Lecture Notes
in Computational Science
and Engineering

47

Artan Boriçi · Andreas Frommer
Bálint Joó · Anthony Kennedy
Brian Pendleton
Editors

QCD and Numerical Analysis III

Proceedings of the Third International Workshop on Numerical Analysis and Lattice QCD, Edinburgh, June–July 2003

With 29 Figures and 15 Tables

Springer

Editors

Artan Boriçi, Bálint Joó, Anthony Kennedy, Brian Pendleton
University of Edinburgh
School of Physics
Mayfield Road
EH9 3JZ Edinburgh, United Kingdom
email: borici, bj, tony.kennedy, b.pendleton@ph.ed.ac.uk

Andreas Frommer
Bergische Universität Wuppertal
Fachbereich C – Mathematik und Naturwissenschaften
Gaußstr. 20
42097 Wuppertal, Germany
email: frommer@math.uni-wuppertal.de

Library of Congress Control Number: 2005926502

Mathematics Subject Classification (2000): 82B80, 81T80, 65F10, 65F30, 65F15

ISSN 1439-7358
ISBN-10 3-540-21257-4 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-21257-7 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media
springeronline.com

Printed in The Netherlands

Typesetting: by the authors using a Springer TEX macro package
Cover design: *design & production*, Heidelberg

Printed on acid-free paper SPIN: 10966556 41/TechBooks - 5 4 3 2 1 0

Preface

The Third International Workshop on Numerical Analysis and Lattice QCD took place at the University of Edinburgh from June 30th to July 4th, 2003. It continued a sequence which started in 1995 at the University of Kentucky and continued in 1999 with a workshop at the University of Wuppertal. The aim of these workshops is to bring together applied mathematicians and theoretical physicists to stimulate the exchange of ideas between leading experts in the fields of lattice QCD and numerical analysis. Indeed, the last ten years have seen quite a substantial increase in cooperation between the two scientific communities, and particularly so between numerical linear algebra and lattice QCD.

The workshop was organised jointly by the University of Edinburgh and the UK National e-Science Centre. It promoted scientific progress in lattice QCD as an e-Science activity that encourages close collaboration between the core sciences of physics, mathematics, and computer science.

In order to achieve more realistic computations in lattice quantum field theory substantial progress is required in the exploitation of numerical methods. Recently, there has been much progress in the formulation of lattice chiral symmetry satisfying the Ginsparg–Wilson relation. Methods for implementing such chiral fermions efficiently were the principal subject of this meeting, which, in addition, featured several tutorial talks aiming at introducing the important concepts of one field to colleagues from the other. These proceedings reflect this, being organised in three parts: part I contains introductory survey papers, whereas parts II and III contain latest research results in lattice QCD and in computational methods.

Part I starts with a survey paper by Neuberger on lattice chiral symmetry: it reviews the important mathematical properties and concepts, and related numerical challenges. This article is followed by a contribution of Davies and Higham on numerical techniques for evaluating matrix functions, the matrix sign function being the common link between these two first articles. Then, Boriçi reviews the state-of-the-art for computing the fermion determinant with a focus on Krylov methods. He also shows that another version of fermions

respecting chiral symmetry, so-called domain wall fermions, is very closely related to overlap fermions when it comes to numerical computations. Finally, Peardon addresses aspects of stochastic processes and molecular dynamics in QCD simulations. In particular, he reviews the Hybrid Monte Carlo method (HMC), the work-horse of lattice QCD simulations.

Part II starts with a contribution by Boriçi on statistical aspects of the computation of the quark determinant: he suggests using order-statistics estimators rather than noisy methods to eliminate bias, and illustrates this with results for the Schwinger model. The paper by de Forcrand and Jahn studies Monte Carlo for $SU(N)$ Young-Mills theories: instead of the usual approach of accumulating $SU(2)$ updates, they perform overrelaxation in full $SU(N)$ and show that this approach is more efficient in practical simulations. In the next article, Follana considers various improvements of classical staggered fermions: for the pion spectrum he shows that undesirable doublers at light quark masses can indeed be avoided by such improvements. Drummond *et al.* also consider improved gauge actions, now using twisted boundary conditions as an infrared regulator: as they show, the resulting two-loop Landau-mean-links accurately describe high-β Monte Carlo simulations. The contribution by Joó is devoted to the study of potential instabilities in Hybrid Monte Carlo simulations: a theoretical study is presented for the simple harmonic oscillator; implications for (light quark) QCD simulations are discussed and illustrated by numerical experiments. The paper by Liu discusses a canonical ensemble approach to finite baryon density algorithms: several stochastic techniques are required there, including a new Hybrid Noisy Monte Carlo algorithm to reduce large fluctuations. The article by Young, Leinweber and Thomas presents finite-range regularized effective field theory as an efficient tool to study the quark mass variation of QCD observables: this includes regularisation schemes and extrapolation methods for the nucleon mass about the chiral limit.

Part III starts with a paper by Ashby, Kennedy and O'Boyle on a new software package implementing Krylov subspace solvers in a modular manner: the main idea is to gain flexibility and portability by separating the generation of the basis from the actual computation of the iterates. The paper by van den Eshof, Sleijpen and van Gijzen analyses Krylov subspace methods in the presence of only inexact matrix vector products; as an important practical consequence, they are able to devise strategies on how to tune the accuracy requirements yielding an overall fastest method, recursive preconditioning being a major ingredient. Fleming addresses data analysis and modelling for data resulting from lattice QCD calculations: he shows that the field might highly profit from elaborate techniques used elsewhere, like Baysian methods, constrained fitting or total least squares. The paper by Arnold *et al.* compares various Krylov subspace methods for different formulations of the overlap operator; a less known method (SUMR), having had no practical applications so far, turns out to be extremely relevant here. In the next article, Kennedy discusses theoretical and computational aspects of the Zolotarev approximation. This is a closed formula L_∞ best rational approximation to the sign function

on two intervals left and right of zero, and its efficient matrix evaluation is of crucial importance in simulations of overlap fermions. Finally, Wenger uses a continued fraction expansion of the sign function to show that overlap fermions are intimately related to the 5-dimensional formulation of lattice chiral symmetry: based on this he shows he that equivalence transformations can be used to make the operators involved better conditioned.

We would like to express our gratitude to the authors of the present volume for their effort in writing their contribution. All papers have undergone a strict refereeing process and we would like to extend our thanks to all the referees for their thorough reviewing. AF and ADK gratefully acknowledge support by the Kavli Institute of Theoretical Physics (KITP), Santa Barbara (supported in part by the National Science Foundation under Grant No. PHY99–07949).

The book cover shows a QCD-simulation of quark confinement, a result from a simulation run at Wuppertal University. We are very thankful to Thomas Lippert and Klaus Schilling for providing us with the picture. Special thanks are also due to the LNCSE series editors and to Thanh-Ha Le Thi from Springer for the always very pleasant and efficient cooperation during the preparation of this volume.

Edinburgh, Santa Barbara, Pahrump, Tirana
March 2005

Artan Boriçi
Andreas Frommer
Bálint Joó
Anthony D. Kennedy
Brian Pendleton

Contents

List of Contributors

G. Arnold
Fachbereich Mathematik und Naturwissenschaften, Universität Wuppertal, D-42097 Wuppertal, Germany
arnold@theorie.physik.uni-wuppertal.de

T.J. Ashby
School of Physics, University of Edinburgh, James Clerk Maxwell Building, Mayfield Road, Edinburgh EH9 3JZ, United Kingdom
t.ashby@ed.ac.uk

A. Boriçi
School of Physics, University of Edinburgh, James Clerk Maxwell Building, Mayfield Road, Edinburgh EH9 3JZ, United Kingdom
borici@ph.ed.ac.uk

N. Cundy
Fachbereich Mathematik und Naturwissenschaften, Universität Wuppertal, D-42097 Wuppertal, Germany
cundy@theorie.physik.uni-wuppertal.de

P.I. Davies
School of Mathematics, University of Manchester, Manchester, M13 9PL, England
pdavies@ma.man.ac.uk

Ph. de Forcrand
Institute for Theoretical Physics, ETH Zürich, CH-8093 Zürich, Switzerland
forcrand@phys.ethz.ch

I.T. Drummond
DAMTP, Cambridge University, Wilberforce Road, Cambridge CB3 0WA, United Kingdom
it.drummond@damtp.cam.ac.uk

G. T. Fleming
Jefferson Lab, 12000 Jefferson Ave, Newport News, VA 23606, USA
flemingg@jlab.org

Eduardo Follana
Department of Physics and Astronomy, University of Glasgow, G12 8QQ Glasgow, United Kingdom
e.follana@physics.gla.ac.uk

A. Frommer
Fachbereich Mathematik und Naturwissenschaften, Universität Wuppertal, D-42097 Wuppertal, Germany
frommer@math.uni-wuppertal.de

A. Hart
School of Physics, University of Edinburgh, King's Buildings, Edinburgh EH9 3JZ, United Kingdom
a.hart@ed.ac.uk

N.J. Higham
School of Mathematics, University of Manchester, Manchester, M13 9PL, England
higham@ma.man.ac.uk

R.R. Horgan
DAMTP, Cambridge University, Wilberforce Road, Cambridge CB3 0WA, United Kingdom
r.r.horgan@damtp.cam.ac.uk

O. Jahn
Department of Physics, CERN, Theory Division, CH-1211 Geneva 23, Switzerland
and
LNS, MIT, Cambridge MA 02139, USA
jahn@mit.edu

Bálint Joó
School of Physics, University of Edinburgh, James Clerk Maxwell Building, Mayfield Road, Edinburgh EH9 3JZ, United Kingdom
bj@ph.ed.ac.uk

A.D. Kennedy
School of Physics, University of Edinburgh, James Clerk Maxwell Building, Mayfield Road, Edinburgh EH9 3JZ, United Kingdom
adk@ph.ed.ac.uk

S. Krieg
Fachbereich Mathematik und Naturwissenschaften, Universität Wuppertal, D-42097 Wuppertal, Germany
krieg@theorie.physik.uni-wuppertal.de

D. B. Leinweber
Special Research Centre for the Subatomic Structure of Matter and Department of Physics, University of Adelaide, Adelaide SA 5005, Australia
dleinweb@physics.adelaide.edu.au

Th. Lippert
Fachbereich Mathematik und Naturwissenschaften, Universität Wuppertal, D-42097 Wuppertal, Germany
lippert@theorie.physik.uni-wuppertal.de

Keh-Fei Liu
Deptartment of Physics and Astronomy, University of Kentucky, Lexington, KY 40506, USA
liu@pa.uky.edu

H. Neuberger
Department of Physics and Astronomy, Rutgers University, Piscataway, NJ 08540, USA
neuberg@physics.rutgers.edu

M.F.P. O'Boyle
Institute of Computer Systems Architecture, University of Edinburgh, James Clerk Maxwell Building, Mayfield Road, Edinburgh EH9 3JZ, United Kingdom
mob@inf.ed.ac.uk

M. Peardon
School of Mathematics, Trinity College, Dublin 2, Ireland
mjp@maths.tcd.ie

K. Schäfer
Fachbereich Mathematik und Naturwissenschaften, Universität Wuppertal, D-42097 Wuppertal, Germany
schaefer@math.uni-wuppertal.de

G. L.G. Sleijpen
Department of Mathematics, Utrecht University, P.O. Box 80.010, NL-3508 TA Utrecht, The Netherlands
sleijpen@math.uu.nl

L.C. Storoni
DAMTP, Cambridge University, Wilberforce Road, Cambridge CB3 0WA, United Kingdom
lcs20@cam.ac.uk

A.W. Thomas
Special Research Centre for the Subatomic Structure of Matter and Department of Physics, University of Adelaide, Adelaide SA 5005, Australia
athomas@physics.adelaide.au

J. van den Eshof
Department of Mathematics, University of Düsseldorf, D-40224, Düsseldorf, Germany
eshof@am.uni-duesseldorf.de

M.B. van Guizen
CERFACS, 42 Avenue Gaspard Coriolis, 31057 Toulouse Cedex 01, France
vangijzen@cerfacs.fr

U. Wenger
Theoretical Physics, Oxford University, 1 Keble Road, Oxford OX1 3NP, United Kingdom
and
NIC/DESY Zeuthen, Platanenallee 6, D–15738 Zeuthen, Germany
urs.wenger@desy.de

R. D. Young
Special Research Centre for the Subatomic Structure of Matter and Department of Physics, University of Adelaide, Adelaide SA 5005, Australia
ryoung@physics.adelaide.au

Part I

Surveys

An Introduction to Lattice Chiral Fermions

Herbert Neuberger

Department of Physics and Astronomy, Rutgers University, Piscataway, NJ08540, USA neuberg@physics.rutgers.edu

Summary. This write-up starts by introducing lattice chirality to people possessing a fairly modern mathematical background, but little prior knowledge about modern physics. I then proceed to present two new and speculative ideas.

1 Review

1.1 What are Dirac/Weyl Fermions?

One can think about (Euclidean) Field Theory as of an attempt to define integrals over function spaces [1]. The functions are of different types and are called fields. The integrands consist of a common exponential factor multiplied by various monomials in the fields. The exponential factor is written as $\exp(S)$ where the action S is a functional of the fields. Further restrictions on S are: (1) locality (2) symmetries. Locality means that S can be written as an integral over the base space (space-time) which is the common domain of all fields and the integrand at a point depends at most exponentially weakly on fields at other, remote, space-time points. S is required to be invariant under an all important group of symmetries that act on the fields. In a sense, S is the simplest possible functional obeying the symmetries and generically represents an entire class of more complicated functionals, which are equivalently appropriate for describing the same physics.

Dirac/Weyl fields have two main characteristics: (1) They are Grassmann valued, which means they are anti-commuting objects and (2) there is a form of S, possibly obtained by adding more fields, where the Dirac/Weyl fields, ψ, enter only quadratically. The Grassmann nature of ψ implies that the familiar concept of integration needs to be extended. The definition of integration over Grassmann valued fields is algebraic and for an S where the ψ fields enter quadratically, as in $S = \bar{\psi} K \psi + ...$, requires only the propagator, K^{-1}, and the determinant, $\det K$. Hence, only the linear properties of the operator K come into play, and concepts like a "Grassmann integration measure" are, strictly speaking, meaningless, although they make sense for ordinary, commuting, field integration variables.

Let us focus on a space-time that is a a 4D Euclidean flat four torus, with coordinates x_μ, $\mu = 1, 2, 3, 4$. Introduce the quaternionic basis σ_μ represented by 2×2 matrices:

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \ \sigma_2 = \begin{pmatrix} 0 & -\imath \\ \imath & 0 \end{pmatrix} \ \sigma_3 = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \ \sigma_4 = \begin{pmatrix} \imath & 0 \\ 0 & \imath \end{pmatrix}$$

The ψ fields are split into two kinds, $\bar{\psi}$ and ψ, each being a two component function on the torus. In the absence of other fields the Weyl operators playing the role of the kernel K are $W = \sigma_\mu \partial_\mu$ and $W^\dagger = -\sigma_\mu^\dagger \partial_\mu$. The Dirac operator is made by combining the Weyl operators:

$$D = \begin{pmatrix} 0 & W \\ -W^\dagger & 0 \end{pmatrix} = \begin{pmatrix} 0 & \sigma_\mu \\ \sigma_\mu^\dagger & 0 \end{pmatrix} \partial_\mu \equiv \gamma_\mu \partial_\mu = -D^\dagger$$

The σ_μ obey

$$\sigma_\mu^\dagger \sigma_\nu + \sigma_\nu^\dagger \sigma_\mu = 2\delta_{\mu\nu} \quad \sigma_\mu \sigma_\nu^\dagger + \sigma_\nu \sigma_\mu^\dagger = 2\delta_{\mu\nu}$$

which implies $W^\dagger W = -\partial_\mu \partial_\mu = -\partial^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Thus, one can think about W as a complex square root of the Laplacian. Similarly, one has $D^\dagger D = D D^\dagger = -D^2$, with D^2 being $-\partial^2$ times a 4×4 unit matrix.

When we deal with gauge theories there are other important fields [2]. These are the gauge fields, which define a Lie algebra valued one-form on the torus, denoted by $A \equiv A_\mu dx_\mu$. We shall take $A_\mu(x)$ to be an anti-hermitian, traceless, $N \times N$ matrix. The 1-form defines parallel transport of N-component complex fields Φ by:

$$\Phi(x(1)) = \mathcal{P} e^{\int_{\mathcal{C}} A \cdot dx} \Phi(x(0))$$

where $x_\mu(t), t \in [0,1]$ is a curve $\mathcal{C}$ connecting $x(0)$ to $x(1)$ and $\mathcal{P}$ denotes path ordering, the ordered product of $N \times N$ matrices being implicit in the exponential symbol. Covariant derivatives, $D_\mu = \partial_\mu - A_\mu$, have as main property the transformation rule:

$$g^\dagger(x) D_\mu(A) g(x) = D_\mu(A^g) \quad A^g \equiv A - g^\dagger dg$$

where the $g(x)$ are unitary $N \times N$ matrices with unit determinant. The replacement of ∂_μ by D_μ is known as the principle of minimal substitution and defines A-dependent Weyl and Dirac operators. A major role is played by local gauge transformations, defined by $\psi \to g\psi, \bar{\psi} \to \bar{\psi} g^\dagger$ and $A \to A^g$ where ψ is viewed as a column and $\bar{\psi}$ as a row. The gauge transformations make up an infinite invariance group and only objects that are invariant under this group are of physical interest. In particular, S itself must be gauge invariant and the ψ dependent part of it is of the form $S_\psi = \int_x \bar{\psi} W \psi$ with W possibly replaced by $W^\dagger$ or by D.

Formally, W^{-1} is gauge covariant and $\det W$ is gauge invariant. Both the construction of W and of D meet with some problems: (1) W may have exact "zero modes", reflecting a nontrivial analytical index. The latter is an integer defined as $\dim \mathrm{Ker} W^\dagger(A) - \dim \mathrm{Ker} W(A)$. It is possible for this integer to be non-zero because the form A is required to be smooth only up to gauge transformations. The space of all A's then splits into a denumerable collection of disconnected components, uniquely labeled by the index. The integration over A is split into a sum over components with associated integrals restricted to each component. (2) $\det W$ cannot always be defined in a gauge invariant way, but $\det(W^\dagger W) = |\det W|^2$ can. Thus, $\det W$ is to be viewed as a certain square root of $|\det W|^2$, but, instead of being a function over the spaces of A it is a line bundle. As a line bundle it can be also viewed as a line bundle over the space of gauge orbits of A, where a single orbit is the collection of all elements A^g for a fixed A and all g. The latter bundle may be twisted, and defy attempts to find a smooth gauge invariant section. When this happens we have an anomaly.

1.2 Why is There a Problem on the Lattice?

Lattice field theory [3] tries to construct the desired functional integral by first replacing space-time by a finite, uniform, toroidal square lattice and subsequently constructing a limit in which the lattice spacing, a, is taken to zero. Before the limit is taken functional integration is replaced by ordinary integration producing well defined quantities. One tries to preserve as much as possible of the desired symmetry, and, in particular, there is a symmetry group of lattice gauge transformations given by $\prod_x SU(N)$, where x denotes now a discrete lattice site.

The one-form A is replaced by a collection of elementary parallel transporters, the link matrices $U_\mu(x)$, which are unitary and effect parallel transport from the site x to the neighboring site to x in the positive μ direction. Traversal in the opposite direction goes with $U_\mu^\dagger(x)$. The fields $\bar{\psi}$ and ψ are now defined at lattice sites only. As a result, $W, W^\dagger$ become finite square matrices. Here are the main problems faced by this construction: (1) The space of link variables is connected in an obvious way and therefore the index of W will vanish always. Indeed, W is just a square matrix. (2) $\det W$ is always gauge invariant, implying that anomalies are excluded. In particular, there no longer is any need to stop the construction at the intermediate step of a line bundle. These properties show that no matter how we proceed, the limit where the lattice spacing a goes to zero will not have the required flexibility.

1.3 The Basic Idea of the Resolution

The basic idea of the resolution [4] is to reintroduce a certain amount of indeterminacy by adding to the lattice version a new infinite dimensional space in which ψ is an infinite vector, in addition to its other indices. Other fields do not see this space, and different components of ψ are accordingly referred to as flavors. Among all fields, only the ψ fields come in varying flavors. W shall be replaced by a linear operator that acts nontrivially in the new flavor space in addition to its previous actions. The infinite dimensional structure is chosen as simple as possible to provide for, simultaneously, good mathematical control, the emergence of a non-zero index and the necessity of introducing an intermediary construction of $\det W$ as a line bundle [5].

The structure of the lattice W operator is that of a lattice Dirac type operator. This special lattice Dirac operator, D, has a mass, acting linearly in flavor space. With this mass term, the structure of our lattice D is:

$$D = \begin{pmatrix} aM^\dagger & aW \\ -aW^\dagger & aM \end{pmatrix}$$

Only M acts nontrivially in flavor space. To obtain a single Weyl field relevant for the subspace corresponding to small eigenvalues of $-D^2$, the operator M is required to satisfy: (1) the index of M is unity (2) the spectrum of $MM^\dagger$ is bounded from below by a positive number, Λ^2. $(\Lambda a)^2$ is of order unity and kept finite and fixed as $a \to 0$. In practice it is simplest to set the lattice spacing a to unity and take all other quantities to be dimensionless. Dimensional analysis can always be used to restore the powers of a. In the continuum, we always work in the units in which $c = \hbar = 1$. Numerical integration routines never know what a is in length units. The lower bound on $MM^\dagger$ is taken to be of order unity.

The index structure of M ensures that, for eigenvalues of $-D^2$ that are small relative to unity, the relevant space is dominated by vectors with vanishing upper components. These vectors are acted on by the W sub-matrix of D. Moreover, the main contribution comes from the zero mode of M, so, both the infinite flavor space and the extra doubling implicit in using a Dirac operator, become irrelevant for the small eigenvalues of $-D^2$ and their associated eigenspace.

The standard choice for M stems from a paper by Callan and Harvey [6] which has been ported to the lattice by Kaplan [7]. The matrix M is given by a first order differential (or difference) operator of the form $-\partial_s + f(s)$, where s is on the real line and represents flavor space. $f(s)$ is chosen to be the sign function, but could equally well just have different constant absolute values for s positive and for s negative.

The construction of the lattice determinant line bundle will not be reviewed here and we shall skip ahead directly to the overlap Dirac operator.

1.4 The Overlap Dirac Operator

The continuum Dirac operator combines two independent Weyl operators. The Weyl components stay decoupled so long as there is no mass term, and admit independently acting symmetries. Thus, zero mass Dirac fields have more symmetry than massive ones. In particular, this implies that radiative corrections to small Dirac masses must stay proportional to the original mass, to ensure exact vanishing in the higher symmetry case. A major problem in particle physics is to understand why all masses are so much smaller than the energy at which all gauge interactions become of equal strength and one of the most important examples of a possible explanation is provided by the mechanism of chiral symmetry. Until about six years ago it was believed that one could not keep chiral symmetries on the lattice and therefore lattice work with small masses required careful tuning of parameters.

Once we have a way to deal with individual Weyl fermions, it must be possible to combine them pair-wise just as in the continuum and end up with a lattice Dirac operator that is exactly massless by symmetry. This operator is called the overlap Dirac operator and is arrived at by combining the two infinite flavor spaces of each Weyl fermion into a new single infinite space [8]. However, unlike the infinite space associated with each Weyl fermion, the combined space can be viewed as the limit of a finite space. This is so because the Dirac operator does not have an index – unlike the Weyl operator – nor does it have an ill defined determinant. Thus, there is no major problem if the lattice Dirac operator is approximated by a finite matrix. The two flavor spaces are combined simply by running the coordinate s first over the values for one Weyl component and next over the values for the other Weyl component. Since one Weyl component comes as the hermitian conjugate of the other it is no surprise that the coordinate s will be run in opposite direction when it is continued. Thus, one obtains an infinite circle, with a combined function $f(s)$ which is positive on half of the circle and negative on the other. The circle can be made finite and then one has only approximate chiral symmetry [9]. One can analyze the limit when the circle goes to infinity and carry out the needed projection on the small eigenvalue eigenspaces to restrict one to only the components that would survive in the continuum limit. The net result is a formula for the lattice overlap Dirac operator, D_o [8].

To explain this formula one needs, as a first step, to introduce the original lattice Dirac operator due to Wilson, D_W. That matrix is the most sparse one possible with

the right symmetry properties, excepting chiral symmetry. It is used as a kernel of the more elaborate construction needed to produce produce D_o. Any alternative to D_W will produce, by the same construction, a new D_o, possibly enhancing some of its other properties. The original D_o is still the most popular, because the numerical advantage of maximal sparseness of D_W has proven hard to beat by benefits coming from other improvements. Thus, we restrict ourselves here only to D_W.

$$D_W = m + 4 - \textstyle\sum_\mu V_\mu$$
$$V_\mu = \tfrac{1-\gamma_\mu}{2} T_\mu + \tfrac{1+\gamma_\mu}{2} T_\mu^\dagger$$
$$\langle x|T_\mu|\Phi^i\rangle = U_\mu(x)^{ij}\langle x|\Phi^j\rangle$$
$$U_\mu(x)U_\mu^\dagger(x) = 1 \qquad \gamma_\mu = \begin{pmatrix} 0 & \sigma_\mu \\ \sigma_\mu^\dagger & 0 \end{pmatrix} \qquad \gamma_5 = \gamma_1\gamma_2\gamma_3\gamma_4$$

$|\Phi^i\rangle$ is a vector with components labeled by the sites x. The notation indicates that this is the i-th component of a vector $|\Phi\rangle$ with components labeled by both a site x and a group index, j. It is easy to see that $V_\mu V_\mu^\dagger = 1$, so D_W is bounded. $H_W = \gamma_5 D_W$ is hermitian and sparse. The parameter m must be chosen in the interval $(-2, 0)$, and typically is around -1. For gauge fields that are small, the link matrices are close to unity and a sizable interval around zero can be shown to contain no eigenvalues of H_W [10]. This spectral gap can close for certain gauge configurations, but these can be excluded by a simple local condition on the link matrices. When that condition is obeyed, and otherwise independently on the gauge fields, all eigenvalues of H_W^2 are bigger than some positive number μ^2. This makes it possible to unambiguously define the sign function of H_W, $\epsilon(H_W)$. Moreover, $\epsilon(H_W)$ can be infinitely well approximated by a smooth function so long as $\mu^2 > 0$. Since, in addition, the spectrum of H_W is bounded from above, Weierstrass's approximation theorem applies and one can approximate uniformly $\epsilon(H_W)$ by a polynomial in H_W. Thus, as a matrix, ϵ is no longer sparse, but, for $\mu^2 > 0$, it still is true that entries associated with lattice sites separated by distances much larger than $\frac{1}{\mu}$ are exponentially small.

The exclusion of some configurations ruins the simple connectivity of the space of link variables just as needed to provide for a lattice definition of the integer n, which in the continuum labels the different connected components of gauge orbit space. The appropriate definition of n on the lattice is [11]

$$n = \frac{1}{2} Tr\epsilon(H_W)$$

It is obvious that it gives an integer since H_W must have even dimensions as is evident from the structure of the γ-matrices. Moreover, it becomes very clear why configurations for which H_W could have a zero eigenvalue needed to be excised. These configurations were first found to need to be excised when constructing the lattice version of the $\det W$ line bundle.

The overlap Dirac operator is

$$D_o = \frac{1}{2}(1 + \gamma_5\epsilon(H_W))$$

γ_5 and ϵ make up a so called "Kato pair" with elegant algebraic properties [12].

1.5 What About the Ginsparg-Wilson Relation?

In practice, the inverse of D_o is needed more than D_o itself. Denoting $\gamma_5\epsilon(H_W) = V$, where V is unitary and obeys "γ_5-hermiticity", $\gamma_5 V \gamma_5 = V^\dagger$, we easily prove that $D_o^{-1} = \frac{2}{1+V}$ obeys

$$\{\gamma_5, D_o^{-1} - 1\} = 0$$

Here, we introduced the anti-commutator $\{a, b\} \equiv ab + ba$. In the continuum, the same relation is obeyed by D^{-1} and reflects chiral symmetry. We see that a slightly altered propagator will be chirally symmetric. The above equation, modifying the continuum relation $\{\gamma_5, D^{-1}\} = 0$, was first written down by Ginsparg and Wilson (GW) in 1982 [14] in a slightly different form. By a quirk of history, their paper became famous only after the discovery of D_o. The main point of the GW paper is that shifting an explicitly chirally symmetric propagator by a matrix which is almost diagonal in lattice sites and unity in spinor space does not destroy physical chiral symmetry.

It turns out that the explicitly chirally symmetric propagator, $\frac{1-V}{1+V}$, can be used as the propagator associated with the monomials of the fields that multiply e^S, but in other places where the propagator appears (loops), one needs to use the more subtly chirally symmetric propagator, $D_o^{-1} = \frac{2}{1+V}$. This dichotomy is well understood and leads to no inconsistencies [15].

Any solution of the GW relation, if combined with γ_5 hermiticity, is of the form $\frac{2}{1+V}$, producing a propagator which anti-commutes with γ_5 of the form $\frac{1-V}{1+V}$. V is a unitary, γ_5-hermitian, matrix. Thus the overlap is the general γ_5-hermitian solution to the GW relation, up to an inconsequential generalization which adds a sparse, positive definite, kernel matrix to the GW relation. The overlap goes beyond the GW paper in providing a generic procedure to produce explicit acceptable matrices V starting from explicit matrices of the same type as H_W.

When the GW relation was first presented, in 1982, the condition of γ_5-hermiticity was not mentioned. The solution was not written in terms of a unitary matrix V, and there was no explicit proposal for the dependence of the solution on the gauge fields. For these reasons, the paper fell into oblivion, until 1997, when D_o was arrived at by a different route. With the benefit of hindsight we see now that it was a mistake not to pursue the GW approach further.

In 1982 neither the mathematical understanding of anomalies - specifically the need to find a natural $U(1)$ bundle replacing the chiral determinant - nor the paramount importance of the index of the Weyl components were fully appreciated. Only after these developments became widely understood did it become possible to approach the problem of lattice chirality from a different angle and be more successful at solving it. The convergence with the original GW insight added a lot of credence to the solution and led to a large number of papers based on the GW relation.

Already in 1982 GW showed that if a solution to their relation were to be found, the slight violation of anti-commutativity with γ_5 that it entailed, indeed was harmless, and even allowed for the correct reproduction of the continuum triangle diagram, the key to calculating anomalies. Thus, there was enough evidence in 1982 that should have motivated people to search harder for a solution, but this did not happen. Rather, the prevailing opinion was that chirality could not be preserved on the lattice. This opinion was fed by a an ongoing research project which attempted

to solve the lattice chirality problem by involving extra scalar fields, interacting with the fermions by trilinear (Yukawa) interactions. In this approach one ignored the topological properties of the continuum Dirac operator with respect to the gauge background. The Yukawa models never worked, but the people involved did not attribute this to the failing treatment of topology, and slowly the feeling that chiral symmetry could not be preserved on the lattice took root.

In retrospect, something went wrong in the field's collective thought process, but parallel developments mentioned earlier eventually provided new impetus to deal with the problem correctly. Luckily, this second opportunity was not missed. There was however substantial opposition and even claims that the new approach was not different from the one based on Yukawa interactions, and therefore, was unlikely to be correct [16].

After the discovery of D_o, fifteen years after the GW paper, a flood of new papers, developing the GW approach further, appeared. Because the overlap development already had produced all its new conceptual results by then, no further substantial advance took place. For example, the importance of topology was reaffirmed in a GW framework [17], but the overlap already had completely settled this issue several years earlier. However, this renewed activity generated enough reverberations in the field to finally eradicate the prevailing assumption of the intervening years, that chiral symmetry could not be preserved on the lattice.

1.6 Basic Implementation

Numerically the problem is to evaluate $\epsilon(H_W)$ on a vector, without storing it, basing oneself on the sparseness of H_W. This can be done because, possibly after deflation, the spectrum of H_W has a gap around 0, the point where the sign function is discontinuous. In addition, since H_W is bounded we need to approximate the sign function well only in two disjoint segments, one on the positive real line and the other its mirror image on the negative side. A convenient form is the Higham representation, which introduces $\epsilon_n(x)$ as an approximation to the sign function:

$$\epsilon_n(x) = \begin{cases} \tanh[2n \tanh^{-1}(x)] & \text{for } |x| < 1 \\ \tanh[2n \tanh^{-1}(x^{-1})] & \text{for } |x| > 1 \\ x & \text{for } |x| = 1 \end{cases}$$

Equivalently,

$$\epsilon_n(x) = \frac{(1+x)^{2n} - (1-x)^{2n}}{(1+x)^{2n} + (1-x)^{2n}} = \frac{x}{n} \sum_{s=1}^{n} \frac{1}{x^2 \cos^2\left[\frac{\pi}{2n}\left(s - \frac{1}{2}\right)\right] + \sin^2\left[\frac{\pi}{2n}\left(s - \frac{1}{2}\right)\right]}$$

$$\lim_{n \to \infty} \epsilon_n(x) = \text{sign}(x)$$

$\epsilon_n(H_W)\psi$ can be evaluated using a single Conjugate Gradient (CG) iteration with multiple shifts for all the pole terms labeled by s above [18]. The cost in operations is that of a single CG together with an overhead that is linear in n and eventually dominates. The cost in storage is of $2n$ large vectors. The pole representation can be further improved using exact formulae due to Zolotarev who solved the Chebyshev approximation problem analytically for the sign function, thus eliminating the need to use the general algorithm due to Remez. However, for so called

quenched simulations, where one replaces $\det D_o$ by unity in the functional integration, the best is to use a double pass [19] version introduced a few years ago but fully understood only recently [20]. In the double pass version storage and number of operations become n-independent for large n, which, for double precision calculations means an n larger than 30 or so. Thus, the precise form of the pole approximation becomes irrelevant and storage requirements are modest. In "embarrassingly parallel" simulations this is the method of choice because it simultaneously attains maximal numerical accuracy and allows maximal exploitation of machine cycles.

When one goes beyond the $\det D_o = 1$ approximation, one needs to reconsider methods that employ order n storage. A discussion of the relevant issues in this case would take us beyond the limits of this presentation; these issues will be covered by other speakers who are true experts.

2 Beyond Overlap/GW?

The overlap merged with GW because both ideas exploited a single real extra coordinate. The starting point of the overlap construction however seems more general, since it would allow a mass matrix in infinite flavor space even if the latter were associated with two or more coordinates. Thus, one asks whether using two extra coordinates might lead to a structurally new construction [21]. While this might not be better in practice, it at least has the potential of producing something different, unattainable if one just sticks to the well understood GW track.

The function $f(s)$ from the overlap is replaced now by two functions $f_1(s_1)$ and $f_2(s_2)$ and the single differential operator $\partial_s + f(s)$ by two such operators, $d_\alpha = \partial_\alpha + f_\alpha(s_\alpha)$. Clearly, d_1 and d_2 commute. A mass matrix with the desired properties can be now constructed as follows:

$$M = \begin{pmatrix} d_1 & -id_2^\dagger \\ id_2 & -d_1^\dagger \end{pmatrix}$$

The two dimensional plane spanned by s_α is split into four quadrants according to the pair of signs of f_α and, formally, the chiral determinant can be written as the trace of four Baxter Corner Transfer Matrices,

$$\text{chiral det} = Tr[K^{\mathbf{I}} K^{\mathbf{II}} K^{\mathbf{III}} K^{\mathbf{IV}}]$$

While this structure is intriguing, I have made no progress yet on understanding whether it provides a natural definition of a $U(1)$ bundle with the right properties. If it does, one could go over to the Dirac case, and an amusing geometrical picture seems to emerge. It is too early to tell whether this idea will lead anywhere or not.

3 Localization and Domain Wall Fermions

3.1 What are Domain Wall Fermions?

Before the form of D_o was derived we had a circular s space with $f(s)$ changing sign at the opposite ends of a diameter. One of the semi-circles can be eliminated by

taking $|f(s)|$ to infinity there, leaving us with a half circle that can be straightened into a segment with two approximate Weyl fields localized at its ends. This is known as the domain wall setup, the walls extending into the physical directions of space-time. Keeping the length of the segment finite but large one has approximate chiral symmetry and an operator D_{DW} which acts on many Dirac fields, exactly one of them having a very small effective mass, and the rest having masses of order unity.

The chiral symmetry is only approximate because matrix elements of $\frac{1}{D^\dagger_{DW} D_{DW}}$ connecting entries associated with opposite ends of the segment, L and R, do not vanish exactly. Using a spectral decomposition of $D^\dagger_{DW} D_{DW}$ we have:

$$\langle L|\frac{1}{D^\dagger_{DW} D_{DW}}|R\rangle = \sum_n \frac{1}{\Lambda_n}\langle \Psi_n|R\rangle\langle \Psi_n|L\rangle^* \qquad \langle \Psi_n|\Psi_n\rangle = 1$$

Weyl states are localized at L and R and should not connect with each other. So long as the distance between R and L is infinite and $H_W^2 > \mu^2$ this is exactly proven to be the case. For a finite distance S, the correction goes as $e^{-\mu S}$. Unfortunately, μ can be very small numerically and this would require impractically large values of S. Note that the worse situation occurs if one has simultaneously a relatively large wave-function contribution, $|\langle \Psi_n|R\rangle\langle \Psi_n|L\rangle|$, and a small Λ_n. Unfortunately, this worse case seems to come up in practice.

3.2 The Main Practical Problem

As already mentioned, for the purpose of keeping track of $\det D_o$, one may want to keep in the simulation the dependence on the coordinate s, or, what amounts to a logical equivalent, the n fields corresponding to the pole terms in the sign function truncation. This is the main reason to invest resources in domain wall simulations. In my opinion, if one works in the approximation where $\det D_o = 1$ it does not pay to deal with domain wall fermions because it is difficult to safely assess the magnitude of chirality violating effects in different observables.

The main problem faced by practical domain wall simulations is that in the range of interest for strong interaction (QCD) phenomenology H_W, the kernel of the overlap, has eigenstates with very small eigenvalues in absolute value. It turns out that these states are strongly localized in space-time. However, because of approximate translational invariance in s they hybridize into delocalized bands into the extra dimension. As such, they provide channels by which the Weyl modes at the two locations L and R, where the combined $f(s)$ vanishes, communicate with each other, spoiling the chiral symmetry. To boot, these states have small Λ_n. The one way known to eliminate this phenomenon is to take the separation between the Weyl modes to infinity. This leads to the overlap where the problem becomes only of a numerical nature and is manageable by appropriately deflating H_W to avoid the states for which the reconstruction of the sign function by iterative means is too expensive.

3.3 The New Idea

The new idea is to exploit the well known fact that one dimensional random systems typically always localize. The standard approach uses a homogeneous s coordinate;