Confinement, chiral symmetry, and the lattice

Michael Creutz [1]
Brookhaven National Laboratory
Upton, NY 11973

Two crucial properties of QCD, confinement and chiral symmetry breaking, cannot be understand within the context of conventional Feynman perturbation theory. Non-perturbative phenomena enter the theory in a fundamental way at both the classical and quantum level. Over they years a coherent qualitative picture of the interplay between chiral symmetry, quantum mechanical anomalies, and the lattice has emerged and is reviewed here.

PACS: 11.30.Rd, 12.39.Fe, 11.15.Ha, 11.10.Gh

Contents

1 QCD **4**
1.1 Why quarks 4
1.2 Gluons and confinement 5
1.3 Perturbation theory is not enough 7

2 Path integrals and statistical mechanics **10**
2.1 Discretizing time 10
2.2 The transfer matrix 11

3 Quark fields and Grassmann integration **13**
3.1 Fermionic transfer matrices 15
3.2 Normal ordering and path integrals 16

4 Lattice gauge theory **18**
4.1 Link variables 19
4.2 Group Integration 20
4.3 Gauge invariance 22
4.4 Numerical simulation 23
4.5 Order parameters 24

[1] E-mail address: creutz@bnl.gov

5 Renormalization and the continuum limit **25**
5.1 Coupling constant renormalization . . . 25
5.2 A parameter free theory . . . 27
5.3 Including quark masses . . . 28
5.4 Which beta function? . . . 30
5.5 Flows and irrelevant operators . . . 32

6 Classical gauge fields and topology **34**
6.1 Surface terms . . . 36
6.2 An explicit solution . . . 37
6.3 Zero modes and the Dirac operator . . . 38
6.4 The index theorem . . . 39

7 Chiral symmetry **41**
7.1 Effective potentials . . . 42
7.2 Goldstone Bosons . . . 43
7.3 Pions and spontaneous symmetry breaking . . . 44

8 The chiral anomaly **47**
8.1 What broke the symmetry? . . . 47
8.2 A discrete chiral symmetry . . . 49
8.3 The ’t Hooft vertex . . . 52
8.4 Fermions in higher representations . . . 54

9 Massive quarks and the Theta parameter **56**
9.1 Twisted tilting . . . 56
9.2 Odd N_f . . . 57
9.3 Quark scattering and mass mixing . . . 59
9.4 General masses in two flavor QCD . . . 63
9.5 The strong CP problem and the up quark mass . . . 64
9.6 The two flavor phase diagram . . . 65

10 Lattice fermions **67**
10.1 Hopping and doublers . . . 68
10.2 Wilson Fermions . . . 70
10.3 Lattice versus continuum parameters . . . 73
10.4 Artifacts and the Aoki phase . . . 73
10.5 Twisted mass . . . 76

11 Lattice actions preserving chiral symmetry **79**
11.1 The Nielsen-Ninomiya theorem . . . 79
11.2 Minimal doubling . . . 80
11.3 Domain wall and overlap fermions . . . 83
11.4 Staggered fermions . . . 87
11.5 The rooting issue . . . 88

12 Other issues **90**
12.1 Quantum fluctuations and topology 90
12.2 The standard model 93
12.3 Where is the parity violation? 94
12.4 A lattice model 95

13 Final remarks **97**

1 QCD

Quarks interacting with non-Abelian gauge fields are now widely accepted as the basis the strong nuclear force. This quantum field theory is known under the somewhat whimsical name of Quantum Chromodynamics, or QCD[2]. This system is remarkable in its paucity of parameters. Once the overall scale is set, perhaps by working in units where the proton mass is unity, the only remaining parameters are the quark masses. The quarks represent a new level of substructure within hadronic particles such as the proton.

The viability of this picture relies on some rather unusual features. These include confinement, the inability to isolate a single quark, and the spontaneous breaking of chiral symmetry, needed to explain the lightness of the pions relative to other hadrons. The study of these phenomena requires the development of techniques that go beyond traditional Feynman perturbation theory. Here we concentrate on the interplay of two of these, lattice gauge theory and effective chiral models.

The presentation is meant to be introductory. The aim is to provide a qualitative picture of how the symmetries of this theory work together rather than to present detailed methods for calculation. In this first section we briefly review why this theory is so compelling.

1.1 Why quarks

Although an isolated quark has not been seen, we have a variety of reasons to believe in the reality of quarks as the basis for this next layer of matter. First, quarks provide a rather elegant explanation of certain regularities in low energy hadronic spectroscopy. Indeed, it was the successes of the eightfold way [1] which originally motivated the quark model. Two "flavors' of low mass quarks lie at the heart of isospin symmetry in nuclear physics. Adding the somewhat heavier "strange" quark gives the celebrated multiplet structure described by representations of the group $SU(3)$.

Second, the large cross sections observed in deeply inelastic lepton-hadron scattering point to structure within the proton at distance scales of less than 10^{-16} centimeters, whereas the overall proton electromagnetic radius is of order 10^{-13} centimeters [2]. Furthermore, the angular dependences observed in these experiments indicate that any underlying charged constituents carry half-integer spin.

Yet a further piece of evidence for compositeness lies in the excitations of the low-lying hadrons. Particles differing in angular momentum fall neatly into place along the famous "Regge trajectories" [3]. Families of states group together as orbital excitations of an underlying extended system. The sustained rising of these trajectories with increasing angular momentum points toward strong long-range forces between the constituents.

Finally, the idea of quarks became incontrovertible with the discovery of heavier quark species beyond the first three. The intricate spectroscopy of the charmonium and upsilon families is admirably explained via potential models for non-relativistic bound states. These systems represent what are sometimes thought of as the "hydrogen atoms" of elementary particle physics. The fine details of their structure provides a major testing ground for quantitative predictions from lattice techniques.

[2] If you prefer not to confuse this with the 4000 Angstroms typical of color, you could regard this as an acronym for Quark Confining Dynamics.

1.2 Gluons and confinement

Despite its successes, the quark picture raises a variety of puzzles. For the model to work so well, the constituents should not interact so strongly that they loose their identity. Indeed, the question arises whether it is possible to have objects display point-like behavior in a strongly interacting theory. The phenomenon of asymptotic freedom, discussed in more detail later, turns out to be crucial to realizing this picture.

Perhaps the most peculiar aspect of the theory relates to the fact that an isolated quark has never been observed. These basic constituents of matter do not copiously appear as free particles emerging from high energy collisions. This is in marked contrast to the empirical observation in hadronic physics that anything which can be created will be. Only phenomena prevented by known symmetries are prevented. The difficulty in producing quarks has led to the concept of a principle of exact confinement. Indeed, it may be simpler to have a constituent which can never be produced than an approximate imprisonment relying on an unnaturally small suppression factor. This is particularly true in a theory like the strong interactions, which is devoid of any large dimensionless parameters.

But how can one ascribe any reality to an object which cannot be produced? Is this just some sort of mathematical trick? Remarkably, gauge theories potentially possess a simple physical mechanism for giving constituents infinite energy when in isolation. In this picture a quark-antiquark pair experiences an attractive force which remains non-vanishing even for asymptotically large separations. This linearly rising long distance potential energy forms the basis of essentially all models of quark confinement.

For a qualitative description of the mechanism, consider coupling the quarks to a conserved "gluo-electric" flux. In usual electromagnetism the electric field lines thus produced spread and give rise to the inverse square law Coulombic field. If one can somehow eliminate massless fields, then a Coulombic spreading will no longer be a solution to the field equations. If in removing the massless fields we do not destroy the Gauss law constraint that the quarks are the sources of electric fields, the electric lines must form into tubes of conserved flux, schematically illustrated in Fig. 1. These tubes begin and end on the quarks and their antiparticles. The flux tube is meant to be a real physical object carrying a finite energy per unit length. This is the storage medium for the linearly rising inter-quark potential. In some sense the reason we cannot have an isolated quark is the same as the reason that we cannot have a piece of string with only one end. In this picture a baryon would require a string with three ends. It lies in the group theory of non-Abelian gauge fields that this peculiar state of affairs is allowed.

Of course a length of real string can break into two, but then each piece has itself two ends. In the QCD case a similar phenomenon occurs when there is sufficient energy in the flux tube to create a quark-antiquark pair from the vacuum. This is qualitatively what happens when a rho meson decays into two pions.

One model for this phenomenon is a type II superconductor containing magnetic monopole impurities. Because of the Meissner effect [4], a superconductor does not admit magnetic fields. However, if we force a hypothetical magnetic monopole into the systems, its lines of magnetic flux must go somewhere. Here the role of the "gluo-electric" flux is played by the magnetic field, which will bore a tube of normal material through the superconductor until it either ends on an anti-monopole or it leaves the boundary of the system [5]. Such flux tubes have been experimentally observed in real superconductors [6].

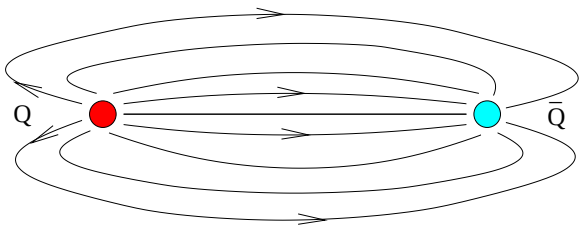

Fig. 1. A tube of gluonic flux connects quarks and anti-quarks. The strength of this string is 14 tons.

Another example of this mechanism occurs in the bag model [7]. Here the gluonic fields are unrestricted in the bag-like interior of a hadron, but are forbidden by *ad hoc* boundary conditions from extending outside. In attempting to extract a single quark from a proton, one would draw out a long skinny bag carrying the gluo-electric flux of the quark back to the remaining constituents.

The above models may be interesting phenomenologically, but they are too arbitrary to be considered as the basis for a fundamental theory. In their search for a more elegant approach, theorists have been drawn to non-Abelian gauge fields [8]. This dynamical system of coupled gluons begins in analogy with electrodynamics with a set of massless gauge fields interacting with the quarks. Using the freedom of an internal symmetry, the action also includes self-couplings of the gluons. The bare massless fields are all charged with respect to each other. The confinement conjecture is that this input theory of massless charged particles is unstable to a condensation of the vacuum into a state in which only massive excitations can propagate. In such a medium the gluonic flux around the quarks should form into the flux tubes needed for linear confinement. While this has never been proven analytically, strong evidence from lattice gauge calculations indicates that this is indeed a property of these theories.

The confinement phenomenon makes the theory of the strong interactions qualitatively rather different from the theories of the electromagnetic and weak forces. The fundamental fields of the Lagrangian do not manifest themselves in the free particle spectrum. Physical particles are all gauge singlet bound states of the underlying constituents. In particular, an expansion about the free field limit is inherently crippled at the outset. This is perhaps the prime motivation for the lattice approach.

In the quark picture, baryons are bound states of three quarks. Thus the gauge group should permit singlets to be formed from three objects in the fundamental representation. This motivates the use of $SU(3)$ as the underlying group of the strong interactions. This internal symmetry must not be confused with the broken $SU(3)$ represented in the multiplets of the eightfold way. Ironically, one of the original motivations for quarks has now become an accidental symmetry, arising only because three of the quarks are fairly light. The gauge symmetry of importance to us now is hidden behind the confinement mechanism, which only permits observation of singlet states.

The presentation here assumes, perhaps too naively, that the nuclear interactions can be considered in isolation from the much weaker effects of electromagnetism, weak interactions, and gravitation. This does not preclude the possible application of the techniques to the other interactions. Indeed, unification may be crucial for a consistent theory of the world. To describe physics

at normal laboratory energies, however, it is only for the strong interactions that we are forced to go beyond well-established perturbative methods. Thus we frame the discussion around quarks and gluons.

1.3 Perturbation theory is not enough

The best evidence we have for confinement in a non-Abelian gauge theory comes by way of Wilson's [9, 10] formulation on a space time lattice. At first this prescription seems a little peculiar because the vacuum is not a crystal. Indeed, experimentalists work daily with highly relativistic particles and see no deviations from the continuous symmetries of the Lorentz group. Why, then, have theorists spent so much time describing field theory on the scaffolding of a space-time lattice?

The lattice should be thought of as a mathematical trick. It provides a cutoff removing the ultraviolet infinities so rampant in quantum field theory. On a lattice it makes no sense to consider momenta with wavelengths shorter than the lattice spacing. As with any regulator, it must be removed via a renormalization procedure. Physics can only be extracted in the continuum limit, where the lattice spacing is taken to zero. As this limit is taken, the various bare parameters of the theory are adjusted while keeping a few physical quantities fixed at their continuum values.

But infinities and the resulting need for renormalization have been with us since the beginnings of relativistic quantum mechanics. The program for electrodynamics has had immense success without recourse to discrete space. Why reject the time-honored perturbative renormalization procedures in favor of a new cutoff scheme?

Perturbation theory has long been known to have shortcomings in quantum field theory. In a classic paper, Dyson [11] showed that electrodynamics could not be analytic in the coupling around vanishing electric charge. If it were, then one could smoothly continue to a theory where like charges attract rather than repel. This would allow creating large separated regions of charge to which additional charges would bind with more energy than their rest masses. This would mean there is no lowest energy state; creating matter-antimatter pairs and separating them into these regions would provide an inexhaustible source of free energy.

The mathematical problems with perturbation theory appear already in the trivial case of zero dimensions. Consider the toy path integral

$$Z(m,g) = \int d\phi \, \exp(-m^2\phi^2 - g\phi^4). \tag{1}$$

Formally expanding and naively exchanging the integral with the sum gives

$$Z(m,g) = \sum c_i g^i \tag{2}$$

with

$$c_i = \frac{(-1)^i}{i!} \int d\phi \, e^{-m^2\phi^2} \phi^{4i} = \frac{(-1)^i (4i)!}{m^{2i+1} i!}. \tag{3}$$

A simple application of Sterling's approximation shows that at large order these coefficients grow faster than any power. Given any value for g, there will always be an order in the series where the terms grow out of control. Note that by scaling the integrand we can write

$$Z(m,g) = g^{-1/4} \int d\phi \, \exp(-m^2\phi^2/g^{-1/2} - \phi^4). \tag{4}$$

This explicitly exposes a branch cut at the origin, yet another way of seeing the non analyticity at vanishing coupling.

Thinking non-perturbatively sometimes reveals somewhat surprising results. For example, the ϕ^3 theory of massive scalar bosons coupled with a cubic interaction seems to have a sensible perturbative expansion after renormalization. However this theory almost certainly doesn't exist as a quantum system. This is because when the field becomes large the cubic term in the interaction dominates and the theory has no minimum energy state. The Euclidean path integral is divergent from the outset since the action is unbounded both above and below.

Perhaps even more surprising, it is widely accepted, although not proven rigorously, that a ϕ^4 theory of bosons interacting with a quartic interaction also does not have a non-trivial continuum limit. The expectation here is that with a cutoff in place, the renormalized coupling will display an upper bound as the bare coupling varies from zero to infinity. If this upper bound then decreases to zero as the cutoff is removed, then the renormalized coupling is driven to zero and we have a free theory.

This issue is sometimes discussed in terms of what is known as the "Landau pole" [12]. In non-asymptotically free theories, such as ϕ^4 and quantum electrodynamics, there is a tendency for the effective coupling to rise with energy. A simple analysis suggests the possibility of the coupling diverging at a finite energy. Not allowing this would force the coupling at smaller energies to zero.

The importance of non-perturbative effects is well understood in a class of two dimensional models that can be solved via a technique known as "bosonization" [13]. This includes massless two dimensional electrodynamics, i.e. the Schwinger model [14], the sine-Gordon model [15], and the Thirring model [16]. These solutions exploit a remarkable mapping between fermionic and bosonic fields in two dimensions. This mapping is also closely related to the solution to the two dimensional Ising model [17].

Returning to the main problem, QCD, we are driven to the lattice by the necessary prevalence of non-perturbative phenomena in the strong interactions. Most predominant of these is confinement, but issues related to chiral symmetry and quantum mechanical anomalies, to be discussed in later sections, are also highly non-perturbative. The theory at vanishing coupling constant has free quarks and gluons and bears no resemblance to the observed physical world of hadrons. Renormalization group arguments, discussed later, explicitly demonstrate essential singularities when hadronic properties are regarded as functions of the gauge coupling. To go beyond the diagrammatic approach, one needs a non-perturbative cutoff. Herein lies the main virtue of the lattice, which directly eliminates all wavelengths less than the lattice spacing. This occurs before any expansions or approximations are begun.

This situation contrasts sharply with the great successes of quantum electrodynamics, where perturbation theory is central. Most conventional regularization schemes are based on the Feynman expansion; some process is calculated is calculated diagrammatically until a divergence is met, and the offending diagram is regulated. Since the basic coupling is so small, only a few terms give good agreement with experiment. While non-perturbative effects are expected, their magnitude is exponentially suppressed in the inverse of the coupling.

On a lattice, a field theory becomes mathematically well-defined and can be studied in various ways. Conventional perturbation theory, although somewhat awkward in the lattice framework, should recover all conventional results of other regularization schemes. Discrete space-time, however, allows several alternative approaches. One of these, the strong coupling expansion,

is straightforward to implement. Remarkably, confinement is automatic in the strong coupling limit because the theory reduces to one of quarks on the end of strings with finite energy per unit length. While this realization of the flux tube picture provides insight into how confinement can work, unfortunately this limit is not the continuum limit. The latter, as we will see later, involves the weak coupling limit. To study this one can turn to numerical simulations, made possible by the lattice reduction of the path integral to a conventional but large many-dimensional integral.

Non-perturbative effects in QCD introduce certain interesting aspects that are invisible to perturbation theory. Most famous of these is the possibility of having an explicit CP violating term in the theory. In the classical theory this involves adding a total derivative term to the action. This can be rotated away in the perturbative limit. However, as we will discuss extensively later, in the quantum theory there are dramatic physical consequences.

Non-perturbative effects also raise subtle questions on the meaning of quark masses. Ordinarily the mass of a particle is correlated with how it propagates over long distances. This approach fails due to confinement and the fact that a single quark cannot be isolated. With multiple quarks, we will also see that there is a complicated dependence of the theory on the number of quark species. As much of our understanding of quantum field theory is based on perturbation theory, several of these effects remain controversial.

This picture has evolved over many years. One unusual result is that, depending on the parameters of the theory, QCD can spontaneously break CP symmetry. This is tied to what is known as Dashen's phenomenon [18], first noted even before the days of QCD. In the mid 1970's, 't Hooft [19] elucidated the underlying connection between the chiral anomaly and the topology of gauge fields. This connection revealed the possible explicit CP violating term, usually called Θ, the dependence on which does not apper in perturbative expansions. Later Witten [20] used large gauge group ideas to discuss the behavior on Θ in terms of effective Lagrangians. Refs. [21–23] represent a few of the early studies of the effects of Θ on effective Lagrangians via a mixing between quark and gluonic operators. The topic continues to appear in various contexts; for example, Ref. [24] contains a different approach to understanding the behavior of QCD at $\Theta = \pi$ via the framework of the two-flavor Nambu Jona-Lasinio model.

All these issues are crucial to understanding certain subtleties with formulating chiral symmetry on the lattice. Much of the picture presented here is implicit in the discussion of Ref. [25]. Since then the topic has raised some controversial issues, including the realization that the ambiguities in defining quark masses precludes a vanishing up quark mass as a solution to the strong CP problem [26]. The non-analytic behavior in the number of quark species reveals an inconsistency with one of the popular algorithms in lattice gauge theory [27]. These conclusions directly follow from the intricate interplay of the anomaly with chiral symmetry. The fact that some of these issues remain disputed is much of the motivation for this review.

The discussion here is based on a few reasonably uncontroversial assumptions. First, QCD with N_f light quarks should exist as a field theory and exhibit confinement in the usual way. Then we assume the validity of the standard picture of chiral symmetry breaking involving a quark condensate $\langle\overline{\psi}\psi\rangle \neq 0$. The conventional chiral perturbation theory based on expanding in masses and momenta around the chiral limit should make sense. We assume the usual result that the anomaly generates a mass for the η' particle and this mass survives the chiral limit. Throughout we consider N_f is taken small enough to avoid any potential conformal phase of QCD [28].

2 Path integrals and statistical mechanics

Throughout this review we will be primarily focussed on the Euclidean path integral formulation of QCD. This approach to quantum mechanics reveals deep connections with classical statistical mechanics. Here we will explore this relationship for the simple case of a non-relativistic particle in a potential. Starting with a partition function representing a path integral on an imaginary time lattice, we will see how a transfer matrix formalism reduces the problem to the diagonalization of an operator in the usual quantum mechanical Hilbert space of square integrable functions [29]. In the continuum limit of the time lattice, we obtain the canonical Hamiltonian. Except for our use of imaginary time, this treatment is identical to that in Feynman's early work [30].

2.1 Discretizing time

We begin with the Lagrangian for a free particle of mass m moving in potential $V(x)$

$$L(x, \dot{x}) = K(\dot{x}) + V(x) \tag{5}$$

where $K(\dot{x}) = \frac{1}{2}m\dot{x}^2$ and $\dot{x}$ is the time derivative of the coordinate x. Note the unconventional relative positive sign between the two terms in Eq. (5). This is because we formulate the path integral directly in imaginary time. This improves mathematical convergence, yet we are left with the usual Hamiltonian for diagonalization.

For a particle traversing a trajectory $x(t)$, we have the action

$$S = \int dt\ L(\dot{x}(t), x(t)). \tag{6}$$

This appears in the path integral

$$Z = \int (dx) e^{-S}. \tag{7}$$

Here the integral is over all possible trajectories $x(t)$. As it stands, Eq. (7) is rather poorly defined. To characterize the possible trajectories we introduce a cutoff in the form of a time lattice. Putting our system into a temporal box of total length T, we divide this interval into $N = \frac{T}{a}$ discrete time slices, where a is the timelike lattice spacing. Associated with the i'th such slice is a coordinate x_i. This construction is sketched in Figure 2. Replacing the time derivative of x with a nearest-neighbor difference, we reduce the action to a sum

$$S = a \sum_i \left[\frac{1}{2} m \left(\frac{x_{i+1} - x_i}{a} \right)^2 + V(x_i) \right]. \tag{8}$$

The integral in Eq. (7) is now defined as an ordinary integral over all the coordinates

$$Z = \int \left(\prod_i dx_i \right)\ e^{-S}. \tag{9}$$

This form for the path integral is precisely in the form of a partition function for a statistical system. We have a one dimensional polymer of coordinates x_i. The action represents the inverse

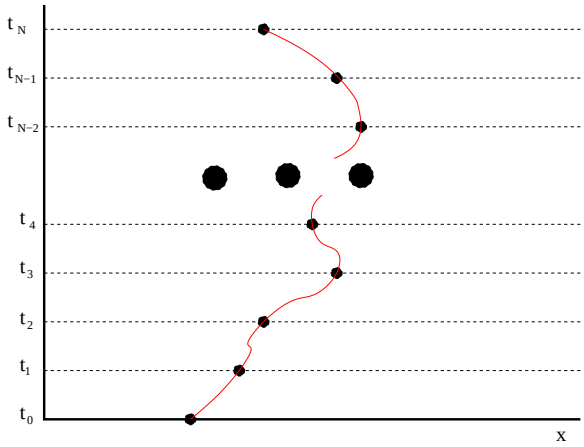

Fig. 2. Dividing time into a lattice of N slices of timestep a.

temperature times the Hamiltonian of the thermal analog. This is a special case of a deep result, a D space-dimensional quantum field theory is equivalent to the classical thermodynamics of a $D+1$ dimensional system. In this example, we have one degree of freedom and D is zero; for the lattice gauge theory of quarks and gluons, D is three and we work with the classical statistical mechanics of a four dimensional system.

We will now show that the evaluation of this partition function is equivalent to diagonalizing a quantum mechanical Hamiltonian obtained from the action via canonical methods. This is done with the use of the transfer matrix.

2.2 The transfer matrix

The key to the transfer-matrix analysis is to note that the local nature of the action permits us to write the partition function as a matrix product

$$Z = \int \prod_i dx_i \, T_{x_{i+1},x_i} \tag{10}$$

where the transfer-matrix elements are

$$T_{x',x} = \exp\left[-\frac{m}{2a}(x'-x) - \frac{a}{2}(V(x')+V(x))\right]. \tag{11}$$

The transfer matrix itself is an operator in the Hilbert space of square integrable functions with the standard inner product

$$\langle\psi'|\psi\rangle = \int dx \psi'^*(x)\psi(x). \tag{12}$$

We introduce the non-normalizable basis states $|x\rangle$ such that

$$|\psi\rangle = \int dx \; \psi(x) \; |x\rangle \tag{13}$$

$$\langle x'|x\rangle = \delta(x'-x) \tag{14}$$

$$1 = \int dx\, |x\rangle\langle x|. \tag{15}$$

Acting on the Hilbert space are the canonically conjugate operators $\hat{p}$ and $\hat{x}$ that satisfy

$$\begin{aligned} &\hat{x}|x\rangle = x|x\rangle \\ &[\hat{p},\hat{x}] = -i \\ &e^{-ipy}|x\rangle = |x+y\rangle. \end{aligned} \tag{16}$$

The operator T is defined via its matrix elements

$$\langle x'|T|x\rangle = T_{x',x} \tag{17}$$

where $T_{x',x}$ is given in Eq. (11). With periodic boundary conditions on our lattice of N sites, the path integral is compactly expressed as as a trace over the Hilbert space

$$Z = \mathrm{Tr}\, T^N. \tag{18}$$

Expressing T in terms of the basic operators $\hat{p}, \hat{x}$ gives

$$T = \int dy\, e^{-y^2/(2a)}\; e^{-aV(\hat{x})/2}\; e^{-i\hat{p}y}\; e^{-aV(\hat{x})/2}. \tag{19}$$

To prove this, check that the right hand side has the appropriate matrix elements. The integral over y is Gaussian and gives

$$T = \left(\frac{2\pi a}{m}\right)^{1/2}\; e^{-aV(\hat{x})/2} e^{-a\hat{p}^2/(2m)} e^{-aV(\hat{x})/2}. \tag{20}$$

The connection with the usual quantum mechanical Hamiltonian appears in the small lattice spacing limit. When a is small, the exponents in the above equation combine to give

$$T = \left(\frac{2\pi a}{m}\right)^{1/2}\; e^{-aH+O(a^2)} \tag{21}$$

with

$$H = \frac{\hat{p}^2}{2m} + V(\hat{x}). \tag{22}$$

This is just the canonical Hamiltonian operator following from our starting Lagrangian.

The procedure for going from a path-integral to a Hilbert-space formulation of quantum mechanics consists of three steps. First define the path integral with a discrete time lattice. Then construct the transfer matrix and the Hilbert space on which it operates. Finally, take the logarithm of the transfer matrix and identify the negative of the coefficient of the linear term in the lattice spacing as the Hamiltonian. Physically, the transfer matrix propagates the system from one time slice to the next. Such time translations are generated by the Hamiltonian.

The eigenvalues of the transfer matrix are related to the energy levels of the quantum system. Denoting the i'th eigenvalue of T by λ_i, the path integral or partition function becomes

$$Z = \sum_i \lambda_i^N. \tag{23}$$

As the number of time slices goes to infinity, this expression is dominated by the largest eigenvalue λ_0

$$Z = \lambda_0^N \times [1 + O(\exp[-N \log(\lambda_0/\lambda_1)])]. \tag{24}$$

In statistical mechanics the thermodynamic properties of a system follow from this largest eigenvalue. In ordinary quantum mechanics the corresponding eigenvector is the lowest eigenstate of the Hamiltonian. This is the ground state or, in field theory, the vacuum $|0\rangle$. Note that in this discussion, the connection between imaginary and real time is trivial. Whether the generator of time translations is H or iH, we have the same operator to diagonalize.

In statistical mechanics one is often interested in correlation functions between the statistical variables at different points. This corresponds to a study of the Green's functions of the corresponding field theory. These are obtained upon insertion of polynomials of the fundamental variables into the path integral.

An important feature of the path integral is that a typical path is non-differentiable [31, 32]. Consider the discretization of the time derivative

$$\dot{x} \sim \frac{x_{i+1} - x_i}{a}. \tag{25}$$

The kinetic term in the path integral controls how close the fields are on adjacent sites. Since this appears as simple Gaussian factor $\exp(-(x_{i+1} - x_i)^2 m/a)$ we see that

$$\frac{1}{2} m \langle \dot{x}^2 \rangle = O(1/ma). \tag{26}$$

This diverges as the lattice spacing goes to zero.

One can obtain the average kinetic energy in other ways, for example through the use of the virial theorem or by point splitting. However, the fact that the typical path is not differentiable means that one should be cautious about generalizing properties of classical fields to typical configurations in a numerical simulation. We will see that such questions naturally arise when considering the topological properties of gauge fields.

3 Quark fields and Grassmann integration

Of course since we are dealing with a theory of quarks, we need additional fields to represent them. There are subtle complications in defining their action on a lattice; we will go into these in some detail later. For now we just assume the quark fields ψ and $\overline{\psi}$ are associated with the sites of the lattice and carry suppressed spinor, flavor, and color indices. Being generic, we take an action which is a quadratic form in these fields $\overline{\psi}(D + m)\psi$. Here we formally separate the kinetic and mass contributions. For the path integral, we are to integrate over ψ and $\overline{\psi}$ as independent Grassmann variables. Thus ψ and $\overline{\psi}$ on any site anti-commutes with ψ and $\overline{\psi}$ on any other site.

Grassmann integration is defined formally as a linear function satisfying a shift symmetry. Consider a single Grassmann variable ψ. Given any function f of ψ, we impose

$$\int d\psi \, f(\psi) = \int d\psi \, f(\psi + \chi) \tag{27}$$

where χ is another fixed Grassmann variable. Since the square of any Grassmann variable vanishes, we can expand f in just two terms

$$f(\psi) = \psi a + b. \tag{28}$$

Assuming linearity on inserting this into Eq. (27) gives

$$\left(\int d\psi\, \psi\right) a + \left(\int d\psi\, 1\right) b = \left(\int d\psi\, \psi\right) a + \left(\int d\psi\, 1\right) (a\chi + b). \tag{29}$$

This immediately tells us $\int d\psi\, 1$ must vanish. The normalization of $\int d\psi\, \psi$ is still undetermined; the convention is to take this to be unity. Thus the basic Grassmann integral of a single variable is completely determined by

$$\int d\psi\, \psi = 1 \tag{30}$$

$$\int d\psi\, 1 = 0. \tag{31}$$

Note that the rule for Grassmann integration seems quite similar to what one would want for differentiation. Indeed, it is natural to define derivatives as anticommuting objects that satisfy

$$\frac{d}{d\psi}\, \psi = 1 \tag{32}$$

$$\frac{d}{d\psi}\, 1 = 0. \tag{33}$$

This is exactly the same rule as for integration. For Grassmann variables, integration and differentiation are the same thing. It is a convention what we call it. For the path integral it is natural to keep the analogy with bosonic fields and refer to integation. On the other hand, for both fermions and bosons we refer to differentiation when using sources in the path integral as a route to correlation functions.

We can make changes of variables in a Grassmann integration in a similar way to ordinary integrals. For example, if we want to change from ψ to $\chi = a\psi$, the above integration rules imply

$$\int d(\psi) f(a\psi) = a \int d(\chi) f(\chi) \tag{34}$$

or simply $d(a\psi) = d\chi = \frac{1}{a} d\psi$. We see that the primary difference from ordinary integration is that the Jacobean is inverted. If we consider a multiple integral and take $\chi = M\psi$ with M a matrix, the transformation generalizes to

$$d\chi = d(M\psi) = \frac{1}{\det(M)}\, d\psi. \tag{35}$$

A particularly important consequence is that we can formally evaluate the Gaussian integrals that appear in the path integral as

$$\int d\psi d\overline{\psi}\, \exp\left(\overline{\psi}(D+m)\psi\right) = \frac{1}{\det(D+m)} = \det\left((D+m)^{-1}\right). \tag{36}$$

The normalization is fixed by the earlier conventions. Note that in the path integral formulation ψ and $\overline{\psi}$ represent independent Grassmann fields; in the next subsection we will discuss the connection between these and the canonical anti-commutation relations for fermion creation and annihilation operators in a quantum mechanical Hilbert space..

In practice Eq. (36) allows one to replace fermionic integrals with ordinary commuting fields ϕ and $\overline{\phi}$ as

$$\int d\psi d\overline{\psi}\ \exp\left(\overline{\psi}(D+m)\psi\right) \propto \int d\phi d\overline{\phi}\ \exp\left(\overline{\phi}(D+m)^{-1}\phi\right). \tag{37}$$

This forms the basis for most Monte Carlo algorithms, although the intrinsic need to invert the large matrix $D+m$ makes such simulations extremely computationally intensive. This approach is, however, still much less demanding than any known way to deal directly with the Grassmann integration in path integrals [33].

3.1 Fermionic transfer matrices

The concept of continuity is lost with Grassman variables. There is no meaning to saying that fermion fields at nearby sites are near each other. This is closely tied to the doubling issues that we will discuss later. But is also raises interesting complications in relating Hamiltonian quantum mechanics with the Euclidian formulation involving path integrals. Here we will go into how this connection is made with an extremely simple zero space-dimensional model.

Anti-commutation is at the heart of fermionic behavior. This is true in both the Hamiltonian operator formalism and the Lagrangian path integral, but in rather complementary ways. Starting with a Hamiltonian approach, if an operator $a^\dagger$ creates a fermion in some normalized state on the lattice or the continuum, it satisfies the basic relation

$$[a, a^\dagger]_+ \equiv aa^\dagger + a^\dagger a = 1. \tag{38}$$

This contrasts sharply with the fields in a path integral, which all anti-commute

$$[\chi, \chi^\dagger]_+ = 0. \tag{39}$$

The connection between the Hilbert space approach and the path integral appears through the transfer matrix formalism. For bosonic fields this is straightforward [29], but for fermions certain subtleties arise related to the doubling issue [34].

To be more precise, consider a single fermion state created by the operator $a^\dagger$, and an antiparticle state created by another operator $b^\dagger$. For an extremely simple model, consider the Hamiltonian

$$H = m(a^\dagger a + b^\dagger b). \tag{40}$$

Here m can be thought of as a "mass" for the particle. What we want is an exact path integral expression for the partition function

$$Z = \mathrm{Tr} e^{-\beta H}. \tag{41}$$

Of course, since the Hilbert space generated by a and b has only four states, this is trivial to work out: $Z = 1 + 2e^{-\beta m} + e^{-2\beta m}$. However, we want this in a form that easily generalizes to many variables.

The path integral for fermions uses Grassmann variables. We introduce a pair of such, χ and $\chi^\dagger$, which will be connected to the operator pair a and $a^\dagger$, and another pair, ξ and $\xi^\dagger$, for b, $b^\dagger$. All the Grassmann variables anti-commute. Integration over any of them is determined by the simple formulas mentioned earlier

$$\int d\chi\, 1 = 0\,; \qquad \int d\chi\, \chi = 1. \tag{42}$$

For notational simplicity combine the individual Grassmann variables into spinors

$$\psi = \begin{pmatrix} \chi \\ \xi^\dagger \end{pmatrix}; \quad \psi^\dagger = (\,\chi^\dagger \quad \xi\,)\,. \tag{43}$$

To make things appear still more familiar, introduce a "Dirac matrix"

$$\gamma_0 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{44}$$

and the usual

$$\overline{\psi} = \psi^\dagger \gamma_0. \tag{45}$$

Then we have

$$\overline{\psi}\psi = \chi^\dagger \chi + \xi^\dagger \xi. \tag{46}$$

where the minus sign from using $\xi^\dagger$ rather than ξ in defining ψ is removed by the γ_0 factor. The temporal projection operators

$$P_\pm = \frac{1}{2}(1 \pm \gamma_0) \tag{47}$$

arise when one considers the fields at two different locations

$$\chi_i^\dagger \chi_j + \xi_i^\dagger \xi_j = \overline{\psi}_i P_+ \psi_j + \overline{\psi}_j P_- \psi_i. \tag{48}$$

The indices i and j will soon label the ends of a temporal hopping term; this formula is the basic transfer matrix justification for the Wilson projection operator formalism that we will return to in later sections.

3.2 Normal ordering and path integrals

For a moment ignore the antiparticles and consider some general operator $f(a, a^\dagger)$ in the Hilbert space. How is this related to an integration in Grassmann space? To proceed we need a convention for ordering the operators in f. We adopt the usual normal ordering definition with the notation $:f(a, a^\dagger):$ meaning that creation operators are placed to the left of destruction operators, with a minus sign inserted for each exchange. In this case a rather simple formula gives the trace of the operator as a Grassmann integration

$$\mathrm{Tr}\ :f(a, a^\dagger): = \int d\chi d\chi^\dagger e^{2\chi^\dagger \chi} f(\chi, \chi^\dagger). \tag{49}$$

To verify, just check that all elements of the complete set of operators $\{1, a, a^\dagger, a^\dagger a\}$ work. However, this formula is actually much more general; given a set of many Grassmann variables with one pair associated with each of several fermion states, this immediately generalizes to the trace of any normal ordered operator acting in a many fermion Hilbert space.

What about a product of several normal ordered operators? This leads to the introduction of multiple sets of Grassmann variables and the general formula

$$\begin{aligned} \mathrm{Tr} \quad & \big(: f_1(a^\dagger, a) : : f_2(a^\dagger, a) : \ldots : f_n(a^\dagger, a) :\big) \\ = \quad & \int d\chi_1 \, d\chi_1^* \ldots d\chi_n \, d\chi_n^* \, e^{\chi_1^*(\chi_1+\chi_n)} e^{\chi_2^*(\chi_2-\chi_1)} \ldots e^{\chi_n^*(\chi_n-\chi_{n-1})} \\ & \qquad \times f_1(\chi_1^*, \chi_1) f_2(\chi_2^*, \chi_2) \ldots f_n(\chi_n^*, \chi_n). \end{aligned} \tag{50}$$

The positive sign on χ_n in the first exponential factor indicates the natural occurrence of anti-periodic boundary conditions; i.e. we can define $x_0 = -x_n$. With just one factor, this formula reduces to Eq. (49). Note how the "time derivative" terms are "one sided;" this is how doubling is eluded.

This exact relationship provides the starting place for converting our partition function into a path integral. The simplicity of our example Hamiltonian allows this to be done exactly at every stage. First we break "time" into a number N of "slices"

$$Z = \mathrm{Tr}\left(e^{-\beta H/N}\right)^N. \tag{51}$$

Now we need normal ordered factors for the above formula. For this we use

$$e^{\alpha a^\dagger a} = 1 + (e^\alpha - 1) a^\dagger a = : e^{(e^\alpha - 1) a^\dagger a} : , \tag{52}$$

which is true for arbitrary α.[3] This is all the machinery we need to write

$$Z = \int (d\psi d\overline{\psi}) e^S \tag{53}$$

where

$$S = \sum_{i=1}^{n} \overline{\psi}_n (e^{-\beta m/N} - 1)\psi_n + \overline{\psi}_n P_+ \psi_{n-1} + \overline{\psi}_{n-1} P_- \psi_n. \tag{54}$$

Note how the projection factors of $P_\pm$ automatically appear for handling the reverse convention of χ versus ξ in our field ψ. Expanding the first term gives the $-\beta m/N$ factor appearing in the Hamiltonian form for the partition function.

It is important to realize that if we consider the action as a generalized matrix connecting fermionic variables

$$S = \overline{\psi} M \psi, \tag{55}$$

the matrix M is not symmetric. The upper components propagate forward in time, and the lower components backward. Even though our Hamiltonian was Hermitian, the matrix appearing in the corresponding action is not. With further interactions, such as gauge field effects, the

[3] The definition of normal ordering gives $: (a^\dagger a)^2 := 0$.

intermediate fermion contributions to a general path integral may not be positive, or even real. Of course the final partition function, being a trace of a positive definite operator, is positive. Keeping the symmetry between particles and antiparticles results in a real fermion determinant, which in turn is positive for an even number of flavors. We will later see that some rather interesting things can happen with an odd number of flavors.

For our simple Hamiltonian, this discussion has been exact. The discretization of time adds no approximations since we could do the normal ordering by hand. In general with spatial hopping or more complex interactions, the normal ordering can produce extra terms going as $O(1/N^2)$. In this case exact results require a limit of a large number of time slices, but this is a limit we need anyway to reach continuum physics.

4 Lattice gauge theory

Lattice gauge theory is currently the dominant path to understanding non-perturbative effects. As formulated by Wilson, the lattice cutoff is quite remarkable in preserving many of the basic ideas of a gauge theory. But just what is a gauge theory anyway? Indeed, there are many ways to think of what is meant by this concept.

At the most simplistic level, a Yang-Mills [8] theory is nothing but an embellishment of electrodynamics with isospin symmetry. Being formulated directly in terms of the underlying gauge group, this is inherent in lattice gauge theory from the start.

At a deeper level, a gauge theory is a theory of phases acquired by a particle as it passes through space time. In electrodynamics the interaction of a charged particle with the electromagnetic field is elegantly described by the wave function acquiring a phase from the gauge potential. For a particle at rest, this phase is an addition to its energy proportional to the scalar potential. The use of group elements on lattice links directly gives this connection; the phase associated with some world-line is the product of these elements along the path in question. For the Yang-Mills theory the concept of "phase" is generalized to a rotation in the internal symmetry group.

A gauge theory is also a theory with a local symmetry. Gauge transformations involve arbitrary functions of space time. Indeed, with QCD we have an independent $SU(3)$ symmetry at each point of space time. With the Wilson action formulated in terms of products of group elements around closed loops, this symmetry remains exact even with the cutoff in place.

In perturbative discussions, the local symmetry forces a gauge fixing to remove a formal infinity coming from integrating over all possible gauges. For the lattice formulation, however, the use of a compact representation for the group elements means that the integration over all gauges becomes finite. To study gauge invariant observables, no gauge fixing is required to define the theory. Of course gauge fixing can still be done, and must be introduced to study more conventional gauge variant quantities such as gluon or quark propagators. But physical quantities should be gauge invariant; so, whether gauge fixing is done or not is irrelevant for their calculation.

One aspect of a continuum gauge theory that the lattice does not respect exact is how a gauge field transforms under Lorentz transformations. In a continuum theory the basic vector potential can change under a gauge transformation when transforming between frames. The lattice, of course, breaks Lorentz invariance, and thus this concept loses meaning.

Here we provide only a brief introduction to the lattice approach to a gauge theory. For more details one should turn to one of the several excellent books on the subject [35–39]. We postpone until later a discussion of issues related to lattice fermions. These are more naturally understood after exploring some of the peculiarities that must be manifest in any non-perturbative formulation.

4.1 Link variables

Lattice gauge theory is closely tied to two of the above concepts; it is a theory of phases and it exhibits an exact local symmetry. Indeed it is directly defined in terms of group elements representing the phases acquired by quarks as they hop around the lattice. The basic variables are phases associated with each link of a four dimensional space time lattice. For non-Abelian case, these variables become an elements of the gauge group, i.e. $U_{ij} \in SU(3)$ for the strong interactions. Here i and j denote the sites being conneted by the link in question. We suppress the group indices to keep the notation under control. These are three by three unitary matrices satisfying

$$U_{ij} = U_{ji}^{-1} = (U_{ij})^{\dagger}. \tag{56}$$

The analogy with continuum vector fields A_μ is

$$U_{i,i+e_\mu} = e^{iag_0 A_\mu} \tag{57}$$

Here a represents the lattice spacing and g_0 is the bare coupling considered at the scale of the cutoff.

In the continuum, a non-trivial gauge field arises when the curl (in a four dimensional sense) of the potential is non zero. This in turn means the phase factor around a small closed loop is not unity. The smallest closed path in the lattice is a "plaquette," or elementary square. Consider the phase corresponding to one such

$$U_P = U_{12}U_{23}U_{34}U_{41} \tag{58}$$

where sites 1 through 4 run around the square in question. In an intuitive sense this measures the flux through this plaquette $U_P \sim \exp(iea^2 g_0^2 F_{\mu,\nu})$. This motivates using this quantity to define an action. For this, look at the real part of the trace of U_P

$$\mathrm{ReTr}U_P = N - a^4 g_0^2 F_{\mu\nu}F_{\mu\nu} + O(a^6). \tag{59}$$

The overall added constant N is physically irrelevant. This leads directly to the Wilson gauge action

$$S(U) = -\sum_P \mathrm{ReTr}U_P. \tag{60}$$

Now we have our gauge variables and an action. To proceed we turn to a path integral as an integral over all fields of the exponentiated action. For a Lie group, there is a natural measure that we will discuss shortly. Using this measure, the path integral is

$$Z = \int (dU) e^{-\beta S} \tag{61}$$

where (dU) denotes integration over all link variables. This leads to the conventional continuum expression $\frac{1}{2}\int d^4x\, F_{\mu\nu}F_{\mu\nu}$ if we choose $\beta = 2N/g_0^2$ for group $SU(N)$ and use the conventionally normalized bare coupling g_0.

Physical correlation functions are obtained from the path integral as expectation values. Given an operator $B(U)$ which depends on the link variables, we have

$$\langle B\rangle = \frac{1}{Z}\int (dU)B(U)e^{-\beta S(U)}. \tag{62}$$

Because of the gauge symmetry, discussed further later, this only makes physical sense if B is invariant under gauge transformations.

4.2 Group Integration

The above path integral involves integration over variables which are elements of the gauge group. For this we use a natural measure with a variety of nice properties. Given any function $f(g)$ of the group elements $g \in G$, the Haar measure is constructed so as to be invariant under "translation" by an arbitrary fixed element g_1 of the group

$$\int dg\; f(g) = \int dg\; f(g_1 g) \tag{63}$$

For a compact group, as for the $SU(N)$ relevant to QCD, this is conventionally normalized so that $\int dg\; 1 = 1$. These simple properties are enough for the measure to be uniquely determined.

An explicit representation for this integration measure is almost never needed, but fairly straightforward to write down formally. Suppose a general group element is parameterized by some variables $\alpha_1, ...\alpha_n$. Considering here the case $SU(N)$, there are $n = N^2 - 1$ such parameters. Then assume we know some region R in this parameter space that covers the group exactly once. Define the n dimensional fully antisymmetric tensor $\epsilon_{i_1,...i_n}$ such that, say, $\epsilon_{1,2,...n} = 1$. Now look at the integral

$$I = A\int_R \{d\alpha\}\; f(g(\vec{\alpha}))\; \epsilon_{i_1,...i_n}\mathrm{Tr}\left((g^{-1}\partial_{i_1}g)...(g^{-1}\partial_{i_n}g)\right). \tag{64}$$

This has the required invariance properties of Eq. (63). The properties of a group imply there should be a set of parameters α' depending on α such that $g_1 g(\vec{\alpha}) = g(\vec{\alpha}')$. If we change the integration variables from α to α', then the epsilon factor generates exactly the Jacobian needed for this variable change. The normalization factor A is fixed by the above condition $\int dg\; 1 = 1$. Once this is done, we have the invariant measure. The above form for the measure will appear again when we discuss topological issues for gauge fields in Section 6.

Several interesting properties of the Haar measure are easily found. If the group is compact, the left and right measures are equal

$$\int d_R g\; f(g) = \int d_R g\; f(g g_1) = \int d_L g_1 \int d_R g_2\; f(g_2 g_1) = \int d_L g\; f(g). \tag{65}$$

This also shows the measure is unique since any left invariant measure could be used. (For a non-compact group the normalization can differ.) A similar argument shows

$$\int dg\; f(g) = \int dg\; f(g^{-1}). \tag{66}$$

For a discrete group, $\int dg$ is simply a sum over the elements. For $U(1) = \{e^{i\theta}|0 \leq \theta < 2\pi\}$ the measure is simply an integral over the circle

$$\int dg\ f(g) = \int_0^{2\pi} \frac{d\theta}{2\pi} f(e^{i\theta}). \tag{67}$$

For $SU(2)$, group elements take the form

$$g = \{a_0 + i\vec{a} \cdot \vec{\sigma} | a_0^2 + \vec{a}^2 = 1\} \tag{68}$$

and the measure is

$$\int dg\ f(g) = \frac{1}{\pi^2} \int d^4a\ f(g)\delta(a^2 - 1). \tag{69}$$

In particular, $SU(2)$ is a 3-sphere.

Some integrals are easily evaluated if we realize that group integration picks out the "singlet" part of a function. Thus

$$\int dg R_{ab}(g) = 0 \tag{70}$$

where $R(g)$ is any irreducible matrix representation other than the trivial one, $R = 1$. For the group $SU(3)$ one can write

$$\int dg\ \mathrm{Tr}g\ \mathrm{Tr}g^\dagger = 1 \tag{71}$$

$$\int dg\ (\mathrm{Tr}g)^3 = 1 \tag{72}$$

from the well known formulae $3 \otimes \overline{3} = 1 \oplus 8$ and $3 \otimes 3 \otimes 3 = 1 \oplus 8 \oplus 8 \oplus 10$.

A simple integral useful for the strong coupling expansion is

$$\int dg\ g_{ij}\ (g^\dagger)_{kl} = I_{ijkl} \tag{73}$$

The group invariance says we can multiply the indices arbitrarily by group element on the left or right. There is only one combination of the indices that can survive for $SU(N)$

$$I_{ijkl} = \delta_{il}\delta_{jk}/N. \tag{74}$$

The normalization here is fixed since tracing over jk should give the identity matrix. Another integral that has a fairly simple form is

$$\int dg\ g_{i_1j_1}\ g_{i_2j_2} \ldots g_{i_Nj_N} = \frac{1}{N!}\epsilon_{i_1\ldots i_N}\epsilon_{j_1\ldots j_N} \tag{75}$$

This is useful for studying baryons in the strong coupling regime.

4.3 Gauge invariance

The action of lattice gauge theory has an exact local symmetry. If we associate an arbitrary group element g_i with each site i of the lattice, the action is unchanged if we replace

$$U_{ij} \to g_i^{-1} U_{ij} g_j \tag{76}$$

One consequence is that no link can have a vacuum expectation value [40].

$$\langle U_{ij} \rangle = g_i^{-1} \langle U_{ij} \rangle g_j = 0. \tag{77}$$

Generalizing this, unless one does some sort of gauge fixing, the correlation between any two separated U matrices is zero. Indeed many things familiar from perturbation theory often vanish without gauge fixing, including such fundamental objects as quark and gluon propagators!

An interesting consequence of gauge invariance is that we can forget to integrate over a tree of links in calculating any gauge invariant observable [34]. An axial gauge represents fixing all links pointing in a given direction.[4] Note that This sort of gauge fixing allows the reduction of two dimensional gauge theories to one dimensional spin models. To see this, pick the tree to be a non-intersecting spiral of links starting at the origin and extending out to the boundary. Links transverse to this spiral interact exactly as a one dimensional system. This also shows that two dimensional gauge theories are exactly solvable. Construct the transfer matrix along this one dimensional system. The partition function is the sum of the eigenvalues of this matrix each raised to the power of the volume of the system.

The trace of any product of link variables around a closed loop is the famous Wilson loop. These quantities are by construction gauge invariant and are the natural observables in the lattice theory. The well known criterion for confinement is whether the expectation of the Wilson loop decreases exponentially in the loop area.

More general gauges can be introduced using an analogue of the Fadeev-Popov factor. If $B(U)$ is gauge invariant, then

$$\langle B \rangle = \frac{1}{Z} \int d(U) e^{-S} B(U) = \frac{1}{Z} \int d(U) e^{-S} B(U) f(U)/\phi(U) \tag{78}$$

where $f(U)$ is an arbitrary gauge fixing function and

$$\phi(U) = \int (dg) f(g_i^{-1} U_{ij} g_j) \tag{79}$$

is the integral of the gauge fixing function f over all gauges. A possible gauge fixing scheme might be to ask that some function h of the links vanishes. In this case we could take $f = \delta(h)$ and then $\phi = \int (dg)\delta(h)$. The integral of a delta function of another function is generically a determinant $\phi = \det(\partial g/\partial h)$. A determinant can generally be written as an integral over a set of auxiliary "ghost" fields. Pursuing this yields the usual Fadeev-Popov picture [41].

Gauge fixing in the continuum raises several subtle issues if one wishes to go beyond perturbation theory. Given some gauge fixing condition $h = 0$ and the corresponding $f = \delta(h)$, it is desirable that this function vanish only once on any gauge orbit. Otherwise one should correct for the over counting due to what are known of as "Gribov copies" [42] This turns out to

[4] Using a tree with small highly-serrated leaves might be called a "light comb gauge."

be non-trivial with most perturbative gauges in practice, such as the Coulomb or Landau gauge. One of the great virtues of the lattice approach is that by not fixing the gauge, these issues are sidestepped.

On the lattice gauge fixing is unnecessary and usually not done if one only cares about measuring gauge invariant quantities such as Wilson loops. But this does have the consequence that the basic lattice fields are far from continuous. The correlation between link variables at different locations vanishes. The locality of the gauge symmetry literally means that there is an independent symmetry at each space time point. If we consider a quark-antiquark pair located at different positions, they transform under unrelated symmetries. Thus concepts such as separating the potential between quarks into singlet and octet parts are meaningless unless some gauge fixing is imposed.

4.4 Numerical simulation

Monte Carlo simulations of lattice gauge theory have come to dominate the subject. Here one uses the analogy to statistical mechanics to generate in a computer memory sets of gauge configurations weighted by the exponentiated action of the path integral. This is accomplished via a Markov chain of small weighted changes to a stored system. Various extrapolations are required to obtain continuum results; the lattice spacing needs to be taken to zero and the lattice size to infinity. Also, such simulations become increasingly difficult as the quark masses become small; thus, extrapolations in the quark mass are generally necessary. It is not the purpose of this review to cover these techniques; indeed, the several books mentioned at the beginning of this section are readily available. In addition, the proceedings of the annual Symposium on Lattice Field Theory are available on-line for the latest results.

While confinement is natural in the strong coupling limit of the lattice theory, we will shortly see that this is not the region of direct physical interest. For this a continuum limit is necessary. The coupling constant on the lattice represents a bare coupling defined at a length scale given by the lattice spacing. Non-Abelian gauge theories possess the property of asymptotic freedom, which means that in the short distance limit the effective coupling goes to zero. This remarkable phenomenon allows predictions for the observed scaling behavior in deeply inelastic processes. The way quarks expose themselves in high energy collisions was one of the original motivations for a non-Abelian gauge theory of the strong interactions.

In addition to enabling perturbative calculations at high energies, the consequences of asymptotic freedom are crucial for numerical studies via the lattice approach. As the lattice spacing goes to zero, the bare coupling must be taken to zero in a well determined way. Because of asymptotic freedom, we know precisely how to adjust our simulation parameters to take take the continuum limit!

In terms of the statistical analogy, the decreasing coupling takes us away from high temperature and towards the low temperature regime. Along the way a general statistical system might undergo dramatic changes in structure if phase transitions are present. Such qualitative shifts in the physical characteristics of a system can only hamper the task of demonstrating confinement in the non-Abelian theory. Early Monte Carlo studies of lattice gauge theory have provided strong evidence that such troublesome transitions are avoided in the standard four dimensional $SU(3)$ gauge theory of the nuclear force [43].

Although the ultimate goal of lattice simulations is to provide a quantitative understanding of

continuum hadronic physics, along the way many interesting phenomena arise which are peculiar to the lattice. Non-trivial phase structure does occur in a variety of models, some of which do not correspond to any continuum field theory. We should remember that when the cutoff is still in place, the lattice formulation is highly non-unique. One can always add additional terms that vanish in the continuum limit. In this way spurious transitions might be alternatively introduced or removed. Physical results require going to the continuum limit.

4.5 Order parameters

Formally lattice gauge theory is like a classical statistical mechanical spin system. The spins U_{ij} are elements of a gauge group G. They are located on the bonds of our lattice. Can this system become "ferromagnetic"? Indeed, as mentioned above, this is impossible since $\langle U \rangle = 0$ follows from the links themselves not being gauge invariant [40].

But we do expect some sort of ordering to occur in the $U(1)$ theory. If this is to describe physical photons, there should be a phase with massless particles. Strong coupling expansions show that for large coupling this theory has a mass gap [9]. Thus a phase transition is expected, and has been observed in numerical simulations [44]. Exactly how this ordering occurs remains somewhat mysterious; indeed, although people often look for a "mechanism for confinement," it might be interesting to rephrase this question to "how does a theory such as electromagnetism avoid confinement."

The standard order parameter for gauge theories and confinement involves the Wilson loop mentioned above. This is the trace of the product of link variables multiplied around a closed loop in space-time. If the expectation of such a loop decreases exponentially with the area of the loop, we say the theory obeys an area law and is confining. On the other hand, a decrease only as the perimeter indicates an unconfined theory. This order parameter by nature is non-local; it cannot be measured without involving arbitrarily long distance correlations. The lattice approach is well known to give the area law in the strong coupling limit of the pure gauge theory. Unfortunately, with dynamical quarks this ceases to be a useful measure of confinement. As a loop becomes large, it will be screened dynamically by quarks "popping" out of the vacuum. Thus we always will have a perimeter law.

Another approach to understanding the confinement phase is to use the mass gap. As long as the quarks themselves are massive, a confining theory should contain no physical massless particles. All mesons, glueballs, and nucleons are expected to gain masses through the dimensional transmutation phenomenon discussed later. As with the area law, the presence of a mass gap is easily demonstrated for the strong coupling limit of the pure glue theory.

If the quarks are massless, this definition also becomes a bit tricky. In this case we expect spontaneous breaking of chiral symmetry, also discussed extensively later. This gives rise to pions as massless Goldstone bosons. To distinguish this situation from the unconfined theory, one could consider the number of massless particles in the spectrum by looking at how the "vacuum" energy depends on temperature using the Stefan-Boltzmann law. With N_f flavors we have $N_f^2 - 1$ massless scalar Goldstone bosons. On the other hand, were the gauge group $SU(N)$ not to confine, we would expect $N^2 - 1$ massless vector gauge bosons plus N_f massless quarks, all of which have two degrees of freedom.

5 Renormalization and the continuum limit

Asymptotic freedom is a signature feature of the theory of the strong interactions. Interactions between quarks decrease at very short distances. From one point of view this allows perturbative calculations in the high energy limit, and this has become an industry in itself. But the concept is also of extreme importance to lattice gauge theory. Indeed, asymptotic freedom tells us precisely how to take the continuum limit. This chapter reviews the renormalization group and this crucial connection to the lattice. When fermions are present their masses must also be renormalized, but the renormalization group also tells us exactly how to do this.

5.1 Coupling constant renormalization

At the level of tree Feynman diagrams, relativistic quantum field theory is well defined and requires no renormalization. However as soon as loop corrections are encountered, divergences appear and must be removed by a regularization scheme. In general the theory then depends on some cutoff, which is to be removed with a simultaneous adjustment of the bare parameters while keeping physical quantities finite.

For example, consider a lattice cutoff with spacing a. The proton mass m_p is a finite physical quantity, and on the lattice it will be some, a priori unknown, function of the cutoff a, the bare gauge coupling g and the bare quark masses. For the quark-less theory we could use the lightest glueball mass for this purpose. The basic idea is to hold enough physical properties constant to determine how the coupling and quark masses behave as the lattice spacing is reduced.

As the quark masses go to zero the proton mass is expected to remain finite; thus, to simplify the discussion, temporarily ignore the quark masses. Thus consider the proton mass as a function of the gauge coupling and the cutoff, $m_p(g,a)$. Holding this constant as the cutoff varies determines how g depends on a. This is the basic renormalization group equation

$$a\frac{d}{da}m_p(g(a),a) = 0 = a\frac{\partial}{\partial a}m_p(a,g) + a\left(\frac{dg}{da}\right)\frac{\partial}{\partial g}m_p(a,g). \tag{80}$$

By dimensional analysis, the proton mass should scale as a^{-1} at fixed bare coupling. Thus we know that

$$a\frac{\partial}{\partial a}m_p(g,a) = -m_p(g,a) \tag{81}$$

The “renormalization group function”

$$\beta(g) = a\frac{\partial g}{\partial a} = \frac{-m_p(g,a)}{\frac{\partial}{\partial g}m_p(g,a)}. \tag{82}$$

characterizes how the bare coupling is to be varied for the continuum limit.

As renormalization is not needed until quantum loops are encountered, $\beta(g)$ vanishes as g^3 when the coupling goes to zero. Define perturbative coefficients from the asymptotic series

$$\beta(g) = \beta_0 g^3 + \beta_1 g^5 + \ldots \tag{83}$$

Politzer [45] and Gross and Wilczek [46, 47] first calculated the coefficient β_0 for non-Abelian gauge theories, with the result

$$\beta_0 = \frac{1}{16\pi^2}(11N/3 - 2N_f/3) \tag{84}$$

where the gauge group is $SU(N)$ and N_f denotes the number of fermionic species. As long as $N_f < 11N/2$ this coefficient is positive. Assuming we can reach a region where this first term dominates, decreasing the cutoff corresponds to decreasing the coupling. This is the heart of asymptotic freedom, which tells us that the continuum limit of vanishing lattice spacing requires taking a limit towards vanishing coupling. The two loop contribution to Eq. (83) is also known [48, 49]

$$\beta_1 = \left(\frac{1}{16\pi^2}\right)^2 \left(34N^2/3 - 10NN_f/3 - N_f(N^2-1)/N\right). \tag{85}$$

In general the function $\beta(g)$ depends on the regularization scheme in use. For example it might depend on what physical property is held fixed as well as details of how the cutoff is imposed. Remarkably, however, these first two coefficients are universal. Consider two different schemes each defining a bare coupling as a function of the cutoff, say $g(a)$ and $g'(a)$. The expansion for one in terms of the other will involve all odd powers of the coupling. In the weak coupling limit each formulation should reduce to the classical Yang-Mills theory, and thus to lowest order they should agree

$$g' = g + cg^3 + O(g^5) \tag{86}$$

We can now calculate the new renormalization group function

$$\begin{aligned} \beta'(g') &= a\frac{dg'}{da} = \frac{\partial g'}{\partial g}\beta(g) \\ &= (1+3cg^3)(\beta_0 g^2 + \beta_1 g^3) + O(g^5) \\ &= \beta_0 {g'}^3 + \beta_1 {g'}^3 + O({g'}^5) \end{aligned} \tag{87}$$

Through order ${g'}^3$ the dependence on the parameter c cancels. This, however, does not continue to higher orders, where alternate definitions of the beta function generally differ. We will later comment further on this non-uniqueness.

The renormalization group function determines how rapidly the coupling decreases with cutoff. Separating variables

$$d(\log(a)) = \frac{dg}{\beta_0 g^3 + \beta_1 g^5 + O(g^7)} \tag{88}$$

allows integration to obtain

$$\log(a\Lambda) = -\frac{1}{2\beta_0 g^2} + \frac{\beta_1}{\beta_0^2}\log(g) + O(g^2) \tag{89}$$

where Λ is an integration constant. This immediately shows that the lattice spacing decreases exponentially in the inverse coupling

$$a = \frac{1}{\Lambda} e^{-1/2\beta_0 g^2} g^{-\beta_1/\beta_0^2}(1 + O(g^2)). \tag{90}$$

Remarkably, although the discussion began with the beta function obtained in perturbation theory, the right hand side of Eq. (90) has an essential singularity at vanishing coupling. The renormalization group provides non-perturbative information from a perturbative result.

Dropping the logarithmic corrections, the coupling as a function of the cutoff reduces to

$$g^2 \sim \frac{1}{2\beta_0 \log(1/\Lambda a)} \tag{91}$$

showing the asymptotic freedom result that the bare coupling goes to zero logarithmically with the lattice spacing in the continuum limit.

The integration constant Λ is defined from the bare charge and in a particular cutoff scheme. Its precise numerical value will depend on details, but once the scheme is chosen, it is fixed relative to the scale of the quantity used define the physical scale. In the above discussion this was the proton mass. The existence of a scheme dependence can be seen by considering two different bare couplings as related in Eq. (86). The relation between the integration constants is

$$\log(\Lambda'/\Lambda) = \frac{c}{2\beta_0}. \tag{92}$$

The mass m of a physical particle, perhaps the proton used above, is connected to an inverse correlation length in the statistical analogue of the theory. Measuring this correlation length in lattice units, we can consider the dimensionless combination $\xi = 1/am$. For the continuum limit, we want this correlation length to diverge. Multiplying Eq. (90) by the mass tells us how this divergence depends on the lattice coupling

$$ma = \xi^{-1} = \frac{m}{\Lambda} e^{-1/2\beta_0 g^2} g^{-\beta_1/\beta_0^2}(1 + O(g^2)). \tag{93}$$

Conversely, if we know how a correlation length ξ of the statistical system diverges as the coupling goes to zero, we can read off the particle mass in units of Λ as the coefficient of the behavior in this equation. This exemplifies the close connection between diverging correlation lengths in a statistical system and the continuum limit of the corresponding quantum field theory.

We emphasize again the exponential dependence on the inverse coupling appearing in Eq. (93). This is a function that is highly non-analytic at the origin. This demonstrates quite dramatically that QCD cannot be fully described by perturbation theory.

5.2 A parameter free theory

This discussion brings us to the remarkable conclusion that, ignoring the quark masses, the strong interactions have no free parameters. The cutoff is absorbed into $g(a)$, which in turn is absorbed into the renormalization group dependence. The only remaining dimensional parameter Λ serves to set the scale for all other masses. In the theory considered in isolation, one may select units such that Λ is unity. After such a choice, all physical mass ratios are determined. Coleman and Weinberg [50] have given this process, wherein a dimensionless parameter g and a dimensionful one a manage to "eat" each other, the marvelous name "dimensional transmutation."

In the theory including quarks, their masses represent further parameters. Indeed, these are the only parameters in the theory of the strong interactions. In the limit where the quark masses vanish, referred to as the chiral limit, we return to a zero parameter theory. In this approximation

to the physical world, the pion mass is expected to vanish and all dimensionless observables should be uniquely determined. This applies not only to mass ratios, such as of the rho mass to the proton, but as well to quantities such as the pion-nucleon coupling constant, once regarded as a parameter for a perturbative expansion. As the chiral approximation has been rather successful in the predictions of current algebra, we expect an expansion in the small quark masses to be a fairly accurate description of hadronic physics. Given a qualitative agreement, a fine tuning of the small quark masses should give the pion its mass and complete the theory.

The exciting idea of a parameter-free theory is sadly lacking from most treatments of the other interactions such as electromagnetism or the weak force. There the coupling $\alpha \sim 1/137$ is treated as a parameter. One might optimistically hope that the inclusion of the appropriate non-perturbative ideas into a unified scheme would ultimately render α and the quark and lepton masses calculable.

5.3 Including quark masses

Above we concentrated on the flow of the bare coupling as one takes the continuum limit. Of course with massive quarks in the theory, the bare quark mass is also renormalized. Here we extend the above discussion to see how the two bare parameters flow together to zero in a well defined way.

Including the mass flow, the renormalization group equations become

$$\begin{aligned} a\frac{dg}{da} &= \beta(g) = \beta_0 g^3 + \beta_1 g^5 + \ldots + \text{non-perturbative} \\ a\frac{dm}{da} &= m\gamma(g) = m(\gamma_0 g^2 + \gamma_1 g^4 + \ldots) + \text{non-perturbative.} \end{aligned} \tag{94}$$

Now we have three perturbative coefficients $\beta_0,\ \beta_1,\ \gamma_0$ which are scheme independent and known [45–49, 51, 52]. For $SU(3)$ we have

$$\begin{aligned} \beta_0 &= \tfrac{11-2N_f/3}{(4\pi)^2} &&= .0654365977 \qquad (N_f = 1) \\ \beta_1 &= \tfrac{102-12N_f}{(4\pi)^4} &&= .0036091343 \qquad (N_f = 1) \\ \gamma_0 &= \tfrac{8}{(4\pi)^2} &&= .0506605918 \end{aligned} \tag{95}$$

For simplicity we work with N_f degenerate quarks, although this is easily generalized to the non-degenerate case. It is important to recognize that the "non-perturbative" parts fall faster than any power of g as $g \to 0$. As we will discuss later, unlike the perturbative pieces, the non-perturbative contributions to γ in general need not be proportional to the quark mass.

As with the pure gauge theory discussed earlier, these equations are easily solved to show

$$\begin{aligned} a &= \frac{1}{\Lambda} e^{-1/2\beta_0 g^2} g^{-\beta_1/\beta_0^2}(1 + O(g^2)) \\ m &= M g^{\gamma_0/\beta_0}(1 + O(g^2)). \end{aligned} \tag{96}$$

The quantities Λ and M are "integration constants" for the renormalization group equations. Rewriting these relations gives the coupling and mass flow in the continuum limit $a \to 0$

$$g^2 \sim \frac{1}{\log(1/\Lambda a)} \to 0 \qquad \text{"asymptotic freedom"}$$

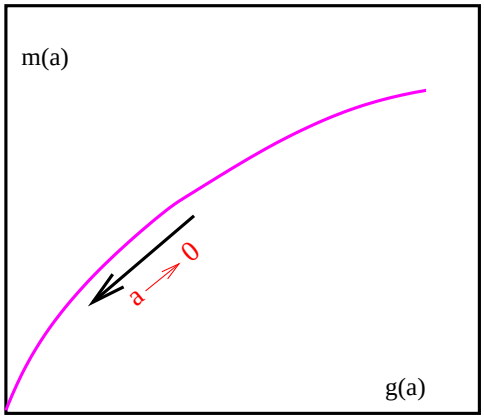

Fig. 3. In the continuum limit both the bare coupling and bare mass for QCD flow to zero.

$$m \sim M \left(\frac{1}{\log(1/\Lambda a)} \right)^{\gamma_0/\beta_0} \to 0 \tag{97}$$

Here Λ is usually regarded as the "QCD scale" and M as the "renormalized quark mass." The resulting flow is sketched in Fig. 3.

The rate of this flow to the origin is tied to the renormalization group constants, which can be obtained from the inverted equations

$$\Lambda = \lim_{a \to 0} \frac{e^{-1/2\beta_0 g^2} g^{-\beta_1/\beta_0^2}}{a} \tag{98}$$

$$M = \lim_{a \to 0} m g^{-\gamma_0/\beta_0}. \tag{99}$$

Of course, as discussed for Λ above, the specific numerical values of these parameters depend on the detailed renormalization scheme.

Defining $\beta(g)$ and $\gamma(g)$ is most naturally done by fixing some physical quantities and adjusting the bare parameters as the cutoff is removed. Because of confinement we can't use the quark mass itself, but we can select several physical particle masses $m_i(g, m, a)$ to hold fixed. This leads to the constraint

$$a \frac{dm_i(g, m, a)}{da} = 0 = \frac{\partial m_i}{\partial g} \beta(g) + \frac{\partial m_i}{\partial m} m\gamma(g) + a \frac{\partial m_i}{\partial a}. \tag{100}$$

For simplicity, continue to work with degenerate quarks of mass m. Then we have two bare parameters (g, m), and we need to fix two quantities. [5] Natural candidates are the lightest baryon mass, denoted here as m_p, and the lightest boson mass, m_π. Then we can explicitly rearrange these relations to obtain a somewhat formal but explicit expression for the renormalization group functions

$$\beta(g) = \left(a \frac{\partial m_\pi}{\partial a} \frac{\partial m_p}{\partial m} - a \frac{\partial m_p}{\partial a} \frac{\partial m_\pi}{\partial m} \right) \Big/ \left(\frac{\partial m_p}{\partial g} \frac{\partial m_\pi}{\partial m} - \frac{\partial m_\pi}{\partial g} \frac{\partial m_p}{\partial m} \right)$$

[5] Actually there is a third parameter related to CP conservation. Here we assume CP is a good symmetry and ignore this complication. This issue will be further discussed in later sections.

$$\gamma(g) = \left(a\frac{\partial m_\pi}{\partial a}\frac{\partial m_p}{\partial g} - a\frac{\partial m_p}{\partial a}\frac{\partial m_\pi}{\partial g} \right) \Big/ \left(\frac{\partial m_p}{\partial m}\frac{\partial m_\pi}{\partial g} - \frac{\partial m_\pi}{\partial m}\frac{\partial m_p}{\partial g} \right). \tag{101}$$

Note that this particular definition includes all perturbative and non-perturbative effects. In addition, this approach avoids any need for gauge fixing.

Once given m_p, m_π, and a renormalization scheme, then the dependence of the bare parameters on the cutoff is completely fixed. The physical masses are mapped onto the integration constants

$$\Lambda = \Lambda(m_p, m_\pi) \tag{102}$$
$$M = M(m_p, m_\pi). \tag{103}$$

Formally these relations can be inverted to express the masses as functions of the integration constants, $m_i = m_i(\Lambda, M)$. Straightforward dimensional analysis tells us that the masses must take the form

$$m_i = \Lambda f_i(M/\Lambda). \tag{104}$$

As we will discuss in more detail in later sections, for the multi-flavor theory we expect the pions to be Goldstone bosons with $m_\pi^2 \sim m_q$. This tells us that the above function for the pion should exhibit a square root singularity $f_\pi(x) \sim x^{1/2}$. This relation removes any additive ambiguity in defining the renormalized quark mass M. As will be discussed in more detail later, this conclusion does not persist if the lightest quark becomes non-degenerate.

5.4 Which beta function?

Thus far our discussion of the renormalization group has been in terms of the bare charge with a cutoff in place. This is the natural procedure in lattice gauge theory; however, there are alternative approaches to the renormalization group that are frequently used in the continuum theory. We now make some comments on connection between the lattice and the continuum approaches.

An important issue is that there are many different ways to define a renormalized coupling; it should first of all be an observable that remains finite in the continuum limit

$$\lim_{a\to 0} g_r(\mu, a, g(a)) = g_r(\mu). \tag{105}$$

Here μ is a dimensionful energy scale introduced to define the renormalized coupling. The subscript r is added to distinguish this coupling from the bare one. For perturbative purposes one might use a renormalized three-gluon vertex in a particular gauge and with all legs at a given scale of momentum proportional to μ. But many alternatives are possible; for example, one might use as an observable the force between two quarks at separation $1/\mu$.

Secondly, to be properly called a renormalization of the classical coupling, g_r should be normalized such that it reduces to the bare coupling in lowest order perturbation theory for the cutoff theory

$$g_r(\mu, a, g) = g + O(g^3). \tag{106}$$

Beyond this, the definition of g_r is totally arbitrary. In particular, given any physical observable H defined at scale μ and satisfying a perturbative expansion

$$H(\mu, a, g) = h_0 + h_1 g^2 + O(g^4) \tag{107}$$

we can define a corresponding renormalized coupling

$$g_H^2(\mu) = (H(\mu) - h_0)/h_1. \tag{108}$$

As the energy scale goes to infinity, this renormalized charge should go to zero. But with a different observable, we will generally obtain a different functional behavior for this flow. From this flow of the renormalized charge we can define a renormalized beta function

$$\beta_r = -\mu \frac{\partial g_r(\mu)}{\partial \mu}. \tag{109}$$

We now draw a remarkable connection between the renormalized renormalization group function $\beta_r(g_r)$ and the function $\beta(g)$ defined earlier for the bare coupling. When the cutoff is still in place, the renormalized coupling is a function of the scale μ of the observable, the cutoff a, and the bare coupling g. Since we are working with dimensionless couplings, g_r can depend directly on μ and a only through their product. This simple application of dimensional analysis implies

$$a \left.\frac{\partial g_r}{\partial a}\right|_g = \mu \left.\frac{\partial g_r}{\partial \mu}\right|_g = -\beta_r. \tag{110}$$

Now, in the continuum limit as we take a to zero and adjust g appropriately, g_r should become a function of the physical scale μ alone. Indeed, we could use $g_r(\mu)$ itself as the physical quantity to hold fixed for the continuum limit. Then we obtain

$$a\frac{\partial g_r}{\partial a} + \frac{\partial g_r}{\partial g} a \frac{\partial g}{\partial a} = 0. \tag{111}$$

Using this in an analysis similar to that in Eq. (87), we find

$$\beta_r(g_r) = \beta_0 g_r^3 + \beta_1 g_r^5 + O(g_r^7). \tag{112}$$

Where β_0 and β_1 are the same coefficients that appear in Eq. (83). Both the renormalized and the bare β functions have the same first two coefficients in their perturbative expansions. Indeed, it was through consideration of the renormalized coupling that β_0 and β_1 were first calculated.

It is important to reiterate the considerable arbitrariness in defining both the bare and the renormalized couplings. Far from the continuum there need be no simple relationship between different formulations. Once one leaves the perturbative region, even such things as zeros in the β functions are not universal. For one extreme example, it is allowed to force the beta function to consist of only the first two terms. In this case, as long as N_f is small enough that $\beta_1 > 0$, there is explicitly no other zero of the beta function except at $g = 0$. On the other hand, one might think it natural to define the coupling from the force between two quarks. When dynamical quarks are present, at large distances this falls exponentially with the pion mass at large distances. In this case the beta function must have another zero in the vicinity of where the screening sets in. Thus, even the existence of zeros in the beta function is scheme dependent. The only exception to this

is if a zero occurs in a region of small enough coupling that perturbation theory can be trusted. This has been conjectured to happen for a sufficient number of flavors [28].

The perturbative expansion of β_R has important experimental consequences. If, as expected, the continuum limit is taken at vanishing bare coupling and the renormalized coupling is small enough that the first terms in Eq. (112) dominate, then the renormalized coupling will be driven to zero logarithmically as its defining scale μ goes to infinity. Not only does the bare coupling vanish, but the effective renormalized coupling becomes arbitrarily weak at short distances. This is the physical implication of asymptotic freedom; phenomena involving only short-distance effects may be accurately described with a perturbative expansion. Indeed, asymptotically free gauge theories were first invoked for the strong interactions as an explanation of the apparently free parton behavior manifested in the structure functions associated with deeply inelastic scattering of leptons from hadrons.

The dependence of the integration constant Λ on the details of the renormalization scheme carries over to the continuum renormalization group as well. Given a particular definition of the renormalized coupling $g_r(\mu)$, its behavior as r goes to zero will involve a scale Λ_r in analogy to the scale in the bare coupling. Hasenfratz and Hasenfratz [53, 54] were the first to perform the necessary one loop calculations to relate Λ from the Wilson lattice gauge theory with Λ_r defined from the three-gluon vertex in the Feynman gauge and with all legs carrying momentum μ^2. They found

$$\frac{\Lambda_r}{\Lambda} = \begin{pmatrix} 57.5 & SU(2) \\ 83.5 & SU(3) \end{pmatrix} \tag{113}$$

for the pure gauge theory. Note that not only is Λ scheme dependent, but that different definitions can vary by rather large factors. The original calculation of these numbers was rather tedious. They have been verified with calculationally more efficient techniques based on quantum fluctuations around a slowly varying classical background field [55].

5.5 Flows and irrelevant operators

We now briefly discuss another way of looking at the renormalization group as relating theories with different lattice spacings. Given one lattice theory, one could imagine generating another with a larger lattice spacing by integrating over all links except those on some super lattice. While this is conceptually possible, to do it exactly in more than one dimension will generate an infinite number of couplings. If we could keep track of such, the procedure would be "exact," but in reality we usually need some truncation. Continuing to integrate out degrees of freedom, the couplings flow and might reach some "fixed point" in this infinite space. With multiple couplings, there can be an attractive "sheet" towards which couplings flow, and then they might continue to flow towards a fixed point, as sketched in Fig. 4. If the fixed point has only one attractive direction, then two different models that flow towards that same fixed point will have the same physics in the large distance limit. This is the concept of universality; *i.e.* exponents are the same for all models with the same attractor.

Some hints on this process come from dimensional analysis, although, in ignoring non-perturbative effects that might occur at strong coupling, the following arguments are not rigorous. In d dimensions a conventional scalar field has dimensions of $M^{\frac{d-2}{2}}$. Thus the coupling constant λ in an interaction of form $\int d^dx\ \lambda\phi^n$ has dimensions of $M^{d-n\frac{d-2}{2}}$. On a lattice of spacing a,

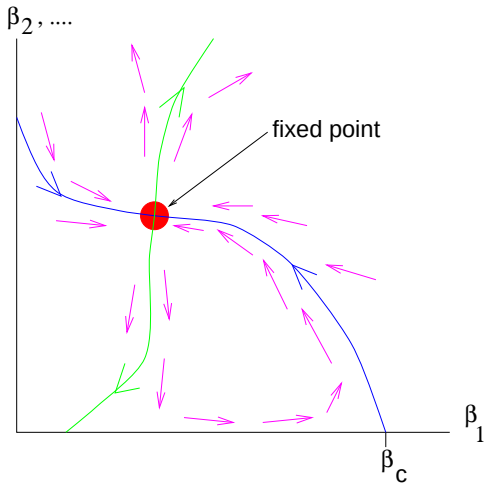

Fig. 4. A generic renormalization group flow. In general this occurs in an infinite dimensional space.

the natural unit of dimension is the inverse lattice spacing. Thus without any special tuning, the renormalized coupling at some fixed physical scale would naturally run as $\lambda \sim a^{n\frac{d-2}{2}-d}$. As long as the exponent in this expression is positive, i.e.

$$n \geq \frac{2d}{d-2}$$

we expect the coupling to become "irrelevant" in the continuum limit. The fixed point is driven towards zero in the corresponding direction. If d exceeds four, this is the case for all interactions. (We ignore ϕ^3 in 6 dimensions because of stability problems.) This suggests that four dimensions is a critical case, with mean field theory giving the right qualitative critical behavior for all larger dimensions. In four dimensions we have several possible "renormalizable" couplings which are dimensionless, suggesting logarithmic corrections to the simple dimensional arguments. Indeed, four-dimensional non-Abelian gauge theories display exactly such a logarithmic flow; this is asymptotic freedom.

This simple dimensional argument applied to the mass term suggests it would flow towards infinity in all dimensions. For a conventional phase transition, something must be tuned to a critical point. In statistical mechanics this is the temperature. In field theory language we usually remap this onto a tuning of the bare mass term, saying that the transition occurs as bare masses go through zero. For a scalar theory this tuning for a continuum limit seems unnatural in and is one of the unsatisfying features of the standard model, driving particle physicists to try to unravel how the Higg's mechanism really works.

In non-Abelian gauge theories with multiple massless fermions, chiral symmetry protects the mass from renormalization, avoiding any special tuning. Indeed, as we have discussed, because of dimensional transmutation, all dimensionless parameters in the continuum limit are completely determined by the basic structure of the initial Lagrangian, without any continuous parameters to tune. In the limit of vanishing pion mass, the rho to nucleon mass ratio should

be determined from first principles; it is the goal of lattice gauge theory to calculate just such numbers.

As we go below four dimensions, this dimensional argument suggests that several couplings can become "relevant," requiring the renormalization group picture of flow towards a non-trivial fixed point. Above two dimensions the finite number of renormalizable couplings corresponds to the renormalization group argument for a finite number of "universality classes," corresponding to different basic symmetries.

One might imagine dimensionality as being a continuously variable parameter. Then just below four dimensions a renormalizable coupling becomes "super-renormalizable" and a new non-trivial fixed point breaks away from vanishing coupling. Near four dimensions this point is at small coupling, forming the basis for an expansion in $4-d$. This has become a major industry, making remarkably accurate predictions for critical exponents in three dimensional systems [56].

An important consequence of this discussion is that a lattice action is in general highly non-unique. One can always add irrelevant operators and expect to obtain the same continuum limit. Alternatively, one might hope to improve the approach the continuum limit by a judicious choice of the lattice action.

The renormalization group is indeed a rich subject. We have only touched on a few issues that are particularly valuable for the lattice theory. Perhaps the most remarkable result of this section is how a perturbative analysis of the renormalization-group equation gives rise to information on the non-perturbative behavior in the particle masses, as exhibited in Eq. (93).

6 Classical gauge fields and topology

The above renormalization group analysis demonstrates that non-perturbative effects are crucial to understanding the continuum limit of QCD on the lattice. However, the importance of going beyond the perturbation expansion for non-Abelian gauge theories was dramatically exposed from a completely different direction with the discovery of non-trivial classical solutions characterized by an essential singularity at vanishing coupling. Here we review these solutions and some of the interesting consequences for the Dirac operator.

We start with some basic definitions to establish notation in continuum language. Being ultimately interested in QCD, we concentrate on the gauge group $SU(N)$. This group has N^2-1 generators denoted λ^α. They are traceless N by N matrices and satisfy the commutation relations

$$[\lambda^\alpha, \lambda^\beta] = i f^{\alpha\beta\gamma}\lambda^\gamma, \tag{114}$$

involving the group structure functions $f^{\alpha\beta\gamma}$. By convention, these generators are orthogonalized and normalized

$$\mathrm{Tr}\lambda^\alpha\lambda^\beta = \frac{1}{2}\delta^{\alpha\beta}. \tag{115}$$

For $SU(2)$ the generators would be the spin matrices $\lambda^\alpha = \sigma^\alpha/2$, and the generators the three indexed antisymmetric tensor $f^{\alpha\beta\gamma} = \epsilon^{\alpha\beta\gamma}$.

Associated with the each of the generators λ^α is a gauge potential $A^\alpha_\mu(x)$. For the classical theory, assume these are differentiable functions of space time and vanish rapidly at infinity. For

the quantum theory this assumption of differentiability is a subtle issue to which we will later return. The notation simplifies a bit by defining a matrix valued field A

$$A_\mu = A_\mu^\alpha \lambda^\alpha. \tag{116}$$

The covariant derivative is a matrix valued differential operator defined as

$$D_\mu = \partial_\mu + igA_\mu. \tag{117}$$

Given the gauge potential, the corresponding matrix valued field strength is

$$F_{\mu\nu} = \frac{-i}{g}[D_\mu, D_\nu] = \partial_\mu A_\nu - \partial_\nu A_\mu + ig[A_\mu, A_\nu] = D_\mu A_\nu - D_\nu A_\mu. \tag{118}$$

We define the dual field strength as

$$\tilde{F}_{\mu\nu} = \frac{1}{2}\epsilon_{\mu\nu\rho\sigma} F_{\rho\sigma} \tag{119}$$

with $\epsilon_{\mu\nu\rho\sigma}$ being the antisymmetric tensor with $\epsilon_{1234} = 1$.

In terms of the field strength, the classical Yang-Mills action is

$$S = \frac{1}{2}\int d^4x \ \mathrm{Tr}\ F_{\mu\nu}F_{\mu\nu}, \tag{120}$$

and the classical equations of motion are

$$D_\mu F_{\mu\nu} = 0. \tag{121}$$

This defines the classical Yang-Mills theory.

The Jacobi identity

$$[A,[B,C]] + [B,[C,A]] + [C,[A,B]] = 0 \tag{122}$$

applied to the covariant derivative implies that

$$\epsilon_{\mu\nu\rho\sigma} D_\nu F_{\rho\sigma} = 0 \tag{123}$$

or $D_\mu \tilde{F}_{\mu\nu} = 0$. This immediately implies that any self-dual or anti-self-dual field with $F = \pm\tilde{F}$ automatically satisfies the classical equations of motion. This is an interesting relation since $F = \tilde{F}$ is linear in derivatives of the gauge potential. This leads to a multitude of known solutions [57], but here we concentrate on just the simplest non-trivial one.

This theory is, after all, a gauge theory and therefore has a local symmetry. We previously discussed this in the lattice context. There it was originally motivated by the continuum gauge transformations of the classical theory, which we now review. Let $h(x)$ be a space dependent element of $SU(N)$ in the fundamental representation. Assume that h is differentiable. Now define the gauge transformed field

$$A_\mu^{(h)} \to h^\dagger A_\mu h - \frac{i}{g} h^\dagger(\partial_\mu h). \tag{124}$$

This transformation takes a simple form for the covariant derivative

$$h^\dagger D_\mu h = D_\mu^{(h)} = \partial_\mu + igA_\mu^{(h)}. \tag{125}$$

Similarly for the field strength we have

$$F_{\mu\nu}^{(h)} = h^\dagger F_{\mu\nu} h. \tag{126}$$

Thus the action is invariant under this transformation, $S(A) = S(A^{(h)})$.

6.1 Surface terms

A remarkable feature of this formalism is that the combination $\mathrm{Tr}\ F\tilde{F}$ is a total derivative. To see this first construct

$$\begin{aligned} F\tilde{F} &= \frac{1}{2}\epsilon_{\mu\nu\rho\sigma}(2\partial_\mu A_\nu + igA_\mu A_\nu)(2\partial_\rho A_\sigma + igA_\rho A_\sigma) \\ &= \frac{1}{2}\epsilon_{\mu\nu\rho\sigma}\left(4\partial_\mu A_\nu \partial_\rho A_\sigma + 4ig(\partial_\mu A_\nu)A_\rho A_\sigma - g^2 A_\mu A_\nu A_\rho A_\sigma\right). \end{aligned} \tag{127}$$

If we take a trace of this quantity, the last term will drop out due to cyclicity. Thus

$$\mathrm{Tr} F\tilde{F} = 2\partial_\mu \epsilon_{\mu\nu\rho\sigma}\mathrm{Tr}\left(A_\nu \partial_\rho A_\sigma + igA_\nu A_\rho A_\sigma\right) = 2\partial_\mu K_\mu \tag{128}$$

where we define

$$K_\mu \equiv \epsilon_{\mu\nu\rho\sigma}\mathrm{Tr}\left(A_\nu \partial_\rho A_\sigma + 2igA_\nu A_\rho A_\sigma\right). \tag{129}$$

Note that although $\mathrm{Tr}\ F\tilde{F}$ is gauge invariant, this is not true for K_μ.

Being a total derivative, the integral of this quantity

$$\int d^4x\ \frac{1}{2}\mathrm{Tr} F\tilde{F} \tag{130}$$

would vanish if we ignore surface terms. What is remarkable is that there exist finite action gauge configurations for which this does not vanish even though the field strengths all go to zero rapidly at infinity. This is because the gauge fields A_μ that appear explicitly in the current K_μ need not necessarily vanish as rapidly as $F_{\mu\nu}$.

These surface terms are closely tied to the topology of the gauge potential at large distances. As we want the field strengths to go to zero at infinity, the potential should approach a pure gauge form $A_\mu \to \frac{-i}{g}h^\dagger \partial_\mu h$. In this case

$$K_\mu \to -\frac{1}{g^2}\epsilon_{\mu\nu\rho\sigma}\mathrm{Tr}(h^\dagger\partial_\nu h)(h^\dagger\partial_\rho h)(h^\dagger\partial_\sigma h). \tag{131}$$

Note the similarity of this form to that for the group measure in Eq. (64). Indeed, it is invariant if we take $h \to h'h$ with h' being a constant group element. The surface at infinity is topologically a three dimensional sphere S_3. If we concentrate on $SU(2)$, this is the same as the topology of the group space. For larger groups we can restrict h to an $SU(2)$ subgroup and proceed similarly. Thus the integral of K_μ over the surface reduces to the integral of h over a sphere with the invariant group measure. This can give a non-vanishing contribution if the mapping of h onto the sphere at infinity covers the entire group in a non-trivial manner. Mathematically, one can map the S_3 of infinite space onto the S_3 of group space an integer number of times, i.e. $\Pi_3(SU(2)) = Z$. Thus we have

$$\int d^4x\ \frac{1}{2}\mathrm{Tr} F\tilde{F} \propto \nu \tag{132}$$

where ν is an integer describing the number of times $h(x)$ wraps around the group as x covers the sphere at infinity. The normalization involves the surface area of a three dimensional sphere and can be worked out with the result

$$\int d^4x\ \frac{1}{2}\mathrm{Tr} F\tilde{F} = \frac{8\pi^2\nu}{g^2}. \tag{133}$$

For groups larger than $SU(2)$ one can deform h to lie in an $SU(2)$ subgroup, and thus this quantization of the surface term applies to any $SU(N)$.

If we were to place such a configuration into the path integral for the quantum theory, we might expect a suppression of these effects by a factor of $\exp(-8\pi^2/g^2)$. This is clearly non-perturbative, however this factor strongly underestimates the importance of topological effects. The problem is that we only need to excite non-trivial fields over the quantum mechanical vacuum, not the classical one. The correct suppression is indeed exponential in the inverse coupling squared, but the coefficient in the exponent can be determined from asymptotic freedom and dimensional transmutation. We will return to this point in Subsection 9.3.

The combination $\mathrm{Tr}F\tilde{F}$ is formally a dimension four operator, the same as the basic gauge theory action density $\mathrm{Tr}FF$. This naturally leads to the question of what would happen if we consider a new action which also includes this parity odd term. Classically it does nothing since it reduces to the surface term described above. However quantum mechanically this is no longer the case. As we will discuss extensively later, the physics of QCD depends quite non-trivially on such a term. An interesting feature of this new term follows from the quantization of the resulting surface term. Because of the above quantization and an imaginary factor in the path integral, physics is periodic in the coefficient of $\frac{1}{2}\mathrm{Tr}F\tilde{F}$. Although discussing the consequences directly with such a term in the action is traditional, we will follow a somewhat different path in later sections and introduce this physics through its effects on fermions.

6.2 An explicit solution

To demonstrate that non-trivial winding solutions indeed exist, specialize to $SU(2)$ and find an explicit example. To start, consider the positivity of the norm of $F \pm \tilde{F}$

$$0 \leq \int d^4x\, (F \pm \tilde{F})^2 = 2\int d^4x\, F^2 \pm 2\int d^4x\, F\tilde{F}. \tag{134}$$

This means that the action is bounded below by $\int d^4x\, \frac{1}{2}\mathrm{Tr}F\tilde{F}$ and this bound is reached only if $F = \pm\tilde{F}$. As mentioned earlier, reaching this is sufficient to guarantee a solution to the equations of motion. We will now explicitly construct such a self dual configuration.

Start with a gauge transformation function which is singular at the origin but maps around the group at a constant radius

$$h(x_\mu) = \frac{t + i\vec{\tau}\cdot\vec{x}}{\sqrt{x^2}} = \frac{T_\mu x_\mu}{|x|}. \tag{135}$$

Here we define the four component object $T_\mu = \{1, i\vec{\tau}\}$. Considering space with the origin removed, construct the pure gauge field

$$B_\mu = \frac{-i}{g}h^\dagger\partial_\mu h = \frac{-i}{g}h^\dagger(T_\mu x^2 - x_\mu T\cdot x)/|x|^3. \tag{136}$$

Because this is nothing but a gauge transformation of a vanishing gauge field, the corresponding field strength automatically vanishes

$$\partial_\mu B_\nu - \partial_\nu B_\mu + ig[B_\mu, B_\nu] = 0. \tag{137}$$

This construction gives a unit winding at infinity. However this gauge field is singular at the origin where the winding unwraps. If we smooth this singularity at $x = 0$ with a field of form

$$A_\mu = f(x^2)B_\mu. \tag{138}$$

where $f(0) = 0$ and $f(\infty) = 1$, this will remove the unwrapping at the origin and automatically leave a field configuration with non-trivial winding. The idea is to find a particular $f(x^2)$ such that A also gives a self dual field strength and thereby is a solution to the equations of motion.

We have set things up symmetrically under space-time rotations about the origin. This connection with $O(4)$ is convenient in that we only need to verify the self duality along a single direction. Consider this to be the time axis, along which self duality requires

$$F_{01}(\vec{x} = 0, t) = \pm F_{23}(\vec{x} = 0, t). \tag{139}$$

A little algebra gives

$$F_{\mu\nu} = (f - f^2)(\partial_\mu B_\nu - \partial_\nu B_\mu) + 2f'(x_\mu B_\nu - x_\nu B_\mu). \tag{140}$$

So along the time axis we have

$$F_{01} \to 2f't\frac{\tau_1}{gt} \qquad\qquad F_{23} \to (f - f^2)\frac{2\tau_1}{gt^2} \tag{141}$$

Thus the self duality condition reduces to a simple first order differential equation

$$zf'(z) = \pm(f - f^2). \tag{142}$$

This is easily solved to give

$$f(z) = \frac{1}{1 + \rho^2 z^{\pm 1}} \tag{143}$$

where ρ is an arbitrary constant of integration. To have the function vanish at the origin we take the minus solution. The resulting form for the gauge field

$$A_\mu = \frac{-ix^2}{g(x^2 + \rho^2)} h^\dagger \partial_\mu h \tag{144}$$

is the self dual instanton. The parameter ρ controls the size of the configuration. Its arbitrary value is a consequence of the conformal invariance of the classical theory. Switching h and $h^\dagger$ gives a solution with the opposite winding.

6.3 Zero modes and the Dirac operator

A particularly important and intriguing aspect of the above field configuration is that it supports an exact zero mode for the classical Dirac operator. We will later discuss the rigorous connection between the gauge field winding and the zero modes of the Dirac operator. Here, however, we will verify this connection explicitly for the above solution. Thus we look for a spinor field $\psi(x)$ satisfying

$$\gamma_\mu D_\mu \psi(x) = \gamma_\mu \left(\partial_\mu + igA_\mu\right)\psi(x) = 0 \tag{145}$$

where we insert the gauge field from Eq. (144). The wave function ψ is a spinor in Dirac space and a doublet in $SU(2)$ space; i.e. it has 8 components. Similarly, $\gamma_\mu A_\mu$ is an 8 by 8 matrix, with a factor of four from spinor space and a factor of two from the internal gauge symmetry. The solution entangles all of these indices in a non-trival manner.

Since we don't want a singularity in ψ at the origin, it is natural to look for a solution of form

$$\psi(x) = p(|x|)V \tag{146}$$

where p is a scalar function of the four dimensional radius and V is a constant vector in spinor and color space. As before, it is convenient to look for the solution along the time axis. There A_0 vanishes and we have

$$\vec{A} = \frac{1}{g}\,\frac{t}{t^2+\rho^2}\vec{\tau} \tag{147}$$

Then the equation of interest reduces to

$$\gamma_0 \frac{d}{dt}\psi(t) = -\frac{t}{t^2+\rho^2}\,\vec{\tau}\cdot\vec{\gamma}\,\psi(t) \tag{148}$$

The 8 by 8 matrix $\vec{\tau}\cdot\vec{\gamma}$ is readily diagonalized giving the eigenvalues $\{-3,1,1,1,1,1,1,3\}$. Only the $+3$ eigenvalue gives a normalizable solution

$$\psi(t) = \psi(0)\exp\left(-3\int_0^t \frac{t\,dt}{t^2+\rho^2}\right). \tag{149}$$

For general x_μ this becomes

$$\psi(x) = \psi(0)\left(\frac{\rho^2}{x^2+\rho^2}\right)^{3/2} \tag{150}$$

At large x this goes at x^{-3} so its square is normalizable. None of the other eigenvalues of $\vec{\tau}\cdot\vec{\gamma}$ go to zero fast enough for normalization; thus, the solution is unique.

We see the appearance of a direct product of two $SU(2)$'s, one from spin and one from isospin. As we rotate around the origin, for the large eigenvalue these rotate together as an overall singlet. The other positive eigenvalues of $\vec{\tau}\cdot\vec{\gamma}$ represent the triplet combination while the negative eigenvalues come from from antiparticle states.

This zero eigenvalue of D is robust under smooth deformations of the gauge field. This is because the anti-commutation of D with γ_5 says that all non-zero eigenvalues of D occur in conjugate pairs. Without bringing in another eigenvalue, the isolated one at zero cannot move. In the next subsection we will demonstrate the general result that for arbitrary smooth gauge fields the number of zero modes of the Dirac operator is directly given by the topological winding number.

6.4 The index theorem

We have seen that associated with a particular topologically non-trivial gauge configuration is a zero action solution to the Dirac equation. Here we will give a simple derivation of the general index theorem relating zero modes of the Dirac operator to the overall topology of the gauge

field. We continue to work directly with the naive continuum Dirac operator. We assume for this section that the gauge fields are smooth and differentiable. While this is unlikely to be true for typical fields in the path integral, the main purpose here is to show that robust zero modes must already exist in the classical theory. We will later see that the generalization of these zero modes to the quantum theory is intimately tied to certain quantum mechanical anomalies crucial to non-perturbative physics.

The combination of the anti-Hermitean character of the classical Dirac operator $D = \gamma_\mu D_\mu$ along with its anti-commutation with γ_5 shows that the non-zero eigenvalues of D all occur in complex conjugate pairs. In particular, if we have

$$D|\psi\rangle = \lambda|\psi\rangle \tag{151}$$

then we immediately obtain the conjugate eigenvector from

$$D\gamma_5\,|\psi\rangle = -\lambda\,\gamma_5|\psi\rangle. \tag{152}$$

Since $|\psi\rangle$ and $\gamma_5|\psi\rangle$ have different eigenvalues under the anti-Hermitean operator D, they must be orthogonal

$$\langle\psi|\gamma_5|\psi\rangle = 0. \tag{153}$$

On the other hand, any exact zero eigenmodes need not be paired. Furthermore, restricted to the space of zero eigenmodes, γ_5 and D commute and can be simultaneously diagonalized. The eigenvalues of γ_5 are all either plus or minus unity. Combining all these ideas together gives a simple method to count the number of zero modes of the Dirac operator weighted by their chirality. In particular we have the relation

$$\nu = n_+ - n_- = \mathrm{Tr}\gamma_5 e^{D^2/\Lambda^2} \tag{154}$$

where $n_\pm$ denotes the number of zero modes with eigenvalue ± 1 under γ_5. Here the parameter Λ is introduced to control the behavior of the trace as the eigenvalues go to infinity. It can be thought of as a regulator, although the above equation is independent of its value.

To proceed, we first write the square of the Dirac operator appearing in the above exponential

$$D^2 = \partial^2 - g^2A^2 + 2igA_\mu\partial_\mu + ig(\partial_\mu A_\mu) - \frac{g}{2}\sigma_{\mu\nu}F_{\mu\nu} \tag{155}$$

where $[\gamma_\mu, \gamma_\nu] = 2i\sigma_{\mu\nu}$. Expanding Eq. (154) for the winding number in powers of the gauge field, the first non-vanishing term appears in the fourth power of the Dirac operator. This involves two powers of the sigma matrices through the relation

$$\mathrm{Tr}\;\gamma_5\sigma_{\mu\nu}\sigma_{\rho\sigma} = 4\epsilon_{\mu\nu\rho\sigma}. \tag{156}$$

Thus our expression for the winding number becomes

$$\nu = \mathrm{Tr}\gamma_5 e^{D^2/\Lambda^2} = \frac{g^2}{2\Lambda^4}\mathrm{Tr}_{x,c}e^{\partial^2/\Lambda^2}\epsilon_{\mu\nu\rho\sigma}F_{\mu\nu}F_{\rho\sigma} + O(\Lambda^{-6}) \tag{157}$$

where $\mathrm{Tr}_{x,c}$ refers to the trace over space and color, the trace over the spinor index having been done to give the factor of the antisymmetric tensor. It is the trace over the space index that will

give a divergent factor removing the Λ^{-4} prefactor. Higher order terms go to zero rapidly enough with Λ to be ignored.

The factor e^{∂^2/Λ^2} serves to mollify traces over position space. Consider some function $f(x)$ as representing a a diagonal matrix in position space $M(x,x') = f(x)\delta(x-x')$. The formal trace would be $\mathrm{Tr}M = \int dx M(x,x)$, but this diverges since it involves a delta function of zero. Writing the delta function in terms of its Fourier transform

$$e^{\partial^2/\Lambda^2}\delta(x-x') = \int \frac{d^4p}{(2\pi)^4} e^{ip\cdot(x-x')} e^{-p^2/\Lambda^2} = \frac{\Lambda^4}{16\pi^2} e^{-(x-x')^2\Lambda^2/4} \tag{158}$$

shows how this "heat kernel" spreads the delta function. This regulates the desired trace

$$\mathrm{Tr}_x f(x) \equiv \frac{\Lambda^4}{16\pi^2}\int d^4x f(x). \tag{159}$$

Using this to remove the spatial trace in the above gives the well known relation

$$\nu = \frac{g^2}{32\pi^2}\mathrm{Tr}_c \int d^4x \epsilon_{\mu\nu\rho\sigma} F_{\mu\nu}F_{\rho\sigma} = \frac{g^2}{16\pi^2}\mathrm{Tr}_c \int d^4x F_{\mu\nu}\tilde{F}_{\mu\nu}. \tag{160}$$

As discussed earlier, this integral involves a total derivative that can be partially integrated into an integral over spatial infinity that counts the topological winding of the gauge field. Thus counting the zero modes of the Dirac operator in a given configuration is an equivalent way to determine this topology. The equivalence of Eq. (133) and Eq. (160) is the index theorem.

7 Chiral symmetry

Much older a tool than the lattice, ideas based on chiral symmetry have historically provided considerable insight into how the strong interactions work. In particular, this concept is crucial to our understanding of why the pion is so much lighter than the rho, despite them both being made of the same quarks. Combining these ideas with the lattice has provided considerable insight into many non-perturbative issues in QCD. Here we review the basic ideas of chiral symmetry for the strong interactions. A crucial aspect of this discussion is the famous anomaly and its consequences for the η' meson.

The classical Lagrangian for QCD couples left and right handed quark fields only through mass terms. Thus naively the massless theory has independent conserved currents associated with each handedness. For N_f massless flavors, this would be an independent $U(N_f)$ symmetry associated with each chirality, giving what is often written in terms of axial and vector fields as an $U(N_f)_V \times U(N_f)_A$ symmetry. As is well known, this full symmetry does not survive quantization, being broken to a $SU(N_f)_V \times SU(N_f)_A \times U(1)_B$, where the $U(1)_B$ represents the symmetry of baryon number conservation. The only surviving axial symmetries of the massless quantum theory are non-singlet under flavor symmetry.

This breaking of the classical $U(1)$ axial symmetry is closely tied to the possibility of introducing into massive QCD a CP violating parameter, usually called Θ. For an extensive review, see Ref. [58]. While such a term is allowed from fundamental principles, experimentally it appears to be extremely small. This raises an unresolved puzzle for attempts to unify the strong

interactions with the weak. Since the weak interactions do violate CP, why is there no residue of this remaining in the strong sector below the unification scale?

One goal of this section is to provide a qualitative understanding of the role of the Θ parameter in meson physics. We concentrate on symmetry alone and do not attempt to rely on any specific form for an effective Lagrangian. We build on the connection between Θ and a flavor-singlet Z_{N_f} symmetry that survives the anomaly. We will see that, when the lightest quarks are made massive and degenerate, a first order transition must occur when Θ passes through π. This transition is quite generic, but can be avoided under limited conditions with one quark considerably lighter than the others. This discussion should also make it clear that the sign of the quark mass is physically relevant for an odd number of flavors. This is a non-perturbative effect that is invisible to naive diagrammatic treatments.

Throughout this section we use the language of continuum field theory. Of course underlying this we must assume some non-perturbative regulator has been imposed so that we can make sense of various products of fields, such as the condensing combination $\sigma = \overline{\psi}\psi$. For a momentum space cutoff, assume that it is much larger than Λ_{QCD}. Correspondingly, for a lattice cutoff imagine that the lattice spacing is much smaller than $1/\Lambda_{QCD}$. In this section we ignore any lattice artifacts that should vanish in the continuum limit. We will return to such issues later when we discuss lattice fermions.

7.1 Effective potentials

We begin with an elementary review of the concept of effective potentials in quantum field theory. In generic continuum field theory language, consider the path integral for a scalar field

$$Z = \int d\phi e^{-S(\phi)}. \tag{161}$$

After adding in some external sources

$$Z(J) = \int d\phi e^{-S(\phi)+J\phi}, \tag{162}$$

general correlation functions can be found by differentiating with respect to J. Here we use a shorthand notation that suppresses the space dependence; *i.e.* $J\phi = \int dx J(x)\phi(x)$ in the continuum, or $J\phi = \sum_i J_i\phi_i$ on the lattice.

One can think of J as an external force pulling on the field. Such a force will tend to drive the the field to have an expectation value

$$\langle\phi\rangle_J = -\frac{\partial F}{\partial J} \tag{163}$$

where the free energy in the presence of the source is defined as $F(J) = -\log(Z(J))$.

Now imagine inverting Eq. (163) to determine what value of the force J would be needed to give some desired expectation value Φ; *i.e.* we want solve

$$\Phi(J) = \langle\phi\rangle_{J(\Phi)} = -\frac{\partial F}{\partial J} \tag{164}$$

for $J(\Phi)$. In terms of this formal solution, construct the "Legendre transform"

$$V(\Phi) = F(J(\Phi)) + \Phi J(\Phi) \tag{165}$$

and look at

$$\frac{\partial V}{\partial \Phi} = -\Phi \frac{\partial J}{\partial \Phi} + J + \Phi \frac{\partial J}{\partial \Phi} = J. \tag{166}$$

If we now turn off the sources, this derivative vanishes. Thus the expectation value of the field in the absence of sources occurs at an extremum of $V(\Phi)$. This quantity V is referred to as the "effective potential."

An interesting formal property of this construction follows from looking at the second derivative of V

$$\frac{\partial^2 V}{\partial \Phi^2} = \frac{\partial J}{\partial \Phi} \tag{167}$$

Actually, it is easier to look at the inverse

$$\frac{\partial \Phi}{\partial J} = -\frac{\partial^2 F}{\partial J^2} = \langle \phi^2 \rangle - \langle \phi \rangle^2 = \langle (\phi - \langle \phi \rangle)^2 \rangle \geq 0. \tag{168}$$

Thus this second derivative is never negative! This first of all shows we are actually looking for a minimum and not a maximum of V, but it also implies that $V(\Phi)$ can only have ONE minimum!

This convexity property is usually ignored in conventional discussions, where phase transitions are signaled by jumps between distinct minima of the potential.. So what is going on? Are phase transitions impossible? Physically, the more you pull on the field, the larger the expectation of Φ will become. It won't go back. The proper interpretation is that we must do Maxwell's construction. If we force the expectation of ϕ to lie between two distinct stable phases, the system will phase separate into a heterogeneous mixture. In this region the effective potential is flat. Note that there is no large volume limit required in the above discussion. However other definitions of V can allow a small barrier at finite volume due to surface tension effects. A mixed phase must contain interfaces, and their energy represents a small barrier.

7.2 Goldstone Bosons

Now we turn to a brief discussion on some formal aspects of Goldstone Bosons. Suppose we have a field theory containing a conserved current

$$\partial_\mu j_\mu = 0 \tag{169}$$

so the corresponding charge $Q = \int d^3x j_0(x)$ is a constant

$$\frac{dQ}{dt} = -i[H, Q] = 0. \tag{170}$$

Here H is the Hamiltonian for the system under consideration. Suppose, however, that for some reason the vacuum is not a singlet under this charge

$$Q|0\rangle \neq 0 \tag{171}$$

Then there must exist a massless particle in the theory. Consider the state

$$\exp(i\theta \int d^3x j_0(x) e^{-\epsilon x^2})|0\rangle \tag{172}$$

where ϵ is a convenient cutoff and θ some parameter. As epsilon goes to zero this state by assumption is not the vacuum, but since the Hamiltonian commutes with Q, the expectation value of the Hamiltonian goes to zero (normalize so the ground state energy is zero). We can thus find a state that is not the vacuum but with arbitrarily small energy. The theory has no mass gap. This situation of having a symmetry under which the vacuum is not invariant is referred to as "spontaneous symmetry breaking." The low energy states represent massless particles called Goldstone bosons [59].

Free massless field theory is a marvelous example where everything can be worked out. The massless equation of motion

$$\partial_\mu \partial_\mu \phi = 0 \tag{173}$$

can be written in the form

$$\partial_\mu j_\mu = 0 \tag{174}$$

where

$$j_\mu = \partial_\mu \phi. \tag{175}$$

The broken symmetry is the invariance of the Lagrangian $L = \int d^4x (\partial_\mu \phi)^2/2$ under constant shifts of the field

$$\phi \to \phi + c. \tag{176}$$

Note that $j_0 = \partial_0 \phi = \pi$ is the conjugate variable to ϕ. Since it is a free theory, one could work out explicitly

$$\langle 0| \exp(i\theta \int d^3x j_0(x) e^{-\epsilon x^2/2})|0\rangle. \tag{177}$$

We can, however, save ourselves the work using dimensional analysis. The field ϕ has dimensions of inverse length, while j_0 goes as inverse length squared. Thus θ above has units of inverse length. These are the same dimensions as ϵ^2. Now for a free theory, by Wick's theorem, the answer must be Gaussian in θ. We conclude that the above overlap must go as

$$\exp(-C\theta^2/\epsilon^4) \tag{178}$$

where C is some non-vanishing dimensionless number. This expression rapidly goes to zero as epsilon becomes small, showing that the vacuum is indeed not invariant under the symmetry. In the limit of ϵ going to zero, we obtain a new vacuum that is not even in the same Hilbert space. The overlap of this new state with any local polynomial of fields on the original vacuum vanishes.

7.3 Pions and spontaneous symmetry breaking

We now extend the effective potential to a function of several relevant meson fields in QCD. Intuitively, V represents the energy of the lowest state for a given field expectation, as discussed more formally earlier via a Legendre transformation. Here we will ignore the result that effective potentials must be convex functions of their arguments. As discussed, this issue is easily

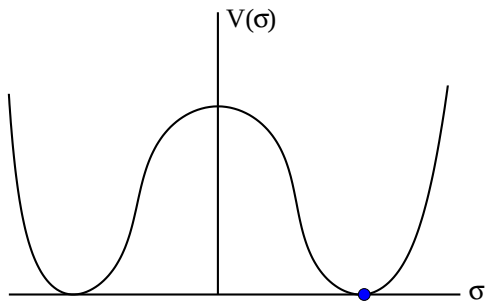

Fig. 5. Spontaneous chiral symmetry breaking is represented by a double well effective potential with the vacuum settling into one of two possible minima. In this minimum chiral symmetry is broken by the selection of a specific value for the quark condensate.

understood in terms of a Maxwell construction involving the phase separation that will occur if one asks for a field expectation in what would otherwise be a concave region. Thus we will use the traditional language of spontaneous symmetry breaking corresponding to having an effective potential with more than one minimum. When the underlying theory possesses some symmetry but the individual minima do not, spontaneous breaking comes about when the vacuum selects one of the minima arbitrarily. The discussion here closely follows that in Ref. [60].

We work here with the composite scalar and pseudoscalar fields

$$\begin{array}{ccc} \sigma & \sim & \overline{\psi}\psi \\ \pi_\alpha & \sim & i\overline{\psi}\lambda_\alpha\gamma_5\psi \\ \eta' & \sim & i\overline{\psi}\gamma_5\psi \end{array} \tag{179}$$

Here the λ_α are the structure constants for the flavor group $SU(N_f)$. They are generalizations of the usual Gell-Mann matrices from $SU(3)$; however, now we are concerned with the flavor group, not the internal symmetry group related to confinement. As mentioned earlier, we must assume that some sort of regulator, perhaps a lattice, is in place to define these products of fields at the same point. Indeed, most of the quantities mentioned in this section are formally divergent, although we will concentrate on those aspects that survive the continuum limit.

To simplify the discussion, consider degenerate quarks with a small common mass m. Later we will work out in some detail the two flavor case for non-degenerate quarks. It is also convenient to initially restrict N_f to be even, saving for later some interesting subtleties arising with an odd number of flavors. And we assume N_f is small enough to maintain asymptotic freedom as well as avoiding any possible conformal phases.

The conventional picture of spontaneous chiral symmetry breaking in the limit of massless quarks assumes that the vacuum acquires a quark condensate with

$$\langle\overline{\psi}\psi\rangle = \langle\sigma\rangle = v \neq 0. \tag{180}$$

In terms of the effective potential, $V(\sigma)$ should acquire a double well structure, as sketched in Fig. 5. The symmetry under $\sigma \leftrightarrow -\sigma$ is associated with the invariance of the action under a flavored chiral rotation. For example, with two flavors the change of variables

$$\begin{array}{l} \psi \to e^{i\pi\tau_3\gamma_5/2}\psi = i\tau_3\gamma_5\psi \\ \overline{\psi} \to \overline{\psi}e^{i\pi\tau_3\gamma_5/2} = \overline{\psi}\, i\tau_3\gamma_5 \end{array} \tag{181}$$

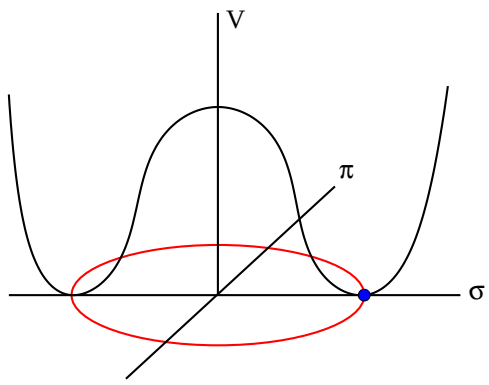

Fig. 6. The flavor non-singlet pseudoscalar mesons are Goldstone bosons corresponding to flat directions in the effective potential.

leaves the massless action invariant but takes σ to its negative. Here τ_3 is the conventional Pauli matrix corresponding to the third component of isospin.

Extending the effective potential to a function of the non-singlet pseudoscalar fields gives the standard picture of Goldstone bosons. These are massless when the quark mass vanishes, corresponding to $N_f^2 - 1$ "flat" directions for the potential. One such direction is sketched schematically in Fig. 6. For the two flavor case, these rotations represent a symmetry mixing the sigma field with the pions

$$\begin{aligned} \sigma &\to \quad \sigma\cos(\phi) + \pi^\alpha \sin(\phi) \\ \pi^\alpha &\to -\sigma\sin(\phi) + \pi^\alpha\cos(\phi). \end{aligned} \tag{182}$$

In some sense the pions are waves propagating through the non-vanishing sigma condensate. The oscillations of these waves occur in a direction "transverse" to the sigma expectation. They are massless because there is no restoring force in that direction.

If we now introduce a small mass for the quarks, this will effectively tilt the potential $V(\sigma) \to V(\sigma) - m\sigma$. This selects one minimum as the true vacuum. The tilting of the potential breaks the global symmetry and gives the Goldstone bosons a small mass proportional to the square root of the quark mass, as sketched in Fig. 7. The standard chiral Lagrangean approach is a simultaneous expansion in the masses and momenta of these light particles.

As discussed earlier, Goldstone bosons are associated with conserved currents with charges that do not leave the vacuum invariant. In the present case these are the axial currents formally given by the quark bilinears

$$A_\mu^\alpha = \overline{\psi}\lambda^\alpha\gamma_\mu\gamma_5\psi. \tag{183}$$

Combined with the vector fields

$$V_\mu^\alpha = \overline{\psi}\lambda^\alpha\gamma_\mu\psi, \tag{184}$$

these give rise to the famous algebra of currents. Indeed, it was this algebra that motivated Bjorken to propose the idea of scaling in deep inelastic lepton scattering [61, 62].

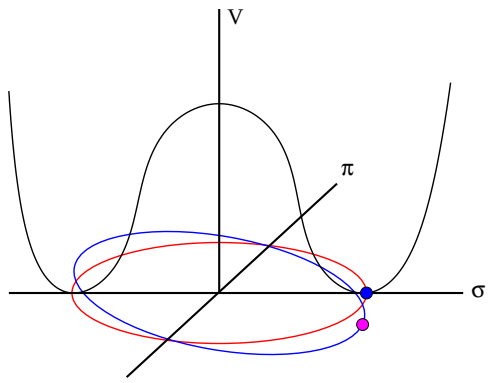

Fig. 7. A small quark mass term tilts the effective potential, selecting one direction for the true vacuum and giving the Goldstone bosons a mass proportional to the square root of the quark mass.

8 The chiral anomaly

This picture is, of course, completely standard. It is also common lore that the anomaly prevents the η' from being a Goldstone boson and leaves it with a mass of order Λ_{QCD}, even in the massless quark limit. The issue is that the effective potential V must not be symmetric under an anomalous rotation between η' and σ

$$\begin{aligned} \sigma &\to \quad \sigma\cos(\phi) + \eta'\sin(\phi) \\ \eta' &\to -\sigma\sin(\phi) + \eta'\cos(\phi). \end{aligned} \tag{185}$$

In the next subsection we discuss how this symmetry disappears and its connection to the zero modes of the Dirac operator.

If we consider the effective potential as a function of the fields σ and η', it should have a minimum at $\sigma \sim v$ and $\eta' \sim 0$. Expanding about that point we expect a qualitative form

$$V(\sigma,\eta') \sim m_\sigma^2(\sigma - v)^2 + m_{\eta'}^2 {\eta'}^2 + O((\sigma - v)^3, {\eta'}^4). \tag{186}$$

We expect both m_σ and $m_{\eta'}$ to remain of order Λ_{QCD}, even in the chiral limit. And, at least with an even number of flavors as we are currently considering, there should be a second minimum with $\sigma \sim -v$.

At this point one can ask whether we know anything else about the effective potential in the (σ, η') plane. We will shortly see that indeed we do, and the potential has a total of N_f equivalent minima in the chiral limit. But first we review how the above minima arise in quark language.

8.1 What broke the symmetry?

The classical QCD Lagrangian has a symmetry under a rotation of the underlying quark fields

$$\begin{aligned} \psi &\to e^{i\phi\gamma_5/2}\psi \\ \overline{\psi} &\to \overline{\psi}e^{i\phi\gamma_5/2} \end{aligned} \tag{187}$$

This corresponds directly to the transformation of the composite fields given in Eq. 185. This symmetry is "anomalous" in that any regulator must break it with a remnant surviving in the continuum limit [63–65].

The specifics of how the anomaly works depend on the details of the regulator. Here we will follow Fujikawa [66] and consider the fermionic measure in the path integral. If we make the above rotation on the field ψ, the measure changes by the determinant of the rotation matrix

$$d\psi \to |e^{-i\phi\gamma_5/2}|d\psi = e^{-i\phi \mathrm{Tr}\gamma_5/2}d\psi. \tag{188}$$

Here is where the subtlety of the regulator comes in. Naively γ_5 is a simple four by four traceless matrix. If it is indeed traceless, then the measure would be invariant. However, in the regulated theory this is not the case. This is intimately tied with the index theorem for the Dirac operator in topologically non-trivial gauge fields.

A typical Dirac action takes the form $\overline{\psi}(D+m)\psi$ with the kinetic term D a function of the gauge fields. In the naive continuum theory D is anti-Hermitian, $D^\dagger = -D$, and anti-commutes with γ_5, i.e. $[D,\gamma_5]_+ = 0$. What complicates the issue with fermions is the index theorem discussed earlier and reviewed in Ref. [67]. If a background gauge field has winding ν, then there must be at least ν exact zero eigenvalues of D. Furthermore, on the space spanned by the corresponding eigenvectors, γ_5 can be simultaneously diagonalized with D. The net winding number equals the number of positive eigenvalues of γ_5 minus the number of negative eigenvalues. In this subspace the trace of γ_5 does not vanish, but equals ν.

What about the higher eigenvalues of D? We discussed these earlier when we formulated the index theorem. Because $[D,\gamma_5]_+ = 0$, non-vanishing eigenvalues appear in opposite sign pairs; i.e. if $D|\psi\rangle = \lambda|\psi\rangle$ then $D\gamma_5|\psi\rangle = -\lambda\gamma_5|\psi\rangle$. For an anti-Hermitean D, these modes are orthogonal with $\langle\psi|\gamma_5|\psi\rangle = 0$. As a consequence, γ_5 is traceless on the subspace spanned by each pair of eigenvectors.

So what happened to the opposite chirality states to the zero modes? In a regulated theory they are in some sense "above the cutoff." In a simple continuum discussion they have been "lost at infinity." With a lattice regulator there is no "infinity"; so, something more subtle must happen. With the overlap [68, 69] or Wilson [10] fermions, discussed in more detail later, one gives up the anti-Hermitean character of D. Most eigenvalues still occur in conjugate pairs and do not contribute to the trace of γ_5. However, in addition to the small real eigenvalues representing the zero modes, there are additional modes where the eigenvalues are also real but large. With Wilson fermions these appear as massive doubler states. With the overlap, the eigenvalues are constrained to lie on a circle. In this case, for every exact zero mode there is another mode with the opposite chirality lying on the opposite side of the circle. These modes are effectively massive and break chiral symmetry. The necessary involvement of both small and large eigenvalues warns of the implicit danger in attempts to separate infrared from ultraviolet effects. When the anomaly is concerned, going to short distances is not sufficient for ignoring non-perturbative effects related to topology.

So with the regulator in place, the trace of γ_5 does not vanish on gauge configurations of non-trivial topology. The change of variables indicated in Eq. 188 introduces into the path integral a modification of the weighting by a factor

$$e^{-i\phi \mathrm{Tr}\gamma_5} = e^{-i\phi N_f \nu}. \tag{189}$$

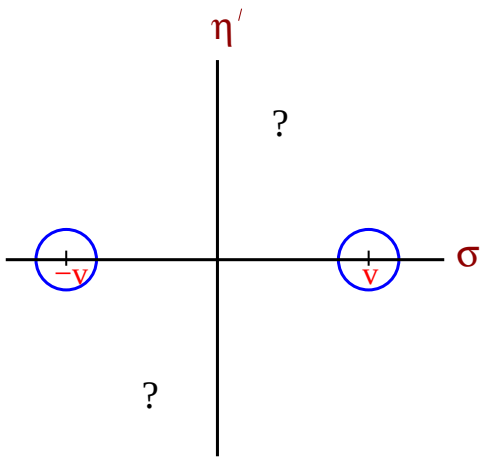

Fig. 8. We have two minima in the σ, η' plane located at $\sigma = \pm v$ and $\eta' = 0$. The circles represent that the fields will fluctuate in a small region about these minima. Can we find any other minima?

Here we have applied the rotation to all flavors equally, thus the factor of N_f in the exponent. The conclusion is that gauge configurations that have non-trivial topology receive a complex weight after the anomalous rotation. Although not the topic of discussion here, note that this factor introduces a sign problem if one wishes to study this physics via Monte Carlo simulations. Here we treat all N_f flavors equivalently; this corresponds to dividing the conventionally defined CP violation angle, to be discussed later, equally among the flavors, i.e. effectively $\phi = \Theta/N_f$.

8.2 A discrete chiral symmetry

We now return to the effective Lagrangian language of before. For the massless theory, the symmetry under $\sigma \leftrightarrow -\sigma$ indicates that we expect at least two minima for the effective potential considered in the σ, η' plane. These are located as sketched in Fig. 8. Do we know anything about the potential elsewhere in this plane? The answer is yes, indeed there are actually N_f equivalent minima.

It is convenient to separate the left- and right-hand parts of the fermion field

$$\begin{aligned} \psi_{L,R} &= \frac{1}{2}(1 \pm \gamma_5)\psi \\ \overline{\psi}_{L,R} &= \overline{\psi}\frac{1}{2}(1 \mp \gamma_5). \end{aligned} \tag{190}$$

The mass term is thus

$$m\overline{\psi}\psi = m(\overline{\psi}_L\psi_R + \overline{\psi}_R\psi_L) \tag{191}$$

and mixes the left and right components.

Using this notation, due to the anomaly the singlet rotation

$$\psi_L \to e^{i\phi}\psi_L \tag{192}$$

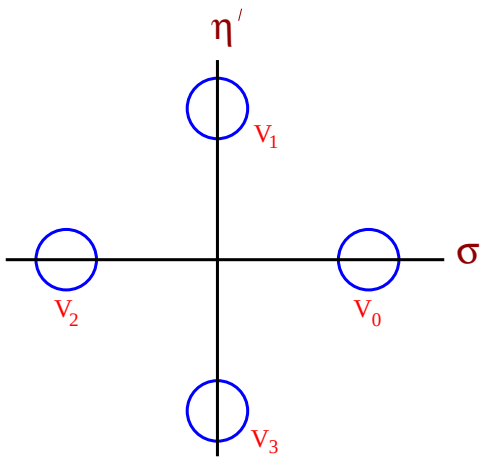

Fig. 9. For four massless flavors we have four equivalent minima in the σ, η' plane. This generalizes to N_f minima with N_f flavors.

is not a valid symmetry of the theory for generic values of the angle ϕ. On the other hand, flavored chiral symmetries should survive, and in particular

$$\psi_L \to g_L \psi_L = e^{i\phi_\alpha \lambda_\alpha} \psi_L \tag{193}$$

is expected to be a valid symmetry for any set of angles ϕ_α. The point of this subsection is that, for special special discrete values of the angles, the rotations in Eq. 192 and Eq. 193 can coincide. At such values the singlet rotation becomes a valid symmetry. In particular, note that

$$g = e^{2\pi i \phi / N_f} \in Z_{N_f} \subset SU(N_f). \tag{194}$$

Thus a valid discrete symmetry involving only σ and η' is

$$\begin{aligned} \sigma &\to \quad \sigma \cos(2\pi/N_f) + \eta' \sin(2\pi/N_f) \\ \eta' &\to -\sigma \sin(2\pi/N_f) + \eta' \cos(2\pi/N_f). \end{aligned} \tag{195}$$

The potential $V(\sigma, \eta')$ has a Z_{N_f} symmetry manifested in N_f equivalent minima in the (σ, η') plane. For four flavors this structure is sketched in Fig. 9.

This discrete flavor singlet symmetry arises from the trivial fact that Z_N is a subgroup of both $SU(N)$ and $U(1)$. At the quark level the symmetry is easily understood since the quark measure receives an additional phase proportional to the winding number from every flavor. With N_F flavors, these multiply together making

$$\psi_L \to e^{2\pi i / N_f} \psi_L \tag{196}$$

a valid symmetry even though rotations by smaller angles are not.

The role of the Z_N center of $SU(N)$ is illustrated graphically in Fig. 10, taken from Ref. [25]. Here we plot the real and the imaginary parts of the traces of 10,000 $SU(3)$ matrices drawn randomly with the invariant group measure. The region of support only touches the $U(1)$ circle at the elements of the center. All elements lie on or within the curve mapped out by elements of form $\exp(i\phi\lambda_8)$. Fig. 11 is a similar plot for the group $SU(4)$.

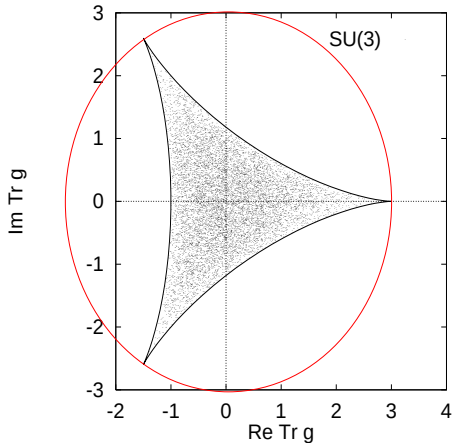

Fig. 10. The real and imaginary parts for the traces of 10,000 randomly chosen $SU(3)$ matrices. All points lie within the boundary representing matrices of the form $\exp(i\phi\lambda_8)$. The tips of the three points represent the center of the group. The outer curve represents the boundary that would be found if the group was the full $U(1)$. Taken from Ref. [25].

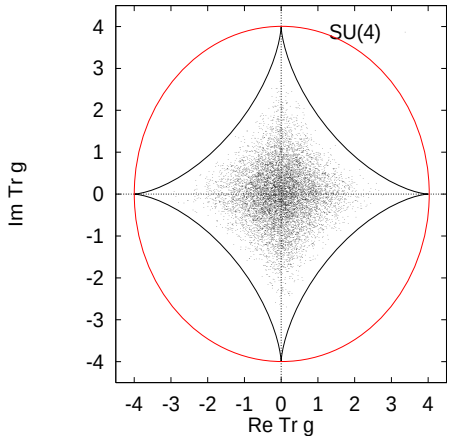

Fig. 11. The generalization of Fig. 10 to $SU(4)$. The real and imaginary parts for the traces of 10,000 randomly chosen $SU(4)$ matrices. Taken from Ref. [25].

8.3 The 't Hooft vertex

The consequences of non-trivial gauge topology and the connections to the anomaly are often described in terms of an effective multi fermion interaction referred to as the " 't Hooft vertex." To understand the 't Hooft interaction in path integral language, we begin with a reminder of the underlying strategy of lattice simulations. Consider the generic path integral, or "partition function," for quarks and gluons

$$Z = \int (dA)(d\psi\, d\overline{\psi}) \exp\left(-S_g(A) - \overline{\psi} D(A)\psi\right). \tag{197}$$

Here A denotes the gauge fields and $\overline{\psi}, \psi$ the quark fields. The pure gauge part of the action is $S_g(A)$ and the matrix describing the fermion part of the action is $D(A)$. Since direct numerical evaluation of the fermionic integrals appears to be impractical, the Grassmann integrals are conventionally evaluated analytically, reducing the partition function to

$$Z = \int (dA)\, e^{-S_g(A)}\, |D(A)|. \tag{198}$$

Here $|D(A)|$ denotes the determinant of the Dirac matrix evaluated in the given gauge field. Thus motivated, the basic lattice approach is to generate a set of random gauge configurations weighted by $\exp(-S_g(A))\, |D(A)|$. Given an ensemble of such configurations, one then estimates physical observables by averages over this ensemble.

This procedure seems innocent enough, but it can run into trouble when one has massless fermions and corresponding zero modes associated with topology. To see the issue, write the determinant as a product of the eigenvalues λ_i of the matrix D. In general D may not be a normal matrix; so, one should pick either left or right eigenvectors at one's discretion. This is a technical detail that will not play any further role here. In order to control infrared issues with massless quarks, introduce a small explicit mass m and reduce the path integral to

$$Z = \int (dA)\, e^{-S_g(A)} \prod (\lambda_i + m). \tag{199}$$

Now suppose we have a configuration where one of the eigenvalues of $D(A)$ vanishes, *i.e.* assume that some $\lambda_i = 0$. This, of course, is what happens with non-trivial topology present. As we take the mass to zero, any configurations involving such an eigenvalue will drop out of the ensemble. At first one might suspect this would be a set of measure zero in the space of all possible gauge fields. However, as discussed above, the index theorem ties gauge field topology to such zero modes. In general these modes are robust under small deformations of the fields. Under the traditional lattice strategy the corresponding configurations would then have zero weight in the massless limit. The naive conclusion is that such configurations are irrelevant to physics in the chiral limit.

It was this reasoning that 't Hooft showed to be incorrect. Indeed, he demonstrated that it is natural for some observables to have $1/m$ factors when zero modes are present. These can cancel the terms linear in m from the determinant, leaving a finite contribution.

As a simple example, consider the quark condensate in one flavor QCD

$$\langle \overline{\psi}\psi \rangle = \frac{1}{VZ} \int (dA)\, e^{-S_g}\, |D|\ \mathrm{Tr} D^{-1}. \tag{200}$$

Here V represents the system volume, inserted to give an intensive quantity. Expressing the fermionic factors in terms of the eigenvalues of D reduces this to

$$\langle\overline{\psi}\psi\rangle = \frac{1}{VZ}\int (dA)\, e^{-S_g}\, \prod(\lambda_i + m)\ \sum_i \frac{1}{\lambda_i + m}. \tag{201}$$

Now if there is a mode with $\lambda_i = 0$, the factor of m is canceled by a $1/m$ piece in the trace of D^{-1}. Configurations containing a zero mode give a constant contribution to the condensate and this contribution survives in the massless limit. Note that this effect is unrelated to spontaneous breaking of chiral symmetry and appears even with finite volume.

This contribution to the condensate is special to the one-flavor theory. Because of the anomaly, this quark condensate is not an order parameter for any symmetry. With more fermion species there will be additional factors of m from the determinant. Then the effect of the 't Hooft vertex is of higher order in the fermion fields and does not appear directly in the condensate. For two or more flavors the standard Banks-Casher picture [70] of an eigenvalue accumulation leading to the spontaneous breaking of chiral symmetry should apply.

The conventional discussion of the 't Hooft vertex starts by inserting fermionic sources into the path integral

$$Z(\eta,\overline{\eta}) = \int (dA)\ (d\psi)\ (d\overline{\psi})\ e^{-S_g - \overline{\psi}(D+m)\psi + \overline{\psi}\eta + \overline{\eta}\psi}. \tag{202}$$

Differentiation, in the Grasmannian sense, with respect to these sources will generate the expectation for an arbitrary product of fermionic operators. Integrating out the fermions reduces this to

$$Z = \int (dA)\ e^{-S_g + \overline{\eta}(D+m)^{-1}\eta}\ \prod(\lambda_i + m). \tag{203}$$

Consider a zero mode ψ_0 satisfying $D\psi_0 = 0$. In general there is also a left zero mode satisfying $\overline{\psi}_0 D = 0$. If the sources have an overlap with the mode, that is $(\overline{\eta}|\psi_0) \neq 0$, then a factor of $1/m$ in the source term can cancel the m from the determinant. Although non-trivial topological configurations do not contribute to Z, their effects can survive in correlation functions. For the one-flavor theory the effective interaction is bilinear in the fermion sources and is proportional to

$$(\overline{\eta}|\psi_0)(\overline{\psi}_0|\eta). \tag{204}$$

As discussed earlier, the index theorem tells us that in general the zero mode is chiral; it appears in either $\overline{\eta}_L\eta_R$ or $\overline{\eta}_R\eta_L$, depending on the sign of the gauge field winding.

With $N_f \geq 2$ flavors, the cancellation of the mass factors in the determinant requires source factors from each flavor. This combination is the 't Hooft vertex. It is an effective $2N_f$ fermion operator. In the process, every flavor flips its spin, as sketched in Fig. 12. Indeed, this is the chiral anomaly; left and right helicities are not separately conserved.

Because of Pauli statistics, the multi-flavor vertex can be written in the form of a determinant. This clarifies how the vertex preserves flavored chiral symmetries. With two flavors, call their sources u and d, Eq. 204 generalizes to

$$\begin{vmatrix} (\overline{u}|\psi_0)(\overline{\psi}_0|u) & (\overline{u}|\psi_0)(\overline{\psi}_0|d) \\ (\overline{d}|\psi_0)(\overline{\psi}_0|u) & (\overline{d}|\psi_0)(\overline{\psi}_0|d) \end{vmatrix}. \tag{205}$$

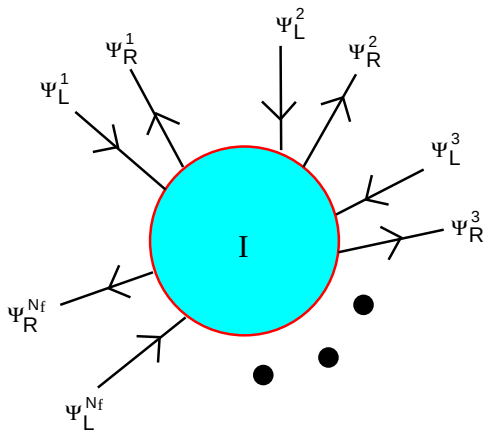

Fig. 12. The 't Hooft vertex for N_f flavors is a $2N_f$ effective fermion operator that flips the spin of every flavor.

Note that the effect of the vertex is non-local. In general the zero mode ψ_0 is spread out over the finite region of the "instanton", *i.e.* the size parameter ρ from the explicit solution given earlier. This means there is an inherent position space uncertainty on where the fermions are interacting. A particular consequence is that fermion conservation is only a global symmetry. In Minkowski space language, this non-locality can be thought of in terms of states sliding in and out of the Dirac sea at different locations.

8.4 Fermions in higher representations

When the quarks are massless, the classical field theory corresponding to the strong interactions has a $U(1)$ axial symmetry under the transformation

$$\psi \to e^{i\theta\gamma_5}\psi \qquad \overline{\psi} \to \overline{\psi}e^{i\theta\gamma_5}. \tag{206}$$

It is the 't Hooft vertex that explains how this symmetry does not survive quantization. In this subsection we discuss how when the quarks are in non-fundamental representations of the gauge group, discrete subgroups of this symmetry can remain because of additional zeros in the Dirac operator.

While these considerations do not apply to the usual theory of the strong interactions where the quarks are in the fundamental representation, there are several reasons to study them anyway. At higher energies, perhaps as being probed at the Large Hadron Collider, one might well discover new strong interactions that play a substantial role in the spontaneous breaking of the electroweak theory. Also, many grand unified theories involve fermions in non-fundamental representations. As one example, we will see that massless fermions in the 10 representation of $SU(5)$ possess a Z_3 discrete chiral symmetry. Similarly the left handed 16 covering representation of $SO(10)$ gives a chiral gauge theory with a surviving discrete Z_2 chiral symmetry. Understanding these symmetries may eventually play a role in an eventual discretization of chiral gauge theories on the lattice.

Here we are generalizing the index theorem relating gauge field topology to zero modes of the Dirac operator. In particular, fermions in higher representations can involve in multiple zero modes for a given winding. Being generic, consider representation X of a gauge group G. Denote by N_X the number of zero modes that are required per unit of winding number in the gauge fields. That is, suppose the index theorem generalizes to

$$n_r - n_l = N_X \nu \tag{207}$$

where n_r and n_l are the number of right and left handed zero modes, respectively, and ν is the winding number of the associated gauge field. The basic 't Hooft vertex receives contributions from each zero mode, resulting in an effective operator which is a product of $2N_X$ fermion fields. Schematically, the vertex is modified along the lines $\overline{\psi}_L \psi_R \longrightarrow (\overline{\psi}_L \psi_R)^{N_X}$. While this form still breaks the $U(1)$ axial symmetry, it is invariant under $\psi_R \to e^{2\pi i/N_X}\psi_R$. In other words, there is a Z_{N_X} discrete axial symmetry.

There are a variety of convenient tools for determining N_X. Consider building up representations from lower ones. Take two representations X_1 and X_2 and form the direct product representation $X_1 \otimes X_2$. Let the matrix dimensions for X_1 and X_2 be D_1 and D_2, respectively. Then for the product representation we have

$$N_{X_1 \otimes X_2} = N_{X_1} D_{X_2} + N_{X_2} D_{X_1}. \tag{208}$$

To see this, start with X_1 and X_2 representing two independent groups G_1 and G_2. With G_1 having winding, there will be a zero mode for each of the dimensions of the matrix index associated with X_2. Similarly there will be multiple modes for winding in G_2. These modes are robust and all should remain if we now constrain the groups to be the same.

As a first example, denote the fundamental representation of $SU(N)$ as F and the adjoint representation as A. Then using $\overline{F} \otimes F = A \oplus 1$ in the above gives $N_A = 2N_F$, as noted some time ago [71]. With $SU(3)$, fermions in the adjoint representation will have six-fold degenerate zero modes.

For another example, consider $SU(2)$ and build up towards arbitrary spin $s \in \{0, \frac{1}{2}, 1, \frac{3}{2}, \ldots\}$. Recursing the above relation gives the result for arbitrary spin

$$N_s = s(2s+1)(2s+2)/3. \tag{209}$$

Another technique for finding N_X in more complicated groups begins by rotating all topological structure into an $SU(2)$ subgroup and then counting the corresponding $SU(2)$ representations making up the larger representation of the whole group. An example to illustrate this procedure is the antisymmetric two indexed representation of $SU(N)$. This representation has been extensively used in [72–75] for an alternative approach to the large gauge group limit. The basic $N(N-1)/2$ fermion fields take the form

$$\psi_{ab} = -\psi_{ba}, \qquad a, b \in 1, 2, ...N. \tag{210}$$

Consider rotating all topology into the $SU(2)$ subgroup involving the first two indices, i.e. 1 and 2. Because of the anti-symmetrization, the field ψ_{12} is a singlet in this subgroup. The field pairs $(\psi_{1,j}, \psi_{2,j})$ form a doublet for each $j \geq 3$. Finally, the $(N-2)(N-3)/2$ remaining fields do not transform under this subgroup and are singlets. Overall we have $N-2$ doublets under the $SU(2)$ subgroup, each of which gives one zero mode per winding number. We conclude

that the 't Hooft vertex leaves behind a Z_{N-2} discrete chiral symmetry. Specializing to the 10 representation of $SU(5)$, this is the Z_3 mentioned earlier.

Another example is the group $SO(10)$ with fermions in the 16 dimensional covering group. This forms the basis of a rather interesting grand unified theory, where one generation of fermions is placed into a single left handed 16 multiplet [76]. This representation includes two quark species interacting with the $SU(3)$ subgroup of the strong interactions, Rotating a topological excitation into this subgroup, we see that the effective vertex will be a four fermion operator and preserve a Z_2 discrete chiral symmetry.

It is unclear whether these discrete symmetries are expected to be spontaneously broken. Since they are discrete, such breaking is not associated with Goldstone bosons. But the quark condensate does provide an order parameter; so, when $N_X > 1$, any such breaking would be conceptually meaningful. This could be checked in numerical simulations.

9 Massive quarks and the Theta parameter

As discussed earlier and illustrated in Fig. 9, a quark mass term $-m\overline{\psi}\psi \sim -m\sigma$ is represented by a "tilting" of the effective potential. This selects one of the multiple minima in the σ, η' plane as the true vacuum. For masses small compared to the scale of QCD, the other minima will persist as extrema, although due to the flat flavor non-singlet directions, some of them will become unstable under small fluctuations. Counting the minima sequentially with the true vacuum having $n = 0$, each is associated with small excitations in the pseudo-Goldstone directions having an effective mass of $m_\pi^2 \sim m\cos(2\pi n/N_f)$. Note that when N_f exceeds four, there will be more than one meta-stable state. However, in the usual case of considering two or three quarks as light, only one minimum remains locally stable.

9.1 Twisted tilting

Conventionally the mass tilts the potential downward in the positive σ direction. However, it is an interesting exercise to consider tilts in other directions in the σ, η' plane. This is accomplished with an anomalous rotation on the mass term

$$\begin{aligned} -m\overline{\psi}\psi \quad &\rightarrow \quad -m\cos(\phi)\overline{\psi}\psi - im\sin(\phi)\overline{\psi}\gamma_5\psi \\ &\sim \quad -m\cos(\phi)\sigma + m\sin(\phi)\eta'. \end{aligned} \tag{211}$$

Were it not for the anomaly, this would just be a redefinition of fields. However the same effect that gives the η' its mass indicates that this new form for the mass term gives an inequivalent theory. As $i\overline{\psi}\gamma_5\psi$ is odd under CP, this theory is explicitly CP violating.

The conventional notation for this effect involves the angle $\Theta = N_f\phi$. Then the Z_{N_f} symmetry amounts to a 2π periodicity in Θ. As Fig. 13 indicates, at special values of the twisting angle ϕ, there will exist two degenerate minima. This occurs, for example, at $\phi = \pi/N_f$ or $\Theta = \pi$. As the twisting increases through this point, there will be a first order transition as the true vacuum jumps from the vicinity of one minimum to the next.

Because of the Z_{N_f} symmetry of the massless theory, all the N_f separate minima are physically equivalent. This means that if we apply our mass term in the direction of any of them, we obtain the same theory. In particular, for four flavors the usual mass term $m\overline{\psi}\psi$ is equivalent

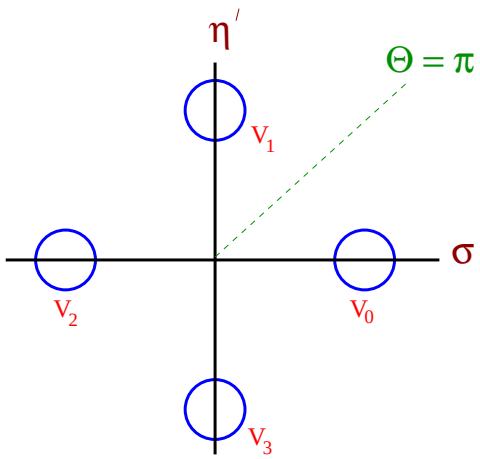

Fig. 13. With massive quarks and a twisting angle of $\phi = \pi/N_f$, two of the minima in the σ, η' plane become degenerate. This corresponds to a first order transition at $\Theta = \pi$.

to using the alternative mass term $im\overline{\psi}\gamma_5\psi$. This result, however, is true if and only if N_f is a multiple of four.

9.2 Odd N_f

One interesting consequence of this picture concerns QCD with an odd number of flavors. The group $SU(N_f)$ with odd N_f does not include the element -1. In particular, the Z_{N_f} structure is not symmetric under reflections about the η' axis. Fig. 14 sketches the situation for $SU(3)$. One immediate conclusion is that positive and negative mass are not equivalent. Indeed, a negative mass with three degenerate flavors corresponds to the $\Theta = \pi$ case and a spontaneous breaking of CP is expected. In this case there is no symmetry under taking $\sigma \sim \overline{\psi}\psi$ to its negative. The simple picture sketched in Fig. 5 no longer applies.

At $\Theta = \pi$ the theory lies on top of a first order phase transition line. A simple order parameter for this transition is the expectation value for the η' field. As this field is odd under CP symmetry, this shows that negative mass QCD with an odd number of flavors spontaneously breaks CP.[6] This does not contradict the Vafa-Witten theorem [77] because in this regime the fermion determinant is not positive definite.

Note that the asymmetry in the sign of the quark mass is not easily seen in perturbation theory. Any quark loop in a perturbative diagram can have the sign of the quark mass flipped by a γ_5 transformation. It is only through the subtleties of regulating the divergent triangle diagram [63–65] that the sign of the mass enters.

A remarkable conclusion of these observations is that two physically distinct theories can have identical perturbative expansions. For example, with flavor $SU(3)$ the negative mass theory has spontaneous CP violation, while the positive mass theory does not. Yet both cases have

[6] Dashen's original paper [18] speculates that this might be related to the parity breaking seen in nature. This presumably requires a new "beyond the standard model" interaction rather than QCD.

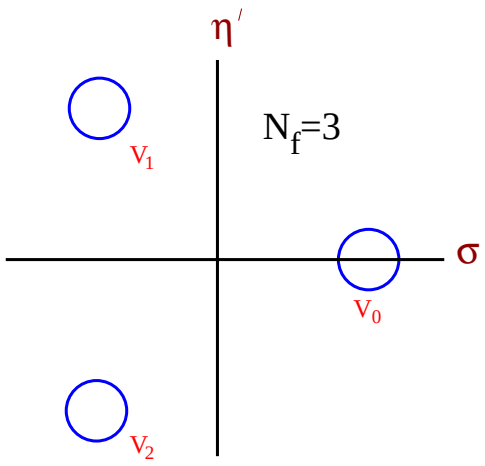

Fig. 14. For odd N_f, such as the three flavor case sketched here, QCD is not symmetric under changing the sign of the quark mass. Negative mass corresponds to taking $\Theta = \pi$.

exactly the same perturbation theory. This dramatically demonstrates what we already knew: non-perturbative effects are essential to understanding QCD.

A special case of an odd number of flavors is one-flavor QCD [78]. In this case the anomaly removes all chiral symmetry and there is a unique minimum in the σ, η' plane, as sketched in Fig. 15. This minimum does not occur at the origin, being shifted to $\langle\overline{\psi}\psi\rangle > 0$ by the 't Hooft vertex, which for one flavor is just an additive mass shift [79]. Unlike the case with more flavors, the resulting expectation value for σ is not from a spontaneous symmetry breaking; indeed, there is no chiral symmetry to break in one flavor QCD. Any regulator that preserves a remnant of chiral symmetry must inevitably fail [27]. Note also that for one-flavor QCD there is no longer the necessity of a first order phase transition at $\Theta = \pi$. It has been argued [78] that for finite quark mass such a transition should still occur if the mass is sufficiently negative, but the region around vanishing mass is not expected to show any singularity.

An unusual feature of one-flavor QCD is that the renormalization of the quark mass is not multiplicative when non-perturbative effects are taken into account. The additive mass shift is generally scheme dependent since the details of the instanton effects depend on scale. This is the basic reason that a massless up quark is not a possible solution to the strong CP problem [26]. Later we will discuss this in more detail in the context of the two flavor theory with non-degenerate masses.

Because of this shift in the mass, the conventional parameters Θ and m are singular coordinates for the one-flavor theory. A cleaner set of variables would be the coefficients of the two possible mass terms $\overline{\psi}\psi$ and $i\overline{\psi}\gamma_5\psi$ appearing in the Lagrangian. The ambiguity in the quark mass is tied to rough gauge configurations with ambiguous winding number. This applies even to the formally elegant overlap operator that we will discuss later; when rough gauge fields are present, the existence of a zero mode can depend on the detailed fermion operator in use. Smoothness conditions imposed on the gauge fields to remove this ambiguity appear to conflict with fundamental principles, such as reflection positivity [80].

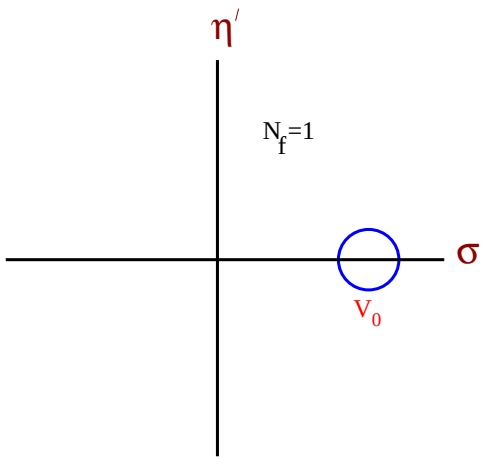

Fig. 15. The effective potential for one-flavor QCD with small quark mass has a unique minimum in the σ, η' plane. The minimum is shifted from zero due to the effect of the 't Hooft vertex.

The Z_{N_f} symmetry discussed here is a property of the fermion determinant and is independent of the gauge field dynamics. In Monte Carlo simulation language, this symmetry appears configuration by configuration. With N_f flavors, we always have $|D| = |e^{2\pi i/N_f} D|$ for any gauge field. This discrete chiral symmetry is inherently discontinuous in N_f. This non-continuity lies at the heart of the issues with the rooted staggered quark approximation. We will return to this topic in a later section.

9.3 Quark scattering and mass mixing

So far we have worked with degenerate quarks. In general each species introduces another complex mass parameter. Using flavored chiral rotations we can move the phases of the masses around arbitrarily, leaving only one overall phase, the Theta parameter. Thus once the overall scale has been set, QCD depends on $N_f + 1$ parameters.

Here we explore the rich phase diagram of two flavor QCD as a function of the most general quark masses, including the Θ parameter. This section closely follows the discussion in Ref. [81]. This theory involves three independent parameters. One is CP violating; its strong experimental limit is the strong CP problem. Here we will characterize the parameters by distinguishing their transformations under various symmetries. As we define them, the resulting variables are each multiplicatively renormalized. However non-perturbative effects are not universal, leaving individual quark mass ratios with a renormalization scheme dependence. This exposes ambiguities in matching lattice results with perturbative schemes and the tautology involved in approaches that attack the strong CP problem via a vanishing mass for the lightest quark.

Before turning on the masses, we reemphasize the qualitative properties expected in massless two-flavor QCD. Of course, being an interacting quantum field theory, nothing has been proven rigorously. While the classical theory is conformally invariant, as discussed earlier, confinement and dimensional transmutation generate a non-trivial mass scale. The theory should, of course,

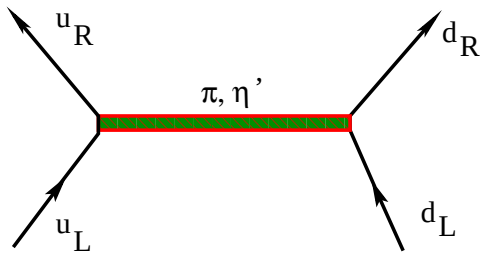

Fig. 16. Both pion and eta-prime exchange contribute to spin flip scattering between up and down quarks. Figure from Ref. [81].

contain massive stable nucleons. On the other hand, spontaneous chiral symmetry breaking should give rise to three massless pions as Goldstone bosons. Bound states of glue in general will acquire a width due to decays into pions. In addition, the two flavor analog of the eta-prime meson should acquire its mass from the anomaly.

In this standard picture, the eta-prime and neutral pion involve distinct combinations of quark-antiquark bound states. In the simple quark model the neutral pseudoscalars involve the combinations

$$\begin{aligned} \pi_0 &\sim \overline{u}\gamma_5 u - \overline{d}\gamma_5 d \\ \eta' &\sim \overline{u}\gamma_5 u + \overline{d}\gamma_5 d + \text{glue}. \end{aligned} \tag{212}$$

Here we include a gluonic contribution from mixing between the η' and glueball states. When the quarks are degenerate, isospin forbids such mixing for the pion.

Projecting out helicity states for the quarks $q_{L,R} = (1 \pm \gamma_5)q/2$, the pseudoscalars are combinations of left with right handed fermions, *i.e.* $\overline{q}_L q_R - \overline{q}_R q_L$. Thus, as shown schematically in Fig. 16, meson exchange will contribute to a a spin flip process in a hypothetical quark scattering experiment. More precisely, the four point function $\langle \overline{u}_R u_L \overline{d}_R d_L \rangle$ is not expected to vanish. Scalar meson exchange will also contribute to this process, but this is not important for the qualitative argument below. Of course we must assume that some sort of gauge fixing has been done to eliminate a trivial vanishing of this function from an integral over gauges. We also consider this four point function at a scale before confinement sets in.

It is important that the π_0 and η' are not degenerate. This is due to the anomaly and the fact that the η' is not a Goldstone boson. As we discussed earlier, the π_0-η' mass difference can be ascribed to topological structures in the gauge field. Because the mesons are not degenerate, their contributions to the above diagram cannot cancel. The conclusion of this simple argument is that helicity-flip quark-quark scattering is not suppressed in the chiral limit.

Now consider turning on a small down quark mass while leaving the up quark massless. Formally such a mass allows one to connect the ingoing and outgoing down quark lines in Fig. 16 and thereby induce a mixing between the left and right handed up quark. Such a process is sketched in Fig. 17. Here we allow for additional gluon exchanges to compensate for turning the pseudoscalar field into a traditional mass term.

So the presence of a non-zero d-quark mass will induce an effective mass for the u quark, even if the latter initially vanishes. As a consequence, non-perturbative effects renormalize

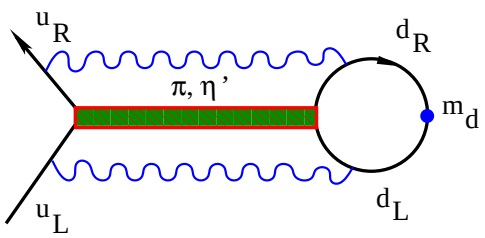

Fig. 17. Through physical meson exchange, a down quark mass can induce an effective mass for the up quark. The gluon exchanges can compensate for the pseudoscalar nature of the meson fields. Figure from Ref. [81].

m_u/m_d. If this ratio is zero at some scale, it cannot remain so for all scales. Only in the isospin limit are quark mass ratios renormalization group invariant. As lattice simulations include all perturbative and non-perturbative effects, this phenomenon is automatically included in such an approach.

This cross talk between the masses of different quark species is a relatively straightforward consequence of the chiral anomaly and has been discussed several times in the past, usually in the context of gauge field topology and the index theorem [26, 82–84]. This result is, however, frequently met with consternation from the community well versed in perturbation theory. Indeed, Feynman diagrams tend to suppress spin-flip processes as the quark masses go to zero. The above argument shows that this lore need not apply when anomalous processes come into play. In particular, mass renormalization is not flavor blind and the concept of mass independent regularization is problematic. Since the quark masses influence each other, there are inherent ambiguities defining $m_u = 0$. This has consequences for the strong CP problem, discussed further below. Furthermore, a traditional perturbative regulator such as $\overline{MS}$ is not complete when $m_u \neq m_d$. Because of this, the practice of matching lattice calculations to $\overline{MS}$ is also problematic.

Given the simplicity of the above argument, it is perhaps somewhat surprising that it continues to receive criticism. The first complaint sometimes made is that one should work directly with bare quark masses. This ignores the fact that the bare masses all vanish under renormalization. We discussed earlier the renormalization group equation for a quark mass

$$a\frac{dm_i}{da} = \gamma(g)m_i = \gamma_0 g^2 + O(g^4). \tag{213}$$

As asymptotic freedom drives the bare coupling to zero, the bare masses behave as

$$m \sim g^{\gamma_0/\beta_0}(1 + O(g^2)) \to 0 \tag{214}$$

where β_0 is the first term in the beta function controlling the vanishing of the bare coupling in the continuum limit. Since all bare quark masses are formally zero, one must address these questions in terms of a renormalization scheme at a finite cutoff.

A second frequent objection is that in a mass independent regularization scheme, mass ratios are automatically constant. Such an approach asks that the renormalization group function $\gamma(g)$ in Eq. (213) be chosen to be independent of the quark species and mass. This immediately

implies the constancy of all quark mass ratios. As only the first term in the perturbative expansion of $\gamma(g)$ is universal, a mass independent scheme is indeed an allowed procedure. However, such a scheme obscures the off-diagonal m_d effect on m_u discussed above. In particular, by forcing constancy of bare mass ratios, the ratios of physical particle masses must vary as a function of cutoff. This will be in a manner that cancels the flow from the process discussed above. The fact that physical particle mass ratios are not just a function of quark mass ratios is shown explicitly in subsection 9.6, where we observe that in the chiral limit the combination $1 - m_{\pi_0}^2/m_{\pi_\pm}^2$ is proportional to $\frac{(m_d-m_u)^2}{(m_d+m_u)\Lambda_{qcd}}$.

From a non-perturbative point of view, having physical mass ratios vary with the cutoff seems rather peculiar; indeed, the particle masses are physical quantities that would be natural to hold fixed. And, even though a mass independent approach is theoretically possible, there is no guarantee that any given quark mass ratio will be universal between schemes. Finally, the lattice approach itself is usually implemented with physical particle masses as input. As such it is not a mass independent regulator, making a perturbative matching to lattice results rather subtle.

A third complaint against the above argument is that one should simply do the matching at some high energy, say 100 GeV, where "instanton" effects are exponentially suppressed and irrelevant. This point of view has several problems. First, current lattice simulations are not done at miniscule scales and non-perturbative effects are present and substantial. Furthermore, the exponential suppression of topological effects is in the inverse coupling, which runs logarithmically with the scale. As such, the non-perturbative suppression is a power law in the scale and straightforward to estimate.

Since the eta-prime mass is expected to be of order Λ_{qcd}, we know from the previous renormalization group discussion how it depends on the bare coupling in the continuum limit

$$m_{\eta'} \propto \frac{1}{a} e^{-1/(2\beta_0 g_0^2)} g_0^{-\beta_1/\beta_0^2}. \tag{215}$$

While this formula indeed shows an exponential suppression in $1/g_0^2$, this is cancelled by the inverse cutoff factor in just such a way that the mass of this physical particle remains finite. The ambiguity in the quark mass splitting is controlled by the mass splitting $m_{\eta'} - m_{\pi_0}$ as well as being proportional to $m_d - m_u$. Considering $m_d = 5$ MeV at a scale of $\mu = 2$ GeV, a rough estimate of the order of the u quark mass shift is

$$\Delta m_u(\mu) \sim \left(\frac{m_{\eta'} - m_{\pi_0}}{\Lambda_{qcd}}\right) (m_d - m_u) = O(1 \text{ MeV}), \tag{216}$$

a number comparable to typical phenomenological estimates. This result depends on the scale μ, but that dependence is only logarithmic and given by Eq. (214). Additional flavors will reduce the size of this effect; with the strange quark present, it should be proportional to $m_d m_s$.

A particularly important observation is that the exponent controlling the coupling constant suppression in Eq. 215 is substantially smaller than the classical instanton action

$$\frac{1}{2\beta_0 g_0^2} = \frac{8\pi^2}{(11 - 2n_f/3)g_0^2} << \frac{8\pi^2}{g_0^2}. \tag{217}$$

This difference arises because one needs to consider topological excitations above the quantum, not the classical, vacuum. Zero modes of the Dirac operator are still responsible for the bulk of the eta prime mass, but naive semi-classical arguments strongly underestimate their effect.

9.4 General masses in two flavor QCD

Given the confusion over the meaning of quark masses, it is interesting to explore how two flavor QCD behaves as these quantities are varied, including the possibility of explicit CP violation through the Theta parameter. The full theory has a rather rich phase diagram, including first and second order phase transitions, some occuring when none of the quark masses vanish.

We consider the quark fields ψ as carrying implicit isospin, color, and flavor indices. Assume as usual that the theory in the massless limit maintains the $SU(2)$ flavored chiral symmetry under

$$\begin{aligned} \psi &\longrightarrow e^{i\gamma_5\tau_\alpha\phi_\alpha/2}\psi \\ \overline{\psi} &\longrightarrow \overline{\psi}e^{i\gamma_5\tau_\alpha\phi_\alpha/2}. \end{aligned} \tag{218}$$

Here τ_α represents the Pauli matrices generating isospin rotations. The angles ϕ_α are arbitrary rotation parameters.

We wish to construct the most general two-flavor mass term to add to the massless Lagrangian. Such should be a dimension 3 quadratic form in the fermion fields and should transform as a singlet under Lorentz transformations. For simplicity, only consider quantities that are charge neutral as well. This leaves four candidate fields, giving the general form for consideration

$$m_1\overline{\psi}\psi + m_2\overline{\psi}\tau_3\psi + im_3\overline{\psi}\gamma_5\psi + im_4\overline{\psi}\gamma_5\tau_3\psi. \tag{219}$$

The first two terms are naturally interpreted as the average quark mass and the quark mass difference, respectively. The remaining two are less conventional. The m_3 term is connected with the CP violating parameter of the theory. The final m_4 term has been used in conjunction with the Wilson discretization of lattice fermions, where it is referred to as a "twisted mass" [85, 86]. Its utility in this context is the ability to reduce lattice discretization errors. We will return to this term later when we discuss the effect of lattice artifacts on chiral symmetry.

These four terms are not independent. Indeed, consider the above flavored chiral rotation in the τ_3 direction, $\psi \to e^{i\theta\tau_3\gamma_5/2}\psi$. Under this the composite fields transform as

$$\overline{\psi}\psi \longrightarrow \cos(\theta)\overline{\psi}\psi + \sin(\theta)i\overline{\psi}\gamma_5\tau_3\psi \tag{220}$$

$$\overline{\psi}\tau_3\psi \longrightarrow \cos(\theta)\overline{\psi}\tau_3\psi + \sin(\theta)i\overline{\psi}\gamma_5\psi \tag{221}$$

$$i\overline{\psi}\tau_3\gamma_5\psi \longrightarrow \cos(\theta)i\overline{\psi}\tau_3\gamma_5\psi - \sin(\theta)\overline{\psi}\psi \tag{222}$$

$$i\overline{\psi}\gamma_5\psi \longrightarrow \cos(\theta)i\overline{\psi}\gamma_5\psi - \sin(\theta)\overline{\psi}\tau_3\psi \tag{223}$$

This rotation mixes m_1 with m_4 and m_2 with m_3. Using this freedom, we can select any one of the m_i to vanish and a second to be positive.

The most common choice is to set $m_4 = 0$ and use m_1 as controlling the average quark mass. Then m_2 gives the quark mass difference, while CP violation appears in m_3. This, however, is only a convention. The alternative "twisted mass" scheme [85, 86], makes the choice $m_1 = 0$. This uses $m_4 > 0$ for the average quark mass and m_3 becomes the up-down mass difference. In this case m_2 becomes the CP violating term. It is amusing to note that an up down quark mass difference in such a formulation involves the naively CP odd $i\overline{\psi}\gamma_5\psi$. The strong CP problem has been rotated into the smallness of the $\overline{\psi}\tau_3\psi$ term, which with the usual conventions is the mass difference. But because of the flavored chiral symmetry, both sets of conventions are physically equivalent.

For the following, take the arbitrary choice $m_4 = 0$, although one should remember that this is only a convention and we could have chosen any of the four parameters in Eq. (219) to vanish. With this choice, two flavor QCD, after scale setting, depends on three mass parameters

$$m_1\overline{\psi}\psi + m_2\overline{\psi}\tau_3\psi + im_3\overline{\psi}\gamma_5\psi. \tag{224}$$

It is the possible presence of m_3 that represents the strong CP problem. As all the parameters are independent and transform differently under the symmetries of the problem, there is no connection between the strong CP problem and m_1 or m_2.

As discussed extensively above, the chiral anomaly is responsible for the iso-singlet rotation

$$\begin{aligned} \psi &\longrightarrow e^{i\gamma_5\phi/2}\psi \\ \overline{\psi} &\longrightarrow \overline{\psi}e^{i\gamma_5\phi/2} \end{aligned} \tag{225}$$

not being a valid symmetry, despite the fact that γ_5 naively anti-commutes with the massless Dirac operator. Subection 8.1 showed this anomaly is nicely summarized via Fujikawa's [66] approach where the fermion measure in the path integral picks up a non-trivial factor. In any given gauge configuration only the zero eigenmodes of $\not{D}$ contribute, and by the index theorem they are connected to the winding number of the gauge configuration. The conclusion is that the above rotation changes the fermion measure by an amount depending non-trivially on the gauge field configuration.

Note that this anomalous rotation allows one to remove any topological term from the gauge part of the action. Naively this would have been yet another parameter for the theory, but by including all three mass terms for the fermions, this can be absorbed. For the following we consider that any topological term has thus been rotated away. After this one is left with the three mass parameters above, all of which are independent and relevant to physics.

These parameters are a complete set for two flavor QCD; however, this choice differs somewhat from what is often discussed. Formally we can define the more conventional variables as

$$m_u = m_1 + m_2 + im_3 \tag{226}$$

$$m_d = m_1 - m_2 + im_3 \tag{227}$$

$$e^{i\Theta} = \frac{m_1^2 - m_2^2 - m_3^2 + 2im_1m_3}{\sqrt{m_1^4 + m_2^4 + m_3^4 + 2m_1^2m_3^2 + 2m_2^2m_3^2 - 2m_1^2m_2^2}.} \tag{228}$$

Particularly for Θ, this is a rather complicated change of variables. For non-degenerate quarks in the context of the phase diagram discussed below, the variables $\{m_1, m_2, m_3\}$ are more natural.

9.5 The strong CP problem and the up quark mass

Strong interactions preserve CP to high accuracy [87]. Thus only two of the three possible mass parameters seem to be needed. With the above conventions, it is natural to ask why is m_3 so small? It is the concept of unification that brings this question to the fore. We know that the weak interactions violate CP. Thus, if the electroweak and the strong interactions separate at some high scale, shouldn't some remnant of this breaking survive? How is CP recovered for the strong force?

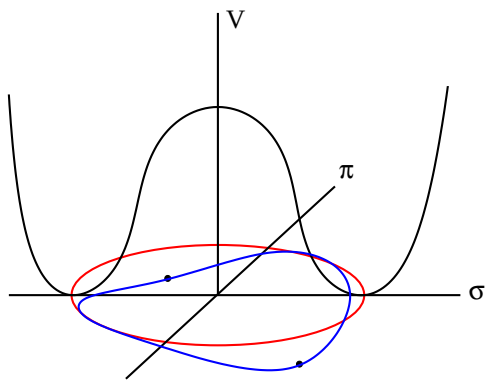

Fig. 18. The m_2 and m_3 terms warp the Mexican hat potential into two separate minima. The direction of the warping is determined by the relative size of these parameters. Figure taken from Ref. [81].

One possible solution is that there is no unification and one should just consider the weak interactions as a small perturbation. Another approach involves adding a new dynamical "axion" field that couples to the quarks through a coupling to $i\overline{\psi}\gamma_5\psi$. Shifts in this field make m_3 essentially dynamical, and potentially the theory could relax to $m_3 = 0$.

There is a third proposed solution, being criticized here, that the up quark mass might vanish. This would naively allow a flavored chiral rotation to remove any phases from the quark mass matrix. Why is a vanishing up quark mass not a sensible approach? From the above, one can define the up quark mass as a complex number

$$m_u \equiv m_1 + m_2 + im_3. \tag{229}$$

But the quantities m_1, m_2, and m_3 are independent parameters with different symmetry properties. With our conventions, m_1 represents an iso-singlet mass contribution, m_2 is isovector in nature, and m_3 is CP violating. And, as discussed earlier, the combination $m_1 + m_2 = 0$ is scale and scheme dependent. The strong CP problem only requires small m_3. So while it may be true formally that

$$m_1 + m_2 + im_3 = 0 \;\Rightarrow\; m_3 = 0, \tag{230}$$

this would depend on scale and might well be regarded as "not even wrong."

9.6 The two flavor phase diagram

As a function of the three mass parameters, QCD has a rather intricate phase diagram that we now discuss. Using simple chiral Lagrangian arguments, this can be qualitatively mapped out. To begin we consider the composite fields similar to those used in the earlier discussion of pions as Goldstone bosons

$$\begin{aligned} \sigma &\sim \overline{\psi}\psi & \qquad \eta &\sim i\overline{\psi}\gamma_5\psi \\ \vec{\pi} &\sim i\overline{\psi}\gamma_5\vec{\tau}\psi & \qquad \vec{a}_0 &\sim \overline{\psi}\vec{\tau}\psi. \end{aligned} \tag{231}$$

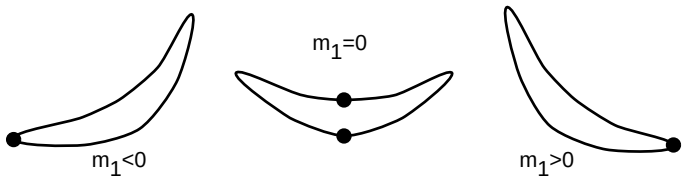

Fig. 19. In the m_1, m_2 plane, $m^2_{\pi_0}$ can pass through zero, giving rise to pion condensation at an Ising-like transition. Figure taken from [25].

In terms of these, a natural model for a starting effective potential is

$$\begin{aligned} V = \quad & \lambda(\sigma^2 + \vec{\pi}^2 - v^2)^2 - m_1\sigma - m_2 a_{03} - m_3\eta \\ & +\alpha(\eta^2 + \vec{a}_0^2) - \beta(\eta\sigma + \vec{a}_0 \cdot \vec{\pi})^2. \end{aligned} \tag{232}$$

Here α and β are "low energy constants" that bring in a chirally symmetric coupling of $(\sigma, \vec{\pi})$ with $(\eta, \vec{a}_0)$. As discussed in Ref. [25], the sign of the β term is suggested so that $m_\eta < m_{a_0}$.

This potential augments the famous "Mexican hat" or "wine bottle" potential discussed earlier, in which the Goldstone pions are associated with the flat directions running around at constant $\sigma^2 + \vec{\pi}^2 = v^2$. The m_2 and m_3 terms do not directly affect the σ and π fields, but induce an expectation value for a_{03} and η, respectively. This in turn results in the α and β terms inducing a warping of the Mexican hat into two separate minima, as sketched in Fig. 18. The direction of this warping is determined by the relative size of m_2 and m_3; m_2 (m_3) warps downward in π_0 (σ) direction. If we now turn on m_1, this will select one of the two minimum as favored. This gives rise to a generic first order transition at $m_1 = 0$.

There is additional structure in the m_1, m_2 plane when m_3 vanishes. In this situation the quadratic warping is downward in the σ direction. For large $|m_1|$ only σ will have an expectation, with sign determined by the sign of m_1. The pion will be massive, but with m_2 reducing the neutral pion mass below that of the charged pions. If now m_1 is decreased in magnitude at fixed m_2, eventually the neutral pion becomes massless and condenses. How this occurs is sketched in Fig. 19. An order parameter for the transition is the expectation value of the π_0 field, with the transition being Ising-like. It occurs in a region where the two flavors have opposite signs for their masses, i.e. $|m_1| < |m_2|$.

This structure can also be understood in terms of the expectation values for the pion and sigma fields as functions of the average quark mass while holding the quark mass difference fixed. This is sketched in Fig. 20. The jump in σ as we go from large positive to large negative masses is split into two transitions with the pion field acquiring an expectation value in the intermediate region.

Note that this transition occurs when both m_u and m_d are non-vanishing but of opposite sign. At the transition the correlation length diverges. This shows that it is possible to have significant long distance physics without small Dirac eigenvalues. In contrast, we see that there is no transition at the point where only one of the quark masses vanishes. In this situation there is no long distance physics despite the possible existence of small Dirac eigenvalues.

Putting this all together, we obtain the phase diagram sketched in Fig. 21. There are two intersecting first order transition surfaces, one at $(m_1 = 0,\, m_3 \neq 0)$ and the second at $(m_1 < m_2,\, m_3 = 0)$. These each occur where $\Theta = \pi$. However, note that with non-degenerate quarks

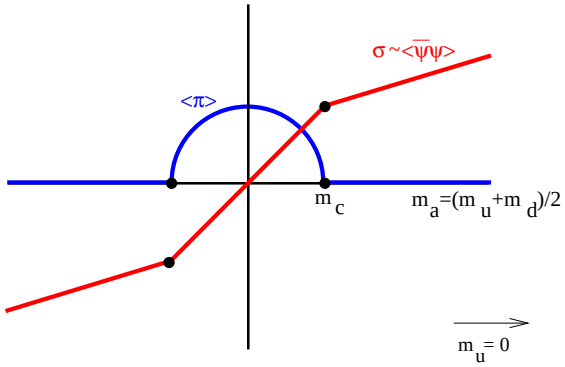

Fig. 20. With a constant up-down quark mass difference, the jump in the chiral condensate splits into two second order transitions. The order parameter distinguishing the intermediate phase is the expectation value of the neutral pion field. Figure taken from Ref. [88].

there is also a $\Theta = \pi$ region at $m_2 = m_1 + \epsilon$ for small but non-vanishing ϵ where there is no transition. The absence of a physical singularity at $m_u = 0$ when $m_d \neq 0$ lies at the heart of the problem in defining a vanishing up quark mass.

In the next section we will see that the structure in the m_1, m_2 plane is closely related to an interesting lattice artifact in the degenerate quark limit. Aoki [89] discussed a possible phase with spontaneous parity violation with the Wilson fermion formulation. Indeed, lattice artifacts can modify the effective potential in a similar way to the m_2 term and allow the CP violating phase at finite cutoff to include part of the m_1 axis as well.

10 Lattice fermions

We now have a fairly coherent picture of how the spectrum of pseudoscalar mesons is connected with chiral symmetry in the continuum theory. The anomaly plays a crucial role in introducing the Theta parameter into the theory and contributing to the η' mass. Throughout we have assumed that we have in hand a regulator to define the various composite fields, but we have not been specific in how that regulator is formulated. Early sections indicated the lattice should provide a natural route to a non-perturbative formulation, but we have postponed the details until some of the desired continuum features were elucidated.

The lattice can be regarded as a fully non-perturbative definition of a quantum field theory. As such, the entire structure explored in previous sections should follow as we approach the continuum limit. But there are a variety of interesting and subtle issues concerning how this comes about. When the lattice is in place, all infinities in the theory are automatically removed. However we have argued that the anomaly is closely tied to the divergences in the theory. As such the physics associated with the anomaly must appear somewhere in any valid lattice formulation. If we try to formulate a lattice version of QCD with all classical symmetries in place, there is no way for this to happen. In particular, this imposes subtleties for how the action for the quarks is formulated. Here we go into this problem in some detail and explore some of the methods for dealing with it.

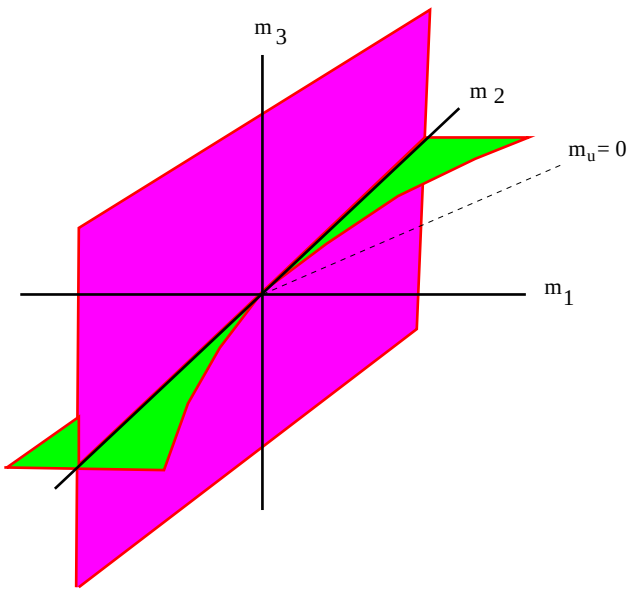

Fig. 21. The full phase diagram for two flavor QCD as a function of the three mass parameters. It consists of two intersecting first order surfaces with a second order edge along curves satisfying $m_3 = 0$, $|m_1| < |m_2|$. There is no structure along the $m_u = 0$ line except when both quark masses vanish. Figure from Ref. [81]

10.1 Hopping and doublers

The essence of lattice doubling already appears in the quantum mechanics of the simplest Fermion Hamiltonian in one space dimension

$$H = iK \sum_j a_{j+1}^\dagger a_j - a_j^\dagger a_{j+1}. \tag{233}$$

Here j is an integer labeling the sites of an infinite chain and the a_j are Fermion annihilation operators satisfying standard anti-commutation relations

$$\left[a_j, a_k^\dagger\right]_+ \equiv a_j a_k^\dagger + a_k^\dagger a_j = \delta_{j,k}. \tag{234}$$

The fermions hop from site to neighboring site with amplitude K; thus, we refer to K as the "hopping parameter" and by convention take it to be positive. The bare vacuum $|0\rangle$ satisfies $a_j|0\rangle = 0$. This vacuum is not the physical one, which requires constructing a filled Dirac sea. Energy eigenstates in the single Fermion sector

$$|\chi\rangle = \sum_j \chi_j a_j^\dagger |0\rangle \tag{235}$$

can be easily found in momentum space

$$\chi_j(q) = e^{iqj}\chi_0 \tag{236}$$

where we can restrict $-\pi < q \leq \pi$. The resulting energy is

$$E(q) = 2K\sin(q). \tag{237}$$

The physical vacuum fills all the negative energy states, *i.e.* those with $-\pi < q < 0$.

On this vacuum, consider constructing a Fermionic wave packet by exciting a few modes of small momentum q. This packet will have a group velocity $dE/dq \sim 2K$ that is positive. Thus it moves to the right and represents a right-moving fermion. On the other hand, a wave packet of low energy can also be produced by exciting momenta in the vicinity of $q \sim \pi$. This packet will have group velocity $\left.\frac{dE}{dq}\right|_{q=\pi} \sim -2K$ and therefore be left moving. The essence of the Nielsen Ninomiya theorem [90] is that we must have both types of excitation. We will go into this in more detail later, but for this one dimensional case the periodicity in q requires the dispersion relation to have an equal number of zeros with positive and negative slopes. If we now consider a two component spinor to describe the fermion, we will have independent states corresponding to each component. This is the so called "doubling" issue.

As a preliminary to later discussion, here we concentrate on a Hamiltonian version of the Wilson approach to remove the doublers. Continuing to work in one dimension, consider a two component spinor

$$\psi = \begin{pmatrix} a \\ b \end{pmatrix}. \tag{238}$$

where a and b are distinct fermion annihilation operators on the lattice sites. The so called "naive" lattice Hamiltonian begins with the simple hopping case of above and adds in the lower components and a mass term that mixes the upper and lower components

$$H = iK\sum_j a^\dagger_{j+1}a_j - a^\dagger_j a_{j+1} - b^\dagger_{j+1}b_j + b^\dagger_j b_{j+1} + M\sum_j a^\dagger_j b_j + b^\dagger_j a_j. \tag{239}$$

Introducing two by two Dirac matrices

$$\gamma_0 = \sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \;\; \gamma_1 = \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \;\; \gamma_5 = \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \tag{240}$$

and defining $\overline{\psi} = \psi^\dagger\gamma_0$, we write the Hamiltonian compactly as

$$H = \sum_j K(\overline{\psi}_{j+1}\gamma_1\psi_j - \overline{\psi}_j\gamma_1\psi_{j+1}) + M\sum_j \overline{\psi}_j\psi_j. \tag{241}$$

This looks very much like the continuum Dirac Hamiltonian with the derivative term represented on the lattice by a nearest neighbor difference. Chiral symmetry is manifest in the possibility of independent rotations of the a and b type particles when the mass term is absent. The latter mixes these components and opens a gap in the spectrum.

As before, the single particle states are easily found by Fourier transformation and satisfy

$$E^2 = 4K^2\sin^2(q) + M^2 \tag{242}$$

At each momentum there is one positive and one negative energy state. Again, we are to fill the negative energy sea to form the physical vacuum. The doubling issue is that there are low energy

excitations that satisfy the Dirac equation appearing both at $q \sim 0$ and $q \sim \pi$. The physical momenta k of the latter excitations appear in expanding about pi, i.e $k = q - \pi$. These states have a smooth spatial dependence in a redefined field $\chi_j = (-1)^j \psi_j$. The doublers at $q \sim \pi$ are still with us.

10.2 Wilson Fermions

One way to remove the degeneracy of the doublers is to make the mixing of the upper and lower components momentum dependent. A simple way of doing this was proposed by Wilson [10]. To our Hamiltonian model, we add one more term

$$\begin{aligned} H &= iK \sum_j a^\dagger_{j+1} a_j - a^\dagger_j a_{j+1} - b^\dagger_{j+1} b_j + b^\dagger_j b_{j+1} + \\ &\quad M \sum_j a^\dagger_j b_j + b^\dagger_j a_j - rK \sum_j a^\dagger_j b_{j+1} + b^\dagger_j a_{j+1} + b^\dagger_{j+1} a_j + a^\dagger_{j+1} b_j \\ &= \sum_j K(\overline{\psi}_{j+1}(\gamma_1 - r)\psi_j - \overline{\psi}_j(\gamma_1 + r)\psi_{j+1}) + \sum_j M \overline{\psi}_j \psi_j. \end{aligned} \tag{243}$$

Now the spectrum satisfies

$$E^2 = 4K^2 \sin^2(q) + (M - 2rK\cos(q))^2. \tag{244}$$

The doublers at $q \sim \pi$ are increased in energy relative to the states at $q \sim 0$. The physical particle mass is now $m = M - 2rK$ while that of the doubler is at $M + 2rK$.

When r becomes large, the dip in the spectrum of near $q = \pi$ actually becomes a maximum. This is irrelevant for our discussion, although we note that the case $r = 1$ is somewhat special. For this value, the matrices $(\gamma_1 \pm 1)/2$, which determine how the Fermions hop along the lattice, are projection operators. In a sense, the doubler is removed because only one component can hop. This choice $r = 1$ has been the most popular in practice.

The hopping parameter has a critical value at

$$K_c = \frac{M}{2r} \tag{245}$$

At this point the gap in the spectrum closes and one species of Fermion becomes massless. The Wilson term, proportional to r, still mixes the a and b type particles; so, there is no exact chiral symmetry. Nevertheless, in the continuum limit this represents a candidate for a chirally symmetric theory. Before the limit, chiral symmetry does not provide a good order parameter.

Now we generalize this approach to the Euclidean path integral formulation in four space-time dimensions. In the continuum, one usually writes for the free fermion action density

$$\overline{\psi} D \psi = \overline{\psi}(\not{\partial} + m)\psi \tag{246}$$

or in momentum space

$$\overline{\psi}(i\not{p} + m)\psi. \tag{247}$$

By convention we use Hermitean gamma matrices. Note that D is the sum of Hermitian and anti-Hermitian parts. In the continuum the former is just a constant, the mass. A hermitian operator appears in the combination $\gamma_5 D$, but we don't need that just now.

A matrix can be diagonalized when it commutes with its adjoint; then it is called "normal." For the naive continuum operator this is the case, and we see that all eigenvalues of D lie along a line parallel to the imaginary axis intersecting the real axis at m. This simple structure will be lost on the lattice.

As discussed earlier, a simple transcription of derivatives onto the lattice replaces factors of p with trigonometric functions. Thus the naive lattice action becomes

$$\overline{\psi}\left(\frac{i}{a}\sum_{\mu}\gamma_{\mu}\sin(p_{\mu}a)+m\right)\psi \tag{248}$$

where we have explicitly included the lattice spacing a. For small momentum this reduces to the continuum result $\overline{\psi}(i\gamma_{\mu}p_{\mu}+m)\psi$. Now let one component of p get large and be near π/a. Then we again have small eigenvalues and a nearby pole in the propagator. As any of the four components of momentum can be near 0 or π there are a total of 16 places in momentum space that give rise to a Dirac like behavior. The naive fermion action gives rise to 16 doublers.

As in the earlier example, the Wilson solution adds a momentum dependent mass. Wishing to maintain only nearest neighbor terms, it also involves trigonometric functions. To maintain hyper-cubic symmetry, we put in the Wilson term symmetrically for all space-time directions. For simplicity we set the Wilson parameter r from before to unity. Explicitly for free fields we consider the momentum space form

$$\overline{\psi}D_W\psi=\overline{\psi}\left(\frac{1}{a}\sum_{\mu}(i\gamma_{\mu}\sin(p_{\mu}a)+1-\cos(p_{\mu}a))+m\right)\psi \tag{249}$$

Now for a momentum component near π the eigenvalues are of order $1/a$ in size. Note that the lattice artifacts in the propagator start at order p^2a, rather than $O(a^2)$ as for naive fermions. The eigenvalue structure of D_W is rather interesting. The eigenvalues for the free Wilson theory occur at

$$\lambda=\pm\frac{i}{a}\sqrt{\sum_{\mu}\sin^2(p_{\mu}a)}+\frac{1}{a}\sum_{\mu}1-\cos(p_{\mu}a)+m \tag{250}$$

The eigenvalues of this free operator lie on a set of "nested circles," as sketched in Fig. 22. Note that $m\leftrightarrow -m$ is not a symmetry. Naively it would be in the continuum, but as we discussed earlier, it cannot be so in the quantum theory when one has an odd number of flavors.

Note that to obtain real eigenvalues in Eq. (250), each component of the momentum must be an integer multiple of π. There are actually several critical values that can give rise to massless fermions. For $m=0,-2,-4,-6,-8$ we have $1,4,6,4,1$ massless species. When interactions are present these values of the mass will also be renormalized. [7] Whether a continuum limit at these alternative points is useful has not been investigated.

Rescaling to lattice units and restoring the hopping parameter, the Wilson fermion action with the site indices explicit becomes

$$D_{W\,ij}=\delta_{i,j}+K\sum_{\mu}(1-\gamma_{\mu})\delta_{i,j+e_{\mu}}+(1+\gamma_{\mu})\delta_{i,j-e_{\mu}}. \tag{251}$$

[7] Actually the 6 flavor case at m=-4 does have a discrete symmetry that will protect against additive mass renormalization.

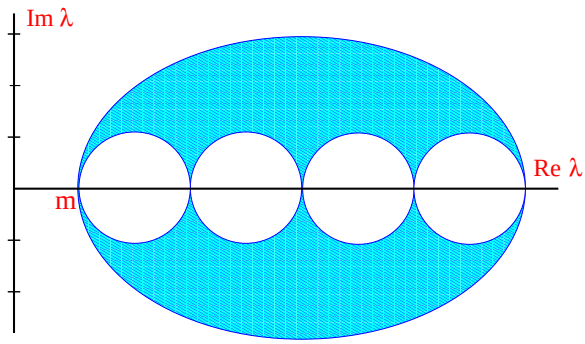

Fig. 22. The eigenvalue spectrum of the free Wilson fermion operator is a set of nested circles. On turning on the gauge fields, some eigenvalues drift into the open regions. Some complex pairs can collide and become real. These are connected to gauge field topology. Figure taken from Ref. [91].

By taking the coefficient r of the Wilson term as unity we have projection operators in the hoppings. The physical fermion mass is read off from the small momentum behavior as $m = \frac{1}{2a}(1/K - 8)$. This vanishes at at $K = K_c = 1/8$.

Here we consider that the gauge fields are formulated as usual with group valued matrices on the lattice links. These are to be inserted into the above hopping terms. One could use the simple Wilson gauge action as a sum over plaquettes

$$S_g = \frac{\beta}{3} \sum_p \mathrm{Re\ Tr}\, U_p \tag{252}$$

although this specific form is not essential the qualitative nature of the phase diagram. When the gauge fields are turned on, the dynamics will move the fermion eigenvalues around, partially filling the holes in eigenvalue pattern of Fig. 22. Some eigenvalues can become real and are related to gauge field topology [78].

For the free theory the Hermitian and anti-Hermitian parts of the action commute. This ceases to be true in the interacting case since both terms contain gauge matrices that themselves do not commute. Thus the left eigenvalues are generally different from right ones. Nevertheless, it is still true that the eigenvalues either appear in complex conjugate pairs or they are real. This follows from γ_5 Hermiticity, $D^\dagger = \gamma_5 D \gamma_5$. Since γ_5 has unit determinant, $|D - \lambda| = 0$ implies $|D^\dagger - \lambda| = |D - \lambda^*|^* = 0$.

A technical difficulty with this approach is that gauge interactions will renormalize the parameters. To obtain massless pions one must finely tune K to K_c, an *a priori* unknown function of the gauge coupling. Despite the awkwardness of such tuning, this is how numerical simulations with Wilson quarks generally proceed. The hopping parameter is adjusted to get the pion mass right, and one assumes that the remaining predictions of current algebra reappear in the continuum limit.

Note that the basic lattice theory has two parameters, β and K. These are related to bare coupling, $\beta \sim 6/g_0^2$, and quark mass, $(1/K - 1/K_c) \sim m_q$. We will now turn to a discussion of this relation in more detail.

10.3 Lattice versus continuum parameters

As emphasized earlier, QCD is a remarkably economical theory in terms of the number of adjustable parameters. First of these is the overall strong interaction scale, Λ_{qcd}. This is scheme dependent, but once a renormalization procedure has been selected, it is well defined. It is not independent of the coupling constant, the connection being fixed by asymptotic freedom. In addition, the theory depends on the renormalized quark masses m_i, or more precisely the dimensionless ratios m_i/Λ_{qcd}. As with the overall scale, the definition of m_i is scheme dependent. The two flavor theory with degenerate quarks and $\Theta = 0$ has one such mass parameter. As we wish to formulate the theory with a lattice cutoff in place, there is a scale for this cutoff. As with everything else, it is convenient to measure this in units of the overall scale; so, a third parameter for the cutoff theory is $a\Lambda_{qcd}$, where one can regard a as the lattice spacing.

How the bare parameters behave as the continuum limit is taken was discussed rather abstractly in Section 5. The goal here is to explore some of the the lattice artifacts that arise with Wilson fermions [10]. On the lattice it is generally easier to work directly with lattice parameters. One of these is the plaquette coupling β, which, with the usual conventions, is related to the bare coupling $\beta = 6/g_0^2$. For the quarks, the natural lattice quantity is the "hopping parameter" K. And finally, the connection with physical scales appears via the lattice spacing a.

The set of physical parameters and the set of lattice parameters are, of course, equivalent, and there is a well understood non-linear mapping between them

$$\left\{ \frac{m}{\Lambda_{qcd}}, a\Lambda_{qcd} \right\} \longleftrightarrow \{\beta, K\}. \tag{253}$$

Of course, to extract physical predictions we are interested in the continuum limit $a\Lambda_{qcd} \to 0$. For this, asymptotic freedom tells us we must take $\beta \to \infty$ at a rate tied to Λ_{qcd}. Simultaneously we must take the hopping parameter to a critical value. With normal conventions, this takes $K \to K_c \to 1/8$ at a rate tied to desired quark mass m. Fig. 23 sketches how the continuum limit is taken in the β, K plane. Here we wish to further explore this phase diagram with particular attention to hopping parameters larger than K_c. This discussion is adapted from Ref. [91] and adds the possible twisted mass term to the exposition from Ref. [92].

10.4 Artifacts and the Aoki phase

We previously made extensive use of an effective field theory to describe the interactions of the pseudo-scalar mesons. Here we will begin with the simplest form for the two flavor theory and then add terms to mimic possible lattice artifacts. The language is framed as before in terms of the isovector pion field $\vec{\pi} \sim i\overline{\psi}\gamma_5\vec{\tau}\psi$ and the scalar sigma $\sigma \sim \overline{\psi}\psi$. the starting point for this discussion is the canonical "Mexican hat" potential

$$V_0 = \lambda(\sigma^2 + \vec{\pi}^2 - v^2)^2 \tag{254}$$

schematically sketched earlier in Fig. 6. The potential has a symmetry under $O(4)$ rotations amongst the pion and sigma fields expressed as the four vector $\Sigma = (\sigma, \vec{\pi})$. This represents the axial symmetry of the underlying quark theory.

As discussed before, the massless theory is expected to spontaneously break chiral symmetry with the minimum for the potential occurring at a non-vanishing value for the fields. As usual,

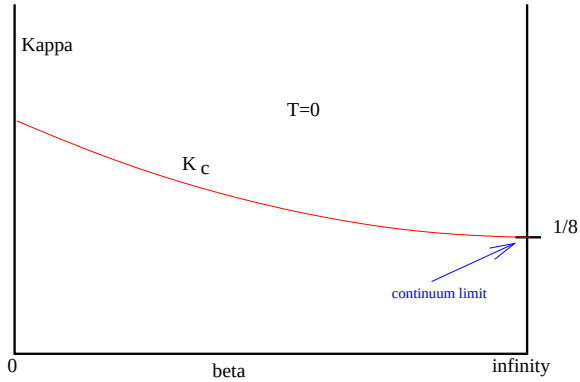

Fig. 23. The continuum limit of lattice gauge theory with Wilson fermions occurs at $\beta \to \infty$ and $K \to 1/8$. Coming in from this point to finite beta is the curve $K_c(\beta)$, representing the lowest phase transition in K for fixed beta. The nature of this phase transition is a delicate matter, discussed in the text. Figure taken from Ref. [91].

we take the vacuum to lie in the sigma direction with $\langle\sigma\rangle > 0$. The pions are then Goldstone bosons, being massless because the potential provides no barrier to oscillations of the fields in the pion directions. Also as discussed before, we include a quark mass by adding a constant times the sigma field

$$V_1 = -m\sigma \tag{255}$$

This explicitly breaks the chiral symmetry by "tilting" the potential as sketched in Fig. 7. That selects a unique vacuum which, for $m > 0$, gives a positive expectation for sigma. In the process the pions gain a mass, with $m_\pi^2 \sim m$.

Because of the symmetry of V_0, it does not matter physically in which direction we tilt the vacuum. In particular, a mass term of form

$$m\sigma \to m\cos(\theta)\sigma + m\sin(\theta)\pi_3 \tag{256}$$

should give equivalent physics for any θ. In the earlier continuum discussion we used this freedom to rotate the second term away. However, as we will see, lattice artifacts can break this symmetry, introducing the possibility of physics at finite lattice spacing depending on this angle. As mentioned before, the second term in this equation is what is usually called a "twisted mass."

The Wilson term inherently breaks chiral symmetry. This will give rise to various modifications of the effective potential. The first correction is expected to be an additive contribution to the quark mass, i.e. an additional tilt to the potential. This means that the critical kappa, defined as the smallest kappa where a singularity is found in the β, K plane, will move away from the limiting value of $1/8$. Thus we introduce the function $K_c(\beta)$ and imagine that the mass term is modeled with the form

$$m \to c_1(1/K - 1/K_c(\beta)). \tag{257}$$

In general the lattice modification of the effective potential will have further corrections of higher order in the effective fields. A natural way to include such is as an expansion in the chiral

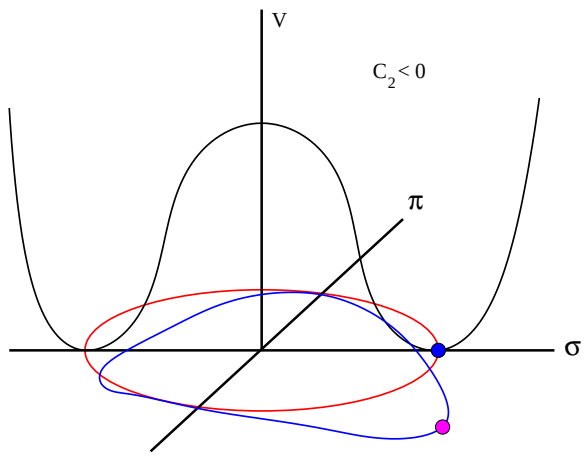

Fig. 24. Lattice artifacts could quadratically warp the effective potential. If this warping is downward in the sigma direction, the chiral transition becomes first order without the pions becoming massless. Figure taken from Ref. [91].

fields. With this motivation we include a term in the potential of form $c_2\sigma^2$. Including these ideas in the effective model, we are led to

$$V(\vec{\pi},\sigma) = \lambda(\sigma^2 + \vec{\pi}^2 - v^2)^2 - c_1(1/K - 1/K_c(\beta))\sigma + c_2\sigma^2. \tag{258}$$

Such a term was considered in Refs. [92,93]. The predicted phase structure depends qualitatively on on the sign of c_2, but a priori we have no information on this.[8] Indeed, as it is a lattice artifact, it is expected that this sign might depend on the choice of gauge action. Note that we could have added a term like $\vec{\pi}^2$, but this is essentially equivalent since $\vec{\pi}^2 = (\sigma^2 + \vec{\pi}^2) - \sigma^2$, and the first term here can be absorbed, up to an irrelevant constant, into the starting Mexican hat potential.

First consider the case when c_2 is less than zero, thus lowering the potential energy when the field points in the positive or negative sigma direction. This quadratic warping helps to stabilize the sigma direction, as sketched in Fig. 24, and the pions cease to be true Goldstone bosons when the quark mass vanishes. Instead, as the mass passes through zero, we have a first order transition as the expectation of σ jumps from positive to negative. This jump occurs without any physical particles becoming massless.

Things get a bit more complicated if $c_2 > 0$, as sketched in Fig. 25. In that case the chiral transition splits into two second order transitions separated by phase with an expectation for the pion field, i.e. $\langle\vec{\pi}\rangle \neq 0$. The behavior is directly analogous to that shown in Fig. 20, the main difference being that now the two quarks are degenerate. Since the pion field has odd parity and charge conjugation as well as carries isospin, all of these symmetries are spontaneously broken in the intermediate phase. As isospin is a continuous group, this phase will exhibit Goldstone bosons. The number of these is two, representing the two flavor generators orthogonal to the direction of the expectation value. If higher order terms do not change the order of the transitions, there will be a third massless particle exactly at the transition endpoints. In this way the

[8] Ref. [94] has argued that c_2 should be positive. We will return to this argument a bit later in this section.

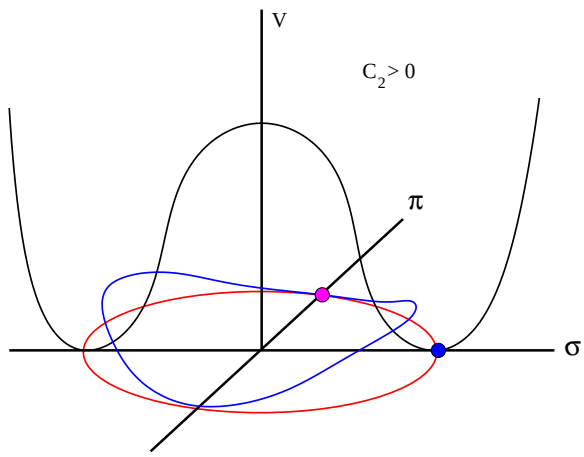

Fig. 25. If the lattice artifacts warping the potential upward in the sigma direction, the chiral transition splits into two second order transitions separated by a phase where the pion field has an expectation value. Figure taken from Ref. [91].

theory reveals three massless pions exactly at the transitions, as discussed by Aoki [89]. The intermediate phase is usually referred to as the "Aoki phase." Assuming this $c_2 > 0$ case, Fig. 26 shows the qualitative phase diagram expected.

Note the similarity of this discussion to that leading to the phase diagram in Fig. 21. Indeed, lattice artifacts can lead to the spontaneously broken CP region found there for the (m_1, m_2) plane to open up and remain present for degenerate quarks. The Aoki phase is closely related to the possibility of CP violation at $\Theta = \pi$ for unequal mass quarks. Note also that this connection with the earlier continuum discussion shows that with an odd number of flavors, the spontaneous breaking of parity is the normal expectation whenever the hopping parameter exceeds its critical value. Indeed, in this case the Aoki phase is less a lattice artifact than a direct consequence of the CP violation expected in the continuum theory at $\theta = \pi$.

10.5 Twisted mass

The c_2 term breaks the equivalence of different chiral directions. This means that physics will indeed depend on the angle θ if one takes a mass term of the form in Eq. (256). Consider complexifying the fermion mass in the usual way

$$m\overline{\psi}\psi \to m\overline{\psi}_L\psi_R + m^*\overline{\psi}_R\psi_L = \mathrm{Re}\, m\, \overline{\psi}\psi + i\mathrm{Im}\, m\, \overline{\psi}\gamma_5\psi. \tag{259}$$

The rotation of Eq. (256) is equivalent to giving the up and down quark masses opposite phases

$$m_u \to e^{+i\theta} m_u \tag{260}$$

$$m_d \to e^{-i\theta} m_d \tag{261}$$

Thus motivated, we can consider adding a new mass term to the lattice theory

$$\mu\, i\overline{\psi}\tau_3\gamma_5\psi \sim \mu\pi_3 \tag{262}$$

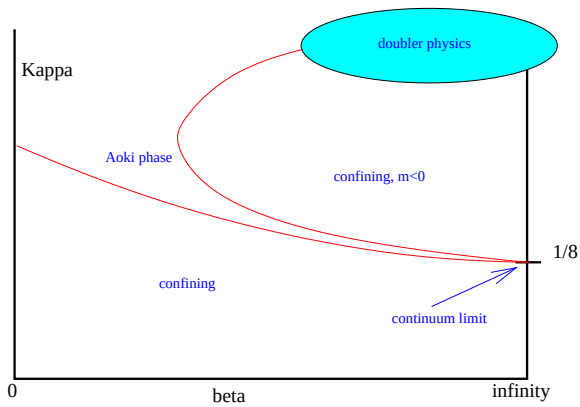

Fig. 26. The qualitative structure of the β, K plane including the possibility of an Aoki phase.

This extends our effective potential to

$$V(\vec{\pi}, \sigma) = \lambda(\sigma^2 + \vec{\pi}^2 - v^2)^2 - c_1(1/K - 1/K_c(\beta))\sigma + c_2\sigma^2 - \mu\pi_3 \tag{263}$$

The twisted mass represents the addition of a "magnetic field" conjugate to the order parameter for the Aoki phase.

There are a variety of motivations for adding such a term to our lattice action [95, 96]. Primary among them is that $O(a)$ lattice artifacts can be arranged to cancel. With two flavors of conventional Wilson fermions, these effects change sign on going from positive to negative mass, and if we put all the mass into the twisted term we are half way between. It should be noted that this cancellation only occurs when all the mass comes from the twisted term; for other combinations with a traditional mass term, some lattice artifacts of $O(a)$ will survive. Also, although it looks like we are putting phases into the quark masses, these cancel between the two flavors. The resulting fermion determinant remains positive and we are working at $\Theta = 0$. Furthermore, the algorithm is considerably simpler and faster than either overlap [69, 97] or domain wall [98, 99] fermions while avoiding the diseases of staggered quarks [100]. Another nice feature of adding a twisted mass term is that it allows a better understanding of the Aoki phase and shows how to continue around it. Figs. 27 and 28 show how this works for the case $c_2 > 0$ and $c_2 < 0$, respectively.

Of course some difficulties come along with these advantages. First, one needs to know K_c. Indeed, with the Aoki phase present, the definition of this quantity is a bit fuzzy. And second, the mass needs to be larger than the c_2 artifacts. Indeed, as Figs. 27 and 28 suggest, if it is not, then one is really studying the physics of the Aoki phase, not the correct continuum limit. This also has implications for how close to the continuum one must be to study this structure; in particular, one must have β large enough so the Aoki phase does not extend into the doubler region.

The question of the sign of c_2 remains open. Simulations suggest that the usual Aoki phase with $c_2 > 0$ is the situation with the Wilson gauge action. Recently Ref. [94] has pointed out that with the twisted mass term present, all eigenvalues of the product of gamma five with the Dirac operator will have non-zero imaginary parts. Thus to have $c_2 < 0$, the phase transition at non-vanishing twisted mass must occur where the fermion determinant does not vanish on any

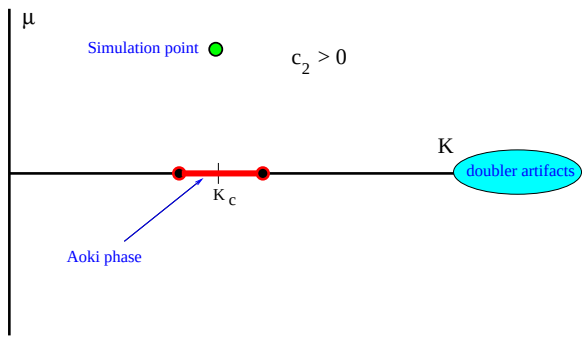

Fig. 27. Continuing around the Aoki phase with twisted mass. This sketch considers the case $c_2 > 2$ where the parity broken phase extends over a region along the kappa axis. Figure taken from Ref. [91].

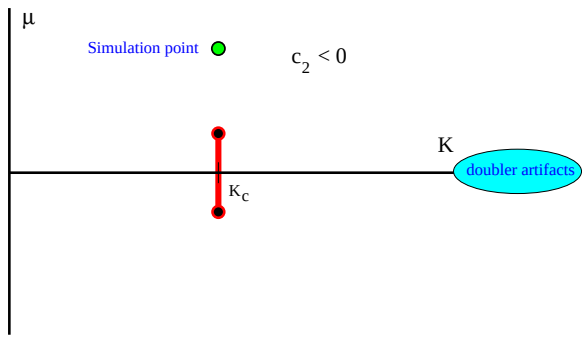

Fig. 28. As in Fig. 27, but for the case $c_2 < 0$ so the chiral transition on the kappa axis becomes first order. Figure taken from Ref. [91].

configuration. This contrasts with the $C_2 > 0$ case where small eigenvalues of D are expected occur in the vicinity of the critical hopping. This at first sight makes $c_2 < 0$ seem somewhat unnatural; however, this is not a proof since we saw in Subsection 9.6 that phase transitions without small eigenvalues of the Dirac operator do occur in the continuum theory for two flavors with non-degenerate quarks.

This picture of the artifacts associated with Wilson fermions raises some interesting questions. One concerns the three flavor theory. As discussed previously, in this case a parity broken phase becomes physical with negative mass. Indeed, three degenerate quarks of negative mass represent QCD with a strong CP angle $\theta = \pi$, for which spontaneous breaking of CP is expected. In some sense the Aoki phase becomes physical. Also, with three flavors the twisting process is not unique, with possible twists in the λ_3 or λ_8 directions. For example, using only λ_3 would suggest a possible twisted mass of form $m_u \sim e^{2\pi i/3}$, $m_d \sim e^{-2\pi i/3}$, $m_s \sim 1$. Whether there is an optimum twisting procedure for three flavors is unclear.

Another special case is one flavor QCD [78]. In this situation the anomaly removes all chiral symmetry, and the quark condensate loses meaning as an order parameter. The critical value of kappa where the mass gap disappears is decoupled from the point of zero physical quark mass. There is a parity broken phase, but it occurs only at sufficiently negative mass. And from the point of view of twisting the mass, without chiral symmetry there is nothing to twist other than turning on the physical parameter Θ.

11 Lattice actions preserving chiral symmetry

11.1 The Nielsen-Ninomiya theorem

As discussed some time ago [90], the doubling issue is closely tied to topology in momentum space. To see how this works, let us first establish a gamma matrix convention

$$\vec{\gamma} = \sigma_1 \otimes \vec{\sigma} = \begin{pmatrix} 0 & \vec{\sigma} \\ \vec{\sigma} & 0 \end{pmatrix} \tag{264}$$

$$\gamma_0 = \sigma_2 \otimes I = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \tag{265}$$

$$\gamma_5 = \sigma_3 \otimes I = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{266}$$

Now suppose we have an anti-Hermitian Dirac operator D that anti-commutes with γ_5

$$D = -D^\dagger = -\gamma_5 D \gamma_5. \tag{267}$$

Considering this quantity in momentum space, its most general form is

$$D(p) = \begin{pmatrix} 0 & z(p) \\ -z^*(p) & 0 \end{pmatrix} \tag{268}$$

where $z(p)$ is a quaternion

$$z(p) = z_0(p) + i\vec{\sigma} \cdot \vec{z}(p). \tag{269}$$

Thus we see that any chirally symmetric Dirac operator maps momentum space onto the space of quaternions.

The Dirac equation is obtained by expanding the momentum space operator around a zero, i.e. $D(p) \simeq \not{p} = \gamma_\mu p_\mu$. Now consider a three dimensional sphere embedded in four-dimensional momentum-space and surrounding the zero with a constant $D^2 \sim p^2$. The above quaternion wraps non-trivially about the the origin as we cover this sphere. Here is where the topology comes in [90, 101]. Momentum space is periodic over Brillouin zones. We must have $z(p) = z(p + 2\pi n)$ where n is an arbitrary four vector with integer components. Because of this, we can restrict the momentum components to lie in the range $-\pi < p_\mu \leq \pi$, and we cannot have any non-trivial topology on the surface of this zone. Any mapping associated with a zero in $z(p)$ must unwrap somewhere else before we get to the surface. Assuming $D(p)$ remains finite, any zero must be accompanied by another wrapping in the opposite sense. Because of doubling, the 16 species with naive fermions split up into 8 zeros of each sense.

The above argument only tells us that a chiral lattice theory must have an even number of species. The case of a minimal doubling with only two species is in fact possible, although all methods presented so far [102] appear to involve a breaking of hyper-cubic symmetry. This breaking is associated with the direction between the zeros; this makes one direction special, although it might be possible to avoid it by having the zeros form a symmetric lattice using the periodicity of momentum space. This has not yet been demonstrated.

In earlier sections we discussed how an odd number of flavors raised some interesting issues; in particular the sign of the mass becomes relevant. In spite of this, there seems to be no contradiction with having, say, three light flavors in the continuum with a well defined chiral limit. The above lattice argument, however, seems to indicate troubles with maintaining an exact chiral symmetry with an odd number of flavors. Whether this apparent conflict is serious is unclear. One could always start with a multiple fermion theory and then, with something like a Wilson term, give a few species masses while leaving behind an odd number of massless fermions. This will involve some parameter tuning, but presumably can give a reasonable chiral limit for odd $N_f > 2$. This does not obviate the fact emphasized earlier that with only one flavor there must not be any remaining chiral symmetry even in the continuum.

11.2 Minimal doubling

Several chiral lattice actions satisfying the minimal condition of $N_f = 2$ flavors are known. Some time ago Karsten [103] presented a simple form by inserting a factor of $i\gamma_4$ into a Wilson like term for space-like hoppings. A slight variation appeared in a discussion by Wilczek [104] a few years later. More recently, a new four-dimensional action was motivated by the analogy with two dimensional graphene [101]. Since then numerous variations have been presented [102, 105–108].

The main potential advantage with these approaches lies in their ultra-locality. They all involve only nearby neighbor hoppings for the fermions. Thus they should be extremely fast in simulations while still protecting masses from additive renormalization and helping control mixing of operators with different chirality. The approach also avoids the uncontrolled errors associated with the rooting approximation discussed later [60, 100, 109, 110]. On the other hand, all minimally-doubled actions presented so far have the above mentioned disadvantage of breaking lattice hyper-cubic symmetry. With interactions, this will lead to the necessity of renormaliza-

tion counter-terms that also violate this symmetry [111]. The extent to which this will complicate practical simulations remains to be investigated.

Minimally-doubled chiral fermions have the unusual property of a single local field creating two distinct fermionic species. Here we discuss a point-splitting method for separating the effects of the two flavors which can be created by a single fermion field. For this we will work with one specific form for the fermion action, but the method should be easily extended to other minimally-doubled formulations.

We concentrate on a minimally-doubled fermion action which is a slight generalization of those presented by Karsten [103] and Wilczek [104]. The fermion term in the lattice action takes the form $\overline{\psi}D\psi$. For free fermions we start in momentum space with

$$D(p) = i\sum_{i=1}^{3}\gamma_i \sin(p_i) + \frac{i\gamma_4}{\sin(\alpha)}\left(\cos(\alpha) + 3 - \sum_{\mu=1}^{4}\cos(p_\mu)\right) \tag{270}$$

This includes a Wilson like term for the space-like hoppings but containing an extra factor of $i\gamma_4$. As a function of the momentum p_μ, the propagator $D^{-1}(p)$ has two poles, located at $\vec{p} = 0$, $p_4 = \pm\alpha$. Relative to the naive fermion action, the other doublers have been given a large "imaginary chemical potential" by the Wilson like term. The parameter α allows adjusting the relative positions of the poles. The original Karsten/Wilczek actions correspond to $\alpha = \pi/2$.

This action maintains one exact chiral symmetry, manifested in the anti-commutation relation $[D, \gamma_5]_+ = 0$. The two species, however, are not equivalent, but have opposite chirality. To see this, expand the propagator around the two poles and observe that one species, that corresponding to $p_4 = +\alpha$, uses the usual gamma matrices, but the second pole gives a proper Dirac behavior using another set of matrices $\gamma'_\mu = \Gamma^{-1}\gamma_\mu\Gamma$. The Karsten/Wilczek formulation uses $\Gamma = i\gamma_4\gamma_5$, although other minimally-doubled actions may involve a different transformation. After this transformation $\gamma'_5 = -\gamma_5$, showing that the two species rotate oppositely under the exact chiral symmetry, and this symmetry should be regarded as "flavored." One can think of the physical chiral symmetry as that generated in the continuum theory by $\tau_3\gamma_5$.

It is straightforward to transform the momentum space action in Eq. (270) to position space and insert gauge fields $U_{ij} = U^\dagger_{ji}$ on the links connecting lattice sites. Explicitly indicating the site indices, the Dirac operator becomes

$$\begin{aligned} D_{ij} &= U_{ij}\sum_{\mu=1}^{3}\gamma_i\frac{\delta_{i,j+e_\mu} - \delta_{i,j-e_\mu}}{2} \\ &+ \frac{i\gamma_4}{\sin(\alpha)}\left((\cos(\alpha)+3)\delta_{ij} - U_{ij}\sum_{\mu=1}^{4}\frac{\delta_{i,j+e_\mu} + \delta_{i,j-e_\mu}}{2}\right). \end{aligned} \tag{271}$$

Again we see analogy with Wilson fermions [10] for the space directions but augmented with an $i\gamma_4$ inserted in the Wilson term.

Perturbative calculations [111] have shown that interactions with the gauge fields can shift the relative positions of the poles along the direction between them. In other words, the parameter α receives an additive renormalization. Furthermore, the form of the action treats the fourth direction differently than the spatial coordinates, this is the breaking of hyper-cubic symmetry mentioned above. There arise three possible new counter-terms for the renormalization of the

theory. First there is a possible renormalization of the on-site contribution to the action proportional to $i\overline{\psi}\gamma_4\psi$. This provides a handle on the shift of the parameter α. Secondly, the breaking of the hyper-cubic symmetry indicates one may need to adjust the fermion "speed of light." This involves a combination of the above on-site term and the strength of temporal hopping proportional to $\delta_{i,j+e_4} + \delta_{i,j-e_4}$. Finally, the breaking of hyper-cubic symmetry can feed back into the gluonic sector, suggesting a possible counter-term of form $F_{4\mu}F_{4\mu}$ to maintain the gluon "speed of light." In lattice language, this corresponds to adjusting the strength of time-like plaquettes relative to space-like ones.

Of these counter-terms, $i\overline{\psi}\gamma_4\psi$ is of dimension 3 and is probably the most essential. Quantum corrections induce the dimension 4 terms, suggesting they may be small and could partially be absorbed by accepting a lattice asymmetry. How difficult these counter-terms are to control awaits simulations.

Note that all other dimension 3 counter-terms are forbidden by basic symmetries. For example, chiral symmetry forbids $\overline{\psi}\psi$ and $i\overline{\psi}\gamma_5\psi$ terms, and spatial cubic symmetry removes $\overline{\psi}\gamma_i\psi$, $\overline{\psi}\gamma_i\gamma_5\psi$, and $\overline{\psi}\sigma_{ij}\psi$ terms. Finally, commutation with γ_4 plus space inversion eliminates $\overline{\psi}\gamma_4\gamma_5\psi$.

The fundamental field ψ can create either of the two species. For a quantity that creates only one of them, it is natural to combine fields on nearby sites in such a way as to cancel the other. In other words, one can point split the fields to separate the poles which occur at distinct "bare momenta." For the free theory, one construction that accomplishes this is to consider

$$\begin{aligned} u(q) &= \frac{1}{2}\left(1 + \frac{\sin(q_4+\alpha)}{\sin(\alpha)}\right)\psi(q+\alpha e_4) \\ d(q) &= \frac{1}{2}\,\Gamma\left(1 - \frac{\sin(q_4-\alpha)}{\sin(\alpha)}\right)\psi(q-\alpha e_4) \end{aligned} \tag{272}$$

where $\Gamma = i\gamma_4\gamma_5$ for the Karsten/Wilczek formulation. Here we have inserted factors containing zeros cancelling the undesired pole. This construction is not unique, and specific details will depend on the particular minimally-doubled action in use. The factor of Γ inserted in the d quark field accounts for the fact that the two species use different gamma matrices. This is required since the chiral symmetry is flavored, corresponding to an effective minus sign in γ_5 for one of the species.

It is now straightforward to proceed to position space and insert gauge field factors to keep gauge transformation properties simple

$$\begin{aligned} u_x &= \frac{1}{2}e^{i\alpha x_4}\left(\psi_x + i\,\frac{U_{x,x-e_4}\psi_{x-e_4} - U_{x,x+e_4}\psi_{x+e_4}}{2\sin(\alpha)}\right) \\ d_x &= \frac{1}{2}\Gamma e^{-i\alpha x_4}\left(\psi_x - i\,\frac{U_{x,x-e_4}\psi_{x-e_4} - U_{x,x+e_4}\psi_{x+e_4}}{2\sin(\alpha)}\right). \end{aligned} \tag{273}$$

The various additional phase factors serve to remove the oscillations associated with the bare fields having their poles at non-zero momentum.

Given the basic fields for the individual quarks, one can go on to construct mesonic fields, which then also involve point splitting. To keep the equations simpler, we now consider the case $\alpha = \pi/2$. For example, the neutral pion field becomes

$$\pi_0(x) = \tfrac{i}{2}(\overline{u}_x\gamma_5 u_x - \overline{d}_x\gamma_5 d_x) =$$

$$\frac{i}{16}\Big(4\overline{\psi}_x\gamma_5\psi_x + \overline{\psi}_{x-e_4}\gamma_5\psi_{x-e_4} + \overline{\psi}_{x+e_4}\gamma_5\psi_{x+e_4} \\ -\overline{\psi}_{x+e_4}UU\gamma_5\psi_{x-e_4} - \overline{\psi}_{x-e_4}UU\gamma_5\psi_{x+e_4}\Big). \tag{274}$$

Note that this involves combinations of fields at sites separated by either 0 or 2 lattice spacings. In contrast, the η' takes the form

$$\eta'(x) = \tfrac{i}{2}(\overline{u}_x\gamma_5 u_x + \overline{d}_x\gamma_5 d_x) = \\ \tfrac{1}{8}\Big(\overline{\psi}_{x-e_4}U\gamma_5\psi_x - \overline{\psi}_x U\gamma_5\psi_{x-e_4} - \overline{\psi}_{x+e_4}U\gamma_5\psi_x + \overline{\psi}_x U\gamma_5\psi_{x+e_4}\Big) \tag{275}$$

where all terms connect even with odd parity sites. In a recent paper, Tiburzi [112] has discussed how the anomaly, which gives the η' a mass of order Λ_{qcd}, can be understood in terms of the necessary point splitting.

11.3 Domain wall and overlap fermions

The overlap fermion was originally developed [113] as a limit of a fermion formulation using four dimensional surface modes on a five dimensional lattice. This effectively amounts to using Shockley surface states as the basis for a theory maintaining chiral symmetry [98]. For a review see Ref. [114]. The idea is to set up a theory in one extra dimension so that surface modes exist, and our observed world is an interface with our quarks and leptons being these surface modes. Particle hole symmetry naturally gives the basic Fermions zero mass. In the continuum limit the extra dimension becomes unobservable due to states in the interior requiring a large energy to create. In this picture, opposing surfaces carry states of opposite helicity, and the anomalies are due to a tunnelling through the extra dimension.

Ref. [115] discussed the general conditions for surface modes to exist. Normalized solutions are bound to any interface separating a region with supercritical from sub-critical hopping. Kaplan's original paper [98] considered not a surface, but an interface with $M = M_{cr} + m\epsilon(x)$, where M_{cr} is the critical value for the mass parameter where the five dimensional fermions would be massless. Shamir [116] presented a somewhat simpler picture where the hopping vanishes on one side, which then drops out of the problem and we have a surface.

To couple gluon fields to this theory without adding unneeded degrees of freedom, the gauge fields are taken to lie in the four physical space-time directions and be independent of the fifth coordinate. In this approach, the extra dimension is perhaps best thought of as a flavor space [117]. With a finite lattice this procedure gives equal couplings of the gauge field to the Fermion modes on opposing walls in the extra dimension. Since the left and right handed modes are separated by the extra dimension, they only couple through the gauge field. The result is an effective light Dirac Fermion. In the case of the strong interactions, this provides an elegant scheme for a natural chiral symmetry without the tuning inherent in the Wilson approach. The breaking of chiral symmetry arises only through finiteness of the extra dimension.[9]

The name "overlap operator" comes from the overlap of eigenstates of the different five dimensional transfer matrices on each side of the interface. Although originally derived from the

[9]The anomaly, however, shows that some communication between the surfaces survives even as the extra dimension becomes infinite. This is possible since the same gauge fields are on each surface.

infinite limit of the five dimensional formalism, one can formulate the overlap operator directly in four dimensions. We begin with the Fermionic part of some generic action as a quadratic form $S_f = \overline{\psi} D \psi$. The usual "continuum" Dirac operator $D = \sum \gamma_\mu D_\mu$ naively anti-commutes with γ_5, i.e. $[\gamma_5, D]_+ = 0$. Then the change of variables $\psi \to e^{i\theta\gamma_5}\psi$ and $\overline{\psi} \to \overline{\psi} e^{i\theta\gamma_5}$ would be a symmetry of the action. This, however, is inconsistent with the chiral anomalies. The conventional continuum discussion presented earlier maps this phenomenon into the Fermionic measure [66].

On the lattice we work with a finite number of degrees of freedom; thus, the above variable change is automatically a symmetry of the measure. To parallel the continuum discussion, it is necessary to modify the symmetry transformation on the action so that the measure is no longer invariant. Remarkably, it is possible to construct a modified symmetry under which corresponding actions are exactly invariant.

To be specific, one particular variation [69, 118–121] modifies the change of variables to

$$\begin{aligned} \psi &\longrightarrow e^{i\theta\gamma_5}\psi \\ \overline{\psi} &\longrightarrow \overline{\psi} e^{i\theta(1-aD)\gamma_5} \end{aligned} \tag{276}$$

where a represents the lattice spacing. Note the asymmetric way in which the independent Grassmann variables ψ and $\overline{\psi}$ are treated. Requiring the action to be unchanged gives the relation [68, 122, 123].

$$D\gamma_5 = -\gamma_5 D + aD\gamma_5 D = -\hat{\gamma}_5 D \tag{277}$$

with $\hat{\gamma}_5 = (1 - aD)\gamma_5$. To proceed, we also assume the Hermeticity condition $\gamma_5 D \gamma_5 = D^\dagger$. We see that the naive anticommutation relation receives a correction of order the lattice spacing. The above "Ginsparg-Wilson relation" along with the Hermeticity condition is equivalent to the unitarity of the combination $V = 1 - aD$.

Neuberger [118, 119] and Chiu and Zenkin [120] presented an explicit operator with the above properties. They first construct V via a unitarization of an undoubled chiral violating Dirac operator, such as the Wilson operator D_w. This operator should also satisfy the above Hermeticity condition $\gamma_5 D_w \gamma_5 = D_w^\dagger$. Specifically, they consider

$$V = -D_w (D_w^\dagger D_w)^{-1/2}. \tag{278}$$

The combination $(D_w^\dagger D_w)^{-1/2}$ is formally defined by finding a unitary operator to diagonalize the Hermitian combination $D_w^\dagger D_w$, taking the square root of the eigenvalues, and then undoing the unitary transformation.

Directly from V we construct the overlap operator as

$$D = (1 - V)/a. \tag{279}$$

The Ginsparg-Wilson relation of Eq. (277) is most succinctly written as the unitarity of V coupled with its γ_5 Hermeticity

$$\gamma_5 V \gamma_5 V = 1. \tag{280}$$

The basic projection process is illustrated in Fig. 29.

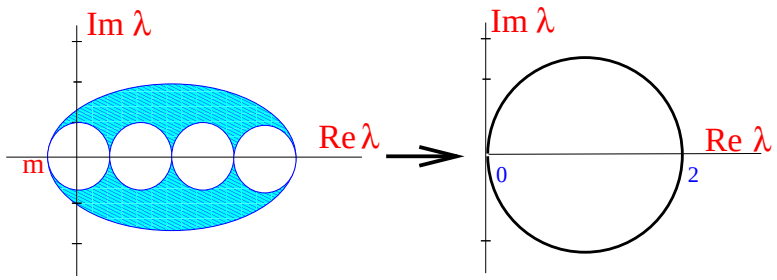

Fig. 29. The overlap operator is constructed by projecting the eigenvalues of the Wilson operator onto a circle. Figure taken from Ref. [124].

The overlap operator has several nice properties. Being constructed from a unitary operator, the normality of D is guaranteed. But, most important, it exhibits a lattice version of an exact chiral symmetry. [125] The fermionic action $\overline{\psi}D\psi$ is invariant under the transformation

$$\begin{aligned} \psi &\to e^{i\theta\gamma_5}\psi \\ \overline{\psi} &\to \overline{\psi}e^{i\theta\hat{\gamma}_5} \end{aligned} \tag{281}$$

where

$$\hat{\gamma}_5 = V\gamma_5. \tag{282}$$

As with γ_5, this quantity which appeared in Eq. (277) is Hermitean and its square is unity. Thus its eigenvalues are all plus or minus unity. The trace defines an index

$$\nu = \frac{1}{2}\mathrm{Tr}\hat{\gamma}_5 \tag{283}$$

which plays exactly the role of the index in the continuum. If the gauge fields are smooth, this counts the topology of the gauge configuration. The factor of $1/2$ in Eq. (283) appears because the exact zero modes of the overlap operator have partners on the opposite side of the unitarity circle that also contribute to the trace.

At this point the hopping parameter in D_w is a parameter. To have the desired single light Fermion per flavor of the theory, the hopping parameter should be appropriately adjusted to lie above the critical value where D_w describes a massless flavor, but not so large that additional doublers come into play [126]. There are actually two parameters to play with, the hopping parameter of D_w, and the lattice spacing. When the latter is finite and gauge fields are present, the location of the critical hopping parameter in D_w is expected to shift from that of the free Fermion theory. As we saw when discussing the Aoki phase, there is potentially a rather complex phase structure in the plane of these two parameters, with various numbers of doublers becoming massless as the hopping is varied. The Ginsparg-Wilson relation in and of itself does not in general determine the number of physical massless Fermions.

Although the Wilson operator entering this construction is local and quite sparse, the resulting overlap action is not. Because of the inversion in Eq. (278), it involves direct couplings between arbitrarily separated sites [127–129]. How rapidly these couplings fall with distance depends on the gauge fields and is not fully understood. The five dimensional domain-wall theory is local

in the most naive sense; all terms in the action only couple nearest neighbor sites. However, were one to integrate out the heavy modes, the resulting low energy effective theory would also involve couplings with arbitrary range. Despite these non-localities, encouraging studies [119, 130–133] show that it may indeed be practical to implement the required inversion in large scale numerical simulations. The overlap operator should have memory advantages over the domain wall approach since a large number of fields corresponding to the extra dimension do not need to be stored.

The overlap approach hides the infinite sea of heavy Fermion states in the extra dimension of the domain wall approach. This tends to obscure the possible presence of singularities in the required inversion of the Wilson kernel. Detailed analysis [134, 135] shows that this operator is particularly well behaved order by order in perturbation theory. This has led to hopes that this may eventually lead to a rigorous formulation of chiral models, such as the standard model.

Despite being the most elegant known way to have an exact remnant of chiral symmetry on the lattice, the overlap operator raises several issues. These complications probably become insignificant as the continuum limit is approached, but should be kept in mind given the high computational cost of this approach. To begin with, the overlap is highly non-unique. It explicitly depends on the kernel being projected onto the unitary circle. Even after choosing the Wilson kernel, there is a dependence on the input mass parameter. One might want to define topology in terms of the number of exact zero modes of the overlap operator. However the non-uniqueness leaves open the question of whether the winding number of a gauge configuration might depend on this choice. Later we will return to the question of possible ambiguities in defining topological susceptibility in the continuum limit.

In this connection, it is possible to make a bad choice for the mass parameter. In particular, if it is chosen below the continuum kappa critical value of $1/8$, no low modes will survive. This us true despite the fact that the corresponding operator will still satisfy the Ginsparg-Wilson condition. This explicitly shows that just satisfying the Ginsparg-Wilson condition is not a sufficient condition for a chiral theory. Conversely, if one chooses the mass parameter too far in the supercritical region, additional low modes will be produced from the doublers. As mentioned earlier, the Ginsparg-Wilson condition does not immediately determine the number of flavors in the theory.

Another issue concerns the one flavor case, discussed earlier. Because of the anomaly, this theory is not supposed to show any chiral symmetry and has no Goldstone bosons. Nevertheless, one can construct the overlap operator and it will satisfy the Ginsparg-Wilson condition. This shows that the consequences of this condition are weaker than for the usual continuum chiral symmetry. With a conventional chiral symmetry, the spectrum cannot show a gap. Either we have the Goldstone bosons of spontaneous chiral breaking or we have massless fermions [136].

Finally the overlap behaves peculiarly for fermions in higher representations than the fundamental. As we discussed earlier, the number of zero modes associated with a non-trivial topology in the continuum theory depends on the fermion representation being considered. It has been observed in numerical simulations that the appropriate multiplicity is not always seen for the overlap operator constructed on rough gauge configurations [137].

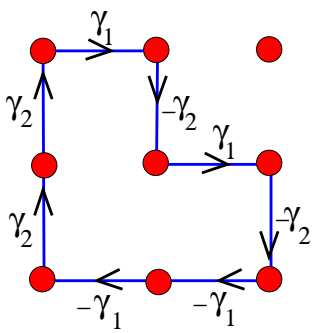

Fig. 30. When a fermion circumnavigates a loop in the naive formulation, it picks up a factor that always involves an even power of any particular gamma matrix. Figure from Ref. [109].

11.4 Staggered fermions

Another fermion formulation that has an exact chiral symmetry is the so called "staggered" approach. To derive this it is convenient to begin with the "naive" discretization of the Dirac equation from before. This considers fermions hopping between nearest neighbor lattice sites while picking up a factor of $\pm i\gamma_\mu$ for a hop in direction $\pm\mu$. Going to momentum space, the discretization replaces powers of momentum with trigonometric functions, for example

$$\gamma_\mu p_\mu \to \gamma_\mu \sin(p_\mu). \tag{284}$$

Here we work in lattice units and thus drop factors of a. As discussed before, this formulation reveals the famous "doubling" issue, arising because the fermion propagator has poles not only for small momentum, but also whenever any component of the momentum is at π. The theory represents not one fermion, but sixteen. And the various doublers have differing chiral properties. This arises from the simple observation that

$$\frac{d}{dp}\sin(p)|_{p=\pi} = -\frac{d}{dp}\sin(p)|_{p=0}. \tag{285}$$

The consequence is that the helicity projectors $(1 \pm \gamma_5)/2$ for a travelling particle depend on which doubler one is observing.

Now consider a fermion traversing a closed loop on the lattice. As illustrated in Fig. 30, the corresponding gamma matrix factors will always involve an even number of any particular γ_μ. Thus the resulting product is proportional to the identity. If a fermion starts off with a particular spinor component, it will wind up in the same component after circumnavigating the loop. This means that the fermion determinant exactly factorizes into four equivalent pieces. The naive theory has an exact $U(4)$ symmetry, as pointed out some time ago by Karsten and Smit [138]. Indeed, for massless fermions this is actually a $U(4) \otimes U(4)$ chiral symmetry. This symmetry does not contradict any anomalies since it is not the full naive $U(16) \otimes U(16)$ of 16 species. The chiral symmetry generated by γ_5 remains exact, but is allowed because it is actually a flavored symmetry. As mentioned above, the helicity projectors for the various doubler species use different signs for γ_5.

The basic idea of staggered fermions is to divide out this $U(4)$ symmetry [139–141] by projecting out a single component of the fermion spinor on each site. Taking $\psi \to P\psi$, one projector that accomplishes this is

$$P = \frac{1}{4}\left(1 + i\gamma_1\gamma_2(-1)^{x_1+x_2} + i\gamma_3\gamma_4(-1)^{x_3+x_4} + \gamma_5(-1)^{x_1+x_2+x_3+x_4}\right) \tag{286}$$

where the x_i are the integer coordinates of the respective lattice sites. This immediately reduces the doubling from a factor of sixteen to four. It is the various oscillating sign factors in this formula that give staggered fermions their name.

At this stage the naive $U(1)$ axial symmetry remains. Indeed, the projector used above commutes with γ_5. This symmetry is allowed since four species, often called "tastes," remain. Among them the symmetry is a taste non-singlet; under a chiral rotation, two rotate one way and two the other.

The next step taken by most of the groups using staggered fermions is the rooting trick. In the hope of reducing the multiplicity down from four, the determinant is replaced with its fourth root, $|D| \to |D|^{1/4}$. With several physical flavors this trick is applied separately to each. In simple perturbation theory this is correct since each fermion loop gets multiplied by one quarter, cancelling the extra factor from the four "tastes."

At this point one should be extremely uneasy: the exact chiral symmetry is waving a huge red flag. Symmetries of the determinant survive rooting, and thus the exact $U(1)$ axial symmetry for the massless theory remains. For the unrooted theory this was a flavored chiral symmetry. But, having reduced the theory to one flavor, how can there be a flavored symmetry without multiple flavors? We will now show why this rooting trick fails non-perturbatively when applied to the staggered quark operator.

11.5 The rooting issue

In previous sections we have seen that the chiral symmetry with N_f fermion flavors has a rather complicated dependence on N_f. With only one flavor there is no chiral symmetry at all, while in general if the fermions are massless, there are $N_f^2 - 1$ Goldstone bosons. We have also seen a qualitative difference in the mass dependence between an even and an odd number of species. Physics does not behave smoothly in the number of flavors and this raises issues for fermion formulations that inherently have multiple flavors, such as staggered fermions.

Starting with four flavors, the basic question is whether one can adjust N_f down to one using the formal expression

$$\begin{vmatrix} D+m & 0 & 0 & 0 \\ 0 & D+m & 0 & 0 \\ 0 & 0 & D+m & 0 \\ 0 & 0 & 0 & D+m \end{vmatrix}^{\frac{1}{4}} = |D+m|? \tag{287}$$

This has been proposed and is widely used as a method for eliminating the extra species appearing with staggered fermion simulations.

At this point it is important to emphasize that asking about the viability of Eq. (287) is a vacuous question outside the context of a regulator. Field theory has divergences that need to be controlled, and, as we have seen above, the appearance of anomalies requires care. In particular,

the regulated theory must explicitly break all anomalous symmetries in a way that survives in the continuum limit.

So we must apply Eq. (287) before removing the regulator. This is generally expected to be okay as long as the regulator breaks any anomalous symmetries appropriately on each of the four factors. For example, we expect rooting to be valid for four copies of the overlap operator. This satisfies a modified chiral symmetry $D\gamma_5 = -\hat{\gamma}_5 D$ where the gauge winding ν appears in the gauge dependent matrix $\hat{\gamma}$ through $\mathrm{Tr}\hat{\gamma}_5 = 2\nu$.

Section 9 showed that in the continuum with N_f degenerate flavors there is a Z_{N_f} symmetry in mass parameter space corresponding to taking $m \to e^{2\pi i\gamma_5/N_f} m$. Suppose we try to force the Z_4 symmetry in the regulated theory before we root. This is easily accomplished by considering the determinant

$$\begin{vmatrix} D + me^{\frac{i\pi\gamma_5}{4}} & 0 & 0 & 0 \\ 0 & D + me^{\frac{-i\pi\gamma_5}{4}} & 0 & 0 \\ 0 & 0 & D + me^{\frac{3i\pi\gamma_5}{4}} & 0 \\ 0 & 0 & 0 & D + me^{\frac{-3i\pi\gamma_5}{4}} \end{vmatrix} . \tag{288}$$

This maintains the above symmetry through a permutation of the four flavors. This modification of the determinant still gives a valid formulation of the four flavor theory at vanishing Θ because the imposed phases cancel. But expressed in this way, we start with four one-flavor theories each with a different value of Θ. Were we to root this form, we would be averaging over four inequivalent theories. This is not expected to be correct, much as we would not expect rooting two different masses to give a theory of the average mass; *i.e.*

$$(|D + m_1||D + m_2|)^{1/2} \neq |D + \sqrt{m_1 m_2}| \, . \tag{289}$$

So we have presented both a correct and an incorrect way to root a four flavor theory down to one. What is the situation with staggered fermions, the primary place where rooting has been applied? The problem is that the kinetic term of the staggered action maintains one exact chiral symmetry even at finite lattice spacing. Without rooting this is flavor non-singlet amongst the "tastes." As discussed earlier, there are two tastes of each chirality. But, because of this exact symmetry, which contains a Z_4 subgroup, rooting to reduce the theory to one flavor is similar to the second case above and is not expected to be valid. In particular, rooting does not remove the Z_4 discrete symmetry in the mass parameter, a symmetry which must not be present in the one flavor theory. Thus, just as in the above example, the tastes are not equivalent and rooting averages inequivalent theories.

The conclusion is that rooted staggered fermions are not QCD. So, what is expected to go wrong? The unbroken Z_4 symmetry will give rise to extra minima in the effective potential as a function of σ and η'. In particular, for one flavor QCD one will get an effective potential with minima along the lines of Fig. 13 instead of the desired structure of Fig. 15. Forcing the extra minima would most likely drive the η' mass down from its physical value. This shift should be rather large, of order Λ_{QCD}. This is testable, but being dominated by disconnected diagrams, may be rather difficult to verify in practice. In addition, if we vary the quark masses, the extra minima will result in phase transitions occurring whenever any single quark mass passes through zero. The previous discussion of the one flavor theory and the two flavor theory with non-degenerate quarks both show that this is unphysical; no structure is expected when only a single mass passes through zero.

This problem is admittedly subtle. Formula (287) seems intuitively obvious and does work if the individual factors take care of the possible anomalies, as with four copies of the overlap operator. [10] It is also correct perturbatively, since the rooting factor reduces any fermion loop by the correct amount. However, the basic structure built up in earlier sections makes it indisputable that the dependence of QCD on the parameter Theta is real. With the staggered action, the distinct tastes are not equivalent due to their different behavior under chiral rotations. It is this inequivalence that is at the heart of the failure of rooting for this particular action.

Despite these problems, several lattice collaborations continue to pursue staggered fermions using the rooting trick [142–144]. The justification is partly because the simulations are slightly faster than using Wilson fermions, and partly because the exact chiral symmetry simplifies operator mixing. The success of a variety of calculations which are not strongly dependent on the anomaly shows the approach, while technically incorrect, is often a good approximation. On the other hand, if one's goal is to test QCD as the theory of the strong interactions or to estimate QCD corrections to standard model processes, then one must be extremely wary of any discrepancies found using this method.

12 Other issues

12.1 Quantum fluctuations and topology

We have seen how zero modes of the Dirac operator are closely tied to the anomaly. And we have seen that for smooth classical fields, configurations that give zero modes for the classical Dirac operator do indeed exist. However, when getting into more detail with defining a lattice Dirac operator, we found subtle issues about which operator to use. And way back in Section 2 we saw that typical fields in path integrals are non-differentiable. This leads to the question of uniqueness for the winding number of a given gauge configuration. Indeed, is something like the topological susceptibility of the vacuum a true physical observable?

In [145] a definition of topological charge was constructed using the naive fermion operator as a regulator for the trace of γ_5 as in the earlier derivation of the index theorem. This operator does not generally give an integer value for a typical gauge configuration in simulations. However, it does reduce to such after a cooling procedure is used to remove short distance fluctuations. The results of such are shown in Fig. 31. On the other hand, the gauge field space in lattice gauge theory is simply connected. Empirically with enough cooling, any $SU(2)$ gauge configuration appears to eventually decay to a state of zero action, gauge equivalent to the vacuum.

Since configurations appear to cool ultimately to trivial winding, using a cooling algorithm to define topology requires an arbitrary selection for cooling time. Modifying the Wilson action can prevent the winding decay. For example, forbidding the lattice action on any given plaquette from becoming larger than a small enough number can prevent instanton decay [146]. Such an "admissibility" condition, however, violates reflection positivity [80] and arbitrarily selects a special instanton size where the action is minimum.

Cooling time is not the only issue here. While attaining an integer winding requires cooling, note in Fig. 31 that the initial cooling stages seem quite chaotic. This raises the question of whether the discrete stages reached after some cooling might depend rather sensitively on the

[10] A few still hide behind this wall so frail,
So blind to chiral twists that made it fail.

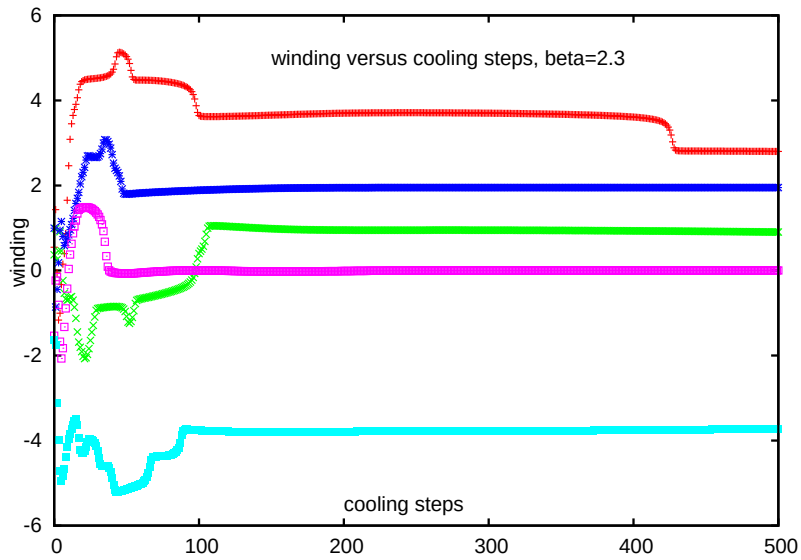

Fig. 31. The winding number as a function of cooling steps for a set of 5 lattices of size 16^4 at $\beta = 2.3$. Note how it settles into approximately integer values with occasional jumps between different windings. Figure from Ref. [145].

cooling algorithm. In Fig. 32 I show the evolution of a single lattice with three different relaxation algorithms. One algorithm consists of sweeping over the lattice using checkerboard ordering and replacing each link with the group element that minimizes the action associated with the given link. This is done by projecting the sum of staples that interact with the link onto the group. For the second approach, an under-relaxed algorithm adds the old link to the sum of the neighborhood staples before projecting onto the new group element. Finally, an over-relaxed approach subtracts the old element from the staple sum. The resulting windings not only depend on cooling time, but also on the specific algorithm chosen.

In an extensive analysis, Ref. [147] has compared a variety of filtering methods to expose topological structures in gauge configurations. All schemes have some ambiguities, but when the topological structures are clear, the various approaches when carefully tuned give similar results. Nevertheless the question remains of whether there is a rigorous and unambiguous definition of topology that applies to all typical configurations arising in a simulation. Luscher has discussed using a differential flow with the Wilson gauge action to accomplish the cooling [148]. This corresponds to the limit of maximal under-relaxation. This approach still allows the above topology collapse unless prevented by something like the admissibility condition or the selection of an arbitrary flow time. In addition, if one wishes to determine the topological charge of a configuration obtained in some large scale dynamical simulation, it is unclear why one should take the particular choice of the Wilson gauge action for the cooling procedure.

The high sensitivity to the cooling algorithm on rough gauge configurations suggests that there may be an inherent ambiguity in defining the topological charge of typical gauge configu-

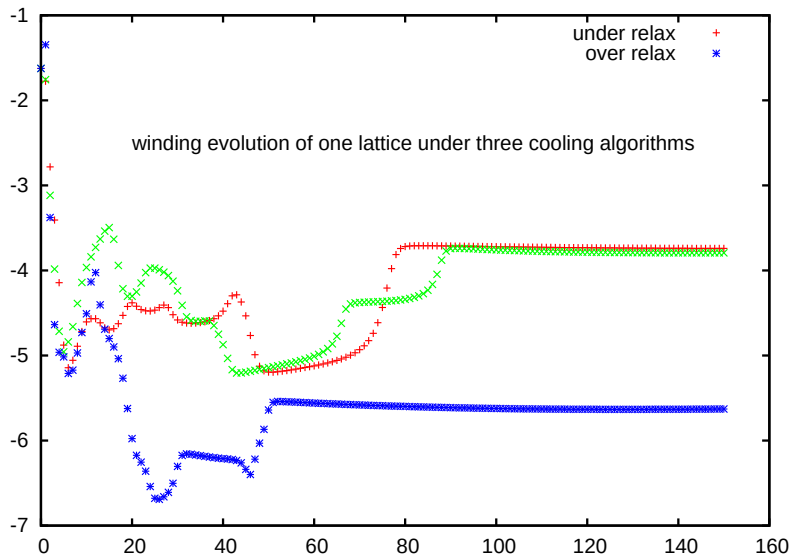

Fig. 32. The topological charge evolution for three different cooling algorithms on a single $\beta = 2.3$ lattice configuration for $SU(2)$ on a 16^4 lattice. Figure from Ref. [145].

rations and consequently a small ambiguity in the definition of topological susceptibility. It also raises the question of how smooth is a given definition of topological charge as the gauge fields vary; how much correlation is there between nearby gauge configurations? Although such issues are quite old [149], they continue to be of considerable interest [150–152].

As topological charge is suppressed by light dynamical quarks, this is connected to the question discussed earlier of whether the concept of a single massless quark is well defined [26]. Dynamical quarks are expected to suppress topological structures, and the chiral limit with multiple massless quarks should give zero topological susceptibility with a chiral fermion operator, such as the overlap. However, with only a single light quark, the lack of chiral symmetry indicates that there is no physical singularity in the continuum theory as this mass passes through zero. Any scheme dependent ambiguity in defining the quark mass would then carry through to the topological susceptibility.

One might argue that the overlap operator solves this problem by defining the winding number as the number of zero eigenvalues of this quantity. Indeed, it has been shown [152, 153] that this definition gives a finite result in the continuum limit. As one is using the fermion operator only as a probe of the gluon fields, this definition can be reformulated directly in terms of the underlying Wilson operator [154]. While the result may have a finite continuum limit, the earlier discussion showed that the overlap operator is not unique. In particular it depends on the initial Dirac operator being projected onto the overlap circle. For the conventional Wilson kernel, there is a dependence on the parameter commonly referred to as the domain-wall height. Whether there is an ambiguity in the index defined this way depends on the density of real eigenvalues of the kernel in the vicinity of the point from which the projection is taken. Numerical

evidence [130] suggests that this density decreases with lattice spacing, but it is unclear if this decrease is rapid enough to give a unique susceptibility in the continuum limit. The admissibility condition also successfully eliminates this ambiguity; however, as mentioned earlier, this condition is inconsistent with reflection positivity.

Whether topological susceptibility is well defined or not seems to have no particular phenomenological consequences. Indeed, this is not a quantity directly measured in any scattering experiment. It is only defined in the context of a technical definition in a particular non-perturbative simulation. Different valid schemes for regulating the theory might well come up with different values; it is only physical quantities such as hadronic masses that must match between approaches. The famous Witten-Veneziano relation [155, 156] does connect topological susceptibility of the pure gauge theory with the eta prime mass in the large number of colors limit. This mass, of course, remains well defined in the physical case of three colors, but the finite N_c corrections to topology can depend delicately on gauge field fluctuations, which are the concern here.

12.2 The standard model

Throughout the above we have concentrated on the strong interactions. It is only for this sector of the standard model that perturbation theory fails so spectacularly. But the weak and electromagnetic interactions are crucial parts of the full standard model (gravitation is ignored here since it has even more serious unsolved problems). And for these interactions it is also true that the perturbative expansion does not converge. Because the underlying couplings are so small, this does not appear to be of any practical importance; however, conceptually it is quite desirable to have a lattice formulation for these interactions as well. From a purist point of view, the continuum limit of a lattice theory defines a continuum field theory. Thus without a lattice description of the other interactions it is unclear whether we can say they are well founded field theories.

In this context we note that the general picture of the standard model has changed dramatically over the years. Originally it was the successes of quantum electrodynamics that made it the model relativistic field theory. Before QCD, the strong interactions were a mystery. But now we see that because of asymptotic freedom, QCD on its own is likely to be a well defined and self contained theory. It is the electroweak theory, where both the electric charge and the Higgs couplings are not asymptotically free, for which we lack a non-perturbative formulation. Indeed, a speculative topic such as the possibility of emergent gravity may be intimately tied to these issues with the weaker forces.

For the electromagnetic interactions, a lattice formulation at first seems straightforward, involving the introduction of an additional $U(1)$ gauge field for the photons. Unlike the strong interaction case, however, for electrodynamics we do not have asymptotic freedom to tell us how to take the continuum limit. And the physical coupling $\alpha \sim 1/137$ seems to be an unnaturally small number. Perhaps electrodynamics on its own does not actually exist as a field theory, much as believed for the scalar ϕ^4 theory. But photons and electrons are essential components to the world around us. One interesting possibility is that electromagnetism is actually only a part of a higher level theory, perhaps in some unification with the strong interactions.

With the weak interactions we hit a more serious snag in that they are known to violate parity. The W bosons appear to interact only to left handed fermions. As such we need to couple the fermions in a chiral manner, and it is not known how to do this in any non-perturbative

scheme. The problem here is closely tied to the anomaly and the fact that not all currents can be simultaneously conserved. Indeed, when applied to the weak interactions, the 't Hooft vertex gives rise to effective interactions that do not conserve baryon number. Any complete non-perturbative formulation must allow for such processes [157]. Some attempts to include such in a domain wall formulation have been presented [115, 158], but these generally involve heavy additional states such as "mirror" fermions [159]. While potentially viable, such approaches so far lack the theoretical elegance of the original Wilson lattice gauge theory. Indeed, it is the problem of chiral gauge theories that encourages studies of chiral symmetry from all possible angles.

Perhaps a lattice formulation more intimately tied to unification ideas could help here. The group $SO(10)$ looks quite interesting in this context [76]. Here a single generation of fermions fits nicely into a single 16 dimensional representation of this group. And in this picture anomalies are automatically cancelled. This would seem to indicate that there should be no obvious requirement for doublers as an obstacle to a lattice construction. However, the usual Wilson approach seems to require a term that is not a singlet under this group. This could be overcome with some added Higgs-like scalar fields, but then we get closer to the above mentioned models with the doublers playing the role of mirror fermions.

12.3 Where is the parity violation?

The standard model of elementary particle interactions is based on the product of three gauge groups, $SU(3) \otimes SU(2) \otimes U(1)_{em}$. Here the $SU(3)$ represents the strong interactions of quarks and gluons, the $U(1)_{em}$ corresponds to electromagnetism, and the $SU(2)$ gives rise to the weak interactions. We ignore here the technical details of electroweak mixing. The full model is, of course, parity violating, as necessary to describe observed helicities in beta decay. This violation is normally considered to lie in the $SU(2)$ of the weak interactions, with both the $SU(3)$ and $U(1)_{em}$ being parity conserving. We will show here that this is actually a convention, adopted primarily because the weak interactions are small compared to the others. We show below that reassigning degrees of freedom allows a reinterpretation where the $SU(2)$ gauge interaction is vector-like. Since the full model is parity violating, this process shifts the parity violation into the strong, electromagnetic, and Higgs interactions. The resulting theory pairs the left handed electron with a right handed anti-quark to form a Dirac fermion. With a vector-like weak interaction, the chiral issues which complicate lattice formulations now move to the other gauge groups. Requiring gauge invariance for the re-expressed electromagnetism then clarifies the mechanism behind one proposal for a lattice regularization of the standard model [160, 161].

For simplicity we consider here only the first generation, which involves four left handed doublets. These correspond to the neutrino/electron lepton pair plus three colors for the up/down quarks

$$\begin{pmatrix} \nu \\ e^- \end{pmatrix}_L, \begin{pmatrix} u^r \\ d^r \end{pmatrix}_L, \begin{pmatrix} u^g \\ d^g \end{pmatrix}_L, \begin{pmatrix} u^b \\ d^b \end{pmatrix}_L. \tag{290}$$

Here the superscripts from the set $\{r, g, b\}$ represent the internal $SU(3)$ index of the strong interactions, and the subscript L indicates left-handed helicities.

If we ignore the strong and electromagnetic interactions, leaving only the weak $SU(2)$, each of these four doublets is equivalent and independent. We now arbitrarily pick two of them and

do a charge conjugation operation, thus switching to their antiparticles

$$\begin{aligned} \begin{pmatrix} u^g \\ d^g \end{pmatrix}_L &\longrightarrow \begin{pmatrix} \overline{d^g} \\ \overline{u^g} \end{pmatrix}_R \\ \begin{pmatrix} u^b \\ d^b \end{pmatrix}_L &\longrightarrow \begin{pmatrix} \overline{d^b} \\ \overline{u^b} \end{pmatrix}_R . \end{aligned} \tag{291}$$

In four dimensions anti-fermions have the opposite helicity; so, we label these new doublets with R representing right handedness.

With two left and two right handed doublets, we can combine them into two Dirac doublets

$$\begin{pmatrix} \begin{pmatrix} \nu \\ e^- \end{pmatrix}_L \\ \begin{pmatrix} \overline{d^g} \\ \overline{u^g} \end{pmatrix}_R \end{pmatrix} \qquad \begin{pmatrix} \begin{pmatrix} u^r \\ d^r \end{pmatrix}_L \\ \begin{pmatrix} \overline{d^b} \\ \overline{u^b} \end{pmatrix}_R \end{pmatrix} . \tag{292}$$

Formally in terms of the underlying fields, the construction takes

$$\begin{aligned} \psi &= \frac{1}{2}(1-\gamma_5)\psi_{(\nu,e^-)} + \frac{1}{2}(1+\gamma_5)\psi_{(\overline{d^g},\overline{u^g})} \\ \chi &= \frac{1}{2}(1-\gamma_5)\psi_{(u^r,d^r)} + \frac{1}{2}(1+\gamma_5)\psi_{(\overline{d^b},\overline{u^b})} . \end{aligned} \tag{293}$$

From the conventional point of view, these fields have rather peculiar quantum numbers. For example, the left and right parts have different electric charges. Electromagnetism now violates parity. The left and right parts also have different strong quantum numbers; the strong interactions violate parity as well. Finally, the components have different masses; parity is violated in the Higgs mechanism.

The different helicities of these fields also have variant baryon number. This is directly related to the known baryon violating processes through weak "instantons" and axial anomalies [19]. When a topologically non-trivial weak field is present, the axial anomaly arises from a level flow out of the Dirac sea [162]. This generates a spin flip in the fields, *i.e.* $e^-_L \to (\overline{u^g})_R$. Because of the peculiar particle identification, this process does not conserve charge, with $\Delta Q = -\frac{2}{3} + 1 = \frac{1}{3}$. This would be a disaster for electromagnetism were it not for the fact that simultaneously the other Dirac doublet also flips $d^r{}_L \to (\overline{u^b})_R$ with a compensating $\Delta Q = -\frac{1}{3}$. This is anomaly cancellation, with the total $\Delta Q = \frac{1}{3} - \frac{1}{3} = 0$. Only when both doublets are considered together is the $U(1)$ symmetry restored. In this anomalous process baryon number is violated, with $L + Q \to \overline{Q} + \overline{Q}$. This is the famous " 't Hooft vertex" [19] discussed earlier in the context of the strong interactions.

12.4 A lattice model

The above discussion on twisting the gauge groups has been in the continuum. Now we turn to the lattice and show how this picture leads to a possible lattice model for the strong interactions, albeit with an unusual added coupling that make the treatment quite difficult to make rigorous [160, 161]. Whether this model is viable remains undecided, but it does incorporate many of the required features.

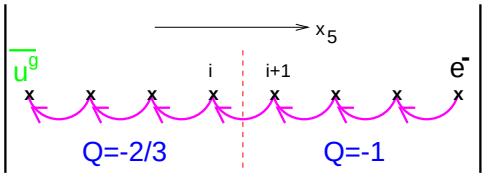

Fig. 33. Pairing the electron with the anti-green-up-quark. Figure taken from [161].

For this we use the domain wall approach for the fermions [98, 116]. As discussed earlier, in this picture, our four dimensional world is a "4-brane" embedded in 5-dimensions. The complete lattice is a five dimensional box with open boundaries, and the parameters are chosen so the physical quarks and leptons appear as surface zero modes. The elegance of this scheme lies in the natural chirality of these modes as the size of the extra dimension grows. With a finite fifth dimension one doubling remains, coming from interfaces appearing as surface/anti-surface pairs. It is natural to couple a four dimensional gauge field equally to both surfaces, giving rise to a vector-like theory.

We now insert the above pairing into this five dimensional scheme. In particular, consider the left handed electron as a zero mode on one wall and the right handed anti-green-up-quark as the partner zero mode on the other wall, as sketched in Fig. 33. This provides a lattice regularization for the $SU(2)$ of the weak interactions.

However, since these two particles have different electric charge, $U(1)_{EM}$ must be broken in the interior of the extra dimension. We now proceed in analogy to the "waveguide" picture [163] and restrict this charge violation to ΔQ to one layer at some interior position $x_5 = i$. Using Wilson fermions, the hopping term from $x_5 = i$ to $i + 1$

$$\overline{\psi}_i P \psi_{i+1} \qquad (P = (\gamma_5 + r)/2) \tag{294}$$

is a $Q = 1/3$ operator. At this layer, electric charge is not conserved. This is unacceptable and needs to be fixed.

To restore the $U(1)$ symmetry one must transfer the charge from ψ to the compensating doublet χ. For this we replace the sum of hoppings with a product on the offending layer

$$\overline{\psi}_i P \psi_{i+1} + \overline{\chi}_i P \chi_{i+1} \longrightarrow \overline{\psi}_i P \psi_{i+1} \times \overline{\chi}_i P \chi_{i+1} \tag{295}$$

This introduces an electrically neutral four-Fermi operator. Note that it is baryon violating, involving a "lepto-quark/diquark" exchange, as sketched in Fig. 34. One might think of the operator as representing a "filter" at $x_5 = i$ through which only charge compensating pairs of fermions can pass.

In five dimensions there is no chiral symmetry. Even for the free theory, combinations like $\overline{\psi}_i P \psi_{i+1}$ have non-vanishing vacuum expectation values. We use such as a "tadpole," with χ generating an effective hopping for ψ and *vice versa*.

Actually the above four fermion operator is not quite sufficient for all chiral anomalies, which can also involve right handed singlet fermions. To correct this we need explicitly include the right handed sector, adding similar four fermion couplings (also electrically neutral). The main difference is that this sector does not couple to the weak bosons.

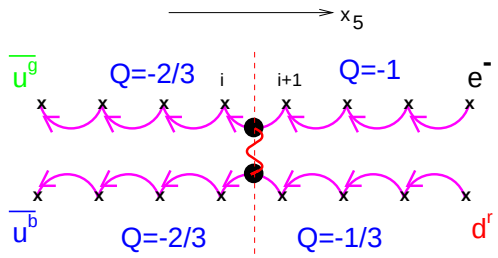

Fig. 34. Transferring charge between the doublets. Figure taken from [161].

Having fixed the $U(1)$ of electromagnetism, we restore the strong $SU(3)$ with an anti-symmetrization of the quark color indices in the new operator, $Q^r Q^g Q^b \longrightarrow \epsilon^{\alpha\beta\gamma} Q^\alpha Q^\beta Q^\gamma$. Note that similar left-right inter-sector couplings are needed to correctly obtain the effects of topologically non-trivial strong gauge fields.

An alternative view of this picture folds the lattice about the interior of the fifth dimension, placing all light modes on one wall and having the multi-fermion operator on the other. This is the model of Ref. [160], with the additional inter-sector couplings correcting a technical error [164].

Unfortunately the scheme is still not completely rigorous. In particular, the non-trivial four-fermion coupling represents a new defect and we need to show that this does not give rise to unwanted extra zero modes. Note, however, that the five dimensional mass is the same on both sides of defect; thus there are no topological reasons for such.

A second worry is that the four fermion coupling might induce an unwanted spontaneous symmetry breaking of one of the gauge symmetries. We need to remain in a strongly coupled paramagnetic phase without spontaneous symmetry breaking. Ref. [160] showed that strongly coupled zero modes do preserve the desired symmetries, but the analysis ignored contributions from heavy modes in the fifth dimension.

Assuming all works as desired, the model raises several other interesting questions. As formulated, we needed a right handed neutrino to provide all quarks with partners. Is there some variation that avoids this particle, which decouples from all gauge fields in the continuum limit? Another question concerns possible numerical simulations; is the effective action positive? Finally, we have used the details of the usual standard model, leaving open the question of whether this model is somehow special. Can we always find an appropriate multi-fermion coupling to eliminate undesired modes in other chiral theories where anomalies are properly cancelled.

13 Final remarks

We have seen how many features of QCD are influenced by non-perturbative physics. This is particularly important to various aspects of chiral symmetry breaking. Taken as a whole, these fit together into a rather elegant and coherent picture. In particular, chiral symmetry is broken in three rather different ways. We have concentrated on the interplay of these mechanisms.

The primary and most important effect is the dynamical symmetry breaking that leads to the pions being light pseudo-Goldstone bosons. Their dynamics represents the most important physics for QCD at low energies. The popular and useful chiral expansion is a natural expansion

in the momenta and masses of these particles.

In addition to the basic dynamical breaking is the anomaly, which eliminates the flavor-singlet axial $U(1)$ symmetry of the classical theory. Thus the η' meson is not a Goldstone boson and acquires a mass of order Λ_{qcd}. Understanding this breaking requires non-perturbative physics associated with the zero modes of the Dirac operator.

Finally, we have the explicit symmetry breaking from the quark masses. This is responsible for the pseudo-scalar mesons not being exactly massless. Using the freedom to redefine fields using chiral rotations, the number of independent mass parameters is $N_f + 1$ where N_f is the number fermion species under consideration. This includes the possibility of CP violation coming from the interplay of the mass term with the anomaly.

Throughout we have used only a few widely accepted assumptions, such as the existence of QCD as a field theory and standard ideas about chiral symmetry. Thus it is perhaps somewhat surprising that several of the conclusions remain controversial. The first of these is that chiral symmetry is lost in a theory with only one light quark. The resulting additive non-perturbative renormalization of the mass precludes using a massless up quark to solve the strong CP problem. Tied to this is the issue of whether topological susceptibility is well defined when non-differentiable fields dominate the path integral. Finally, probably the most bitter controversies revolve about the symmetries inherent in the staggered formulation and how these invalidate the use of rooting to remove unwanted degeneracies.

As simple as the overall picture is, it requires understanding effects that go well beyond perturbation theory. We need aspects of the Dirac spectrum that rely on gauge fields of non-trivial topology. Such appear already in the classical theory, although their true importance only appears in the context of the anomaly. Including this rich physics properly in a lattice formulation is a rich and sometimes controversial topic of active research.

Acknowledgement: This manuscript has been authored by employees of Brookhaven Science Associates, LLC under Contract No. DE-AC02-98CH10886 with the U.S. Department of Energy. The publisher by accepting the manuscript for publication acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

References

[1] M. Gell-Mann: *Phys. Lett.* **8** (1964) 214

[2] S. R. Mishra, F. Sciulli: *Ann. Rev. Nucl. Part. Sci.* **39** (1989) 259

[3] P. D. B. Collins: *An introduction to Regge theory and high-energy physics*: Cambridge (1977)

[4] W. Meissner, R. Ochsenfeld: *Naturwissenschaften* **21** (1933) 787788

[5] A. Abrikosov: *Sov.Phys.JETP* **5** (1957) 1174

[6] H. Hess, R. Robinson, R. Dynes, J. Valles, J. Waszczak: *Phys.Rev.Lett.* **62** (1989) 214

[7] A. Chodos, R. L. Jaffe, K. Johnson, C. B. Thorn, V. F. Weisskopf: *Phys. Rev.* **D9** (1974) 3471

[8] C.-N. Yang, R. L. Mills: *Phys.Rev.* **96** (1954) 191

[9] K. G. Wilson: *Phys. Rev.* **D10** (1974) 2445

[10] K. G. Wilson: *Erice Lectures 1975* (1977): New Phenomena In Subnuclear Physics. Part A. Proceedings of the First Half of the 1975 International School of Subnuclear Physics, Erice, Sicily, July 11 - August 1, 1975, ed. A. Zichichi, Plenum Press, New York, 1977, p. 69, CLNS-321

[11] F. Dyson: *Phys.Rev.* **85** (1952) 631

[12] L. Landau, I. Pomeranchuk: *Dokl.Akad.Nauk Ser.Fiz.* **102** (1955) 489: Also published in Collected Papers of L.D. Landua. Edited by D. Ter Haar. Pergamon Press, 1965. pp. 654-658

[13] S. R. Coleman: *Subnucl.Ser.* **13** (1977) 297

[14] J. S. Schwinger: *Phys.Rev.* **128** (1962) 2425

[15] S. R. Coleman: *Phys.Rev.* **D11** (1975) 2088

[16] W. E. Thirring: *Annals Phys.* **3** (1958) 91

[17] C. Itzykson, J. Drouffe: *Statistical field theory. Vol. 1: From Brownian motion to renormalization and lattice gauge theory*: Cambridge (1989)

[18] R. F. Dashen: *Phys.Rev.* **D3** (1971) 1879

[19] G. 't Hooft: *Phys.Rev.* **D14** (1976) 3432

[20] E. Witten: *Annals Phys.* **128** (1980) 363

[21] C. Rosenzweig, J. Schechter, C. Trahern: *Phys.Rev.* **D21** (1980) 3388

[22] R. L. Arnowitt, P. Nath: *Nucl.Phys.* **B209** (1982) 234

[23] P. Nath, R. L. Arnowitt: *Nucl.Phys.* **B209** (1982) 251

[24] D. Boer, J. K. Boomsma: *Phys.Rev.* **D78** (2008) 054027

[25] M. Creutz: *Phys. Rev.* **D52** (1995) 2951

[26] M. Creutz: *Phys. Rev. Lett.* **92** (2004) 162003

[27] M. Creutz: *PoS* **CONFINEMENT8** (2008) 016

[28] T. Banks, A. Zaks: *Nucl.Phys.* **B196** (1982) 189

[29] M. Creutz: *Phys. Rev.* **D15** (1977) 1128

[30] R. Feynman: *Rev.Mod.Phys.* **20** (1948) 367

[31] R. P. Feynman, A. R. Hibbs: *Quantum mechanics and path integrals*: McGraw Hill (1965)

[32] M. Creutz, B. Freedman: *Ann. Phys.* **132** (1981) 427

[33] M. Creutz: *Phys. Rev. Lett.* **81** (1998) 3555

[34] M. Creutz: *Found. Phys.* **30** (2000) 487

[35] M. Creutz: *Quarks, gluons and lattices*: Cambridge, Uk: Univ. Pr. (1983): Cambridge, Uk: Univ. Pr. (1983) 169 P. (Cambridge Monographs On Mathematical Physics)

[36] C. Gattringer, C. B. Lang: *Lect.Notes Phys.* **788** (2010) 1

[37] I. Montvay, G. Munster: *Quantum fields on a lattice*: Cambridge, UK: Univ. Pr. (1994)

[38] T. DeGrand, C. E. Detar: *Lattice methods for quantum chromodynamics*: World Scientific (2006)

[39] H. Rothe: *World Sci.Lect.Notes Phys.* **74** (2005) 1

[40] S. Elitzur: *Phys.Rev.* **D12** (1975) 3978

[41] L. Faddeev, V. Popov: *Phys.Lett.* **B25** (1967) 29

[42] I. Singer: *Commun.Math.Phys.* **60** (1978) 7

[43] M. Creutz: *Phys. Rev.* **D21** (1980) 2308

[44] M. Creutz, L. Jacobs, C. Rebbi: *Phys. Rev.* **D20** (1979) 1915

[45] H. Politzer: *Phys.Rev.Lett.* **30** (1973) 1346

[46] D. Gross, F. Wilczek: *Phys.Rev.Lett.* **30** (1973) 1343

[47] D. Gross, F. Wilczek: *Phys.Rev.* **D8** (1973) 3633
[48] W. E. Caswell: *Phys.Rev.Lett.* **33** (1974) 244
[49] D. Jones: *Nucl.Phys.* **B75** (1974) 531
[50] S. R. Coleman, E. J. Weinberg: *Phys.Rev.* **D7** (1973) 1888
[51] J. Vermaseren, S. Larin, T. van Ritbergen: *Phys.Lett.* **B405** (1997) 327
[52] K. G. Chetyrkin, M. Misiak, M. Munz: *Nucl.Phys.* **B518** (1998) 473
[53] A. Hasenfratz, P. Hasenfratz: *Phys.Lett.* **B93** (1980) 165
[54] A. Hasenfratz, P. Hasenfratz: *Nucl.Phys.* **B193** (1981) 210
[55] R. F. Dashen, D. J. Gross: *Phys.Rev.* **D23** (1981) 2340
[56] K. Wilson, J. B. Kogut: *Phys.Rept.* **12** (1974) 75
[57] R. Rajaraman: *Solitons and instantons. An introduction to solitons and instantons in quantum field theory*: North-holland (1982)
[58] E. Vicari, H. Panagopoulos: *Phys.Rept.* **470** (2009) 93: * Temporary entry *
[59] J. Goldstone, A. Salam, S. Weinberg: *Phys.Rev.* **127** (1962) 965
[60] M. Creutz: *Annals Phys.* **324** (2009) 1573
[61] J. D. Bjorken: *Phys. Rev.* **163** (1967) 1767
[62] J. D. Bjorken: *Phys. Rev.* **148** (1966) 1467
[63] S. L. Adler: *Phys.Rev.* **177** (1969) 2426
[64] S. L. Adler, W. A. Bardeen: *Phys.Rev.* **182** (1969) 1517
[65] J. Bell, R. Jackiw: *Nuovo Cim.* **A60** (1969) 47
[66] K. Fujikawa: *Phys.Rev.Lett.* **42** (1979) 1195
[67] S. R. Coleman: *Subnucl.Ser.* **15** (1979) 805
[68] P. H. Ginsparg, K. G. Wilson: *Phys.Rev.* **D25** (1982) 2649
[69] H. Neuberger: *Phys.Lett.* **B417** (1998) 141
[70] T. Banks, A. Casher: *Nucl.Phys.* **B169** (1980) 103: Revised Version
[71] E. Witten: *Nucl.Phys.* **B202** (1982) 253
[72] E. Corrigan, P. Ramond: *Phys.Lett.* **B87** (1979) 73
[73] A. Armoni, M. Shifman, G. Veneziano: *Phys.Rev.Lett.* **91** (2003) 191601
[74] F. Sannino, M. Shifman: *Phys.Rev.* **D69** (2004) 125004
[75] M. Unsal, L. G. Yaffe: *Phys.Rev.* **D74** (2006) 105019
[76] H. Georgi, D. V. Nanopoulos: *Nucl.Phys.* **B155** (1979) 52
[77] C. Vafa, E. Witten: *Phys.Rev.Lett.* **53** (1984) 535
[78] M. Creutz: *Annals Phys.* **322** (2007) 1518
[79] M. Creutz: *Annals Phys.* **323** (2008) 2349
[80] M. Creutz: *Phys. Rev.* **D70** (2004) 091501
[81] M. Creutz (2010)
[82] H. Georgi, I. N. McArthur: *unpublished* (1981)
[83] T. Banks, Y. Nir, N. Seiberg: *unpublished* (1994)
[84] M. Creutz: *unpublished* (2003)
[85] R. Frezzotti, P. A. Grassi, S. Sint, P. Weisz: *JHEP* **0108** (2001) 058
[86] P. Boucaud, *et al.*: *Phys.Lett.* **B650** (2007) 304
[87] V. Baluni: *Phys.Rev.* **D19** (1979) 2227

[88] M. Creutz: *PoS* **LAT2005** (2006) 119
[89] S. Aoki: *Phys.Rev.* **D30** (1984) 2653
[90] H. B. Nielsen, M. Ninomiya: *Nucl.Phys.* **B185** (1981) 20
[91] M. Creutz: *Phys. Rev.* **D76** (2007) 054501
[92] M. Creutz (1996)
[93] S. R. Sharpe, J. Singleton, Robert L.: *Phys.Rev.* **D58** (1998) 074501
[94] P. Damgaard, K. Splittorff, J. Verbaarschot: *Phys.Rev.Lett.* **105** (2010) 162002
[95] R. Frezzotti, G. Rossi: *JHEP* **0408** (2004) 007
[96] G. Munster, C. Schmidt, E. E. Scholz: *Nucl.Phys.Proc.Suppl.* **140** (2005) 320
[97] R. Narayanan, H. Neuberger: *Nucl.Phys.* **B443** (1995) 305
[98] D. B. Kaplan: *Phys. Lett.* **B288** (1992) 342
[99] V. Furman, Y. Shamir: *Nucl.Phys.* **B439** (1995) 54
[100] M. Creutz: *Phys. Lett.* **B649** (2007) 230
[101] M. Creutz: *JHEP* **04** (2008) 017
[102] T. Misumi, M. Creutz, T. Kimura (2010)
[103] L. H. Karsten: *Phys.Lett.* **B104** (1981) 315
[104] F. Wilczek: *Phys.Rev.Lett.* **59** (1987) 2397
[105] A. Borici: *Phys.Rev.* **D78** (2008) 074504
[106] P. F. Bedaque, M. I. Buchoff, B. C. Tiburzi, A. Walker-Loud: *Phys.Rev.* **D78** (2008) 017502
[107] T. Kimura, T. Misumi: *Prog.Theor.Phys.* **123** (2010) 63
[108] M. Creutz, T. Misumi: *Phys. Rev.* **D82** (2010) 074502
[109] M. Creutz: *PoS* **LAT2007** (2007) 007
[110] M. Creutz: *PoS* **QCD-TNT09** (2009) 008
[111] S. Capitani, M. Creutz, J. Weber, H. Wittig: *JHEP* **09** (2010) 027
[112] B. C. Tiburzi: *Phys.Rev.* **D82** (2010) 034511
[113] R. Narayanan, H. Neuberger: *Nucl. Phys.* **B412** (1994) 574
[114] K. Jansen: *Phys. Rept.* **273** (1996) 1
[115] M. Creutz, I. Horvath: *Phys. Rev.* **D50** (1994) 2297
[116] Y. Shamir: *Nucl. Phys.* **B406** (1993) 90
[117] R. Narayanan, H. Neuberger: *Phys. Lett.* **B302** (1993) 62
[118] H. Neuberger: *Phys.Rev.Lett.* **81** (1998) 4060
[119] H. Neuberger: *Phys.Lett.* **B427** (1998) 353
[120] T.-W. Chiu, S. V. Zenkin: *Phys.Rev.* **D59** (1999) 074501
[121] S. Chandrasekharan: *Phys.Rev.* **D60** (1999) 074503
[122] P. Hasenfratz, V. Laliena, F. Niedermayer: *Phys.Lett.* **B427** (1998) 125
[123] P. Hasenfratz: *Nucl.Phys.* **B525** (1998) 401
[124] M. Creutz: *Nucl. Phys. Proc. Suppl.* **119** (2003) 837
[125] M. Luscher: *Phys.Lett.* **B428** (1998) 342
[126] H. Neuberger: *Chin.J.Phys.* **38** (2000) 533
[127] P. Hernandez, K. Jansen, M. Luscher: *Nucl.Phys.* **B552** (1999) 363
[128] I. Horvath: *Phys.Rev.Lett.* **81** (1998) 4063
[129] I. Horvath: *Phys.Rev.* **D60** (1999) 034510

[130] R. G. Edwards, U. M. Heller, R. Narayanan: *Phys.Rev.* **D59** (1999) 094510
[131] A. Borici: *J.Comput.Phys.* **162** (2000) 123
[132] S. Dong, F. Lee, K. Liu, J. Zhang: *Phys.Rev.Lett.* **85** (2000) 5051
[133] C. Gattringer: *Phys.Rev.* **D63** (2001) 114501
[134] M. Luscher: *JHEP* **0006** (2000) 028
[135] Y. Kikukawa, A. Yamada: *Phys.Lett.* **B448** (1999) 265
[136] G. 't Hooft: *NATO Adv.Study Inst.Ser.B Phys.* **59** (1980) 135
[137] R. G. Edwards, U. M. Heller, R. Narayanan: *Phys.Lett.* **B438** (1998) 96
[138] L. H. Karsten, J. Smit: *Nucl.Phys.* **B183** (1981) 103
[139] J. B. Kogut, L. Susskind: *Phys.Rev.* **D11** (1975) 395
[140] L. Susskind: *Phys.Rev.* **D16** (1977) 3031
[141] H. Sharatchandra, H. Thun, P. Weisz: *Nucl.Phys.* **B192** (1981) 205
[142] Y. Aoki, Z. Fodor, S. Katz, K. Szabo: *Phys.Lett.* **B643** (2006) 46
[143] M. Cheng, N. Christ, S. Datta, J. van der Heide, C. Jung, *et al.*: *Phys.Rev.* **D77** (2008) 014511
[144] A. Bazavov, D. Toussaint, C. Bernard, J. Laiho, C. DeTar, *et al.*: *Rev.Mod.Phys.* **82** (2010) 1349
[145] M. Creutz (2010)
[146] M. Luscher: *Commun.Math.Phys.* **85** (1982) 39
[147] F. Bruckmann, C. Gattringer, E.-M. Ilgenfritz, M. Muller-Preussker, A. Schafer, *et al.*: *Eur.Phys.J.* **A33** (2007) 333
[148] M. Luscher: *JHEP* **1008** (2010) 071
[149] M. Teper: *Phys.Lett.* **B162** (1985) 357
[150] F. Bruckmann, F. Gruber, K. Jansen, M. Marinkovic, C. Urbach, *et al.*: *Eur.Phys.J.* **A43** (2010) 303
[151] P. J. Moran, D. B. Leinweber, J. Zhang (2010)
[152] M. Luscher: *Phys.Lett.* **B593** (2004) 296
[153] L. Giusti, G. Rossi, M. Testa: *Phys.Lett.* **B587** (2004) 157
[154] M. Luscher, F. Palombi: *JHEP* **1009** (2010) 110
[155] E. Witten: *Nucl.Phys.* **B156** (1979) 269
[156] G. Veneziano: *Nucl.Phys.* **B159** (1979) 213
[157] E. Eichten, J. Preskill: *Nucl. Phys.* **B268** (1986) 179
[158] E. Poppitz, Y. Shang: *Int.J.Mod.Phys.* **A25** (2010) 2761
[159] I. Montvay: *Phys. Lett.* **B199** (1987) 89
[160] M. Creutz, M. Tytgat, C. Rebbi, S.-S. Xue: *Phys. Lett.* **B402** (1997) 341
[161] M. Creutz: *Nucl. Phys. Proc. Suppl.* **63** (1998) 599
[162] J. Ambjorn, J. Greensite, C. Peterson: *Nucl.Phys.* **B221** (1983) 381
[163] M. F. Golterman, K. Jansen, D. N. Petcher, J. C. Vink: *Phys.Rev.* **D49** (1994) 1606
[164] H. Neuberger: *Phys.Lett.* **B413** (1997) 387