Lattice Fermions

Frank Wilczek
Institute for Theoretical Physics, University of California, Santa Barbara, Santa Barbara, California 93106
(Received 9 April 1987)

A simple heuristic proof of the Nielsen-Ninomaya theorem is given. A method is proposed whereby the multiplication of fermion species on a lattice is reduced to the minimal doubling, in any dimension, with retention of appropriate chiral symmetries. Also, it is suggested that use of spatially thinned fermion fields is likely to be a useful and appropriate approximation in QCD—in any case, it is a self-checking one.

PACS numbers: 11.15.Ha, 11.30.Rd, 12.38.Ge

The most powerful methods for fully nonperturbative calculation in field theories, notably including QCD,[1] the theory of the strong interaction, involve discretizing the theory on a space-time lattice.[2] While many impressive results have been attained, such as numerical "proofs" of the existence of the theory and the confinement property,[3] progress on many other important questions has been impeded by technical and computational difficulties which arise when fermions are included.[4,5]

The technical difficulties I have in mind are the seeming necessity either to replicate fermion fields[6]—this is commonly called the doubling problem, although generally a higher power of 2 appears—or to break chiral symmetry explicitly at intermediate stages,[7] hoping to recover it in the continuum limit. The former approach, strictly speaking, does not even allow QCD with a fully realistic spectrum of quark masses to be formulated, while the latter is very demanding computationally. In this Letter, a method is proposed which promises to alleviate these technical problems significantly.

The computational problems have to do with the awkwardness of treating particles obeying Fermi statistics by Monte Carlo methods—the action is not given in a manifestly local, positive definite form. No general remedy to this problem is suggested here, but I will make some specific suggestions for easing the computational burden in QCD, with or without a θ term.

Although the methods advocated in Refs. 6 and 7 seem to be the ones most used in practice for latticizing fermions, there are several other proposals in the literature.[8-11] Some of these are nonlocal (Refs. 8 and 9). Nonlocal formulations raise the spectra of nonlocal counterterms, which seem very difficult to control. The others are sufficiently complex that they have not been put to practical test.

It should be mentioned that for nonrelativistic fermions, the doubling problem does not really arise.[12]

The doubling problem is best understood by focusing on zeros of the inverse propagator (poles of the propagator) in momentum space. On a lattice one cannot realize derivative operators $p_\mu = -i\,\partial/\partial x_\mu$ directly; rather one works with translation operators such as $\exp(ip_\mu a)$, where a is the lattice spacing, and the algebra generated by them. (In this Letter I consider only cubic lattices.) The continuum derivative is supposed to be obtained by taking the limit $a \to 0$ of $(1/a)\sin p_\mu a$.

The doubling problem in one dimension is very simple to appreciate.[4] Namely, the lattice inverse propagator, or single-particle Lagrangean

$$\mathcal{L} = \gamma_1 a^{-1} \sin p_1 a \text{ or } \gamma_1 f(p_1) \tag{1}$$

has a zero not only at $p_1 = 0$ but also at $p_1 = \pi/a$. (Note that p_1 is periodic in $2\pi/a$.) Thus one has generated a second, undesired fermion degree of freedom, which survives in the continuum limit. Can this extra fermion be eliminated by use of some other function in place of sine? A simple topological argument shows it cannot be: Since the Lagrangean is periodic, if it crosses zero at one point, say with positive derivative, it must cross zero at another point with negative derivative in order to return. It is possible to eliminate the second zero by adding a term proportional to the unit matrix, e.g.,

$$\mathcal{L} = \gamma_1 a^{-1} \sin p_1 a + \kappa(\sin^2 p_1 a)/a. \tag{2}$$

This is the essence of Wilson's procedure.[7] This modification of $\mathcal{L}$ (or, more precisely, its analog in higher dimensions) has the unfortunate feature of destroying chiral symmetry, and unless κ is tuned to a critical value the fermion will generally acquire a mass of order the "natural" size, i.e., $1/a$, so that passage to the continuum limit is quite delicate.

In two dimensions slightly more sophisticated arguments are necessary. Let us consider the general chiral-invariant Lagrangean

$$\mathcal{L} = \gamma_1 f_1 + \gamma_2 f_2. \tag{3}$$

Here f_1 and f_2 are functions on a torus, i.e., periodic in p_1 and p_2 with period $2\pi/a$. Again I focus on the loci of the zeros of f_1 and f_2. The intersections of these curves locate the poles of the propagator. The straightforward choice $f_1 = (1/a)\sin p_1 a$, $f_2 = (1/a)\sin p_2 a$ leads to the situation shown in Fig. 1. There are four intersections. In D dimensions, there will be 2^D intersections. A more

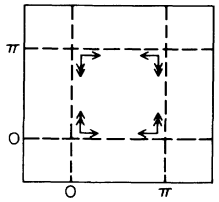

FIG. 1. Loci of zeros in the coefficients of γ matrices for the standard lattice Lagrangean.

general choice might lead to a closed curve for f_1, separating the positive and negative regions as in Fig. 2. But even in this case there must be at least two intersections: The locus of zeros of f_2 cannot end, so that if it enters the region enclosed by the zeros of f_1 it must leave it again. The discreteness of the lattice enters the argument here, in making p space compact, so that the 0 locus cannot run off to infinity. So doubling—in the sense of having poles at two momenta—is unavoidable.

Similar arguments can be made in four dimensions. The intersection of the zeros of f_1, f_2, f_3 (defined in the obvious way) must leave the region enclosed by zeros of f_4 if it enters. Although the argument for doubling presented here is heuristic, I believe it could be made rigorous without great difficulty. It is closely related to the "heuristic roof" of Nielsen and Ninomiya.[13]

It remains to discuss the chirality of the fermions. For definiteness, suppose we started with four-component fermions in 4D. Given a chiral invariant Lagrangean, i.e., of the form

$$\mathcal{L}=\sum_i \gamma_i f_i(p), \tag{4}$$

the fermion fields $\frac{1}{2}(1-\gamma_5)\psi$ and $\bar{\psi}\frac{1}{2}(1+\gamma_5)$ are independent of the other projections; so we may restrict the theory to this set. It might appear then that we have produced a theory with exclusively left-handed particles. This appearance is deceptive, however, for the following reason. The numerical matrix "γ_5" is defined to be the product $\gamma_1\gamma_2\gamma_3\gamma_4$, in this order. Now near different poles the coefficients of the numerical matrices $\gamma_1, \ldots, \gamma_4$ may be associated not with the momenta $p_1, \ldots, p_4$, but with a scrambled set. In other words, the gradients of $f_1, \ldots, f_4$ need not point in the positive $x_1, \ldots, x_4$ directions. A linear transformation of ψ may be necessary to bring the residue at the pole into canonical form. If the linear transformation bringing the scrambled set into canonical form reverses orientation, then the *dynamical* meaning of the *numerical* matrix γ_5 will be reversed: the projection $\frac{1}{2}(1-\gamma_5)\psi$ will correspond to right-handed particles.

Now return to the figures. In each case, the directions of the gradients of f_1 and f_2 are indicated with one- and two-headed arrows, respectively. In Fig. 1, the intersections at the lower left and upper right corners have one orientation, the other two intersections the other. So

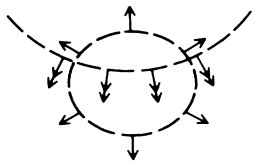

FIG. 2. Loci of zeros in the coefficients, in the generic case.

after projecting with $\frac{1}{2}(1-\gamma_5)$ we will find two left-handed and two right-handed fermions. Similarly, in Fig. 2 we will find one of each. It is clear, intuitively, that successive intersections (following a curve of zeros) will always result in opposite chirality. And so, within this framework one always finds equal numbers of left- and right-handed fermions. When the theory is gauged, it is vectorlike and no anomalies arise. Indeed, it would present a bit of a puzzle if anomalous theories could be regulated in any straightforward way on a lattice, because there would be no ultraviolet divergences to interfere with the formal proof of Ward identities.

The discussion above suggests that while doubling—in the sense of two poles—is unavoidable for chiral Lagrangeans on the lattice, further multiplication of species might be avoided by a judicious choice of f_j. In particular, if one of the loci of zeros is a narrow ellipsoid and the rest are the standard pairs of straight lines, then only two intersections will occur. A specific form which works is the following:

$$f_4=a^{-1}\{\sin p_4 a+\lambda[\sin^2(\tfrac{1}{2}p_1 a)+\sin^2(\tfrac{1}{2}p_2 a)+\sin^2(\tfrac{1}{2}p_3 a)]\}, \tag{5}$$

$$f_j=a^{-1}\sin p_j a, \quad j=1,2,3,$$

where $\lambda>1$. For all the f_j to be zero, we must have $p_j=0$ or π and thus $\sin^2(\frac{1}{2}p_j)=0$ or 1, respectively. But if any $\sin^2(\frac{1}{2}p_j)=1$, then f_4 cannot vanish. Thus the only two poles are at $p=(0,0,0,0)$, $(0,0,0,\pi)$.

Projecting out with $\frac{1}{2}(1-\gamma_5)$, we find a left-handed fermion at the first pole and a right-handed fermion at the second. So we have the degrees of freedom of a single Dirac fermion in the continuum limit.

The proposed form of the Lagrangean destroys the equivalence of all four directions under discrete permutations. This introduces a new relevant operator in the continuum limit, i.e., $F_{0i}F_{0i}$. (In the theory including fermions, additional new relevant operators appear.) So to obtain the correct limit, the lattice version of this will have to be included with the correct coefficient. This delicacy is similar to, but much less severe than, the one mentioned before for Wilson fermions. There is no $1/a$ factor, and the coefficient vanishes in a calculable way for weak coupling.

Since the poles are at distinct points in momentum space, they cannot be connected to one another by a

translation-invariant bilinear mass term. Thus our fermion is massless. Translation through one lattice spacing in the x_4 direction becomes a discrete chiral symmetry in the continuum limit, since it multiplies the left-handed fermion by $+1$ and the right-handed fermion by -1.

Note that the mass term is forbidden by a discrete chiral symmetry. There is no global symmetry in the lattice-regulated theory; indeed if there were, we would have difficulties with the Sutherland-Veltman theorem.

How can we add a mass term? We can isolate the smooth effective fields which become left- and right-handed fermions in the continuum limit by taking appropriate combinations within unit cells, i.e., the two points $(\mathbf{x},na),(\mathbf{x},(n+1)a)$, where $\mathbf{x}$ is arbitrary and n even. The left-handed field is

$$\phi_L(n)=\tfrac{1}{2}(1-\gamma_5)\{\psi(na)+\psi((n+1)a)\}, \tag{6a}$$

and the right-handed field is

$$\phi_R(n)=\gamma_4\gamma_5\tfrac{1}{2}(1-\gamma_5)\{\psi(na)-\psi((n+1)a)\}, \tag{6b}$$

where the unitary transformation necessary to reverse the sign of γ_4, and thus bring the kinetic energy expanded around $p=(0,0,0,\pi)$ into canonical form, has been displayed—and the dependence on $\mathbf{x}$ suppressed.[14] The mass term is then

$$\Delta\mathcal{L}_{m,\theta}=\sum_{n\,\text{even}} e^{i\theta}\bar{\phi}_R(na)\phi_L(na)+\text{H.c.} \tag{7}$$

The possibility of implementing a complex mass is interesting from several points of view. It gives us a method of implementing the P,T-nonconserving θ interaction of QCD, bypassing the rather awkward procedure for doing the same thing in the gluon sector. For example, periodicity in $\theta\to\theta+2\pi$ is entirely trivial to see with fermions, but awkward for gluons, even though both implementations are supposed to lead to the same physics in the continuum limit. It is interesting to study QCD at nonzero θ, not only to test various speculations on QCD dynamics, but also to determine the behavior of the axion mass at high temperature, which is significant for cosmology.[15]

There is also a more subtle advantage to implementing the θ parameter as a complex fermion mass. In the absence of fermions, the dependence of contributions to physical quantities, e.g., the vacuum energy, from different configurations can have violent fluctuations. Specifically, for $\theta=\pi$ addition of just one instanton changes the sign, and so if the instanton density is high the net result will be the result of subtracting two large quantities, which is highly undesirable computationally. In the presence of fermions, the dependence on θ vanishes as $m\to 0$, and can be adjusted to match computational resources.

Even the pure-gluon result, if desired, can be obtained by consideration of a single very massive fermion. As its mass goes to infinity, it decouples, leaving behind only a renormalization of the coupling and the θ parameter.

In conclusion I would like to mention another idea, only tangentially related to the preceding, for speeding up calculations of QCD including fermions.

The standard method for putting QCD on a lattice involves putting the quark fields at lattice sites and gauge fields on the links. In the calculations, much more time is spent calculating the effects of the quarks than the effects of the gauge fields. This is because the measure for gauge fields (i.e., the action) is given in a simple positive-definite local form while no such representation for the measure of the quark fields is known—nor is one likely to exist.

Since the fermion degrees of freedom are much more difficult to deal with than the bosons, it seems good strategy to thin them out as much as possible. And so the following sort of idea suggests itself: Instead of putting fermion fields at every site, let them only be defined at every other site. Of course many variants are possible: One could use every other site in the $\hat{\mathbf{x}}$ direction but all sites in the others, every third site, and so forth. In fact it may be very useful that this flexibility exists—by comparing the results which emerge for different choices, the accuracy of the scheme can be assessed. One would start with some very coarse spacing, and reduce it until successive reductions have no effect on the result, to the desired accuracy.

It is simple to see that this proposal can be implemented in a gauge-invariant way. In a way it is a generalization of the old idea of staggering fermions—pushing that thought further in a systematic way.

Is there reason to think that treating quarks on a coarser spatial scale than gluons is appropriate? Several physical considerations, while far from rigorous, suggest that it may be. First of all, the numerical value of perturbative renormalization-group parameters—the β function and other anomalous dimensions—are typically dominated by gluon contributions. This indicates that the short-distance behavior of QCD is dominated by the gluons, and so it seems appropriate to treat them more carefully at short distances. Related to this is the fact that the glueball spectrum, while sparse at low energies ($\lesssim 2$ GeV), is expected to become extremely dense at higher energies. Plausibly, it is most important to include the quarks at long wavelengths, where the gluons are frozen out by the large glueball mass, but less important to treat them accurately at short wavelengths, where in any case they are dominated by the gluons.

Thanks to V. Nair for early discussions stimulating this work and to U. Heller, C. Pryor, and R. Sugar for helpful comments. This research was supported in part by the National Science Foundation under Grant No. PHY82-17853, supplemented by funds from the National Aeronautics and Space Administration, at the University of California at Santa Barbara.

[1]F. Wilczek, Annu. Rev. Nucl. Sci. **32**, 177 (1982).

[2]A. Hasenfratz and P. Hasenfratz, Annu. Rev. Nucl. Sci. **35**, 559 (1985); M. Creutz, *Quarks, Gluons and Lattices* (Cambridge Univ. Press, New York, 1983).

[3]M. Creutz, Phys. Rev. Lett. **43**, 553 (1979).

[4]L. Karsten and J. Smit, Nucl. Phys. **B192**, 100 (1981).

[5]H. Nielsen and M. Ninomiya, Nucl. Phys. **B185**, 20 (1981).

[6]L. Susskind, Phys. Rev. D **16**, 3031 (1977).

[7]K. Wilson, in *New Phenomena in Subnuclear Physics,* edited by A. Zichichi (Plenum, New York, 1977).

[8]S. Drell, M. Weinstein, and S. Yankelowicz, Phys. Rev. D **14**, 1627 (1976).

[9]C. Rebbi, Boston University Report No. BUHEP 86-6, 1986 (to be published).

[10]L. Jacobs, Phys. Rev. Lett. **51**, 172 (1986).

[11]E. Eichten and J. Preskill, Nucl. Phys. **B268**, 179 (1986).

[12]D. Scalapino and R. Sugar, Phys. Rev. Lett. **46**, 519 (1981).

[13]H. Nielsen and M. Ninomiya, Nucl. Phys. **B193**, 173 (1981); K. Baxter and C. Rebbi, in *Proceedings of the Twenty-Sixth Scottish Universities Summer School in Physics, Edinburgh, 1983,* edited by K. C. Bowler and A. J. McKane (SUSSP, Edinburgh, 1984).

[14]I have checked that in the Schwinger model, the naive axial-vector current constructed from these left- and right-handed fermion fields obeys the anomaly equation in the continuum limit.

[15]J. Preskill, M. Wise, and F. Wilczek, Phys. Lett. **120B**, 127 (1983); L. Abbott and P. Sikivie, Phys. Lett. **120B**, 133 (1983); M. Dine and W. Fischler, Phys. Lett. **120B**, 137 (1983).