$$\bar{\psi}\gamma^\mu D_\mu \psi \to a^4 \sum_{x,\mu} \bar{\psi}(x)\gamma^\mu \left(\frac{U_\mu(x)\psi(x+\hat{\mu}) - \psi(x)}{a} \right) . \tag{89}$$

THIS IS WRONG.

The problems with this implementation are that

1. It is not hypercubic symmetric, because it involves a forward but not a backwards difference. We saw that we need hypercubic symmetry to get $O(4)$ Euclidean invariance in the IR.

2. It is not reflection-Hermitian (the Euclidean analog of Hermitian, where one takes complex conjugates and performs a reflection in some direction; think of the reflection as being in the time direction, and accounting for the fact that e^{-iHt} becomes e^{iHt} and that $\psi^\dagger$ is related to $\bar{\psi}$ by a $\gamma^0_{\rm E}$). When you take its reflection-Hermitian conjugate, $\bar{\psi}$ and ψ change roles and there is a - sign (except for the time component which is unchanged); so

$$\bar{\psi}(x)\gamma^\mu U_\mu(x)\psi(x+\hat{\mu}) \to -\bar{\psi}(x+\hat{\mu})\gamma^\mu U^\dagger_\mu(x)\psi(x) . \tag{90}$$

   The forward difference becomes minus a backwards difference.

Failure of reflection Hermiticity means that the lattice theory doesn't correspond to a unitary Minkowski theory (just as a non-Hermitian Minkowski action means a non-unitary theory).

The point is that one must instead construct a derivative as a symmetric operation; the naive lattice fermion action is

$$\bar{\psi}\gamma^\mu D_\mu \psi \to a^4 \sum_{x,\mu} \bar{\psi}(x)\gamma^\mu \left(\frac{U_\mu(x)\psi(x+\hat{\mu}) - U^\dagger_\mu(x-\hat{\mu})\psi(x-\hat{\mu})}{2a} \right) . \tag{91}$$

(This is forward backward symmetric–remember that $U_\mu(x)$ really lives at $(x+\hat{\mu}/2)$.) Superficially there is nothing wrong with this-in particular it respects chiral symmetry.

Remember what this expression really means. We can rewrite it in an absolutely general way as

$$\sum_{\alpha,\beta} \bar{\psi}_\alpha K_{\alpha\beta} \psi_\beta , \tag{92}$$

where the index α runs over all flavors, Dirac components, and colors at all points (that is, $4N_{\rm c}N_{\rm f}(V/a^4)$ values). The integral

$$\int \mathcal{D}\bar{\psi}\mathcal{D}\psi \ \exp\left(-\sum_{\alpha\beta} \bar{\psi}_\alpha K_{\alpha\beta}\psi_\beta \right) \equiv \ {\rm Det}\, K_{\alpha\beta} , \tag{93}$$

is defined as the determinant of the quadratic kernel $K_{\alpha\beta}$. Insertions of, say, $\bar{\psi}_\alpha\psi_\beta$ (an operator insertion) are defined as the minor of the α row, β column position, *et cetera*. The

matrix K is always square, and reflection Hermiticity ensures that the determinant is real. It depends on the links U. Only the ratio of determinants for different values of the links, Det $K(U_1)$/Det $D(U_2)$, is physically interesting.

Let's evaluate the inverse propagator, for the naive vacuum $U_\mu = \mathbf{1}$. We do it in momentum space by substituting $\psi(x) = e^{ip\cdot x}$ and $\bar{\psi}(x) = e^{-ip\cdot x}$. The result is

$$S^{-1}(p) = m + i\sum_\mu \gamma^\mu \frac{1}{a}\sin(p_\mu a) \equiv m + i\sum_\mu \gamma^\mu \hat{p}_\mu \,. \tag{94}$$

This has, as expected, a zero at $p_\mu = 0$ in the massless case. However, it *also* has zeros at $p_\mu = (0, \pi/a, 0, 0)$, $p_\mu = (\pi/a, \pi/a, 0, \pi/a)$, ... 16 zeros in all. (Note, I define $\hat{p}_\mu = (1/a)\sin(ap_\mu)$.)

The reason for the extra zeros is that the function $\sin p_\mu a$ has a zero not only at $p_\mu = 0$ but at $p_\mu = \pi/a$. That goes for each $\mu = 0, 1, 2, 3$, and so there are 2^4 (generally 2^d) poles in the propagator. Physically, the naive fermion field represents not 1 but 16 fermions, or rather, the IR effective theory which describes it contains 16 fermionic species. If we try to make ψ a Weyl particle by including only 2 spin components, then we get 16 Weyl fermions, of which 8 are left handed and 8 are right handed. Either we couple ψ to the gauge field, or we do not; so we inevitably end up with equal numbers of left and right handed fields with each charge assignment.

[Even though they have very large momenta, these "extra" states still have to be thought of as light states which are included in the infrared description of the theory, because exciting them involves a small amount of energy. The fact that the momentum is π/a also doesn't prevent their creation, since Fermi statistics already ensures that they are created in pairs, and momentum conservation is only valid modulo $2\pi/a$. There is no "good" way to determine, via an infrared experiment, which fermions are the copies and which are the real ones.]

The easy way to see this must be so, is to look at the term multiplying γ^μ (for one μ) in momentum space:

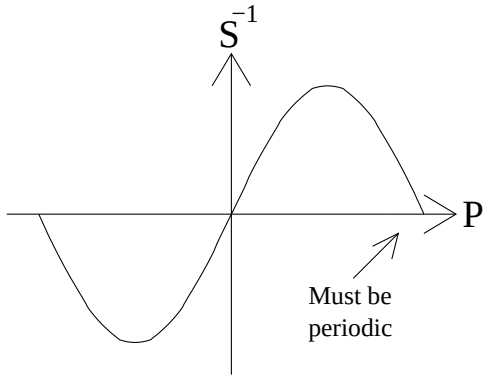


Figure 9: Fourier space inverse propagator: periodicity forces a second zero.

Even if we play with the definition of the single derivative (for instance, including next-nearest neighbors) we always have a second minimum somewhere. We can make the slope at

the second minimum steeper, but not remove its presence. (Making it steeper does not help; the phase space of the particle is smaller, but the gauge coupling is larger in a compensating fashion.) We can only eliminate it by making S^{-1} non-differentiable, which requires a non-local action, bringing in its own set of problems.

The easy way to see that the extra particle is of opposite handedness of the first, is to think about a 1+1 dimensional theory. Then, handedness is just "direction of motion." inspection of Fig. 9 shows that the second minimum has dS/dp of opposite sign and so represents a particle which moves with opposite sense to the one at $p = 0$.

If we cannot eliminate these doublers, or even if we can, but the elimination requires the breaking of chiral symmetry, then our ability to put the Standard Model on the lattice will be destroyed.

6.3 Wilson fermions

Two approaches to putting fermions on the lattice without introducing 16 species are widely used: Wilson fermions and staggered fermions. Both have advantages and disadvantages, and different practitioners consider the disadvantages of one or the other implementation to be more worrisome. Wilson fermions are more popular in Europe and Japan, staggered fermions are more popular in the UK and particularly in the U. S. This is a matter of religious feeling, like whether the Minkowski space metric should be written as $g_{\mu\nu} = \mathrm{Diag}[-+++]$ or as $g_{\mu\nu} = \mathrm{Diag}[+---]$. I will present both ideas, and make comments on the drawbacks and advantages of each.

Wilson fermions are based on the idea of removing the extra "doubler" degrees of freedom at the cost of breaking chiral symmetry at the lattice scale. One adds a high dimension operator which does not vanish at the "extra" points:

$$\begin{aligned} S_{\mathrm{Wilson}} &= -\frac{a^3 r}{2} \sum_{x,\mu} \bar{\psi}(x) \left(U_\mu(x)\psi(x+\hat{\mu}) - 2\psi(x) + U_\mu^\dagger(x-\hat{\mu})\psi(x-\hat{\mu}) \right) \\ &\to (ar) \int d^4x \; -\bar{\psi}\Box\psi \,, \end{aligned} \tag{95}$$

where in the second line I show the continuum ($p \ll 1/a$) limiting behavior.

The bad things are

1. This obviously breaks chiral symmetry, since there is no γ matrix between a $\bar{\psi}$ and a ψ.

2. I cannot write this term for a Weyl fermionic field, since it is contracting the left handed component of ψ with the right handed component of $\bar{\psi}$ and vice versa.

Therefore this is not an option for treating the Standard Model, but it can be used for QCD.

The inverse propagator is now

$$S^{-1}_{\mathrm{Wilson}} = \sum_\mu i\gamma^\mu \hat{p}_\mu + ar\tilde{p}^2 \,. \tag{96}$$

The $\gamma^\mu \hat{p}_\mu$ piece still vanishes at the "doubler points," but the term we have added, $ar\tilde{p}^2 = (4r/a)\sum_\mu \sin^2(p_\mu a/2)$ is large at those points and prevents extra zeros. Therefore there is only the expected single light fermionic field.

The cost is that chiral symmetry has been broken completely, at the lattice spacing scale. There is no accidental IR symmetry which will save us, any chiral symmetry we get in the IR has to come about by explicit tuning.

The lowest order chiral symmetry violating term we are stuck with, is a mass term $\bar{\psi}\psi$, which will now be generated with coefficient $\sim g^2/a$ even if we don't put it in. To see why, remember the discussion about tadpole corrections to links: averaging over UV fields approximately replaces the term $U_\mu(x)$ with $ZU_\mu^{\rm IR}(x)$ with $Z \sim 1 - g^2$ a renormalization and $U^{\rm IR}$ the link when only the IR gauge fields are present. This means

$$\begin{aligned} -U_\mu(x)\psi(x_+) + 2\psi(x) - U_\mu^\dagger(x_-)\psi(x_-) & \\ \simeq & -ZU_\mu^{\rm IR}(x)\psi(x_+) + 2\psi(x) - ZU^{{\rm IR}\,\dagger}_\mu(x_-)\psi(x_-) \\ = & Z\left(-U_\mu^{\rm IR}(x)\psi(x_+) + 2\psi(x) - U^{{\rm IR}\,\dagger}_\mu(x_-)\psi(x_-)\right) + 2(1-Z)\psi(x)\,. \end{aligned} \quad (97)$$

The last piece looks like a mass term, which must be compensated for with a negative mass counterterm. Delicate tuning of the counterterm is needed, because a mass is radiatively generated at every loop order and nonperturbatively. This requires, for instance, determining the mass based on correlation lengths (say, the pion to rho mass ratio) and tuning to get the desired mass, separately at each lattice spacing considered.

Another serious complication is that, since one is integrating over the links U_μ, occasionally a link configuration appears with unusually small UV fluctuations. Then, the mass squared counterterm (which is put into the Lagrangian, it does not vary as we perform the path integral) will over-compensate and the mass will be negative. Such configurations can have negative determinant. This is a problem both for the algorithm generating gauge field configurations (leading to extra numerical costs in performing Wilson quark simulations at small masses), and to the physical interpretation of what we are doing. I believe that this problem is ameliorated in the small a limit, such that the scale where negative eigenvalues set in is $m \sim a\Lambda_{\rm QCD}^2$. Therefore there is no problem in the formal $a \to 0$ limit, but there are problems at values of a which are currently realistic ($a \sim 0.1$ Fermi) and m small enough to be in the chiral limit.

The other problem with the Wilson action is that violating chiral symmetry introduces unwanted corrections in physical correlation functions at the $O(a)$ [dimension 5 operator] level. This problem can be patched up by much hard work. One must add other dimension 5, chiral symmetry breaking operators, such as

$$\frac{iac_{\rm SW}}{4}\bar{\psi}\sigma_{\mu\nu}G^{\mu\nu}\psi\,, \quad (98)$$

to the action, and tune the coefficient $c_{\rm SW}$ to make $O(a)$ effects vanish in physically interesting operators. The tuning is done at the nonperturbative level by varying $c_{\rm SW}$ until axial Ward identities, which would be valid in a chirally symmetric theory, are satisfied for infrared

observables. For instance, defining the axial current as $A_a^\mu = \bar{\psi}\gamma^\mu\gamma^5\tau_a\psi$ and the pseudoscalar current as $A_a = \bar{\psi}\gamma^5\tau_a\psi$ (where τ_a acts in flavor space), a Ward identity is that, for any operator $\mathcal{O}$,

$$\partial_\mu\langle A_a^\mu\mathcal{O}\rangle = 2m\langle P_a\mathcal{O}\rangle\,. \tag{99}$$

This also demands nonperturbative improvement of the currents A_a^μ and P_a. Recently this program has been pushed through, mostly by the ALPHA Collaboration, to the point that one can now nonperturbatively remove all $O(a)$ errors in Wilson fermion numerical simulations.

6.4 Staggered quarks

Staggered quarks were developed by Kogut and Susskind in the mid 1970's. The idea is to go back to naive quarks and notice that the formulation presented at the beginning was 4-fold degenerate. That is, the operator $K_{\alpha\beta}$ implicitly defined in Eq. (91) is block diagonalizable into 4 blocks, so its determinant is $\mathrm{Det}K_{\mathrm{block}}^4$. By finding K_{block} and taking the single power of its determinant, we could get 4, rather than 16, species. Then we just live with having 4 species, or if we want 2, we do something brutal like taking the square root of a determinant.

To show that the naive quark action breaks into 4 blocks, define (denoting the position x as $(x_1, x_2, x_3, x_4) = a(n_1, n_2, n_3, n_4)$ with n_μ integers)

$$\psi(n_1, n_2, n_3, n_4) \equiv \gamma_1^{n_1}\gamma_2^{n_2}\gamma_3^{n_3}\gamma_4^{n_4}\psi'(n_1, n_2, n_3, n_0)\,, \tag{100}$$

where the n_i in the RHS are exponents on the γ matrices. The transformation of $\gamma^\mu\psi(x+\hat{\mu})$ is

$$\gamma^\mu\psi(x+\hat{\mu}) = \gamma^\mu\gamma_1^{n_1+\delta_{1\mu}}\gamma_2^{n_2+\delta_{2\mu}}\gamma_3^{n_3+\delta_{3\mu}}\gamma_4^{n_4+\delta_{4\mu}}\psi'(x+\hat{\mu})\,. \tag{101}$$

There are two extra powers of γ^μ. We can anticommute one to the location of the other and multiply them to get 1, at the cost of (-1) to some power. That means that, if we change basis from ψ to ψ', that the kinetic term becomes diagonal, with the γ matrices replaced by ±1's. Then the 4 components of ψ' do not mix, and we can just keep the first component. (The 4 components of a fermion are then spread in some complicated way over a 2^4 block of the lattice.)

This is a theory of 4 Dirac fermions, with an exact $U(1)\times U(1)$ chiral symmetry–less than the $U(4)\times U(4)$ symmetry we hoped for, but enough to forbid additive mass renormalization and dimension 5 chiral symmetry breaking operators. The 4 distinct Dirac fermions are called "tastes," to distinguish them from distinct "flavors," since the latter usually have different masses and the "tastes" always have the same mass.

If we could really think of the resulting theory as a theory of four "tastes" of fermions, that is, a modification of the path integral to include a $(\mathrm{Det}\not{D})^4$, then there would be no conceptual problems associated with taking the fourth root of the resulting determinant and claiming that we were including only a single fermionic species. The problem is, that the four fermions we get are coupled to each other, since they are arising from a single determinant. A large momentum gluon can "bounce" one taste of fermion into another taste. Integrating out the high momentum gluons, this leads to taste-violating, four-fermion (dimension 6)

operators in the effective description. Taking the fourth root of the determinant when such taste violation is present is not mathematically sound and it leaves some ambiguities in which observables are to be interpreted as the "right" ones for QCD. The issues of taste violation are expected to vanish as a^2 in the small lattice spacing limit.

A great deal of work has gone into designing fermion containing operators for this representation of the fermions, and building highly improved actions which minimize dimension 6 operators responsible for taste violation. Some theoretical issues involving taste violation and taking roots of determinants remain unresolved.

7 Topology and the Lattice

By topology I mean, essentially, instanton number. I will not discuss the fact that the periodicity of a torus gives holonomy (globally nontrivial vacua). I will assume a familiarity with instantons in this chapter.

7.1 Instantons and winding number

I begin by reminding you what topology is all about in the continuum, for differentiable group connections.

Recall that

$$\frac{1}{64\pi^2} F^a_{\mu\nu} F^a_{\alpha\beta} \epsilon^{\alpha\beta\mu\nu} \tag{102}$$

is a total derivative; it equals

$$\frac{1}{64\pi^2} F^a_{\mu\nu} F^a_{\alpha\beta} \epsilon^{\alpha\beta\mu\nu} = \frac{1}{32\pi^2} \epsilon^{\mu\nu\alpha\beta} \partial_\mu \left[F^a_{\nu\alpha} A^a_\beta - \frac{1}{3} f_{abc} A^a_\nu A^b_\alpha A^c_\beta \right] \equiv \partial_\mu K^\mu \,. \tag{103}$$

Often $\epsilon^{\mu\nu\alpha\beta} F_{\alpha\beta}/2$ is called $\tilde{F}^{\mu\nu}$, "F dual".

Generally, on a topologically nontrivial space like T^4 the torus, it is not possible to describe the gauge connection by a gauge field globally, just as it is not generally possible to give a nonsingular coordinate system to a manifold; we can only do so on suitable patches. If we "prick" the 4-torus spacetime by removing a tiny ball around one point, then it is topologically trivial enough that we can describe the group connection with a gauge field A_μ which is free of singularities[2].

The problem of determining Eq. (103) becomes the problem of the surface integral over the tiny ball we cut out of the space. For such a tiny ball, the field strength term $F^a_{\nu\alpha} A^a_\beta$ is always subdominant, we can treat the gauge field as a pure gauge $A^a_\mu = (\partial_\mu \Lambda \Lambda^{-1})$. Now think about SU(2), which happens to be isomorphic to the 3-sphere. For a pure gauge configuration, $F^a_{\nu\alpha} = 0$, the gauge parameter Λ constitutes a mapping from the surface of the ball (which is also S^3) into the group manifold:

$$\Lambda \,:\, S^3 \Rightarrow \mathrm{SU}(2) \,. \tag{104}$$

[2]Actually this is false, because of π_1 holonomy; but this technicality is of no importance in the current discussion.

A^a_μ is how fast one is moving across the surface of SU(2) as one moves in the μ direction. The combination $f_{abc}\epsilon^{\mu\nu\alpha\beta}ds_\mu A^a_\nu A^b_\alpha A^c_\beta$ is (6 times) the infinitesimal volume element which the dx_ν, dx_α, dx_β bit of the 3-sphere covers in SU(2). This should be clear since f_{abc} is like the cross product for SU(2), it picks out how much the 3 directions of motion are orthogonal to each other. The factor of 6 is because having both an ϵ and an f_{abc} sums over the orderings of the A^a_ν 6 times.

Since the volume of SU(2) is $(2\pi^2)\times 8 = 16\pi^2$ (the factor $2\pi^2$ is the volume of a 3-sphere, the factor of $8 = 2^3$ is because the 3 group generators are taken to be $\tau^a/2$ rather than τ^a). The mapping Λ of S^3 into SU(2) must wrap over SU(2) an integer number of times; $\pi_3(\mathrm{SU}(2)) = \mathbf{Z}$. Accounting for the factor of 6 mentioned, the map Λ covers the group manifold SU(2) once if

$$\int ds_\mu \epsilon^{\mu\nu\alpha\beta} f_{abc} A^a_\nu A^b_\alpha A^c_\beta = 96\pi^2\,, \qquad \text{or} \qquad \int ds_\mu K^\mu = 1\,. \tag{105}$$

Therefore, the integral

$$\frac{1}{64\pi^2}\int d^4x \epsilon^{\mu\nu\alpha\beta} F^a_{\mu\nu} F^a_{\alpha\beta} \in \mathbf{Z}\,, \tag{106}$$

is always an integer, called the second Chern class or instanton number of the connection. The integer value of the integral over a surface for a pure gauge configuration is called the winding number of the pure gauge configuration. The specialization to a time surface is called Chern-Simons number. The generalization to SU(N_c) is simple, because the topological properties used above depend only on the ways 3-spheres wrap in the manifold ($\pi_3(\mathrm{SU}(N_\mathrm{c}))$), and $\pi_3(\mathrm{SU}(N_\mathrm{c})) = \mathbf{Z}$ for every $N_\mathrm{c} > 1$.

7.2 $F\tilde{F}$ on the lattice

How should we write $F\tilde{F}$ on the lattice?

Naively,

$$\text{naive}\;\; F^{\mu\nu}_a = \mathrm{Tr}\,\frac{i\tau^a}{2a^2}\Box_{\mu\nu}\,, \tag{107}$$

with $\Box_{\mu\nu}$ the elementary plaquette we met before. The complication is that the plaquette most naturally lives at $(x+\hat{\mu}/2+\hat{\nu}/2)$. We want to contract it with $\tilde{F}^a_{\mu\nu}$, which is the plaquette in the 2 orthogonal directions, call them α,β, and lives most naturally at the point $(x+\hat{\alpha}/2+\hat{\beta}/2)$.

The simplest hypercubic invariant version of $F\tilde{F}$ defines

$$F_{\mu\nu}(x) \equiv \frac{1}{4}\sum_{4\Box}\mathrm{Tr}\,\frac{i\tau^a}{2}\Box_{\mu\nu}\,, \tag{108}$$

where the sum is over the 4 plaquettes in the μ,ν plane which have x as a corner, each oriented with the same handedness. $\tilde{F}$ is a similar sum in the other two directions.

The problem with this definition is that $F\tilde{F}$ is not a total derivative. Intuitively, the reason is because our construction only coincides with $F\tilde{F}$ in the slowly varying field limit;

there are high dimension operator corrections, such as $\epsilon^{\mu\nu\alpha\beta}F_{\mu\nu}D^2_\alpha F_{\alpha\beta}$; the property of being a total derivative is special to the dimension 4 operator and doesn't hold for the high dimension add-ons. It is very easy to find lattice configurations which give nonvanishing contributions at one site only and are flat vacuum fields at large distance.

In fact there is a simple proof that *any* implementation of the continuum expression which is of the form of

$$F^{\mu\nu}_a \tilde{F}^a_{\mu\nu}(\text{lattice}) = \sum_x (\text{local expression}) \,, \tag{109}$$

will *not* be topological. The proof is that any such expression must be a smooth function of the link variables $U_\mu(x)$. However, the space of all possible lattice gauge field configurations is

$$\mathcal{G} = \otimes_{\alpha=(x,\mu)} \mathrm{SU}(N_\mathrm{c}) = [\mathrm{SU}(N_\mathrm{c})]^{4(L/a)^4} \,, \tag{110}$$

which is a compact, path connected, simply connected space. The range of any real valued, continuous function on a path connected space is always path connected, that is, there must be a continuous range of values which the imputed $F\tilde{F}$ implementation takes on. Lattice sums over local quantities are never quantized, unless they are identically zero. Is there no topology on the lattice?

7.3 Instantons on the lattice

The physical reason for our problem is that instantons can come in a continuous range of sizes down to zero size. When they become of order the lattice spacing or smaller, the lattice cannot resolve them; they "fall between the lattice points."

Write down the continuum field of an instanton of size λ at point y;

$$A_\mu(x) = -i\frac{(x-y)^2}{(x-y)^2+\lambda^2}(\partial_\mu\Omega)\Omega^{-1} \,, \quad \Omega = \frac{x_4 - y_4 + i\vec{\sigma}\cdot(\vec{x}-\vec{y})}{\sqrt{(x-y)^2}} \,. \tag{111}$$

Now define a lattice instanton configuration via $U_\mu(x) = \mathrm{Pexp} i\int_x^{x+\hat{\mu}} T^a A^a_\mu dl_\mu$. For $\lambda \gg a$, this corresponds to an instanton at a scale so big that the fact we have a lattice must not matter. Any reasonable thing we write down for $F\tilde{F}$ will give 1 with $O(a^2/\lambda^2)$ corrections (until we add in fluctuations, which will renormalize the lattice $F\tilde{F}$ and give us something smaller...).

Now consider taking λ smoothly to zero, and choosing y not to lie at a lattice site or on any link. When $\lambda << a$, the link variables are all almost exactly the identity, and so the integral gives approximately zero. For intermediate $\lambda \sim a$, it smoothly goes from one to the other. Basically this is why we can't define topology on the lattice.

7.4 Defining topology on the lattice

A truly topological implementation of $\int F\tilde{F}$ on the lattice must *not* be an integral over a local charge density; it must be

- nonlocal, and
- non-smooth on the space $\mathcal{G}$ of all possible link fields (gauge connections).

It should also be gauge invariant. Such things exist. The answer they will give you is not unique, that is, different perfectly acceptable "topological" implementations of instanton number can disagree on the instanton number of a particular configuration. The key is that topological measures have the properties we expect if we *restrict* ourselves to smooth configurations, meaning, say, configurations with $\mathrm{Tr}\,\Box_{\mu\nu}$ always exceeding some value. The smaller an instanton, the smaller a region the field strength has to get squeezed into; a lattice sized instanton bumps up into this restriction and is removed from consideration.

This isn't a problem in practice because, the smaller our lattice spacing, the fewer such non-smooth configurations there are. For instance, the small sized tail of the instanton size distribution in pure glue theories behaves like

$$d\lambda\lambda^{-5}e^{-8\pi^2/g^2(\lambda^{-1})}\,, \tag{112}$$

but $g^2(\lambda^{-1})$ satisfies

$$\frac{1}{g^2(\lambda^{-1})} = \frac{1}{g^2(\lambda'^{-1})} + \beta\ln\frac{\lambda}{\lambda'}\,, \qquad \beta = -\frac{1}{8\pi^2}\frac{11N_{\rm c}-2N_{\rm f}}{3}\,. \tag{113}$$

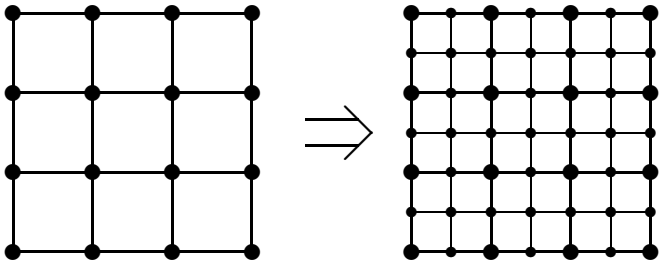
Substituting, the small instanton tail behaves like $d\lambda\lambda^{-5}\lambda^{11N_{\rm c}/3}$, which is perfectly well behaved. If we were speaking of a theory with fermions, then the beta function would be less negative; but the Dirac operator, provided chiral symmetry is properly implemented, would have a set of zero modes in the instanton background, $2N_{\rm f}$ in number, which would contribute an extra $(\lambda m_f)^{2N_{\rm f}}$ suppression, more than making up for the shift in the beta function. What happens to these zero modes of the Dirac operator as the instanton shrinks away, should give us a sinking feeling that there really will be some problems with the implementation of chiral symmetry on the lattice. But the point is that there is already a good mechanism which suppresses non-smooth configurations in the small spacing limit, so imposing that they be thrown out will change the path integral in a way which vanishes as a high power of a and is therefore acceptable.

Specializing to smooth configurations, topology can be defined successfully in a number of ways, which all give the same answer for the instanton number of particular configurations. (They will give different answers on non-smooth configurations but we should not expect otherwise.)

Let me give some examples, since there is some beautiful math in some of them.

7.4.1 Inverse blocking

The idea of inverse blocking is to take a lattice and to construct a lattice which describes the same continuum fields and has half the lattice spacing;



The links of the "refined" or "inverse blocked" lattice are determined by minimizing the action subject to the constraint that

$$U_\mu^{1/2}(x)U_\mu^{1/2}(x+\hat{\mu}/2) = U_\mu^{\text{full}}(x)\,, \tag{114}$$

that is, that the product of two links on the inverse blocked lattice gives the corresponding link of the original lattice. This gives the smoothest possible interpolation of the original lattice configuration. If we iterate we get something which is even smoother (on a lattice with even more links). One evaluates $\int F\tilde{F}$ on a many times inverse blocked lattice, or in principle on the limit of infinite inverse blockings. Except on a measure 0, codimension 1 surface in the space of gauge configurations, this procedure converges to an integer winding number. Typically it approaches very fast so one can "round off" the number after, say, one inverse blocking to determine the right continuum limit.

The method is nonlocal and discontinuous on the space of lattice configuration because the inverse blocking procedure is nonlocal and discontinuous; the minimization of the action of the inverse blocked lattice does not depend strictly locally on the unblocked lattice fields. When an instanton shrinks to about 1 lattice spacing, the inverse blocking action minimum switches from one which reconstructs an instanton to one which does not (discontinuous behavior on the space of lattice configurations).

7.4.2 Overimproved cooling

Overimproved cooling is an algorithm which starts by defining an "overimproved" lattice action. This is an action which involves some traces of Wilson loops larger than the elementary plaquette. The high dimension terms in the standard Wilson action *lower* the action of a small instanton. One chooses instead an action where, say, the dimension 6 terms vanish but dimension 8 terms make a small instanton cost *more* action than a larger one. Then, one takes a field configuration and iteratively applies a "cooling" procedure, where one updates link by link to locally minimize or reduce the action. A boring, perturbative fluctuation is removed. A large instanton, which is a solution of the equations of motion, is preserved. A smallish instanton actually *grows*, because of our choice of an action with positive dimension 8 terms. A very small or "would-be" instanton shrinks away to nothing. One continues to apply cooling until the gauge field configuration is so smooth that a (highly improved) lattice implementation of $\int F\tilde{F}$ will give very close to an integer. Then one rounds off. There is some special size of instanton which "gets stuck" between wanting to grow and shrink

away; this is where the procedure is nonlocal and discontinuous on the space of gauge field configurations.

7.4.3 Lüscher's, Woit's, Phillips and Stone's methods

There are a number of other, elegant, very mathematical, and not very practically useful, methods. You don't want me to describe any of these in detail.

8 Nielsen Ninomiya No-Go Theorem

When we studied fermions on the lattice we found that it was difficult to preserve an exact chiral symmetry, and that the "doubler problem" seemed also to make it difficult to have different numbers of left and right handed particles. In 1981 Nielsen and Ninomiya showed that, under fairly general assumptions, these are actually *impossible*. One *always* gets the same number of right and left handed particles (though, see the next section). Since an exact chiral symmetry would permit us to "just throw away" one handedness of fermions, this also precludes an exact chiral symmetry on the lattice, in the sense of a lattice Hamiltonian which commutes with γ^5.

8.1 Framework and Assumptions

Nielsen and Ninomiya worked in a peculiar framework. They took the lattice time coordinate to be continuous, so there is a Hamiltonian framework for the problem. They also effectively ignored the gauge fields; if you cannot get chiral fermions and different right and left handed numbers of species in the absence of gauge fields, it is very unlikely that you can when the gauge fields are "turned on." Also, as presented their theorem applies for a cubic lattice– though the extension to general periodic lattices is completely straightforward. It is not obvious to me what would happen on a random lattice, and I won't discuss it further.

The assumption that there is a continuous Hamiltonian is actually not problematic, because if the theory is well enough behaved for a continuation to Minkowski time to exist, there has to be a positive transfer matrix[3] which can be generated by exponentiating a Hamiltonian; in other words, even if time is discrete, I have to be able to define the operation of "advancing by some discrete amount of time," and the time evolution by larger amounts of time must be given by iterating this time-advancing operation. The Hamiltonian is the logarithm of the time advancing operation, and I can apply the theorem to that Hamiltonian.

We take there to be some number N of fermionic fields (corresponding for instance to different flavors and/or spin states; I will suppress the index for the fermionic field number),

[3] A quantum state is a wave function on the set of field values available at one time slice, $\Psi(\phi(\mathbf{x}))$. The transfer matrix T is the operator which time evolves this wave function by a finite Euclidean amount, say, one lattice spacing. It has to be positive, so that it can be written $\exp(-H\tau)$ for some Hermitian H. The way you continue to Minkowski time is by extracting this H and using it as the Minkowski time evolution operator on states.

with action

$$S = \int dt \left[\sum_x \bar{\psi} \partial_0 \psi + \sum_{xy} \bar{\psi}(x) H(x,y) \psi(y) \right], \tag{115}$$

and we ask the Hamiltonian and the fields to have the following properties;

1. H is local, meaning that $H(x,y)$ falls as a suitably large power of the separation $(x-y)$, so that its Fourier transform $\tilde{H}$ is differentiable;
2. H is translation invariant, $H(x,y) = H(x-y)$;
3. H is Hermitian;
4. The fields ψ carry a charge Q which is
   - locally defined, $Q = \sum_x j^0(x)$ with $j^0(x)$ a local function of ψ in the same sense that H is local;
   - exactly conserved;
   - quantized (taking integer values);
   - bilinear in the fermionic field.

Under these hypotheses, there will always be equal numbers of right and left handed fields with any given nonzero charge assignment. This does not preclude constructing lattice gauginos, but it does preclude particles of discrete Abelian charges. If we wanted to study the Standard Model, then once we attempt to implement SU(2) as a gauge theory, then even assigning hypercharge as a global charge, we run into trouble.

8.2 Weyl fermions

To start the proof, we have to figure out what a continuum Weyl fermion is, in the context of a lattice theory. At a general momentum $\mathbf{p}$, the set of eigenvalues of H is discrete, call them $\lambda_0, \lambda_1, \ldots \lambda_{N-1}$, with splittings of order the lattice scale $1/a$. Exciting an excitation would require popping a fermion from one level to another, requiring a large energy. Again, think of what an IR observer can probe; they can only put small amounts of energy and momentum into the fermions. At generic momenta, what we have are "UV degrees of freedom." An IR vacuum Weyl fermion is a point where two sheets of allowed solutions touch at a point, see Fig. 10; and the "lower" states are filled while the "higher" states are empty. Then a particle hole pair can be produced by popping one of the "valence band" fermions into the "conduction band," at a small cost of energy and momentum. Because the energy and momentum cost of making an excitation is small, the presence of such a point in the dispersion curve corresponds to an IR flavor of Weyl fermion.

To see what happens near the degeneracy point $\mathbf{p}_{\rm deg}$, where $\lambda_n(\mathbf{p}_{\rm deg}) = \lambda_{n+1}(\mathbf{p}_{\rm deg})$, consider the Hamiltonian, keeping only the vector space spanned by the eigenvectors ψ_n,

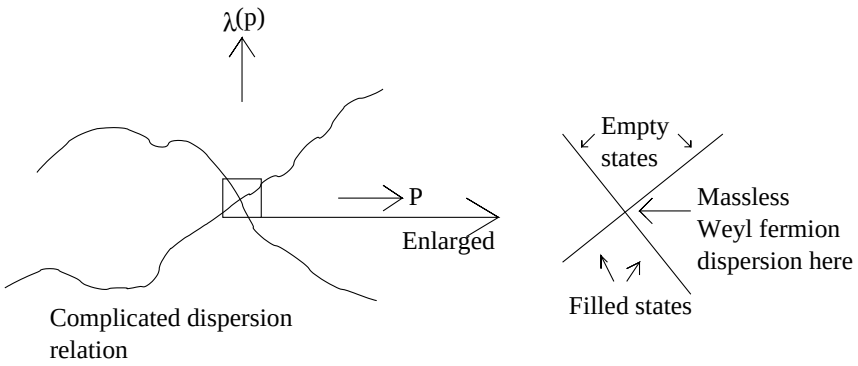


Figure 10: Crossing point of two eigenvalue surfaces, in momentum space, gives a Weyl fermion.

ψ_{n+1} of the two eigenvalues;

$$\begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} \begin{bmatrix} \psi_1 \\ \psi_2 \end{bmatrix} = \left(\sum_{i=1}^{3} c_i(\mathbf{p})\sigma^i + d(\mathbf{p})\mathbf{1} \right) \psi \,, \tag{116}$$

where we used the Hermiticity of H to write it in terms of 4 independent, real functions of $\mathbf{p}$, $c_i(\mathbf{p})$ and $d(\mathbf{p})$. The condition for a degeneracy at $\mathbf{p}_{\rm deg}$ is $c_i(\mathbf{p}_{\rm deg}) = 0$ for each i. This is 3 conditions, which is why, in 3 spatial dimensions, it occurs at points, rather than along lines or planes. There are never three degenerate eigenvalues at once, because that would require the coefficients on all 8 of the Gell-Mann matrices to vanish at once, which is 8 conditions–generically not satisfied in 3 dimensions.

Taylor expand around the degeneracy point, defining $\mathbf{p}' = \mathbf{p} - \mathbf{p}_{\rm deg}$, and $\lambda' = \lambda - d(\mathbf{p}_{\rm deg})$; the Hamiltonian is

$$H = \sigma^i V_{ik} \mathbf{p}'_k + d_i \mathbf{p}'_i \,, \tag{117}$$

which is a Hamiltonian for a Weyl spinor with a potentially unconventional dispersion relation, which we can "canonicalize" by replacing $\mathbf{p}'$ with $\pm V^{-1}\mathbf{p}'$ and λ with $\lambda - d \cdot \mathbf{p}$. The handedness of the Weyl field is determined by whether the resulting equation looks like

$$\sigma^0 \partial_t \psi = \pm i \sigma \cdot \mathbf{p}' \psi \,; \qquad \pm \text{ is } +, \text{ right-handed; -, left-handed} \,. \tag{118}$$

We can make it + by replacing $\mathbf{p}'$ with $+V^{-1}\mathbf{p}'$, but this is an axis-inverting (and therefore handedness flipping) operation if $\mathrm{Det}(V) < 0$. Therefore the handedness of the fermion is determined by

$$\text{Handedness} = \text{sign Det } V_{ik} \,. \tag{119}$$

8.3 First proof: homotopy theory

The Nielsen-Ninomiya no-go theorem then consists of proving that the sum, over all degeneracy points, of this sign of determinant, is always zero. They offered two proofs, and I will sketch, very lightly, each proof.

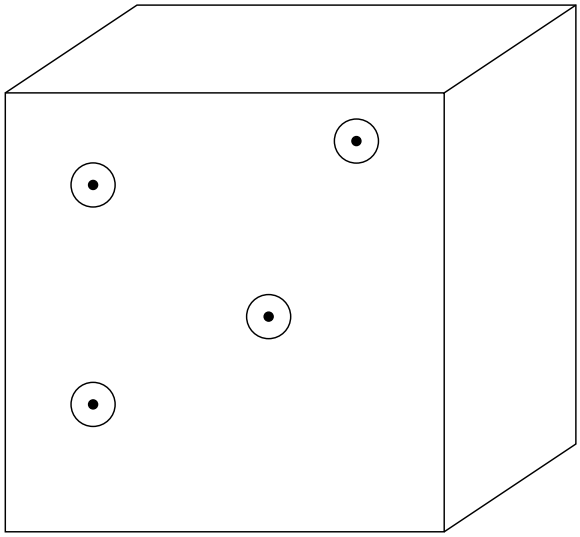


Figure 11: The Brillouin zone with tiny balls around each degeneracy point excised.

The first proof uses homotopy theory. Consider the space of possible momenta $\mathbf{p}$. It forms a compact space without boundary, the Brillouin zone. Excise a tiny ball around each of the degeneracy points and consider the region outside of these balls, see Fig. 11.

Think about the particular eigenvalue λ_n and its eigenvector ψ_n in this region. Since we took out the places where "all the action is," the function ψ_n is continuous (here we use the locality of H). A function is really a map from the domain (in our case, the Brillouin zone with the balls removed) into the range of the function. Naively, we can think of ψ_n as taking N component, complex values, that is, $\psi_n : \mathrm{BZ}_{\mathrm{excised}} \to \mathbf{C}^N$. But really we know that the magnitude of ψ_n has to be of unit length; and the phase of ψ_n is unphysical; the space in which ψ_n (a complex spinor eigenvector of H) lives, is $\mathbf{C}^N$ restricted to unit length and modded out by the overall phase; this space is called CP^{N-1}. The eigenvector ψ_n is a map

$$\psi_n \; : \; \Big\{ \mathrm{BZ} - \text{“balls”} \Big\} \to \mathrm{CP}^{N-1} \, . \tag{120}$$

The piece of information we need to know about CP^{N-1} is, that $\pi_2(\mathrm{CP}^{N-1}) = \mathbf{Z}$, that is, there are topologically different ways that the 2-sphere S^2 can be mapped into CP^{N-1}, classified by a winding number which can be any integer. For instance, CP^1 is the sphere itself; when mapping the sphere into the sphere, you can do so with vanishing winding number (having the whole "domain" sphere land on the "range" north pole), winding number one (the identity map), winding number -1 (the antipode map), or any larger winding number.

Near the degeneracy point, the map is basically controlled by what is happening in the subspace of the two nearly-degenerate eigenvectors; so it effectively reduces to a two-component problem. It turns out that, for a right handed degeneracy point, the map from the surface of the little excised ball (which is a two-sphere), to CP^{N-1}, is winding number 1, as you can convince yourself in about an hour by thinking about the direction that the

positive energy spinor points as a function of the direction of the momentum $\hat{\mathbf{p}}$. Since a left-handed spinor looks like a right handed spinor under space inversion, the winding number around a left handed degeneracy point is -1.

The sum of the winding numbers over each little excised sphere, is the difference in the number of right and left handed Weyl fermions. This is what we want to show is zero. To do that, slowly expand each excised sphere. Because ψ_n is smooth, the sum of winding numbers must change smoothly; because the winding number is quantized to be an integer, it must not change. It also doesn't change when two of our spheres get big enough to touch and merge into one ball. Eventually the surface expands until it fills the Brillouin zone. Then we use the periodicity to show that the winding number must be zero. For complete details you will have to read Nielsen and Ninomiya's original paper.

8.4 Second proof: topology

The second proof is cuter and maybe easier to wrap your brain around, so let's go through it.

Choose some "default" N component spinor, α. Consider the condition $\alpha^\dagger \psi_n = 0$; this is two real conditions. Therefore, in momentum space, it defines some complicated curve, in the Brillouin zone. The curve can have multiple segments, it can explore the periodicity of the Brillouin zone; but it turns out it will have three important properties;

1. The curve $\alpha^\dagger \psi = 0$ goes through every degeneracy point;

2. The curve can be given an orientation; doing so, it goes through right-handed degeneracy points from the *lower* to the *higher* eigenvalue surface; and through left-handed degeneracy points from the *higher* to the *lower* eigenvalue sheet;

3. The curve cannot terminate, but must form some set of closed loops.

First, let us see that the curve can be oriented. At a generic point on the curve, the eigenvector ψ_n is a continuous function. Draw a tiny loop around the curve; as we go around the loop, the phase of $\alpha^\dagger \psi_n$ must wind around in a circle. (You can see this by drawing the plane, perpendicular to the loop, and Taylor expanding ψ_n on this plane, about the point where the loop pierces the plane.) We define the orientation of the loop, from the direction in which the phase winds, by the right hand rule.

Next, let us see that the loop goes through each degeneracy point. At the degeneracy point, there are two eigenvectors ψ_n, ψ_{n+1} with the same eigenvalue. Each has some dot product with α; $c_n = \alpha^\dagger \psi_n$, $c_{n+1} = \alpha^\dagger \psi_{n+1}$. The linear combination $(c_{n+1}\psi_n - c_n \psi_{n+1})/\sqrt{|c_n|^2 + |c_{n+1}|^2}$ vanishes against α; therefore the curve goes through the degeneracy point. Physically, the reason that the curve goes through each degeneracy point is that, near the degeneracy point, the value of the eigenvector is changing very rapidly, and all possible two-component spinors are being explored, at given $|\mathbf{p}|$, over the sphere of possible directions $\hat{\mathbf{p}}$.

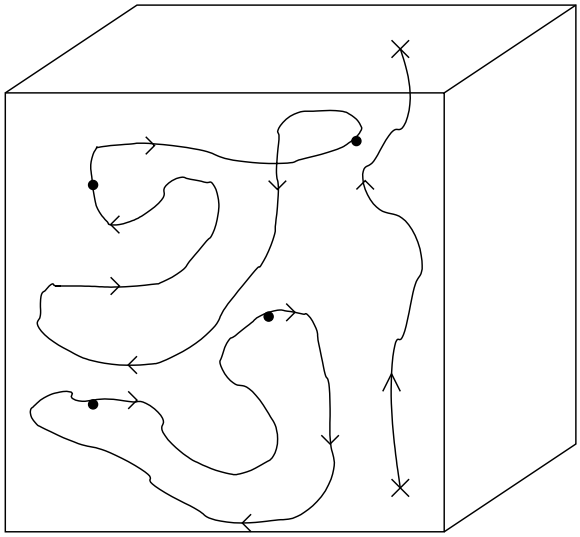


Figure 12: Curve defined by $\alpha^\dagger \phi = 0$, in the Brillouin zone. The curve can have several components, and they can explore the periodicity of the zone (shown by the $\times$ symbol, where the curve leaves the top and appears again at the bottom); but they must form closed, oriented loops.

If the curve comes into a degeneracy point on the lower surface, it must go out on the upper surface, since, for a Weyl fermion, $\psi_-(-\mathbf{p}) = \psi_+(\mathbf{p})$ (where ψ_- is the negative and ψ_+ the positive eigenspinor). Since different right handed degeneracy points are related by a shift and rescaling of the momentum variable, all right handed degeneracy points will either have the curve go from below to above, or vice versa. (One can figure out which it is, but you don't even need to bother.) A left handed degeneracy point is obtained either by spatial inversion (which flips our procedure for determining the line orientation), or if you prefer, by flipping the sign of the energy; so it is obvious that, if the curve goes upwards through right-handed degeneracy points, it goes downwards through left handed ones, and vice versa. Since the curve is closed, it must go up as many times as it goes down; so again, the number of right and left handed Weyl fermions is seen to be equal.

8.5 Meaning in terms of ABJ anomaly

Why do right and left handed fermions have to have equal abundance? We can see this, for the interesting case of a field coupled to an SU(N_c) gauge field, by thinking about the chiral (Adler-Bell-Jackiw) anomaly. An instanton causes *spectral flow*; it makes fermionic states move from negative energy (below the pointlike contact between "conduction" and "valence" dispersion bands) to positive energy for one handedness, and the other direction for the opposite handedness. But the total number of empty and filled states, in a *finite* system like the lattice, must stay the same. Therefore, the flow of states, upwards from negative energy, must be compensated by a flow downwards somewhere else. This implies

that either

1. there are an equal number of left and right handed Weyl fermions, or
2. the chiral anomaly is violated,

on the lattice. If the chiral anomaly is violated, we have not correctly formulated the theory we wanted.

In fact, Nielsen and Ninomiya proved a stronger result; for *ANY* regulation scheme, a regulated theory satisfying

1. The theory is invariant under at least the global subset of gauge transformations,
2. There are different numbers of right and left handed species with a given charge assignment,
3. The theory exhibits the correct ABJ anomaly,
4. The action is bilinear in the fermion fields,

cannot exist. The reason is that the first and fourth conditions can be used to prove that the axial current vanishes, while the second and third conditions demand that it does not vanish.

The way in which well known regulation schemes fail these criteria is,

- Lattice: equal number of right and left handed species.
- Pauli-Villars: breaks gauge invariance (for nonabelian theories).
- Dimensional regularization, $\gamma^5 = (\gamma_1\gamma_2 \ldots \gamma_{4+\epsilon})$: incorrect ABJ anomaly.
- Dimensional regularization, $\gamma^5 = (\gamma_1\gamma_2\gamma_3\gamma_4)$: breaks gauge invariance.

9 Ginsparg-Wilson fermions

The summary of the last section could be, that one cannot construct a lattice Hamiltonian D (which we called H in the last section) satisfying $D\gamma^5 + \gamma^5 D = 0$. (To see that this is a consequence of the Nielsen-Ninomiya theorem, suppose to the contrary that you can. Then, the right and left handed fermions never "talk to" each other, and we can just not bother to include one set of them in the theory, giving us a way to get the chiral theory we want.) Since a chirally symmetric theory in the continuum has this property, $D = D_\mu\gamma^\mu$ and $\gamma^\mu\gamma^5 + \gamma^5\gamma^\mu = 0$, it looks like we cannot get chiral symmetry on the lattice.

9.1 Ginsparg-Wilson equation

How close can we get? Ginsparg and Wilson asked this in a long-forgotten but now famous paper, in 1982. They considered starting with a chirally symmetric theory in the continuum, and arriving at a lattice theory by blocking; essentially, if the continuum field is ϕ and the intended lattice field is ψ, we define the lattice field to follow the continuum field in some way,

$$\psi(x) = \int d^4y \alpha(y-x)\phi(y)\,, \tag{121}$$

with x a lattice site, y a continuous parameter, and α some (unimportant) weighting function, which could be as simple as $\delta(x-y)$ (forcing the lattice thing to equal the continuum one at the same point). Then, considering the integration over ϕ, done in continuous space, one sees how close the induced action for ψ can come to being chirally invariant. Ginsparg and Wilson concluded that the best you could do was

$$D\gamma^5 + \gamma^5 D = (a) D\gamma^5 D\,. \tag{122}$$

Here I have explicitly included the lattice spacing to show that, for suitably infrared fields (or for small enough a, same thing), this becomes the desired continuum relation. Henceforth I generally set $a=1$, you can figure out where it goes from dimension counting.

It turns out that this is much better than just "a discrepancy which vanishes with the lattice spacing." The propagator is defined by

$$S = D^{-1}\,, \quad \text{meaning} \quad D_{\alpha\gamma}(x,z)S_{\gamma\beta}(z,y) = \delta_{\alpha\beta}\delta^4(x-y)\,; \tag{123}$$

acting with S on both sides of Eq. (122), we find that it satisfies

$$\gamma^5 S + S\gamma^5(x,y) = a\gamma^5\delta^4(x-y)\,. \tag{124}$$

The point is that, not only does this vanish as the lattice spacing is made small; it is also *ultralocal*. At all finite distances, we observe a propagator with exact chiral symmetry. This turns out to be enough to give us all the benefits we want of chiral symmetry, such as the absence of additive mass renormalization.

The first such operator was constructed by Neuberger and Narayanan, based on an idea developed by David B. Kaplan. I will skip discussing the idea of Kaplan and jump to the solution. One Dirac operator which satisfies the Ginsparg-Wilson relation, called (for historical reasons) the "overlap operator," is

$$D = 1 - \frac{A}{[A^\dagger A]^{1/2}}\,, \qquad A \equiv 1 - D_W\,, \qquad D_W = \text{Wilson Hamiltonian}\,, \tag{125}$$

where the Wilson Hamiltonian was discussed in subsection 6.3, see particularly Eq. (96).

To see the magic of this combination, consider the infrared behavior, and imagine that D_W has gotten an additive mass renormalization that we didn't want, that is,

$$D_W \sim m + \not{\partial}\,, \quad \text{so} \quad A \simeq 1 - m - \not{\partial}\,. \tag{126}$$

The derivative only appears in $A^\dagger A$ at quadratic order; if we work to first order in derivatives (remember, we are asking about IR behavior), we find $A^\dagger A \simeq (1-m)^2$, so

$$\frac{A}{[A^\dagger A]^{1/2}} \simeq \frac{1-m}{1-m} - \frac{1}{1-m}\not{\partial}\,; \qquad 1 - \frac{A}{[A^\dagger A]^{1/2}} \simeq \frac{1}{1-m}\not{\partial}\,. \tag{127}$$

The renormalization of the kinetic term is harmless–it can be undone by rescaling ψ. The mass correction, though, has disappeared.

The flip side is that the overlap operator involves the inverse of an operator. It isn't obvious that it will be local. It will *not* be local in the sense of $D(x-y)$ vanishing beyond some finite range; it *will* be local, for vacuum gauge fields, in the sense of $D(x-y)$ vanishing as an exponential of $(x-y)$, so that as the lattice spacing is taken to zero, the operator becomes exactly local. To see that this is true, look at the Fourier transform of D, which turns out to be

$$\tilde{D}(p) = 1 - \frac{1 - i\gamma^\mu \tilde{p}_\mu - \frac{1}{2}\tilde{p}^2}{\sqrt{1 + \frac{1}{2}\sum_{\mu<\nu}(\tilde{p}_m u)^2 (\tilde{p}_\nu)^2}}\,. \tag{128}$$

This function is analytic everywhere within the Brillouin zone; therefore its Fourier transform, which is the spatial dependence, shows exponential tails. Exponential locality is sufficient for most purposes.

What about locality when the gauge fields are turned on? This is no longer obvious, since $A^\dagger A$ contains gauge fields, and must be inverted to get D. It has been proven by Hernandes, Jansen, and Lüscher, that for smooth configurations–in the sense of subsection 7.4, with a rather severe restriction on the action per plaquette–that D remains exponentially local. But exponential locality will *fail* for sufficiently "rough" lattice gauge fields. This is no surprise, since we expect a chirally symmetric lattice fermion definition to identify the exact zero modes of an instanton, and we have already seen that nonlocality on "rough" configurations is a precondition to being able to do this. It is also not a fundamental obstacle in an asymptotically free theory, as discussed in the section on topology.

9.2 An exact invariance

It follows trivially from the Ginsparg Wilson relation that the lattice action,

$$\bar{\psi} D \psi\,, \tag{129}$$

has an exact invariance under the infinitesimal transformation,

$$\begin{aligned} \delta\bar{\psi} &= \bar{\psi}\gamma^5\,, \\ \delta\psi &= \gamma^5(1-D)\psi \equiv \hat{\gamma}^5\psi\,. \end{aligned} \tag{130}$$

This exact invariance is enough to give the desired properties of a chirally symmetric theory, such as the vanishing of additive mass renormalization and of dimension 5, chiral symmetry breaking operators. However, the transformation under this symmetry looks different

between ψ and $\bar{\psi}$. Therefore, the *measure* of the path integral is not necessarily invariant under this transformation. In fact, it is not; if we ask about the expectation value of an operator $\mathcal{O}$ under the fermionic part of the action,

$$\langle \mathcal{O} \rangle \equiv \int d\bar{\psi} d\psi \; \mathcal{O} \; e^{-\bar{\psi} D \psi} \,, \tag{131}$$

its variation turns out to be

$$\langle \delta \mathcal{O} \rangle = -a \mathrm{Tr} \, \{\gamma^5 D\} \langle \mathcal{O} \rangle \,. \tag{132}$$

The free $\gamma^5 D$ is traceless; but at nonzero gauge field, it turns out that $\mathrm{Tr}\, \gamma^5 D$ is determined by the index of D, that is, the number of chiral zero modes; and that this equals (twice) the topological number. This is exactly what is demanded by the ABJ anomaly. This ability of D to measure topology is expected of a Dirac operator with an exact chiral invariance, and as discussed in the section on topology, it tells us that D had to be a nonlocal operator on sufficiently "rough" fields.

We see that "ordinary" chiral transformations of ψ are not a symmetry of the theory; but a modified chiral transformation is an exact symmetry, except under the measure. Its failure under the measure gives exactly the axial (ABJ) anomaly. This is Fujikawa's way of seeing the origin of the anomaly.

9.3 Chiral theories?

An exact chiral invariance whets our appetite to write down a chiral theory. Define projection operators,

$$P_\pm \equiv \frac{1 \pm \gamma^5}{2} \qquad \hat{P}_\pm \equiv \frac{1 \pm \hat{\gamma}^5}{2} \,, \qquad \hat{\gamma}^5 \equiv \gamma^5 (1 - D) \;\; \text{as before} \,. \tag{133}$$

The latter is also a projection operator, because

$$\hat{\gamma}^5 \hat{\gamma}^5 = \gamma^5 (1-D) \gamma^5 (1-D) = \gamma^5 \gamma^5 - \gamma^5 (D\gamma^5 + \gamma^5 D - D\gamma^5 D) = \gamma^5 \gamma^5 = 1 \,, \tag{134}$$

using the Ginsparg Wilson relation in the next to last equality.

Then, since $\bar{\psi} P_+$ only talks to $\hat{P}_- \psi$ and $\bar{\psi} P_-$ only talks to $\hat{P}_+ \psi$, why don't we just throw out $\bar{\psi} P_-$ and $\hat{P}_+ \psi$ in the integration, and only integrate over $\bar{\psi} P_+$ and $\hat{P}_- \psi$?

That will be fine, *IF* we can figure out what the measure of the path integration should be. The measure depends on the gauge fields in a nontrivial way, because $\hat{P}_\pm$ do, and $P_\pm$ do not; so there is not a cancellation of the gauge field dependence between the $\mathcal{D}\bar{\psi}$ and $\mathcal{D}\psi$ integration measures.

The problem is that the measure has a gauge field dependent phase, which cannot in general be determined in an unambiguous way. A quantum operator ψ is the contraction of a set of Grassman variables c_i with a set of spinors u_i;

$$\psi(x) = \sum_i c_i(x) u_i(x) \,. \tag{135}$$

The set of all ψ is a sum over an index α, which ranges over location and spinorial index;

$$\psi = \sum_\alpha c_\alpha u_\alpha \,. \tag{136}$$

Under a unitary change of variables,

$$u \to u\mathcal{U}^{-1}\,, \quad c \to \mathcal{U}c \tag{137}$$

the Grassman determinant $\mathrm{Det}\, D$, will get rotated by the determinant of the unitary transformation,

$$\mathrm{Det}\, D \to \mathrm{Det}\,\mathcal{U}\ \mathrm{Det}\, D\,. \tag{138}$$

Therefore there is a phase which depends on our choice of "canonical" spinors u_i in terms of which we write the path integral measure.

Normally we don't worry about this, because, if we perform a rotation on ψ, we should perform the same rotation on $\bar{\psi}$, and the phase will be opposite between them and will cancel. Now, however, we have to perform a unitary transformation on ψ, into the basis where $\hat{P}_-$ is diagonal; and we do *not* want to perform the same transformation on P_+, since it will *not* be diagonal in that basis. Therefore, we pick up a phase. If we wanted a chiral theory with no gauge couplings, this again would not be a problem, as the phase would be common and would factor out of the path integral. Now, however, the phase is gauge field dependent, since $\hat{\gamma}^5$ is. It is not obvious, whether there is an unambiguous–or even sensible–way to choose this gauge field dependent phase.

If we are perverse, we can view a vectorlike but massless theory as a chiral theory which happens to have an equal number of right and left handed degrees of freedom. We are sure that such theories exist. The key, in such theories, is that this phase is exactly the opposite between the right and left handed species of fermions, and so it cancels between species. This suggests that chiral theories can be constructed, but only if the phases in the definition of the fermion measure cancel between species. Martin Lüscher has shown that, to all orders in perturbation theory, this is exactly what happens, and that the criterion that the phase ambiguity vanishes between species, is precisely the condition that the theory is free of gauge anomalies. This means that the lattice can be used as an all-orders regularization of chiral theories satisfying anomaly cancellation.

One expects the nonperturbative analysis to be more challenging; for instance, Witten has shown that SU(2) theory with a single chiral fermion in the fundamental representation is anomalous and cannot exist as a theory, even though it has vanishing perturbative anomalies. Lüscher has made some progress towards nonperturbative construction of chiral theories; he has shown the existence of certain abelian chiral theories, nonperturbatively. However no results exist yet for nonabelian theories, which are more interesting because of asymptotic freedom. This is an open and very interesting problem in lattice gauge theory.