CERN-PH-TH/2010-047

Computational Strategies in Lattice QCD

Martin Lüscher

Physics Department, CERN, 1211 Geneva 23, Switzerland

Summer School on "Modern perspectives in lattice QCD"
Les Houches, August 3–28, 2009

Preface

Numerical lattice QCD has seen many important innovations over the years. In this course an introduction to some of the basic techniques is provided, emphasizing their theoretical foundation rather than their implementation and latest refinements.

The development of computational strategies in lattice QCD requires physical insight to be combined with an understanding of modern numerical mathematics and of the capabilities of massively parallel computers. When a new method is proposed, it should ideally be accompanied by a theoretical analysis that explains why it is expected to work out. However, in view of the complexity of the matter, some experimenting is often required. The field thus retains a certain empirical character.

At present numerical lattice QCD is still in a developing phase to some extent. The baryon spectrum, for example, remains to be difficult to compute reliably, because the signal-to-noise ratio of the associated two-point functions decreases exponentially at large distances. There is certainly ample room for improvements and it may also be necessary to radically depart from the known techniques in some cases. Hopefully these lectures will encourage some of the problems to be studied and to be solved eventually.

Acknowledgements

I am indebted to the organizers for inviting me to lecture at this Summer School and for providing a very pleasant and stimulating atmosphere. During my lectures many questions were asked that helped to clarify some subtle points. I have rarely had such an attentive audience and would like to thank the students for patiently going with me through the rather technical material covered in the course. Finally, I wish to thank the University of Grenoble, the director of the school and the support staff for running this wonderful place in a perfect manner.

Contents

1
Computation of quark propagators

Quark propagators in presence of a specified SU(3) gauge field are fundamental building blocks in lattice QCD. Their computation amounts to solving the Dirac equation

$$D\psi(x) = \eta(x) \tag{1.1}$$

a number of times, where D denotes the massive lattice Dirac operator, $\eta(x)$ a given quark field (the source field) and $\psi(x)$ the desired solution.

Although the subject is nearly as old as lattice QCD itself, there have been important advances in the last few years that allow the quark propagators to be calculated much more rapidly than was possible before. As a consequence, the "measurement" of hadronic quantities and the simulation of the theory with light sea quarks are both accelerated significantly.

1.1 Preliminaries

1.1.1 Accuracy & condition number

The Dirac equation (1.1) is a large linear system that can only be solved iteratively, i.e. through some recursive procedure that generates a sequence $\psi_1, \psi_2, \psi_3, \ldots$ of increasingly accurate approximate solutions. A practical measure for the accuracy of an approximate solution ϕ is the norm of the associated residue

$$\rho = \eta - D\phi. \tag{1.2}$$

If, say, $\|\rho\| < \epsilon\|\eta\|$ for some small value ϵ, an important question is then by how much ϕ deviates from the exact solution ψ of the equation. Using standard norm estimates, the deviation is found to be bounded by

$$\|\psi - \phi\| < \epsilon\kappa(D)\|\psi\|, \tag{1.3}$$

where

$$\kappa(D) = \|D\|\|D^{-1}\| \tag{1.4}$$

is referred to as the condition number of D. The relative error of ϕ thus tends to be larger than ϵ by the factor $\kappa(D)$ (in this chapter, the standard scalar product of quark fields is used as well as the field and operator norms that derive from it).

The extremal eigenvalues of $D^\dagger D$, $\alpha_{\min}$ and $\alpha_{\max}$, are proportional to the square of the quark mass m and the square of the inverse of the lattice spacing a, respectively. In particular, the condition number

$$\kappa(D) = (\alpha_{\max}/\alpha_{\min})^{1/2} \propto (am)^{-1} \tag{1.5}$$

can be very large at small quark masses and lattice spacings. One says that the Dirac operator is "ill-conditioned" in this case. The important point to keep in mind is that the accuracy of the solution of the Dirac equation which can be attained on a given computer is limited by the condition number.

1.1.2 Iterative improvement

If the residual error ϵ of the approximate solution ϕ is still well above the limit set by the machine precision and the condition number of the Dirac operator, an improved solution

$$\tilde{\phi} = \phi + \chi \tag{1.6}$$

may be obtained by approximately solving the residual equation

$$D\chi = \rho. \tag{1.7}$$

It is straightforward to show that the residue of the solution is reduced by the factor δ in this way if χ satisfies $\|\rho - D\chi\| < \delta\|\rho\|$. Note that χ is approximately equal to the deviation of ϕ from the exact solution of the Dirac equation and is therefore usually a small correction to ϕ.

Iterative improvement is used by all solvers that need to be restarted after a while. The GCR algorithm discussed below is an example of such a solver. Another application of iterative improvement, known as "single-precision acceleration", exploits the fact that modern processors perform 32 bit arithmetic operations significantly faster than 64 bit operations. The idea is to solve the Dirac equation to 64 bit precision by going through a few cycles of iterative improvement, where, in each cycle, the residual equation is solved to a limited precision using 32 bit arithmetic, while the residue ρ and the sum (1.6) are evaluated using 64 bit arithmetic (Giusti *et al.*, 2003).

1.2 Krylov-space solvers

The Krylov space $\mathcal{K}_n$ of dimension n is the complex linear space spanned by the fields

$$\eta,\ D\eta,\ D^2\eta, \ldots,\ D^{n-1}\eta. \tag{1.8}$$

Many popular solvers, including the CG, BiCGstab and GCR algorithms, explicitly or implicitly build up a Krylov space and search for the solution of the Dirac equation within this space. The very readable book of Saad (2003) describes these solvers in full detail. A somewhat simpler discussion of the CG (conjugate gradient) algorithm is given in the book of Golub and van Loan (1989) and useful additional references for the BiCGstab algorithm are van der Vorst (1992) and Frommer *et al.* (1994). Here the GCR (generalized conjugate residual) algorithm is discussed as a representative case.

1.2.1 The GCR algorithm

The approximate solutions of the Dirac equation (1.1) generated by the GCR algorithm are the fields $\psi_k \in \mathcal{K}_k$, $k = 1, 2, 3, \ldots$, that minimize the norm of the residues

$$\rho_k = \eta - D\psi_k. \tag{1.9}$$

An equivalent requirement is that $D\psi_k$ coincides with the orthogonal projection of the source field η to the k-dimensional linear space $D\mathcal{K}_k$ (see Fig. 1.1). The algorithm proceeds somewhat indirectly by first constructing an orthonormal basis $\chi_0, \chi_1, \chi_2, \ldots$ of these spaces through a recursive process. Independently of the details of the construction, the orthogonality property mentioned above then implies that the fields

$$\rho_k = \eta - \sum_{l=0}^{k-1} c_l \chi_l, \qquad c_l = (\chi_l, \eta), \tag{1.10}$$

are the residues of the approximate solutions ψ_k. The residues are thus obtained before the latter are known.

In each iteration of the recursion, the next basis field χ_k is constructed from the previous fields $\chi_0, \ldots, \chi_{k-1}$ and the source η. First the residue ρ_k is computed through eqn (1.10) and χ_k is then taken to be a (properly orthonormalized) linear combination of $D\rho_k$ and the previous fields. Since χ_{k-1} is a linear combination of $D\rho_{k-1}$ and the fields $\chi_0, \ldots, \chi_{k-2}$, and so on, it is clear that the recursion also yields the coefficients a_{lj} in the equations

$$\chi_l = \sum_{j=0}^{l} a_{lj} D\rho_j, \qquad l = 0, 1, 2, \ldots, \tag{1.11}$$

in which $\rho_0 = \eta$.

Once the fields $\chi_0, \ldots, \chi_{n-1}$ are known for some n, the last solution ψ_n is obtained starting from the orthogonality condition

$$D\psi_n = \sum_{l=0}^{n-1} c_l \chi_l. \tag{1.12}$$

After substituting eqn (1.11), the equation may then be divided by D and one finds that the solution is given by

$$\psi_n = \sum_{l=0}^{n-1} \sum_{j=0}^{l} c_l a_{lj} \rho_j. \tag{1.13}$$

Note that the right-hand side of this equation can be evaluated straightforwardly since all entries are known at this point.

In total the computation of ψ_n requires n applications of the Dirac operator and the evaluation of some $\frac{1}{2}n^2$ linear combinations and scalar products of quark fields. Moreover, memory space for about $2n$ fields is needed. Choosing values of n from, say, 16 to 32 proves to be a reasonable compromise in practice, where the computational effort must be balanced against the reduction in the residue that is achieved. If the last solution is not sufficiently accurate, the algorithm can then always be restarted following the rules of iterative improvement.

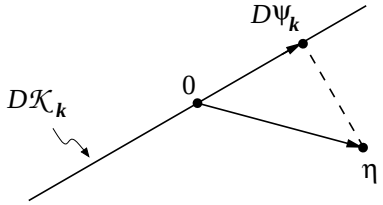


Fig. 1.1 The approximate solution $\psi_k \in \mathcal{K}_k$ of the Dirac equation constructed by the GCR algorithm is such that the distance $\|\eta - D\psi_k\|$ is minimized. The field $D\psi_k$ therefore coincides with the orthogonal projection of the source η to the space $D\mathcal{K}_k$.

1.2.2 Convergence properties

The convergence of the GCR algorithm and related Krylov-space solvers can be proved rigorously if the (complex) spectrum of the Dirac operator is contained in the half-plane on the right of the imaginary axis. In the case of the Neuberger–Dirac operator, for example, all eigenvalues of D lie on the circle shown in Fig. 1.2. The spectrum of the Wilson–Dirac operator is rather more complicated, but is usually contained in an ellipsoidal region in the right half-plane.

For simplicity, the Dirac operator is, in the following paragraphs, assumed to be diagonalizable and to have all its eigenvalues in the shaded disk $\mathbb{D}$ shown in Fig. 1.2. The convergence analysis of the GCR algorithm then starts from the observation that the residue ρ_k is given by

$$\rho_k = \eta - D\psi_k = p_k(D)\eta, \tag{1.14}$$

where $p_k(\lambda)$ is a polynomial of degree k that satisfies $p_k(0) = 1$. Since the algorithm minimizes the residue, it follows that

$$\|\rho_k\| = \min_{p_k} \|p_k(D)\eta\| \leq \min_{p_k} \|p_k(D)\| \|\eta\|, \tag{1.15}$$

the minimum being taken over all such polynomials.

Lattice quark fields are large arrays of complex numbers. In this language, the Dirac operator D is just a complex square matrix. The assumption that D is diagonalizable then implies the existence of a diagonal matrix Λ and of an invertible matrix V such that $D = V\Lambda V^{-1}$. As a consequence

$$\|p_k(D)\| = \|V p_k(\Lambda) V^{-1}\| \leq \kappa(V) \|p_k(\Lambda)\| \tag{1.16}$$

and therefore

$$\|\rho_k\| \leq \kappa(V) \max_{\lambda \in \mathbb{D}} |p_k(\lambda)| \|\eta\| \tag{1.17}$$

for any polynomial $p_k(\lambda)$ of degree k satisfying $p_k(0) = 1$. One may, for example, insert

$$p_k(\lambda) = \left(1 - \frac{\lambda}{M + m}\right)^k, \tag{1.18}$$

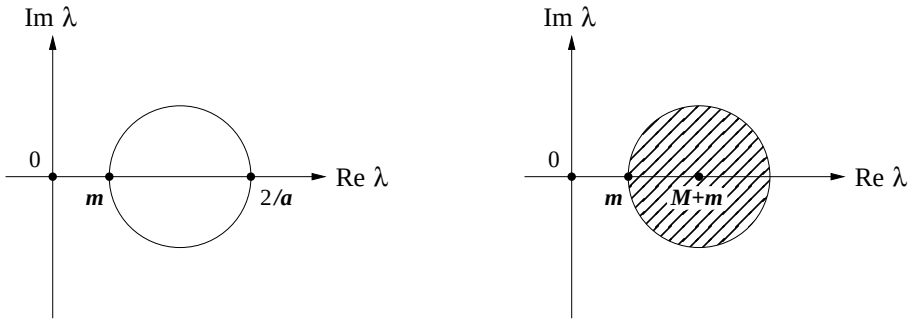


Fig. 1.2 The eigenvalues of the Neuberger–Dirac operator with bare quark mass m lie on a circle in the complex plane (drawing on the left). For the convergence analysis of the GCR algorithm, the spectrum of the Dirac operator is assumed to be contained in a disk $\mathbb{D}$ in the right-half plane with radius M and distance $m > 0$ from the origin (drawing on the right).

in which case the inequality (1.17) leads to the bound [1]

$$\|\rho_k\| \leq \kappa(V) \left(1 + \frac{m}{M}\right)^{-k} \|\eta\|. \tag{1.19}$$

The GCR algorithm thus converges roughly like $\exp(-km/M)$ if $m/M \ll 1$. Note that the convergence rate $m/M \sim 2/\kappa(D)$ can be quite small in practice. For $m = 10$ MeV and $M = 2$ GeV, for example, the estimate (1.19) suggests that values of k as large as 4000 are required for a reduction of the residue by the factor 10^{-10}.

The GCR algorithm can also be applied to the so-called normal equation

$$D^\dagger D\psi = \eta \tag{1.20}$$

and to the Dirac equation

$$(i\gamma_5 D + \mu)\psi = \eta \tag{1.21}$$

in "twisted-mass" QCD. A notable difference with respect to the ordinary Dirac equation is that the operators on the left of these equations can be diagonalized through unitary transformations. Moreover, their spectra are contained in straight-line segments in the complex plane (see Fig. 1.3). Using Chebyshev polynomials in place of the power (1.18), the estimate

$$\|\rho_k\| \lesssim 2\mathrm{e}^{-rk}\|\eta\| \tag{1.22}$$

may be derived in these cases, where $r = m/M$ for the normal equation and $r = \mu/2M$ for the twisted-mass Dirac equation.

The convergence of Krylov-space solvers is thus mainly determined by the properties of the spectrum of the operator considered. In QCD the fact that the masses of

[1] A rigorous mathematical result, known as Zarantello's lemma, asserts that it is not possible to obtain a more stringent bound by choosing a different polynomial, i.e. the polynomial (1.18) is the optimal one.

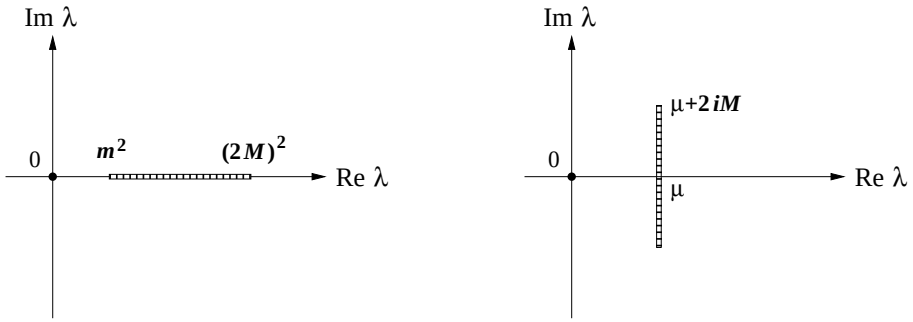


Fig. 1.3 The eigenvalues of the hermitian operator $D^\dagger D$ and the twisted-mass Dirac operator $i\gamma_5 D + \mu$ occupy the shaded line segments shown in the left and right drawings, respectively.

the light quarks are much smaller than the inverse lattice spacing consequently tends to slow down the computations enormously. One can do better, however, by exploiting specific properties of the Dirac operator.

1.2.3 Preconditioning

Preconditioning is a general strategy that allows such properties to be taken into account. Let L and R be some invertible operators acting on quark fields. Instead of the Dirac equation (1.1), one may then consider the so-called preconditioned equation

$$LDR\phi = L\eta. \tag{1.23}$$

Once this equation solved, using a Krylov-space solver for example, the solution of the Dirac equation is obtained by setting $\psi = R\phi$. If $D \approx L^{-1}R^{-1}$, and if the application of L and R to a given quark field is not too time-consuming, the total computer time required for the solution of the equation may be significantly reduced in this way.

In lattice QCD, a widely used preconditioning method for the Wilson–Dirac operator is "even-odd preconditioning". A lattice point $x \in \mathbb{Z}^4$ is referred to as even or odd depending on whether the sum of its coordinates x_μ is even or odd (see Fig. 1.4). If the points are ordered such that the even ones come first, the Dirac operator assumes the block form

$$D = \begin{pmatrix} D_{\mathrm{ee}} & D_{\mathrm{eo}} \\ D_{\mathrm{oe}} & D_{\mathrm{oo}} \end{pmatrix}, \tag{1.24}$$

where D_{eo}, for example, stands for the hopping terms that go from the odd to the even sites. The blocks on the diagonal, D_{ee} and D_{oo}, include the mass term and a Pauli term if the theory is O(a)-improved. Since they do not couple different lattice points, they can be easily inverted and it makes sense to consider the preconditioners

$$L = \begin{pmatrix} 1 & -D_{\mathrm{eo}}D_{\mathrm{oo}}^{-1} \\ 0 & 1 \end{pmatrix}, \qquad R = \begin{pmatrix} 1 & 0 \\ -D_{\mathrm{oo}}^{-1}D_{\mathrm{oe}} & 1 \end{pmatrix}. \tag{1.25}$$

With this choice, the Dirac operator is block-diagonalized,

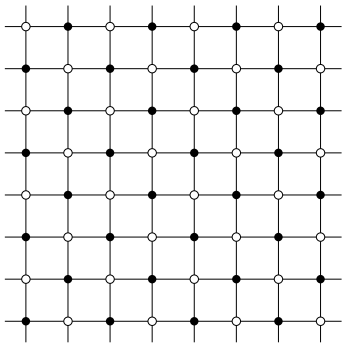


Fig. 1.4 Hyper-cubic lattices may be divided into the sublattices of the even and the odd sites (black and white points, respectively). Even-odd preconditioning effectively amounts to "integrating out" the quark field on the odd sublattice.

$$LDR = \begin{pmatrix} \hat{D} & 0 \\ 0 & D_{\mathrm{oo}} \end{pmatrix}, \qquad \hat{D} = D_{\mathrm{ee}} - D_{\mathrm{eo}} D_{\mathrm{oo}}^{-1} D_{\mathrm{oe}}, \tag{1.26}$$

and the solution of eqn (1.23) thus amounts to solving a system in the space of quark fields on the even sublattice. The condition number of $\hat{D}$ is usually less than half the one of D and even-odd preconditioning consequently leads to an acceleration of the solver by a factor 2 to 3 or so.

Other preconditioners used in lattice QCD are the successive symmetric overrelaxation (SSOR) preconditioner (Frommer *et al.*, 1994) and a domain-decomposition preconditioner based on the Schwarz alternating procedure (Lüscher, 2004). The latter is an example of an "expensive" preconditioner, whose implementation involves an iterative procedure and is therefore inexact to some extent. Inaccuracies at this level of the algorithm are however not propagated to the final results if the GCR solver is used for the preconditioned equation (1.23). This algorithm actually always finds the best approximation to the solution in the space generated by applying the preconditioner to the residues of the previous solutions. For the same reason, the GCR algorithm is also safe of rounding errors.

1.3 Low-mode deflation

The low modes of the Dirac operator are intimately related to the spontaneous breaking of chiral symmetry and therefore play a special rôle in QCD. Treating them separately from the other modes seems appropriate from the physical point of view and is recommended for technical reasons at small quark masses.

1.3.1 Textbook deflation

In the case of the hermitian system,

$$A\psi = \eta, \qquad A = D^\dagger D, \tag{1.27}$$

there exists an orthonormal basis of eigenvectors v_k, $k = 1, 2, 3, \ldots$, such that

$$Av_k = \alpha_k v_k, \qquad 0 \leq \alpha_1 \leq \alpha_2 \leq \ldots \tag{1.28}$$

The action on any quark field ψ of the orthonormal projector P to the N lowest modes is then given by

$$P\psi = \sum_{k=1}^{N} v_k (v_k, \psi). \tag{1.29}$$

Since P commutes with A, the linear system (1.27) splits into the decoupled equations

$$A_\parallel \psi_\parallel = \eta_\parallel, \qquad \psi_\parallel = P\psi, \tag{1.30}$$

$$A_\perp \psi_\perp = \eta_\perp, \qquad \psi_\perp = (1-P)\psi, \tag{1.31}$$

where $A_\parallel = PAP$ and $A_\perp = (1-P)A(1-P)$ are, respectively, referred to as the "little operator" and the "deflated operator".

If the eigenvectors $v_1, \ldots, v_N$ are known, and if there are no zero-modes, the solution of the little system (1.30) can be obtained exactly through

$$\psi_\parallel = \sum_{k=1}^{N} \frac{1}{\alpha_k} v_k (v_k, \eta). \tag{1.32}$$

The deflated system (1.31), on the other hand, can only be solved iteratively using the CG algorithm, for example. With respect to the full system, the associated condition number

$$\kappa(A_\perp) = \frac{\alpha_1}{\alpha_{N+1}} \kappa(A) \tag{1.33}$$

is however reduced and one therefore expects the solver to be accelerated by the factor $(\alpha_{N+1}/\alpha_1)^{1/2}$ or so.

The deflation of the hermitian system (1.27) along these lines is straightforward to implement, but the method tends to be limited to small lattices, because the computer time required for the calculation of the low eigenvectors grows rapidly with the lattice volume. In the past few years, it was nevertheless further developed and improved in various directions (for a review, see Wilcox (2007), for example).

1.3.2 The Banks–Casher relation

In the continuum theory (which is considered here for simplicity), the eigenvalues of the Dirac operator D in presence of a given gauge field are of the form $m + i\lambda_k$, where m denotes the quark mass and $\lambda_k \in \mathbb{R}$, $k = 1, 2, 3, \ldots$, the eigenvalues of the massless operator. The associated average spectral density,

$$\rho(\lambda, m) = \frac{1}{V} \sum_{k=1}^{\infty} \langle \delta(\lambda - \lambda_k) \rangle \,, \tag{1.34}$$

is conventionally normalized by the space-time volume V so that it has a meaningful infinite-volume limit.

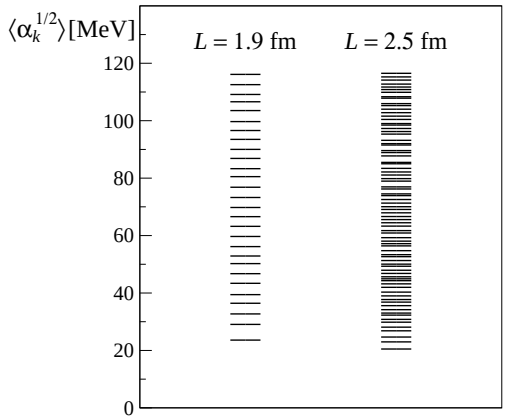


Fig. 1.5 Expectation value of the eigenvalues of $(D^\dagger D)^{1/2}$ below 116 MeV in O(a)-improved two-flavour QCD on a $2L \times L^3$ lattice at two values of L. Both spectra were obtained at lattice spacing $a = 0.08$ fm and renormalized sea-quark mass $m = 26$ MeV. From the smaller to the larger lattice, the number of modes per MeV increases approximately as predicted by the Banks–Casher relation.

In a now famous paper, Banks and Casher (1980) showed many years ago that the density at the origin,

$$\lim_{\lambda\to 0}\lim_{m\to 0}\lim_{V\to\infty}\rho(\lambda, m) = \frac{1}{\pi}\Sigma, \tag{1.35}$$

is proportional to the quark condensate

$$\Sigma = -\lim_{m\to 0}\lim_{V\to\infty}\langle\bar{u}u\rangle, \qquad u\text{: up-quark field}, \tag{1.36}$$

in the chiral limit. The spontaneous breaking of chiral symmetry in QCD is thus linked to the presence of a non-zero density of eigenvalues at the low end of the spectrum of the Dirac operator.

Since the eigenvalues $\alpha_k = m^2 + \lambda_k^2$ of $D^\dagger D$ are simply related to the eigenvalues of D, the Banks–Casher relation (1.35) immediately leads to the estimate

$$\nu(M, m) \simeq \frac{2}{\pi}\Lambda\Sigma V, \qquad \Lambda^2 = M^2 - m^2, \tag{1.37}$$

for the number of low modes of $D^\dagger D$ with eigenvalues $\alpha_k \le M^2$. If one sets $m = 0$, $M = 100$ MeV and $\Sigma = (250\,\text{MeV})^3$, for example, and considers a space-time volume of size $V = 2L^4$, the mode numbers are estimated to be 21, 106 and 336 for $L = 2$, 3 and 4 fm, respectively. As illustrated by Fig. 1.5, the low-mode condensation is readily observed in numerical simulations of lattice QCD.

Since $\nu(M, m)$ increases proportionally to space-time volume V, an effective deflation of the Dirac equation requires O(V) modes to be deflated. The associated computational effort increases like V^2 and straightforward deflation consequently tends to become inefficient or impractical at large volumes. However, as explained in the following, the V^2-problem can be overcome using inexact deflation (Giusti *et al.*, 2003) and domain-decomposed deflation subspaces (Lüscher, 2007*a*).

1.3.3 Deflation w/o eigenvectors

Returning to the lattice Dirac equation (1.1), a more general form of deflation will now be described, which starts from an unspecified set $\phi_1, \ldots, \phi_N$ of N orthonormal quark fields. The linear space $\mathcal{S}$ spanned by these fields will play the rôle of the deflation subspace, but is not assumed to be an eigenspace of the Dirac operator.

The action on any quark field ψ of the orthogonal projector P to $\mathcal{S}$ is given by

$$P\psi = \sum_{k=1}^{N} \phi_k \, (\phi_k, \psi) \, . \tag{1.38}$$

As before, the restriction PDP of the Dirac operator to $\mathcal{S}$ will be referred to as the "little Dirac operator". Its action is encoded in the complex $N \times N$ matrix

$$A_{kl} = (\phi_k, D\phi_l) \, , \quad k, l = 1, \ldots, N, \tag{1.39}$$

through

$$PDP\psi = \sum_{k,l=1}^{N} \phi_k A_{kl} \, (\phi_l, \psi) \, . \tag{1.40}$$

A technical assumption made in the following is that A (and thus PDP as an operator acting in $\mathcal{S}$) is invertible.

The form of inexact deflation discussed below is based on the projectors

$$P_L = 1 - DP(PDP)^{-1}P, \tag{1.41}$$

$$P_R = 1 - P(PDP)^{-1}PD. \tag{1.42}$$

It is not difficult to prove that

$$P_L^2 = P_L, \qquad P_R^2 = P_R, \tag{1.43}$$

$$PP_L = P_R P = (1 - P_L)(1 - P) = (1 - P)(1 - P_R) = 0, \tag{1.44}$$

$$P_L D = D P_R. \tag{1.45}$$

In particular, the operator P_L projects any quark field to the orthogonal complement $\mathcal{S}^{\perp}$ of the deflation subspace. Note, however, that P_L and P_R are not hermitian and therefore not ordinary orthogonal projectors.

The Dirac equation (1.1) may now be split into two decoupled equations,

$$D\psi_{\parallel} = \eta_{\parallel}, \qquad D\psi_{\perp} = \eta_{\perp}, \tag{1.46}$$

where

$$\psi_{\parallel} = (1 - P_R)\psi, \qquad \psi_{\perp} = P_R \psi, \tag{1.47}$$

$$\eta_{\parallel} = (1 - P_L)\eta, \qquad \eta_{\perp} = P_L \eta. \tag{1.48}$$

The use of a different projector for the splitting of the solution ψ and the source η is entirely consistent in view of the commutator relation (1.45) and merely reflects the fact that the Dirac operator is not hermitian.

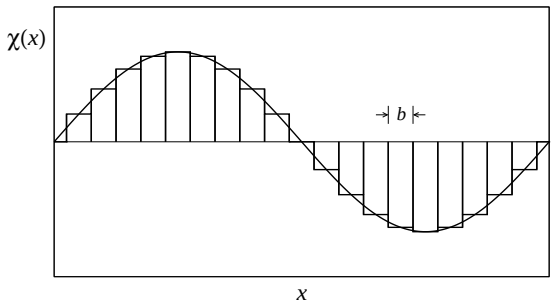


Fig. 1.6 The low modes $\chi(x)$ of the lattice Dirac operator in the free-quark theory are plane waves with small momenta p. These can be well approximated in the norm by functions that are constant on blocks of lattice points if the block size b satisfies $bp \ll 1$. Efficient deflation subspaces can thus be constructed using fields that are discontinuous and therefore far from being approximate eigenmodes of the Dirac operator.

The solution of the little system,

$$\psi_{\|} = P(PDP)^{-1}P\eta = \sum_{k,l=1}^{N} \phi_k (A^{-1})_{kl} \, (\phi_l, \eta) \, , \tag{1.49}$$

is easily found, but the other equation can only be solved using an iterative procedure such as the GCR algorithm. However, the operator on the left of the equation is the deflated operator

$$\hat{D} = DP_R = P_L D(1-P), \tag{1.50}$$

which acts on quark fields in $\mathcal{S}^{\perp}$ and which may have a much smaller condition number than D, particularly so at small quark masses. An acceleration of the computation is then achieved since the solution is obtained in fewer iterations than in the case of the unmodified Dirac equation.

The condition number $\kappa(\hat{D}) = \|\hat{D}\|\|\hat{D}^{-1}\|$ of $\hat{D}$ depends on the quark mass mainly through the factor

$$\|\hat{D}^{-1}\| = \|(1-P)D^{-1}(1-P)\| \leq \|(1-P)(D^{\dagger}D)^{-1}(1-P)\|^{1/2}. \tag{1.51}$$

It is then quite obvious that $\kappa(\hat{D})$ will be much smaller than $\kappa(D)$ if $1-P$ effectively "projects away" the low modes of $D^{\dagger}D$, i.e. if they are well approximated by the deflation subspace. However, contrary to what may be assumed, this requirement is fairly weak and does not imply that the deflation subspace must be spanned by approximate eigenvectors of $D^{\dagger}D$ (see Fig. 1.6).

1.3.4 Domain-decomposed deflation subspaces

In the following, a decomposition of the lattice into rectangular blocks Λ such as the one shown in Fig. 1.7 will be considered. A special kind of deflation subspace $\mathcal{S}$ is then

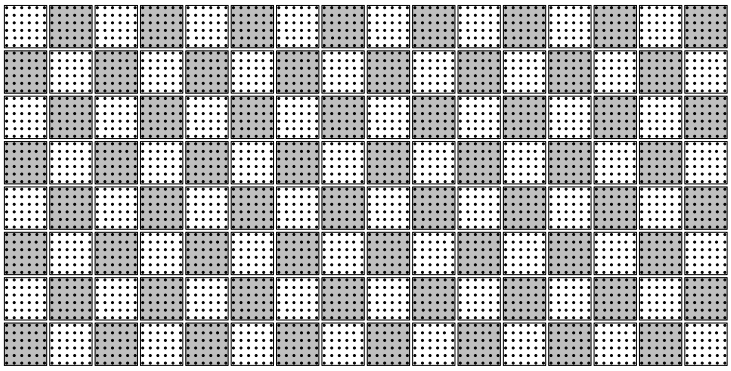


Fig. 1.7 Two-dimensional view of a 96×48^3 lattice, divided into 8192 non-overlapping blocks of size 6^4.

obtained by choosing a set of orthonormal fields $\phi_1^\Lambda, \ldots, \phi_{N_s}^\Lambda$ on each block Λ and by taking $\mathcal{S}$ to be the linear span of all these fields. The associated projector is

$$P = \sum_\Lambda P_\Lambda, \qquad P_\Lambda \psi = \sum_{k=1}^{N_s} \phi_k^\Lambda \left(\phi_k^\Lambda, \psi\right). \tag{1.52}$$

Evidently, this construction fits the general scheme discussed in Section 1.3.3 except perhaps for the labeling of the fields that span the deflation subspace.

The size of the blocks Λ is a tunable parameter of domain-decomposed subspaces. Usually the blocks are taken to be fairly small (4^4, 6^4 or 8×4^3, for example), but the exact choice should eventually be based on the measured performance of the deflated solver. A key feature of domain-decomposed subspaces is the fact that their dimension is proportional to the volume V of the lattice, while the application of the projector P to a given quark field requires only $\mathrm{O}(N_s V)$ floating-point operations (and not $\mathrm{O}(N_s V^2)$ operations as would normally be the case). Such subspaces may thus allow the V^2-problem to be overcome, provided high deflation efficiencies can be achieved for some volume-independent number N_s of block modes.

In the case of the free-quark theory, Fig. 1.6 suggests that this strategy will work out if the block modes are chosen to be constant. Since quark fields have 12 complex components, one needs $N_s = 12$ such modes per block. In presence of an arbitrary gauge field, the choice of the block modes is less obvious, however, because the notion of smoothness ceases to have a well-defined meaning.

1.3.5 Local coherence & subspace generation

It is helpful to note at this point that the deflation deficits

$$\|(1 - P)\chi\|^2 = \sum_\Lambda \|(1 - P_\Lambda)\chi\|^2 \tag{1.53}$$

of the low modes χ of $D^\dagger D$ can only be small if they are small on all blocks Λ. Since the local deflation subspace has fixed dimension N_s, and since the number of low modes is

Fig. 1.8 When restricted to a small block of lattice points, the O(V) low modes of the Dirac operator tend to align to a relatively low-dimensional linear space, a property referred to as local coherence.

proportional to V and thus tends to be much larger than N_s, this condition cannot in general be met unless the low modes happen to collapse to a lower dimensional space on each block, i.e. unless they are "locally coherent" (see Fig. 1.8).

The local coherence of the low modes is numerically well established. On a 64×32^3 lattice with spacing $a = 0.08$ fm, for example, deflation deficits as small as a few percent are achieved when using 4^4 blocks and $N_s = 12$ block fields. Moreover, N_s does not need to be adjusted when the lattice volume increases. So far, however, no theoretical explanation of why the modes are locally coherent has been given. In particular, it is unclear whether the property has anything to do with chiral symmetry.

Efficient domain-decomposed deflation subspaces can now be constructed fairly easily (Lüscher, 2007*a*). One first notes that any quark field ψ satisfying $\|D\psi\| \leq M\|\psi\|$ for some sufficiently small value of M (say, $M = 100$ MeV) is well approximated by a linear combination of low modes and is therefore locally coherent with these. By generating a set $\psi_1, \ldots, \psi_{N_s}$ of independent fields of this kind, using inverse iteration, for example, and by applying the Gram–Schmidt orthonormalization process to the projected fields

$$\psi_k^\Lambda(x) = \begin{cases} \psi_k(x) & \text{if } x \in \Lambda, \\ 0 & \text{otherwise,} \end{cases} \tag{1.54}$$

one thus obtains a basis $\phi_1^\Lambda, \ldots, \phi_{N_s}^\Lambda$ of block fields with large projections to the low modes [2]. The generation of the deflation subspace along these lines requires a modest amount of computer time and certainly far less than would be needed for an approximate computation of the low modes of the Dirac operator. Moreover, since N_s can be held fixed, the computational effort scales like V rather than V^2.

1.3.6 Solving the deflated system

Once the deflation subspace is constructed, the deflated equation $D\psi_\perp = \eta_\perp$ can in principle be solved using the GCR algorithm with D replaced by $\hat{D} = P_L D$. However, the application of the projector P_L requires the little system to be solved for a given source field, which is not a small task in general. Note that the little Dirac operator

[2]The deflation subspace can alternatively be generated "on the fly" while solving the Dirac equation, exploiting the fact that the residue of an approximate solution tends to align to the low modes of the Dirac operator (Brannick *et al.*, 2008; Clark *et al.*, 2008*a*).

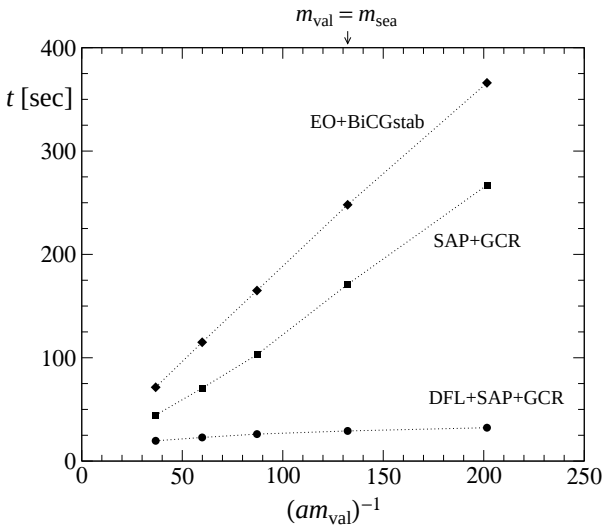


Fig. 1.9 Computer time needed for the solution of the O(a)-improved Wilson–Dirac equation in two-flavour QCD on a 64×32^3 lattice with spacing $a = 0.08$ fm. In these tests, the sea-quark mass $m_{\rm sea}$ was 26 MeV, the valence-quark mass $m_{\rm val}$ ranged from about 15 to 90 MeV and the relative residue of the solution was required to be 10^{-10}. All timings were taken on a PC cluster with 64 (single-core) processors.

acts on fields on the block lattice with N_s complex components. An exact solution of the little system is therefore not practical on large lattices.

In the case of the Wilson–Dirac operator and its relatives, the little Dirac operator has only nearest-neighbour couplings among the blocks. The solution of the little system may then be obtained iteratively using the even-odd preconditioned GCR algorithm, for example. The effort required for the solution of the little system is nevertheless not completely negligible and it is advisable to consider solving the deflated right-preconditioned system

$$P_L D R \phi = \eta_\perp, \qquad \psi_\perp = P_R R \phi, \tag{1.55}$$

instead of the deflated equation directly, the operator R being a suitable preconditioner for D. A preconditioner that has been used in this context is the Schwarz alternating procedure (Lüscher, 2004). The important point to note is that the preconditioner tends to reduce the high-mode components of the residue of the current approximate solution, while the low-mode component of the residue is projected away by the projector P_L. Deflation and right-preconditioning thus tend to complement one another.

The performance figures plotted in Fig. 1.9 show that local deflation works very well in lattice QCD. In this study, the block size was taken to be 4^4 and N_s was set to 20. With respect to the even-odd preconditioned BiCGstab algorithm (points labeled EO+BiCGstab in the figure), the deflated Schwarz-preconditioned GCR algorithm (DFL+SAP+GCR) achieves an acceleration by more than an order of magnitude at the smallest quark masses considered.

On other lattices, the deflated Schwarz-preconditioned solver for the O(a)-improved Wilson–Dirac equation performs as well as in the case reported in Fig. 1.9. In particular, the accumulated experience unambiguously shows that the algorithm overcomes the V^2-problem and that the critical slowing down towards the chiral limit, which previously hampered quark-propagator computations, is nearly eliminated.

2
Simulation algorithms

Lattice QCD simulations are based on Markov chains and the concept of importance sampling. More specifically, most large-scale simulations performed today rely on some variant of the so-called Hybrid Monte Carlo algorithm (Duane *et al.*, 1987). An exception to this rule are simulations of the pure SU(3) gauge theory, where link-update algorithms are usually preferred for reasons of efficiency.

QCD simulation algorithms have a long history and incorporate many ideas and improvements. Some important deficits remain, however, and further progress in algorithms will no doubt be required to be able to perform accurate simulations of a significantly wider range of lattices than is possible at present.

2.1 Importance sampling

2.1.1 Statistical interpretation of the functional integral

The quark fields in the QCD functional integral take values in a Grassmann algebra. So far no practical method has been devised that would allow such fields to be simulated directly. The theory simulated is therefore always the one obtained after integrating out the quark fields.

In the case of two-flavour lattice QCD with mass-degenerate Wilson quarks, the partition function then assumes the form

$$\mathcal{Z} = \int \mathrm{D}[U] \left\{\det D(U)\right\}^2 \mathrm{e}^{-S_\mathrm{g}(U)}, \qquad \mathrm{D}[U] = \prod_{x,\mu} \mathrm{d}U(x,\mu), \tag{2.1}$$

where $D(U)$ denotes the massive Wilson–Dirac operator in presence of the gauge field U, $S_\mathrm{g}(U)$ the gauge action and $\mathrm{d}U(x,\mu)$ the SU(3)-invariant integration measure for the link variable $U(x,\mu)$. Since $D^\dagger = \gamma_5 D \gamma_5$, the quark determinant $\det D$ is real and the product

$$p(U) = \frac{1}{\mathcal{Z}} \left\{\det D(U)\right\}^2 \mathrm{e}^{-S_\mathrm{g}(U)} \tag{2.2}$$

is therefore a normalized probability density on the space of all gauge fields. The physics described by the theory is eventually extracted from expectation values

$$\langle \mathcal{O} \rangle = \int \mathrm{D}[U]\, p(U) \mathcal{O}(U) \tag{2.3}$$

of observables $\mathcal{O}(U)$ such as a Wilson loop or a quark-line diagram. From this point of view, lattice QCD thus looks like a classical statistical system, where the states (the

gauge-field configurations) occur with a certain probability and where one is interested in the expectation values of some properties of the states.

The simulation algorithms discussed in the following depend on the existence of a probabilistic representation of the theory. In particular, they do not apply to two-flavour QCD with non-degenerate quark masses and three-flavour QCD unless the product of the quark determinants is guaranteed to be non-negative (as is the case if the lattice Dirac operator preserves chiral symmetry).

2.1.2 Representative ensembles

Representative ensembles $\{U_1, \ldots, U_N\}$ of gauge fields are obtained by choosing the fields randomly with probability $\mathrm{D}[U]\, p(U)$, i.e. such that

$$\text{no. of fields in } \mathfrak{R} = \int_{\mathfrak{R}} \mathrm{D}[U]\, p(U) + \mathrm{O}(N^{-1/2}) \tag{2.4}$$

for any open region $\mathfrak{R}$ in field space, the term of order $N^{-1/2}$ being a statistical error that depends on $\mathfrak{R}$ and the generated ensemble of fields. The high-probability regions thus contain many fields U_i and are therefore sampled well, while in other areas of field space there may be only a few fields or none at all (if N is not astronomically large).

Given a representative ensemble of fields, the expectation values of the observables of interest can be estimated through

$$\langle \mathcal{O} \rangle = \frac{1}{N} \sum_{i=1}^{N} \mathcal{O}(U_i) + \mathrm{O}(N^{-1/2}). \tag{2.5}$$

A bit surprising may be the fact that results with small statistical errors can often be obtained in this way even if the ensemble contains only 100 or perhaps 1000 field configurations. Naive estimates, taking the dimension of field space into account, actually suggest that an accurate numerical evaluation of the functional integral requires some k^{32n} configurations, where n is the number of lattice points and k at least 10 or so.

The apparent paradox is resolved by noting that small ensembles of field configurations can only capture some aspects of the theory. That is, one should not expect to obtain the expectation values of all possible observables with small statistical errors. Typically any quantity sensitive to the correlations of the field variables at large distances tends to have large statistical errors, sometimes to the extent that the results of the computation are completely useless.

2.1.3 Translation symmetry and the infinite-volume limit

In order to minimize finite-volume effects, periodic boundary conditions are usually imposed on all fields, an exception being the quark fields, which are often taken to be anti-periodic in time. Since the translation symmetry of the theory is preserved by this choice of boundary conditions, representative ensembles of gauge fields are expected "to look the same" in distant regions of a large lattice.

The meaning of this statement is best explained by dividing the lattice into blocks Λ, as in Section 1.3.4, and by considering an extensive quantity

$$\mathcal{E} = \sum_x \mathcal{O}(x), \tag{2.6}$$

where $\mathcal{O}(x)$ is some local gauge-invariant field. Translation symmetry then implies that the contributions $\langle \mathcal{O}_\Lambda \rangle$ to the sum

$$\langle \mathcal{E} \rangle = \sum_\Lambda \langle \mathcal{O}_\Lambda \rangle, \qquad \mathcal{O}_\Lambda = \sum_{x \in \Lambda} \mathcal{O}(x), \tag{2.7}$$

are all equal. Moreover, the statistical fluctuations of $\mathcal{O}_\Lambda$ and $\mathcal{O}_{\Lambda'}$ are practically uncorrelated and tend to cancel one another in the sum (2.7) if the blocks Λ and Λ' are separated by a distance larger than the range of the connected correlation function of $\mathcal{O}(x)$. For a fixed ensemble size, the statistical error of the density $\langle \mathcal{E} \rangle / V$ therefore decreases like $V^{-1/2}$ for $V \to \infty$, i.e. the statistics is effectively multiplied by a factor proportional to V.

The discussion also illustrates the fact that the efficiency of importance sampling depends on the observable considered and on how its expectation value is calculated. Rather than from eqn (2.6), one might actually start from the identity $\langle \mathcal{E} \rangle = V \langle \mathcal{O}(0) \rangle$, in which case the information contained in the field ensemble away from the origin $x = 0$ remains unused. Both calculations yield the correct expectation value, but only the first profits from the available data on the full lattice and consequently obtains the result with much better statistical precision.

2.1.4 Simulating gaussian distributions

Representative ensembles of fields can be easily constructed in the case of a complex field $\phi(x)$ distributed according to the gaussian probability density

$$p_A(\phi) \propto \exp\Big\{ - \sum_x \phi(x)^\dagger (A\phi)(x) \Big\}, \tag{2.8}$$

where A is a hermitian, strictly positive linear operator. If one sets

$$A = -\Delta + m_0^2, \qquad \Delta\text{: lattice laplacian}, \tag{2.9}$$

for example, the theory describes a free scalar field with bare mass m_0. Another possible choice of A is

$$A = (DD^\dagger)^{-1}, \qquad D\text{: lattice Dirac operator}. \tag{2.10}$$

The field $\phi(x)$ must be a pseudo-fermion field in this case, i.e. a complex-valued and therefore bosonic quark field (see Section 2.5.1).

A representative ensemble of fields $\phi_1, \ldots, \phi_N$ may be generated for any gaussian distribution by randomly choosing a set $\chi_1, \ldots, \chi_N$ of fields with normal distribution $p_1(\chi)$ and by setting

$$\phi_i(x) = (B\chi_i)(x), \qquad i = 1, \ldots, N, \tag{2.11}$$

where B is an operator satisfying

$$A = (BB^\dagger)^{-1}. \tag{2.12}$$

In the case of the pseudo-fermion fields, for example, one can simply take $B = D$, while a possible choice in the case of the free scalar field is

$$B = (-\Delta + m_0^2)^{-1/2}. \tag{2.13}$$

It is then straightforward to check that the fields ϕ_i generated in this way are correctly distributed.

Since the fields $\chi_i(x)$ are normally distributed, their components at different lattice points are decoupled and can be drawn randomly one after another. The generation of random numbers on a computer is a complicated subject, however, with many open ends and a vast literature (for an introduction, see Knuth (1997), for example). For the time being, one of the simulation-quality random number generators included in the GNU Scientific Library (`http://www.gnu.org/software/gsl`) may be used, among them the `ranlux` generator (Lüscher, 1994; James, 1994), which is based on a strongly chaotic dynamical system and thus comes with some theoretical understanding of why the generated numbers are random. An efficient ISO C code for this generator can be downloaded from `http://cern.ch/luscher/ranlux`.

2.2 Markov chains

Gaussian distributions are an exceptionally simple case where representative ensembles of fields can be generated instantaneously. In general, however, representative ensembles are generated through some recursive procedure (a Markov process) which obtains the field configurations one after another according to some stochastic algorithm.

This section is devoted to a theoretical discussion of such Markov processes. Rather than QCD or the SU(3) gauge theory, an abstract discrete system will be considered in order to avoid some technical complications, which might obscure the mechanism on which Markov-chain simulations are based. The discrete system is left unspecified, but is assumed to have the following properties:

(a) *There is a finite number n of states s.*

(b) *The equilibrium distribution $P(s)$ satisfies $P(s) > 0$ for all s and $\sum_s P(s) = 1$.*

(c) *The observables are real-valued functions $\mathcal{O}(s)$ of the states s.*

One is then interested in calculating the expectation values

$$\langle \mathcal{O} \rangle = \textstyle\sum_s \mathcal{O}(s) P(s) \tag{2.14}$$

of the observables $\mathcal{O}(s)$.

2.2.1 Transition probabilities

A Markov chain is a random sequence $s_1, s_2, s_3, \ldots, s_N$ of states, where s_k is obtained from s_{k-1} through some stochastic algorithm. The chain thus depends on the initial state s_1 and the transition probability $T(s \to s')$ to go from the current state s to the next state s'. In the following, the basic idea is to choose the latter such that the Markov chain provides a representative ensemble of states for large N. The expectation value of any observable $\mathcal{O}(s)$ is then given by

$$\langle \mathcal{O} \rangle = \frac{1}{N} \sum_{k=1}^{N} \mathcal{O}(s_k) + \mathrm{O}(N^{-1/2}), \tag{2.15}$$

where the error term is dominated by the random fluctuations of the chain.

When trying to construct such transition probabilities, one may be guided by the following plausible requirements:

1. $T(s \to s') \geq 0$ *for all* s, s' *and* $\sum_{s'} T(s \to s') = 1$ *for all* s.
2. $\sum_s P(s) T(s \to s') = P(s')$ *for all* s'.
3. $T(s \to s) > 0$ *for all* s.
4. *If* S *is a non-empty proper subset of states, there exist two states* $s \in S$ *and* $s' \notin S$ *such that* $T(s \to s') > 0$.

Property 1 merely guarantees that $T(s \to s')$ is a probability distribution in s' for any fixed s, while property 2 says that the equilibrium distribution should be preserved by the update process. The other properties ensure that the Markov process does not get trapped in cycles (property 3, referred to as "aperiodicity") or in subsets of states (property 4, "ergodicity").

Later it will be shown that any transition probability $T(s \to s')$ satisfying $1-4$ generates Markov chains that simulate the system in the way explained above. These properties are thus sufficient to guarantee the correctness of the procedure.

2.2.2 The acceptance-rejection method

For illustration, an explicit example of a valid transition probability will now be constructed. In order to simplify the notation a little bit, the states are labeled from 0 to $n-1$ and are thought to be arranged on a circle so that the neighbours of the state i are the states $i \pm 1 \bmod n$. The construction then starts from the transition probability

$$T_0(i \to j) = \begin{cases} \frac{1}{3} & \text{if } j = i \text{ or } j = i \pm 1 \bmod n, \\ 0 & \text{otherwise,} \end{cases} \tag{2.16}$$

which generates a random walk on the circle. This transition probability satisfies $1-4$, the equilibrium distribution being the flat distribution $P_0(i) = 1/n$.

Now if the equilibrium distribution is not flat, as the one shown in Fig. 2.1, a valid transition probability is given by (Metropolis *et al.*, 1953)

$$T(i \to j) = T_0(i \to j) P_{\mathrm{acc}}(i,j) + \delta_{ij} \textstyle\sum_k T_0(i \to k) \left(1 - P_{\mathrm{acc}}(i,k)\right), \tag{2.17}$$

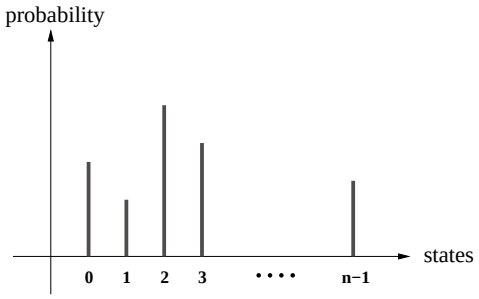


Fig. 2.1 Example of an equilibrium probability distribution of the abstract discrete system considered in this section. The update algorithm implementing the transition probability (2.17) generates a random walk in the space of states, where, in each step, one moves from the current state i to one of its neighbours j with probability $\frac{1}{3}P_{\rm acc}(i,j)$ or else stays at i.

where

$$P_{\rm acc}(i,j) = \min\{1, P(j)/P(i)\} \tag{2.18}$$

is the so-called acceptance probability. In other words, starting from the current state i, the next state j is proposed with probability $T_0(i \to j)$. A random number $r \in [0,1]$ is then chosen, with uniform distribution, and j is accepted if $P(j) \geq rP(i)$. If the proposed state is not accepted, the next state is taken to be the state i.

The acceptance-rejection method is widely used in various incarnations. In general, the challenge is to find an a priori transition probability $T_0(i \to j)$ where the proposed states are accepted with high probability, as otherwise the simulation will be very slow and therefore inefficient. The transition probability (2.17) incidentally satisfies

$$P(i)T(i \to j) = P(j)T(j \to i) \quad \text{for all} \quad i,j, \tag{2.19}$$

a property referred to as "detailed balance". In the literature, detailed balance is sometimes required for acceptable algorithms, but the condition is quite strong and not needed to ensure the correctness of the simulation.

2.2.3 Implications of properties 1–4

In the following, it is assumed that $T(s \to s')$ is a given transition probability satisfying the conditions 1–4 listed in Section 2.2.1. The associated Markov process is then shown to have certain mathematical properties, which will allow, in Section 2.2.4, to determine its asymptotic behaviour at large times.

Let $\mathcal{H}$ be the linear space of real-valued functions $f(s)$ defined on the set of all states s. A useful norm of such functions is given by

$$\|f\|_1 = \textstyle\sum_s |f(s)|. \tag{2.20}$$

The transition probability $T(s \to s')$ defines a linear operator T in $\mathcal{H}$ through

$$(Tf)(s') = \textstyle\sum_s f(s)T(s \to s'). \tag{2.21}$$

Note that, by properties 1 and 2, the equilibrium distribution $P(s)$ has unit norm and is an eigenfunction of T with eigenvalue 1.

Lemma 2.1 *For all $f \in \mathcal{H}$, the bound $\|Tf\|_1 \leq \|f\|_1$ holds. Moreover, if $Tf = f$, there exists $c \in \mathbb{R}$ such that $f(s) = cP(s)$ for all states s.*

Proof: Any given function $f \in \mathcal{H}$ may be decomposed into positive and negative parts according to

$$f(s) = f_+(s) - f_-(s), \qquad f_\pm(s) = \tfrac{1}{2}\{|f(s)| \pm f(s)\} \geq 0. \tag{2.22}$$

Property 1 then implies

$$\|Tf_\pm\|_1 = \textstyle\sum_{s'} \big|\sum_s f_\pm(s) T(s \to s')\big| = \sum_s f_\pm(s) \sum_{s'} T(s \to s') = \|f_\pm\|_1. \tag{2.23}$$

Using the triangle inequality, this leads to

$$\begin{aligned} \|Tf\|_1 &= \|Tf_+ - Tf_-\|_1 \\ &\leq \|Tf_+\|_1 + \|Tf_-\|_1 = \|f_+\|_1 + \|f_-\|_1 = \|f\|_1, \end{aligned} \tag{2.24}$$

which proves the first statement made in the lemma. Moreover, eqn (2.24) shows that the equality $\|Tf\|_1 = \|f\|_1$ holds if and only if Tf_+ and Tf_- have disjoint support.

Now let $f \in \mathcal{H}$ be such that $Tf = f$. The functions Tf_+ and Tf_- must have disjoint support in this case and are both non-negative. Since the decomposition $Tf = f_+ - f_-$ in positive and negative parts is unique, it follows that $Tf_+ = f_+$ and $Tf_- = f_-$. Properties 3 and 4 however imply that the support of a non-negative function grows when T is applied, unless the support is empty or the whole set of states. Either f_+ or f_- must therefore be equal to zero. The function f thus has a definite sign.

Another function $g \in \mathcal{H}$ may now be defined through

$$g(s) = f(s) - cP(s), \qquad c = \textstyle\sum_s f(s). \tag{2.25}$$

This function must also have a definite sign since $Tg = g$. Moreover, $\sum_s g(s) = 0$ by construction, which can only be true if $g = 0$, i.e. if $f(s) = cP(s)$. □

Lemma 2.1 shows that the equilibrium distribution is the only distribution that is stationary under the action of the operator T. Some control over the complementary space

$$\mathcal{H}_0 = \big\{ f \in \mathcal{H} \;\big|\; \textstyle\sum_s f(s) = 0 \big\} \tag{2.26}$$

of non-stationary functions will later be required as well. To this end, it is helpful to introduce a funny scalar product in $\mathcal{H}$,

$$(f, g) = \textstyle\sum_s f(s) P(s)^{-1} g(s), \tag{2.27}$$

and the associated norm $\|f\| = (f, f)^{1/2}$.

Lemma 2.2 *There exists $0 \leq \rho < 1$ such that $\|Tf\| \leq \rho \|f\|$ for all $f \in \mathcal{H}_0$.*

Proof: With respect to the scalar product (2.27), the adjoint $T^\dagger$ of the operator T may be defined and thus also the symmetric operator $\hat{T} = T^\dagger T$. The action of $\hat{T}$ on any function $f \in \mathcal{H}$ is given by

$$(\hat{T}f)(s') = \textstyle\sum_s f(s)\hat{T}(s \to s'), \tag{2.28}$$

$$\hat{T}(s \to s') = \textstyle\sum_r T(s \to r)P(r)^{-1}T(s' \to r)P(s'). \tag{2.29}$$

A moment of thought then shows that $\hat{T}(s \to s')$ is a transition probability satisfying conditions 1–4. In particular, Lemma 2.1 holds for $\hat{T}$ as well.

Since $\hat{T}$ is symmetric, there exists a complete set of eigenfunctions $v_i \in \mathcal{H}$ of $\hat{T}$ with real eigenvalues λ_i $(i = 0, 1, \ldots, n-1)$. Without loss one may assume that

$$\lambda_0 \geq \lambda_1 \geq \ldots \geq \lambda_{n-1}, \qquad (v_i, v_j) = \delta_{ij}. \tag{2.30}$$

Moreover, noting $\lambda_i = (v_i, \hat{T}v_i) = \|Tv_i\|^2 \geq 0$, lemma 2.1 (for $\hat{T}$) implies

$$\lambda_0 = 1, \quad v_0 = P \quad \text{and} \quad \lambda_1 < 1. \tag{2.31}$$

Now since the states v_i, $i \geq 1$, satisfy

$$0 = (v_0, v_i) = \textstyle\sum_s v_i(s), \tag{2.32}$$

they form a basis of the subspace $\mathcal{H}_0$, i.e. $\hat{T}$ maps $\mathcal{H}_0$ into itself and its largest eigenvalue in this subspace is λ_1. As a consequence,

$$\|Tf\|^2 = (f, \hat{T}f) \leq \lambda_1 \|f\|^2 \quad \text{for all} \quad f \in \mathcal{H}_0. \tag{2.33}$$

This proves the lemma and also shows that the smallest possible value of the constant ρ is $\lambda_1^{1/2}$. □

2.2.4 Statistical properties of Markov chains

When analyzing Markov chains, an important conceptual point to note is that one cannot reasonably speak of the statistical properties of a single chain. However, one can ask what the average properties of the generated sequences of states are if many independent chains are considered. In practice, this corresponds to running several simulations in parallel, with different streams of random numbers.

In the following theoretical discussion, it will be assumed that an infinite number of independent simulations of length N have been performed, with the same initial state s_1 and a transition probability $T(s \to s')$ that has properties 1–4. As before, the expectation value of an observable $\mathcal{O}$ will be denoted by $\langle \mathcal{O} \rangle$, while

$$\overline{\mathcal{O}} = \frac{1}{N} \sum_{k=1}^{N} \mathcal{O}(s_k) \tag{2.34}$$

stands for its average over the states $s_1, s_2, \ldots, s_N$ generated in the course of a given simulation. One may then also consider functions $\phi(s_1, \ldots, s_N)$ of these sequences of states and their average $\langle\langle \phi \rangle\rangle$ over the infinitely many parallel simulations.

(a) Probability distribution of the states. The state s_k generated after $k-1$ steps changes randomly from one simulation to another. It is not difficult to show that the probability $P_k(s) = \langle\!\langle \delta_{s s_k} \rangle\!\rangle$ for s_k to coincide with s is given by

$$\begin{aligned} P_k(s) &= \sum_{s_2, s_3, \ldots, s_{k-1}} T(s_1 \to s_2) T(s_2 \to s_3) \ldots T(s_{k-1} \to s) \\ &= (T^{k-1} P_0)(s), \qquad P_0(s) = \delta_{s s_1}. \end{aligned} \tag{2.35}$$

Noting $P_0 = P + f$, $f \in \mathcal{H}_0$, property 2 and lemma 2.2 then imply

$$P_k(s) \underset{k \to \infty}{=} P(s) + \mathrm{O}\big(\mathrm{e}^{-k/\tau}\big), \tag{2.36}$$

where $\tau = -1/\ln \rho > 0$ is the so-called exponential autocorrelation time of the Markov process. The state s_k is thus distributed according to the equilibrium distribution if k is much larger than τ. In particular, after so many steps, there is no memory of the initial state s_1 anymore and one says that the simulation has "thermalized".

(b) Calculation of expectation values. Together with the generated states, the average $\overline{\mathcal{O}}$ of the "measured" values of an observable $\mathcal{O}$ fluctuates randomly about the mean value

$$\langle\!\langle \overline{\mathcal{O}} \rangle\!\rangle = \frac{1}{N} \sum_{k=1}^{N} \langle\!\langle \mathcal{O}(s_k) \rangle\!\rangle = \sum_s \mathcal{O}(s) \frac{1}{N} \sum_{k=1}^{N} P_k(s). \tag{2.37}$$

Recalling eqn (2.36), this formula shows that

$$\langle\!\langle \overline{\mathcal{O}} \rangle\!\rangle = \langle \mathcal{O} \rangle \tag{2.38}$$

if the first $k \gg \tau$ measurements of $\mathcal{O}$ are dropped. Averages of the measured values calculated after thermalization thus coincide with the expectation value $\langle \mathcal{O} \rangle$ up to statistical fluctuations.

(c) Autocorrelation functions. The states s_k in a Markov chain are statistically dependent to some extent, because they are generated one after another according to the transition probability $T(s \to s')$. As a consequence, the measured values of an observable $\mathcal{O}$ are statistically correlated, i.e. the autocorrelation function

$$\Gamma(t) = \langle\!\langle \mathcal{O}(s_k) \mathcal{O}(s_{k+t}) \rangle\!\rangle - \langle\!\langle \mathcal{O}(s_k) \rangle\!\rangle \langle\!\langle \mathcal{O}(s_{k+t}) \rangle\!\rangle \tag{2.39}$$

does not vanish. For $k \gg \tau$ and $t \geq 0$, the autocorrelation function is independent of k and given by

$$\Gamma(t) = \sum_{s_k, s_{k+1}, \ldots, s_{k+t}} P(s_k) \mathcal{O}(s_k) T(s_k \to s_{k+1}) \ldots T(s_{k+t-1} \to s_{k+t}) \mathcal{O}(s_{k+t}) - \langle \mathcal{O} \rangle^2. \tag{2.40}$$

Moreover, noting $P(s)\mathcal{O}(s) = P(s)\langle \mathcal{O} \rangle + f(s)$, $f \in \mathcal{H}_0$, it follows from this expression and lemma 2.2 that $\Gamma(t)$ falls off exponentially, like $\mathrm{e}^{-t/\tau}$, at large separations t. The measured values $\mathcal{O}(s_i)$ and $\mathcal{O}(s_j)$ are thus independently distributed if $|i-j| \gg \tau$.

(d) Statistical fluctuations. The statistical variance of the averages $\overline{\mathcal{O}}$ of an observable $\mathcal{O}$ is, after thermalization, given by

$$\langle\langle(\overline{\mathcal{O}} - \langle\mathcal{O}\rangle)^2\rangle\rangle = \frac{1}{N^2}\sum_{l,j=1}^{N}\Gamma(|l-j|) = \Gamma(0)\frac{2\tau_{\mathcal{O}}}{N} + \mathrm{O}(N^{-2}), \tag{2.41}$$

where

$$\tau_{\mathcal{O}} = \frac{1}{2} + \sum_{t=1}^{\infty}\frac{\Gamma(t)}{\Gamma(0)} \tag{2.42}$$

denotes the so-called integrated autocorrelation time of $\mathcal{O}$. Up to terms of order $N^{-3/2}$, the standard deviation of $\overline{\mathcal{O}}$ from its expectation value $\langle\mathcal{O}\rangle$ is thus

$$\sigma = \sigma_0\left(\frac{2\tau_{\mathcal{O}}}{N}\right)^{1/2}, \qquad \sigma_0 = \langle(\mathcal{O} - \langle\mathcal{O}\rangle)^2\rangle^{1/2}, \tag{2.43}$$

i.e. the statistical error of $\overline{\mathcal{O}}$ decreases proportionally to $N^{-1/2}$. Note that σ_0 is just the standard deviation of $\mathcal{O}$ in equilibrium. In particular, σ_0 is a property of the system rather than of the Markov process. The integrated autocorrelation time $\tau_{\mathcal{O}}$, on the other hand, often strongly depends on the simulation algorithm.

The fact that the measured values of $\mathcal{O}$ are correlated leads to an increase of the variance of $\overline{\mathcal{O}}$ by the factor $2\tau_{\mathcal{O}}$ and thus lowers the efficiency of the simulation. In practice one frequently chooses to measure the observables only on a subsequence of states separated by some fixed distance Δt in simulation time. As long as Δt is not much larger than $2\tau_{\mathcal{O}}$, the depletion of the measurements has no or little influence on the statistical error of $\overline{\mathcal{O}}$ and therefore helps reducing the computational load.

2.3 Simulating the SU(3) gauge theory

The theory of Markov chains developed in Section 2.2 can be extended to non-discrete systems like the pure SU(3) gauge theory on a finite lattice. Markov chains are sequences $U_1, U_2, \ldots, U_N$ of gauge-field configurations in this case, which are generated according to some transition probability. The latter must satisfy certain conditions analogous to those listed in Section 2.2.1 for the discrete system. The mathematics required at this point however tends to be quite heavy and no attempt will be made to prove of the correctness of the procedure (see Tierney (1994), for example).

2.3.1 Transition probability densities

The equilibrium probability density to be simulated is

$$p(U) = \frac{1}{\mathcal{Z}}\mathrm{e}^{-S_{\mathrm{g}}(U)}, \qquad \mathcal{Z} = \int \mathrm{D}[U]\,\mathrm{e}^{-S_{\mathrm{g}}(U)}, \tag{2.44}$$

where the gauge action $S_{\mathrm{g}}(U)$ is assumed to be a bounded function of the gauge field U. Markov processes in this theory are characterized by a transition probability density $T(U \to U')$ that specifies the probability $\mathrm{D}[U']\,T(U \to U')$ for the next configuration to be in the volume element $\mathrm{D}[U']$ at U' when the current configuration is U.

Note that transition probability densities may involve δ-function and other singularities, but the product $\mathrm{D}[U']\,T(U \to U')$ must be a well-defined measure on the field manifold for any given U. The obvious requirements are then

1. $T(U \to U') \geq 0$ *for all* U, U' *and* $\int \mathrm{D}[U']\,T(U \to U') = 1$ *for all* U.
2. $\int \mathrm{D}[U]\,p(U)T(U \to U') = p(U')$ *for all* U'.

Further conditions need to be added, however, in order to guarantee the aperiodicity, the ergodicity and thus the convergence of the Markov process. A sufficient but fairly strong condition is

3. *Every gauge field* V *has an open neighbourhood* $\mathcal{N}$ *in field space such that* $T(U \to U') \geq \epsilon$ *for some* $\epsilon > 0$ *and all* $U, U' \in \mathcal{N}$.

This property ensures that, in every step, the Markov process spreads out in an open neighbourhood of the current field. Moreover, using the compactness of the field manifold, it is possible to show that the process will reach any region in field space in a finite number of steps.

While properties 1–3 guarantee the asymptotic correctness of the simulations, the rigorous upper bounds on the exponential autocorrelation time obtained in the course of the convergence proofs tend to be astronomically large. In practice, simulations of lattice QCD therefore remain an empirical science to some extent, where one cannot claim, with absolute certainty, that the simulation results are statistically correct.

2.3.2 Link-update algorithms

If T_1 and T_2 are two transition probability densities satisfying 1 and 2, so does their composition

$$T(U \to U') = \int \mathrm{D}[V]\,T_1(U \to V)T_2(V \to U'). \tag{2.45}$$

An update step according to the composed transition probability density first obtains the intermediate field V with probability $\mathrm{D}[V]\,T_1(U \to V)$ and then generates U' with probability $\mathrm{D}[U']\,T_2(V \to U')$. Composition allows simulation algorithms for fields to be constructed from elementary transitions, where a single field variable is changed at the time.

Link-update algorithms generate the next gauge field by updating the link variables one after another in some order. The Metropolis algorithm, for example, proceeds as follows (the SU(3) notation is summarized in Section 2.3.6):

(a) *Select a link* (x, μ) *and choose* $X \in \mathfrak{su}(3)$ *randomly in the ball* $\|X\| \leq \epsilon$, *with uniform distribution, where* ϵ *is some fixed positive number.*

(b) *Accept* $U'(x, \mu) = \mathrm{e}^X U(x, \mu)$ *as the new value of the link variable on the selected link with probability* $P_{\mathrm{acc}} = \min\{1, \mathrm{e}^{S_{\mathrm{g}}(U) - S_{\mathrm{g}}(U')}\}$.

(c) *Leave the link variable unchanged if the new value proposed in step (b) is not accepted.*

It is not difficult to write down the transition probability density corresponding to the steps (a)–(c) and to check that it satisfies conditions 1 and 2. Moreover, the complete update cycle, where each link is visited once, satisfies condition 3 as well.

If the gauge action is local, the calculation of the action difference $S_{\rm g}(U') - S_{\rm g}(U)$ in step (b) involves only the field variables residing in the vicinity of the selected link. The computer time required per link update is then quite small. When all factors are taken into account, including the autocorrelation times, a more efficient link-update algorithm is however provided by the combination of the heatbath algorithm discussed below with a number of microcanonical moves (Section 2.3.5).

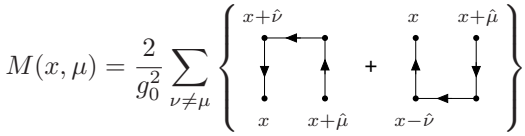
2.3.3 Heatbath algorithm

As a function of the field variable $U(x,\mu)$ residing on a given link (x,μ), the Wilson plaquette action is of the form

$$S_{\rm g}(U) = -\mathrm{Re}\,\mathrm{tr}\{U(x,\mu)M(x,\mu)\} + \ldots, \tag{2.46}$$

where $M(x,\mu)$ and the terms represented by the ellipsis do not depend on $U(x,\mu)$. Up to a constant factor involving the bare gauge coupling g_0, the complex 3×3 matrix

$$M(x,\mu) = \frac{2}{g_0^2}\sum_{\nu\neq\mu}$$



(2.47)

coincides with the "staple sum" of Wilson lines from $x+\hat{\mu}$ to x. Other popular gauge actions, including the $\mathrm{O}(a^2)$-improved Symanzik action, are of the same form except that $M(x,\mu)$ gets replaced by a sum of more complicated Wilson lines.

The heatbath algorithm (Creutz, 1980) is a link-update algorithm, where the new value $U'(x,\mu)$ of the field variable on the selected link is chosen randomly with probability density proportional to $\exp\left(\mathrm{Re}\,\mathrm{tr}\{U'(x,\mu)M(x,\mu)\}\right)$. In other words, the link variable is updated according to its exact distribution in the presence of the other field variables.

While this algorithm fulfills conditions 1–3, it is difficult to implement in practice exactly as described here. However, for gauge group SU(2) (the case considered by Creutz (1980)), the situation is more favourable and there are highly efficient ways to generate random link variables with the required probability distribution (Fabricius and Haan, 1984; Kennedy and Pendleton, 1985).

2.3.4 The Cabibbo–Marinari method

The practical difficulties encountered when implementing the heatbath algorithm in the SU(3) theory can be bypassed as follows (Cabibbo and Marinari, 1982). Let $v \in \mathrm{SU}(2)$ be embedded in SU(3) through

$$v \to V = \begin{pmatrix} v_{11} & v_{12} & 0 \\ v_{21} & v_{22} & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{2.48}$$

A correct one-link update move, $U(x,\mu) \to U'(x,\mu)$, is then obtained by choosing v randomly with probability density proportional to $\exp\left(\mathrm{Re}\,\mathrm{tr}\{VU(x,\mu)M(x,\mu)\}\right)$ and by setting $U'(x,\mu) = VU(x,\mu)$. A moment of thought reveals that the distribution of v is of the same analytic form as the link distribution in the SU(2) theory. The highly efficient methods developed for the latter can thus be used here too.

In order to treat all colour components of the link variables democratically, different embeddings of SU(2) in SU(3) should be used. A popular choice is to perform SU(2) rotations in the $(1,2)$, $(2,3)$ and $(3,1)$ planes in colour space before proceeding to the next link. The ergodicity of the algorithm is then again guaranteed.

2.3.5 Microcanonical moves

The link-update algorithms discussed so far tend to become inefficient when the lattice spacing a is reduced, a phenomenon known as "critical slowing down". Typically the autocorrelation times of physical observables increase approximately like a^{-2}.

Microcanonical algorithms are based on field transformations that preserve the gauge action. Such transformations are valid transitions, satisfying conditions 1 and 2, provided the field integration measure is preserved too. The trajectories in field space generated by a microcanonical algorithm do not involve random changes of direction and are therefore quite different from the random walks performed by the other algorithms. An acceleration of the simulation is then often achieved when microcanonical moves are included in the update scheme.

A microcanonical link-update algorithm for the SU(3) theory is easily constructed following the steps taken in the case of the heatbath algorithm (Creutz, 1987; Brown and Woch, 1987). The link variable $U(x,\mu)$ on the selected link is again updated by applying Cabibbo–Marinari rotations, but the SU(2) matrix v is now set to

$$v = \frac{2w^2}{\mathrm{tr}\{w^\dagger w\}}, \tag{2.49}$$

where w is a 2×2 matrix implicitly defined by

$$\mathrm{Re}\,\mathrm{tr}\{VU(x,\mu)M(x,\mu)\} = \mathrm{tr}\{vw^\dagger\} + \ldots \tag{2.50}$$

and the requirement that w is in SU(2) up to a real scale factor. The existence and uniqueness of w is implied by the reality and linearity properties of the expression on the left of eqn (2.50) (no update is performed if w is accidentally equal to zero).

The transformation $U(x,\mu) \to VU(x,\mu)$ defined in this way preserves the gauge action since $\mathrm{tr}\{vw^\dagger\} = \mathrm{tr}\{w^\dagger\}$, but it is less obvious that it also preserves the link integration measure, because w and therefore v depend on $U(x,\mu)$, i.e. the transformation is non-linear. It is straightforward to show, however, that

$$w|_{U(x,\mu)\to ZU(x,\mu)} = wz^\dagger \tag{2.51}$$

for all $z \in \mathrm{SU}(2)$. Now if $f(U)$ is any integrable function of $U = U(x,\mu)$, the substitution $U \to ZU$ leads to the identity

$$\int \mathrm{d}U f(VU) = \int \mathrm{d}U f(\tilde{V}U), \qquad \tilde{v} = \frac{2wz^\dagger w}{\mathrm{tr}\{w^\dagger w\}}. \tag{2.52}$$

Since this equation holds for any z, it remains valid when integrated over SU(2). Using the invariance of the group integration measures under left- and right-multiplications, one then deduces that

$$\int \mathrm{d}U f(VU) = \int \mathrm{d}z \int \mathrm{d}U f(\tilde{V}U) = \int \mathrm{d}z \int \mathrm{d}U f(Z^\dagger U) = \int \mathrm{d}U f(U), \qquad (2.53)$$

which proves that the transformation $U \to VU$ preserves the link integration measure.

Microcanonical simulation algorithms are not ergodic and must therefore be combined with an ergodic one. A recommended scheme consists in updating all link variables once using the heatbath algorithm and subsequently n times using microcanonical moves. This combination is more efficient than the pure Metropolis or heatbath algorithm and it reportedly has an improved scaling behaviour as a function of the lattice spacing if n is scaled roughly like a^{-1}.

2.3.6 Appendix: SU(3) notation

The Lie algebra $\mathfrak{su}(3)$ of SU(3) consists of all complex 3×3 matrices X that satisfy

$$X^\dagger = -X \quad \text{and} \quad \mathrm{tr}\{X\} = 0. \qquad (2.54)$$

With this convention, the Lie bracket $[X, Y]$ maps any pair X, Y of $\mathfrak{su}(3)$ matrices to another element of $\mathfrak{su}(3)$. Moreover, the exponential series

$$\mathrm{e}^X = 1 + \sum_{k=1}^{\infty} \frac{X^k}{k!} \qquad (2.55)$$

converges to an element of SU(3) (note the absence of factors of i in these and the following formulae).

One can always choose a basis T^a, $a = 1, \ldots, 8$, of $\mathfrak{su}(3)$ such that

$$\mathrm{tr}\{T^a T^b\} = -\tfrac{1}{2}\delta^{ab}. \qquad (2.56)$$

With respect to such a basis, the elements $X \in \mathfrak{su}(3)$ are represented as

$$X = \sum_{a=1}^{8} X^a T^a, \qquad X^a = -2\,\mathrm{tr}\{XT^a\} \in \mathbb{R}. \qquad (2.57)$$

Moreover, the natural scalar product on $\mathfrak{su}(3)$ is given by

$$(X, Y) = \sum_{a=1}^{8} X^a Y^a = -2\,\mathrm{tr}\{XY\}, \qquad (2.58)$$

the associated matrix norm being $\|X\| = (X, X)^{1/2}$. If not specified otherwise, the Einstein summation convention is used for group indices.

2.4 The Hybrid Monte Carlo (HMC) algorithm

The inclusion of the sea quarks in the simulations is difficult, because the quark determinants in the QCD functional integral depend non-locally on the gauge field. In particular, one-link update algorithms would require a computational effort proportional to the square of the lattice volume and are therefore not practical.

The HMC algorithm (Duane *et al.*, 1987) updates all link variables at once and has a much better scaling behaviour with respect to the lattice volume. It will here be explained in general terms for an unspecified (possibly non-local) action $S(U)$, which is assumed to be real and differentiable.

2.4.1 Molecular dynamics

As an intermediate device, the HMC algorithm requires an $\mathfrak{su}(3)$-valued field

$$\pi(x,\mu) = \pi^a(x,\mu)T^a, \qquad \pi^a(x,\mu) \in \mathbb{R}, \tag{2.59}$$

to be added to the theory (the SU(3) notation is as in Section 2.3.6). The new field is interpreted as the canonical momentum of the gauge field, the associated Hamilton function being

$$H(\pi,U) = \tfrac{1}{2}(\pi,\pi) + S(U), \qquad (\pi,\pi) = \sum_{x,\mu} \pi^a(x,\mu)\pi^a(x,\mu). \tag{2.60}$$

Evidently, since

$$\int \mathrm{D}[U]\, \mathcal{O}(U) \mathrm{e}^{-S(U)} = \text{constant} \times \int \mathrm{D}[\pi]\mathrm{D}[U]\, \mathcal{O}(U)\, \mathrm{e}^{-H(\pi,U)}, \tag{2.61}$$

the addition of the momentum field does not affect the physics content of the theory.

In the form (2.61), the theory is reminiscent of the classical statistical systems that describe a gas of molecules. Hamilton's equations[1],

$$\dot{\pi}(x,\mu) = -F(x,\mu), \qquad F^a(x,\mu) = \left.\frac{\partial S(\mathrm{e}^{\omega}U)}{\partial \omega^a(x,\mu)}\right|_{\omega=0}, \tag{2.62}$$

$$\dot{U}(x,\mu) = \pi(x,\mu)U(x,\mu), \tag{2.63}$$

are therefore often referred to as the "molecular-dynamics equations". As usual, the dot on the left of these equations implies a differentiation with respect to time t, which is here a fictitious time unrelated to the time coordinate of space-time. The solutions of the molecular-dynamics equations, $\pi_t(x,\mu)$ and $U_t(x,\mu)$, are uniquely determined by the initial values of the fields at $t = 0$. They may be visualized as trajectories in field space (or, more precisely, in phase space) parameterized by the time t.

[1]The force $F(x,\mu)$ informally coincides with $\partial S(U)/\partial U(x,\mu)$. Derivatives with respect to the link variables however need to be properly defined. According to eqn (2.62), the force field is obtained by substituting $U(x,\mu) \to \exp\{\omega^a(x,\mu)T^a\}U(x,\mu)$, differentiating with respect to the real variables $\omega^a(x,\mu)$ and setting $\omega^a(x,\mu) = 0$ at the end of the calculation.

2.4.2 The HMC strategy

The basic idea underlying the HMC algorithm is to pass from the original theory to the classical system (2.61) and to evolve the fields by integrating the molecular-dynamics equations. Explicitly, the steps leading from the current gauge field $U(x,\mu)$ to the next field $U'(x,\mu)$ are the following:

(a) *A momentum field π is generated randomly with probability density proportional to* $\exp\{-\frac{1}{2}(\pi,\pi)\}$.

(b) *The molecular-dynamics equations are integrated from time $t=0$ to some later time $t=\tau$, taking π and U as the initial values of the fields.*

(c) *The new gauge field U' is set to the field U_τ obtained at time $t=\tau$ through the molecular-dynamics evolution.*

If τ is set to a fixed value, as is usually done, the transition probability density corresponding to the steps (a)–(c) is given by

$$T(U\to U') = \frac{1}{\mathcal{Z}_\pi}\int \mathrm{D}[\pi]\,\mathrm{e}^{-\frac{1}{2}(\pi,\pi)}\prod_{x,\mu}\delta(U'(x,\mu),U_\tau(x,\mu)), \tag{2.64}$$

where the partition function $\mathcal{Z}_\pi$ of the momentum field ensures the correct normalization and the Dirac δ-function is the one appropriate to the gauge-field integration measure.

It is trivial to check that the transition probability density (2.64) satisfies the first of the three conditions listed in Section 2.3.1. The second condition is also fulfilled, but some work is required to show this. An important point to note is that the molecular-dynamics equations are invariant under time reversal $\pi_t, U_t \to -\pi_{\tau-t}, U_{\tau-t}$. The molecular-dynamics evolution $\pi_0, U_0 \to \pi_\tau, U_\tau$ therefore defines an invertible transformation of phase space. Moreover, it preserves the Hamilton function and, by Liouville's theorem, also the phase space integration measure. It follows from these remarks that

$$\begin{aligned}\int \mathrm{D}[U]\,\mathrm{e}^{-S(U)}T(U\to U') &= \frac{1}{\mathcal{Z}_\pi}\int \mathrm{D}[\pi]\mathrm{D}[U]\,\mathrm{e}^{-H(\pi,U)}\prod_{x,\mu}\delta(U'(x,\mu),U_\tau(x,\mu))\\ &= \frac{1}{\mathcal{Z}_\pi}\int \mathrm{D}[\pi_\tau]\mathrm{D}[U_\tau]\,\mathrm{e}^{-H(\pi_\tau,U_\tau)}\prod_{x,\mu}\delta(U'(x,\mu),U_\tau(x,\mu)) = \mathrm{e}^{-S(U')},\end{aligned} \tag{2.65}$$

where the second equation is obtained by performing a change of variables from $\pi, U = \pi_0, U_0$ to π_τ, U_τ. Condition 2 is thus satisfied too.

For sufficiently small τ, the ergodicity of the algorithm (condition 3) can be proved as well. The proof is based on an expansion of π_τ, U_τ in powers of τ, from which one infers that the range of U_τ includes an open neighbourhood of $U = U_0$ when the initial momentum $\pi = \pi_0$ varies over a neighbourhood of the origin. In view of the non-linear nature of the theory, HMC simulations of lattice QCD are however expected to be ergodic at any (non-zero) value of τ, even if one is unable to show this. Ergodicity can,

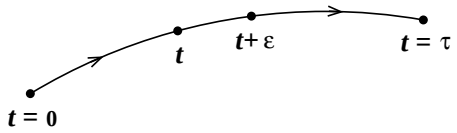


Fig. 2.2 The numerical integration of the molecular-dynamics equations proceeds in time steps of size ϵ. In each step, the fields at time $t+\epsilon$ are computed from the fields at time t and possibly those at earlier times (if a higher-order scheme is used). The integration rule and the step size ϵ should evidently be such that the calculated fields at time $t = n\epsilon$, $n = 0, 1, \ldots, N_0$, closely follow the exact trajectory in phase space.

in any case, always be rigorously ensured by choosing $\tau \in [0, \tau_{\max}]$ randomly from one update step to the next.

2.4.3 Numerical integration of the molecular-dynamics equations

In practice, the molecular-dynamics equations cannot be integrated exactly and one must resort to some numerical integration method. As explained in Section 2.4.4, the integration error can be compensated by including an acceptance-rejection step in the HMC algorithm so that the correctness of the simulation is not compromised.

The numerical integration proceeds by dividing the time interval $[0, \tau]$ in N_0 steps of size ϵ and by applying a discrete integration rule that gives the correct result in the limit $\epsilon \to 0$ (see Fig. 2.2). Considering the Taylor expansions

$$\pi_{t+\epsilon} = \pi_t - \epsilon F|_{U=U_t} + \mathrm{O}(\epsilon^2), \tag{2.66}$$

$$U_{t+\epsilon} = U_t + \epsilon \pi_t U_t + \mathrm{O}(\epsilon^2), \tag{2.67}$$

it is clear that acceptable integration schemes can be built from the elementary operations

$$\mathcal{I}_0(\epsilon): \quad \pi, U \to \pi - \epsilon F, U, \tag{2.68}$$

$$\mathcal{I}_U(\epsilon): \quad \pi, U \to \pi, \mathrm{e}^{\epsilon\pi} U. \tag{2.69}$$

The combination $\mathcal{I}_0(\frac{1}{2}\epsilon)\mathcal{I}_U(\epsilon)\mathcal{I}_0(\frac{1}{2}\epsilon)$, for example, takes π_t, U_t to $\pi_{t+\epsilon}, U_{t+\epsilon}$ up an error of order ϵ^3. The complete integration from time $t = 0$ to time $t = \tau$ then amounts to applying the product

$$\mathcal{J}_0(\epsilon, N_0) = \left\{\mathcal{I}_0(\tfrac{1}{2}\epsilon)\mathcal{I}_U(\epsilon)\mathcal{I}_0(\tfrac{1}{2}\epsilon)\right\}^{N_0}, \qquad \epsilon = \frac{\tau}{N_0}, \tag{2.70}$$

to the initial fields.

The "leap-frog integrator" (2.70) is remarkably simple and has a number of good properties. In particular, the integration is reversible,

$$\mathcal{J}_0(-\epsilon, N_0)\mathcal{J}_0(\epsilon, N_0) = 1, \tag{2.71}$$

and $\mathcal{J}_0(\epsilon, N_0)$ is therefore an invertible mapping of phase space. Starting from the elementary integration steps (2.68) and (2.69), it is also trivial to show that the integrator

preserves the field integration measure $\mathrm{D}[\pi]\mathrm{D}[U]$. Through the numerical integration, these important properties of the molecular-dynamics evolution are thus not lost.

The leap-frog integrator is widely used and appreciated for its simplicity, but there are many other integration schemes that can be employed (Leimkuhler and Reich, 2004; Hairer *et al.*, 2006). In particular, the so-called symplectic integrators all have the good properties mentioned above.

2.4.4 Acceptance-rejection step

For a fixed step size ϵ, the numerical integration of the molecular-dynamics equations normally does not preserve the Hamilton function. In particular, the difference

$$\Delta H(\pi, U) = \{H(\pi_\tau, U_\tau) - H(\pi_0, U_0)\}_{\pi_0=\pi, U_0=U} \tag{2.72}$$

does not vanish in general. The HMC algorithm (as defined through steps (a)–(c) in Section 2.4.2) consequently violates condition 2 if the integration is performed numerically. It is possible to correct for this deficit by replacing step (c) through

(c′) *The new gauge field U' is set to the field U_τ obtained through the integration of the molecular-dynamics equations with probability*

$$P_{\mathrm{acc}}(\pi, U) = \min\bigl\{1, \mathrm{e}^{-\Delta H(\pi,U)}\bigr\}. \tag{2.73}$$

Otherwise, i.e. if the proposed field is rejected, U' is set to U.

The transition probability density of the modified algorithm,

$$T(U \to U') = \frac{1}{\mathcal{Z}_\pi} \int \mathrm{D}[\pi]\, \mathrm{e}^{-\frac{1}{2}(\pi,\pi)} \Big\{ P_{\mathrm{acc}}(\pi, U) \prod_{x,\mu} \delta(U'(x,\mu), U_\tau(x,\mu)) + (1 - P_{\mathrm{acc}}(\pi, U)) \prod_{x,\mu} \delta(U'(x,\mu), U(x,\mu)) \Big\}, \tag{2.74}$$

can then again be shown to have the required properties, for any value of the integration step size ϵ, provided the integrator is reversible and measure-preserving.

The adjustable parameters of the HMC algorithm are then the trajectory length τ, the integration step size ϵ and further parameters of the integration scheme (if any). Evidently, the simulation will be inefficient if the average acceptance rate is low, i.e. if the numerical integration is not very accurate. The acceptance rate must however be balanced against the computer time required for the simulation, which grows roughly linearly with the step number $N_0 = \tau/\epsilon$. In the case of the leap-frog integrator, for example, $\langle P_{\mathrm{acc}} \rangle = 1 - \mathrm{O}(\epsilon^2)$ and the integration step size is then usually tuned so that acceptance rates of $70 - 80$ percent are achieved.

It is more difficult to give a recommendation on the value of τ. Traditionally, the trajectory length is set to 1, but the choice of τ can have an influence on the autocorrelation times and the stability of the numerical integration. Some empirical studies are therefore required in order to determine the optimal value of τ in a given case.

2.5 Application to two-flavour QCD

In QCD with a doublet of mass-degenerate sea quarks, the probability density to be simulated is

$$p(U) = \frac{1}{\mathcal{Z}} \mathrm{e}^{-S(U)}, \qquad S(U) = S_{\mathrm{g}}(U) - \ln |\det D(U)|^2 , \tag{2.75}$$

where $D(U)$ denotes the lattice Dirac operator in presence of the gauge field U (cf. Section 2.1.1). The HMC algorithm can in principle be used to simulate this distribution, but a straightforward application of the algorithm is not possible in practice, because the calculation of the quark determinant and of the force deriving from it would require an unreasonable amount of computer time.

2.5.1 Pseudo-fermion fields

This difficulty can fortunately be overcome using the pseudo-fermion representation

$$|\det D(U)|^2 = \text{constant} \times \int \mathrm{D}[\phi]\, \mathrm{e}^{-S_{\mathrm{pf}}(U,\phi)}, \tag{2.76}$$

$$S_{\mathrm{pf}}(U,\phi) = \big(D(U)^{-1}\phi, D(U)^{-1}\phi\big), \qquad \mathrm{D}[\phi] = \prod_{x,A,\alpha} \mathrm{d}\phi_{A\alpha}(x) \mathrm{d}\phi_{A\alpha}(x)^*, \tag{2.77}$$

of the quark determinant. The auxiliary field $\phi(x)$ introduced here carries a Dirac index A and a colour index α, like a quark field, but its components are complex numbers rather than being elements of a Grassmann algebra. For the scalar product in eqn (2.77) one can take the obvious one for such fields, with any convenient normalization. Step (a) of the HMC algorithm is then replaced by

(a′) *A momentum field π and a pseudo-fermion field ϕ are generated randomly with probability density proportional to* $\exp\big\{-\frac{1}{2}(\pi,\pi) - S_{\mathrm{pf}}(U,\phi)\big\}$.

Note that the pseudo-fermion action is quadratic in ϕ. The field can therefore be easily generated following the lines of Section 2.1.4. Once this is done, the algorithm proceeds as before, where the Hamilton function to be used in steps (b) and (c′) is

$$H(\pi,U) = \tfrac{1}{2}(\pi,\pi) + S_{\mathrm{g}}(U) + S_{\mathrm{pf}}(U,\phi). \tag{2.78}$$

Only the momentum π and the gauge field U are evolved by the molecular-dynamics equations. The pseudo-fermion field ϕ remains unchanged and thus plays a spectator rôle at this point.

The steps (a′), (b) and (c′) implement the transition probability density

$$T(U \to U') = \frac{1}{\mathcal{Z}_\pi \mathcal{Z}_{\mathrm{pf}}(U)} \int \mathrm{D}[\pi]\mathrm{D}[\phi]\, \mathrm{e}^{-\frac{1}{2}(\pi,\pi) - S_{\mathrm{pf}}(U,\phi)}$$
$$\times \Big\{ P_{\mathrm{acc}}(\pi,U) \prod_{x,\mu} \delta(U'(x,\mu), U_\tau(x,\mu)) + (1 - P_{\mathrm{acc}}(\pi,U)) \prod_{x,\mu} \delta(U'(x,\mu), U(x,\mu)) \Big\}, \tag{2.79}$$

where $\mathcal{Z}_{\mathrm{pf}}(U)$ is the partition function of the pseudo-fermion field ϕ in presence of the gauge field U. For simplicity, the dependence of U_τ and P_{acc} on ϕ has been suppressed.

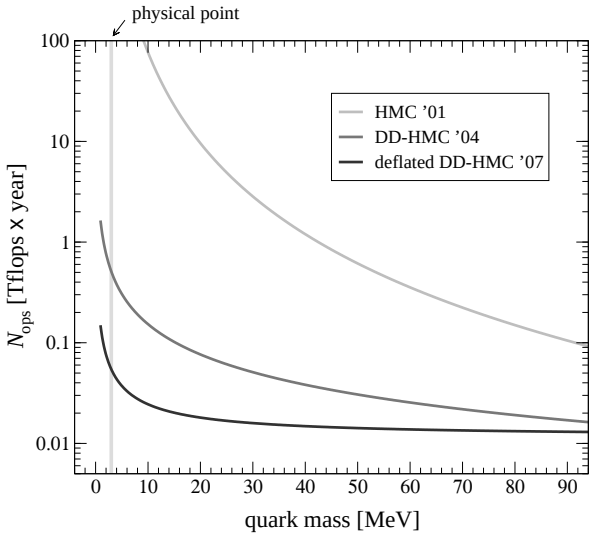


Fig. 2.3 Number of floating-point operations required for the generation of 100 statistically independent gauge-field configurations in O(a)-improved two-flavour QCD on a 64×32^3 lattice with spacing $a = 0.08$ fm. The top curve (Ukawa, 2002) represents the status reported at the memorable Berlin lattice conference in 2001, the middle one was obtained a few years later, using the so-called domain-decomposed HMC algorithm (Lüscher, 2005; Del Debbio *et al.*, 2007), and the lowest curve shows the performance of a recently developed deflated version of the latter (Lüscher, 2007*b*).

Starting from this formula, it is then not difficult to show that the algorithm correctly simulates the distribution (2.75).

2.5.2 Performance of the HMC algorithm

The force F that drives the molecular-dynamics evolution in step (b) of the algorithm has two parts, F_0 and F_1, the first deriving from the gauge action and the other from the pseudo-fermion action in the Hamilton function (2.78). In the case of the Wilson theory, for example, the forces are

$$F_0^a(x,\mu) = -\mathrm{Re}\,\mathrm{tr}\{T^a U(x,\mu) M(x,\mu)\}, \tag{2.80}$$

$$F_1^a(x,\mu) = -2\,\mathrm{Re}\,\big(\gamma_5 D^{-1}\gamma_5\psi, \delta^a_{x,\mu} D\psi\big), \qquad \psi = D^{-1}\phi, \tag{2.81}$$

where $M(x,\mu)$ is the staple sum (2.47) previously encountered in the pure gauge theory and

$$(\delta^a_{x,\mu} D\psi)(y) =$$

$$\delta_{x+\hat\mu,y}\tfrac{1}{2}(1+\gamma_\mu)U(x,\mu)^{-1}T^a\psi(x) - \delta_{x,y}\tfrac{1}{2}(1-\gamma_\mu)T^a U(x,\mu)\psi(x+\hat\mu). \tag{2.82}$$

Note that the computation of the pseudo-fermion force F_1 requires the Dirac equation to be solved two times. The by far largest fraction of the computer time is then usually

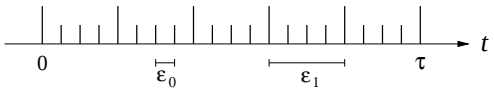


Fig. 2.4 Multiple time-step integration schemes divide the integration range $[0, \tau]$ in a hierarchy of intervals of increasing sizes $\epsilon_0, \epsilon_1, \ldots$ such that ϵ_{k+1} is an integer multiple of ϵ_k.

spent in this part of the simulation program, particularly so at small sea-quark masses where the Dirac operator becomes increasingly ill-conditioned.

Traditionally the performance of QCD simulation algorithms is measured by counting the number of floating-point operations required for the generation of a sample of statistically independent gauge-field configurations. Such performance estimations are quite primitive and tend to be subjective to some extent, because the term "statistically independent" is only loosely defined in this context. Moreover, the specific capabilities of the computers used for the simulation are not taken into account. The performance estimates plotted in Fig. 2.3 nevertheless clearly show that the simulations have become significantly faster since the beginning of the decade. In the following sections, some of the now widely used acceleration techniques are briefly described.

The curves shown in Fig. 2.3 refer to a particular choice of the lattice action and the lattice parameters. If the lattice volume V is increased at fixed lattice spacing and quark mass, and if the leap-frog integrator is used, the numerical effort required for the simulations is known to scale like $V^{5/4}$. As a function of the lattice spacing a, the scaling behaviour of the HMC algorithm is more difficult to determine and is currently a debated issue, but there is little doubt that the required computer time increases at least like a^{-7}.

2.5.3 Multiple time-step integration

An acceleration of the HMC algorithm can often be achieved using adapted integration step sizes for different parts of the force F in the molecular-dynamics equations (Sexton and Weingarten, 1992). The quark force F_1, for example, tends to be significantly smaller than the gauge force F_0 and can therefore be integrated with a larger step size than the latter.

If the integration step sizes for the forces F_0 and F_1 are taken to be

$$\epsilon_0 = \frac{\tau}{N_0 N_1}, \qquad \epsilon_1 = \frac{\tau}{N_1}, \tag{2.83}$$

where N_0, N_1 are some positive integers, the leap-frog integrators

$$\mathcal{J}_0(\epsilon_0, N_0) = \left\{\mathcal{I}_0(\tfrac{1}{2}\epsilon_0)\mathcal{I}_U(\epsilon_0)\mathcal{I}_0(\tfrac{1}{2}\epsilon_0)\right\}^{N_0}, \tag{2.84}$$

$$\mathcal{J}_1(\epsilon_1, N_1) = \left\{\mathcal{I}_1(\tfrac{1}{2}\epsilon_1)\mathcal{J}_0(\epsilon_0, N_0)\mathcal{I}_1(\tfrac{1}{2}\epsilon_1)\right\}^{N_1}, \tag{2.85}$$

integrate the molecular-dynamics equations from time t to $t+\epsilon_1$ and $t+\tau$, respectively (see Fig. 2.4). In these equations,

$$\mathcal{I}_k(\epsilon) : \ \pi, U \to \pi - \epsilon F_k, U \tag{2.86}$$

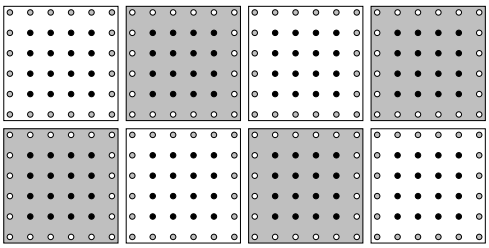


Fig. 2.5 Divisions of the lattice into non-overlapping blocks of lattice points can be chessboard-coloured if there is an even number of blocks along each coordinate axis. The union of the sets of points contained in the black blocks is denoted by Ω and its complement (the points in the white blocks) by Ω^*. Their exterior boundaries, $\partial\Omega$ and $\partial\Omega^*$, consist of all points in the other set with minimal distance from the surfaces separating the blocks (open points).

are the elementary integration steps involving the force F_k. In particular, the application of $\mathcal{J}_0(\epsilon_0, N_0)$ consumes relatively little computer time, because a computation of the quark force is not required.

The hierarchical integration is profitable if the step size ϵ_1 can be set to a value larger than ϵ_0 without compromising the accuracy of the numerical integration too much. An acceleration of the simulation by a factor approximately equal to N_0 is then achieved.

2.5.4 Frequency splitting of the quark determinant

The factorization

$$|\det D|^2 = \det\{DD^\dagger + \mu^2\} \times \det\left\{\frac{DD^\dagger}{DD^\dagger + \mu^2}\right\} \tag{2.87}$$

of the quark determinant separates the contribution of the eigenvalues of $DD^\dagger$ larger than μ^2 from the contribution of the lower ones. In the HMC algorithm, the two factors may be represented by two pseudo-fermion fields, ϕ_1 and ϕ_2, with action

$$S_{\rm pf}(U,\phi) = \big(\phi_1, (DD^\dagger + \mu^2)^{-1}\phi_1\big) + \big(\phi_2, \phi_2 + \mu^2(DD^\dagger)^{-1}\phi_2\big). \tag{2.88}$$

The quark force accordingly splits into two forces, F_1 and F_2, where the first is nearly insensitive to the quark mass while the second involves the inverse of $DD^\dagger$ and is, in this respect, similar to the force derived from the full quark determinant.

When such determinant factorizations were first considered (Hasenbusch, 2001; Hasenbusch and Jansen, 2003), the main effect appeared to be that the fluctuations of the quark force along the molecular-dynamics trajectories were reduced. The integration step size required for a given acceptance rate could consequently be increased by a factor 2 or so. Urbach *et al.* (2006) later noted that F_1 is the far dominant contribution to the quark force at small μ. Using a multiple time-step integrator, and after some tuning of μ, an important acceleration of the simulation was then achieved.

Another factorization of the quark determinant, leading to the DD-HMC algorithm (Lüscher, 2005), is obtained starting from a domain decomposition of the lattice like

the ones previously considered in Section 1.3.4. The rôle of the scale μ that separates the high modes of the Dirac operator from low modes is here played by the block sizes, which are usually chosen to be in the range from 0.5 to 1 fm or so.

For the associated factorization of the quark determinant to work out, one needs to assume that the lattice Dirac operator has only nearest-neighbour hopping terms, as in the case of the (improved) Wilson–Dirac operator, and that the block division of the lattice is chessboard-colourable (see Fig. 2.5). With respect to the subset Ω of points contained in the black blocks and its complement Ω^*, the Dirac operator then naturally decomposes into four parts,

$$D = D_\Omega + D_{\Omega^*} + D_{\partial\Omega} + D_{\partial\Omega^*}, \tag{2.89}$$

where D_Ω includes all (diagonal and hopping) terms acting inside Ω and D_{Ω^*} those acting in Ω^*. The hopping terms from the white to the black blocks and the ones going in the opposite direction are included in $D_{\partial\Omega}$ and $D_{\partial\Omega^*}$, respectively.

Using the fact that Ω and Ω^* are disjoint sets of lattice points, the factorization

$$\det D = \det D_\Omega \det D_{\Omega^*} \det\{1 - D_\Omega^{-1} D_{\partial\Omega} D_{\Omega^*}^{-1} D_{\partial\Omega^*}\}, \tag{2.90}$$

may now be derived, where the first two factors further factorize into the determinants of the block Dirac operators. The factorization corresponds to a splitting of the quark force into an easy high-mode part (the block determinants) and an "expensive" but much smaller low-mode part. No fine-tuning is required in this case and an acceleration of the simulation is again achieved using a multiple time-step integrator. Moreover, the domain decomposition is, as usual, favourable for the parallel processing of large lattices[2].

2.5.5 Chronological inversion and deflation acceleration

In the course of the integration of the molecular-dynamics equations, the Dirac equation must be solved many times. The exact details depend on which integrator and acceleration techniques are used, but the equations to be solved are usually of the form

$$D\psi = \phi, \qquad D\chi = \gamma_5\psi, \tag{2.91}$$

where the source field ϕ does not depend on the integration time t.

For small integration step sizes, the gauge field changes only little from one step to the next and the solutions ψ_t and χ_t obtained at time t therefore tend to evolve smoothly. One may thus attempt to predict them from the previous solutions through a polynomial extrapolation in t, for example (Brower *et al.*, 1997). In general, the predicted solutions of the Dirac equation are not sufficiently accurate, but this deficit can easily be removed through iterative improvement (Section 1.1.2). The total number of iterations performed by the solver program is then often significantly reduced.

An important further reduction of the solver iteration numbers can be achieved using the low-mode deflation method described in Section 1.3 (Lüscher, 2007*b*). Since the

[2]An efficient ISO C program implementing the DD-HMC algorithm for two-flavour QCD can be downloaded from `http://cern.ch/luscher/DDHMC/index.html` under the GNU Public License (GPL). Many of the acceleration techniques discussed here are included in this package.

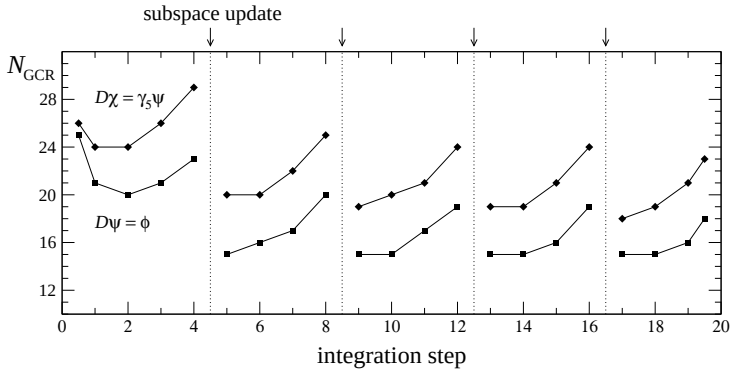


Fig. 2.6 History of the GCR solver iteration numbers required for the solution of the Dirac equation along a molecular-dynamics trajectory, showing the effect of the refreshing of the deflation subspace. The data points were obtained at a sea-quark mass of 26 MeV on a 64×32^3 lattice with spacing $a = 0.08$ fm. On other lattices, the situation is practically the same.

gauge field changes along a molecular-dynamics trajectory, the (domain-decomposed) deflation subspace must be refreshed from time to time in order to preserve its efficiency. An example illustrating this process is shown in Fig. 2.6. The subspace generation at the beginning of the trajectory and the periodic refreshing of the subspace requires some computer time, but this overhead is largely compensated by the fact that the solutions of the Dirac equations (2.91) are obtained much more rapidly than without deflation, particularly so at small quark masses (see Fig. 2.3).

2.5.6 Improved integrators

The use of integration schemes other than the leap-frog integrator can be profitable if fewer evaluations of the quark force are required for a given integration accuracy. Higher-order schemes, for example, where the one-step error is reduced to $\mathrm{O}(\epsilon^5)$, are not difficult to construct (Leimkuhler and Reich, 2004; Hairer *et al.*, 2006), but so far did not prove to be significantly faster than the leap-frog integrator.

Another possibility is to look for $\mathrm{O}(\epsilon^3)$ integrators that minimize the integration error according to some criterion (Omelyan *et al.*, 2002, 2003; Takaishi and de Forcrand, 2006; Clark *et al.*, 2008*b*). An integrator of this kind is, in the case of a single time-step integration, given by

$$\widetilde{\mathcal{J}}_0(\epsilon, N_0) = \left\{\mathcal{I}_0(\tfrac{1}{2}\tilde{\epsilon})\mathcal{I}_U(\tfrac{1}{2}\epsilon)\mathcal{I}_0(\epsilon-\tilde{\epsilon})\mathcal{I}_U(\tfrac{1}{2}\epsilon)\mathcal{I}_0(\tfrac{1}{2}\tilde{\epsilon})\right\}^{N_0}, \tag{2.92}$$

where $\tilde{\epsilon} \propto \epsilon$ is a tunable parameter. Experience suggests that the optimal values of $\tilde{\epsilon}/\epsilon$ are in the range $0.3 - 0.5$, depending a bit on the chosen accuracy criterion. Although the quark force needs to be computed two times per integration step, this integrator achieves a net acceleration by a factor 1.5 or so with respect to the leap-frog integrator, because fewer integration steps are required.

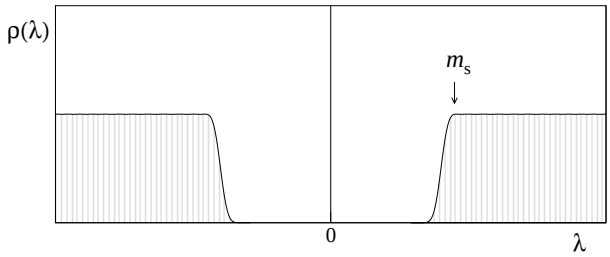


Fig. 2.7 Typical shape of the density $\rho(\lambda)$ of the eigenvalues λ of the hermitian Wilson–Dirac operator Q_s near the origin. At lattice spacings $a \leq 0.1$ fm, and if the strange-quark mass m_s is greater or equal to its physical value, the distribution of the eigenvalues with the smallest magnitude extends only slightly inside the spectral gap of the density in the continuum limit (Del Debbio *et al.*, 2006).

2.6 Inclusion of the strange quark

Unlike the light quarks, the strange and the heavy quarks do not pair up in approximately mass-degenerate doublets. For various technical reasons, single quarks are more difficult to include in the simulations than pairs and a special treatment is required.

2.6.1 Strange-quark determinant

One of the issues that needs to be addressed is the fact that the determinant

$$\det D_\mathrm{s} = \pm\,|\det D_\mathrm{s}| \tag{2.93}$$

of the strange-quark Dirac operator D_s may not have the same sign for all gauge-field configurations. If chiral symmetry is exactly preserved on the lattice, the determinant is guaranteed to be positive, but sign changes from one configuration to another are not excluded in the case of the (improved) Wilson–Dirac operator, for example. The presence of positive and negative contributions to the QCD partition function potentially ruins the foundations on which numerical simulations are based. In particular, importance sampling ceases to have a clear meaning.

In all current simulations of QCD that include the strange quark, the regions in field space, where the strange-quark determinant is negative, are assumed (or can be shown) to have a totally negligible weight in the functional integral. The operator in the determinant may then be replaced by the non-negative hermitian operator

$$|Q_\mathrm{s}| = \left(Q_\mathrm{s}^2\right)^{1/2}, \qquad Q_\mathrm{s} = \gamma_5 D_\mathrm{s}, \tag{2.94}$$

without affecting the simulation results. As explained in the following sections, the so modified theory can again be simulated using the HMC algorithm.

In the Wilson theory, the justification of this procedure rests on the observation that the physical strange quark is relatively heavy and that the hermitian Dirac operator Q_s consequently tends to have a solid spectral gap in presence of the gauge fields that dominate the QCD functional integral (see Fig. 2.7). Since the gap tends to widen when the strange-quark mass is increased, the sign of the determinant of Q_s cannot

change and is therefore the same as the one at large masses, where it can be proved to be positive.

2.6.2 Pseudo-fermion representation

Although well defined, the operator $|Q_\mathrm{s}|$ is not directly accessible and one is forced to use approximations whenever its action on a quark field needs to be computed. In the present context, an approximation of its inverse is required, i.e. a tractable function R of Q_s^2 such that $|Q_\mathrm{s}|R \simeq$ constant. An exact representation of the strange-quark determinant can then be given using two pseudo-fermion fields,

$$\det|Q_\mathrm{s}| = \text{constant} \times \int \mathrm{D}[\phi_1]\mathrm{D}[\phi_2]\,\mathrm{e}^{-S_{\mathrm{pf,s}}(U,\phi_1,\phi_2)} \tag{2.95}$$

where only the second term in the pseudo-fermion action

$$S_{\mathrm{pf,s}} = (\phi_1, (|Q_\mathrm{s}|R)^{-1}\phi_1) + (\phi_2, R\phi_2) \tag{2.96}$$

is included in the molecular-dynamics Hamilton function. The first term is nearly independent of the gauge field and can be taken into account in the acceptance-rejection step at the end of the molecular-dynamics evolution.

So far not many different approximations R were considered. The PHMC algorithm, for example, employs a polynomial approximation (Frezzotti and Jansen, 1997, 1999), while the RHMC algorithm is based on a rational approximation (Horvath *et al.*, 1999; Clark and Kennedy, 2007). There is also a complex version of the PHMC algorithm, where one starts from an approximation of D_s^{-1} by a polynomial in D_s (Takaishi and de Forcrand, 2002; Aoki *et al.*, 2002). None of these algorithms is completely trivial to implement or obviously preferable in view of its simplicity, accessibility to acceleration techniques or speed.

In the following, the RHMC algorithm is worked out as a representative case. The steps to be taken in the PHMC algorithm are similar, the main differences being issues of approximation accuracy and numerical stability.

2.6.3 Optimal rational approximation

The operator norm of the difference of two functions of Q_s^2 is bounded from above by the maximal absolute deviation of the functions in the range covered by the eigenvalues of Q_s^2. For the construction of optimal approximations of $|Q_\mathrm{s}|^{-1}$, some information on the spectral range of Q_s^2 is therefore required as input.

Since Q_s^2 has a solid spectral gap, one can choose some fixed numbers M and $\epsilon > 0$ such that the spectrum of Q_s^2 is fully contained in the interval $[\epsilon M^2, M^2]$ with high probability (i.e. for most configurations in a representative ensemble of gauge fields). Now let

$$R(y) = A\frac{(y+a_1)(y+a_3)\dots(y+a_{2n-1})}{(y+a_2)(y+a_4)\dots(y+a_{2n})}, \tag{2.97}$$

be a rational function of degree $[n,n]$ in y, which approximates the function $1/\sqrt{y}$ in the range $[\epsilon, 1]$. The operator $R(Q_\mathrm{s}^2/M^2)$ then approximates $M/|Q_\mathrm{s}|$ up to a relative error (in the operator norm) less than or equal to

$$\delta = \max_{\epsilon \le y \le 1} |1 - \sqrt{y} R(y)| \tag{2.98}$$

if the spectrum of $Q_{\rm s}^2$ is fully contained in $[\epsilon M^2, M^2]$ and a somewhat larger error if the spectrum extends slightly beyond the limits of this interval.

For a specified degree n, one would evidently like the coefficients a_r in eqn (2.97) to be such that the error δ is minimized. It seems unlikely that the optimal coefficients can be worked out analytically, but the mathematician Zolotarev was able to do that a long time ago by relating the optimization problem to the theory of elliptic functions (see Achiezer (1992), for example; the coefficients are given explicitly in Section 2.6.5)[3].

The optimal rational approximation has a number of remarkable properties. One of them is that the error δ is a very rapidly decreasing function of the degree n. For $\epsilon = 10^{-5}$, for example, the approximation error is 5×10^{-4} if $n = 6$ and decreases to 1×10^{-7} and 8×10^{-15} if $n = 12$ and $n = 24$, respectively. The fact that the minimizing coefficients satisfy

$$a_1 > a_2 > \ldots > a_{2n} > 0, \qquad A > 0, \tag{2.99}$$

is also very important. In particular, the residues in the expansion in partial fractions,

$$R(y) = A \left\{ 1 + \frac{r_2}{y + a_2} + \ldots + \frac{r_{2n}}{y + a_{2n}} \right\}, \tag{2.100}$$

are all positive. The expansion is therefore well suited for the numerical evaluation of the action of the operator $R(Q_{\rm s}^2/M^2)$ on the pseudo-fermion field ϕ_2. Note that this calculation essentially amounts to solving the equations

$$\left(Q_{\rm s}^2 + \mu_{2k}^2\right) \psi_k = M^2 \phi_2, \qquad \mu_r = M\sqrt{a_r}, \tag{2.101}$$

for $k = 1, \ldots, n$. Since the right-hand sides of these equations are the same, it is possible to solve all equations simultaneously using a so-called multi-mass solver (Jegerlehner, 1996). The total computational effort is then not very much larger than what would be required for the sequential solution of the equations that are the most difficult to solve (the ones at the largest values of k).

2.6.4 The RHMC algorithm

Having specified the operator R, the inclusion of the strange quark in the HMC algorithm is now straightforward. The algorithm proceeds according to the steps (a′), (b) and (c′), as before, and one merely has to add the contribution of the strange-quark pseudo-fermion fields. Some specific remarks on what exactly needs to be done may nevertheless be useful.

[3]Zolotarev obtained two different optimal rational approximations $R(y)$ to the function $1/\sqrt{y}$, one of degree $[n, n]$ and the other of degree $[n-1, n]$. Both are used in lattice QCD and are commonly referred to as the Zolotarev rational approximation.

(a) Pseudo-fermion generation. The random generation of the strange-quark pseudo-fermion fields in the first step requires two operators B and C to be found such that

$$BB^\dagger = |Q_\mathrm{s}|R, \quad \text{and} \quad CC^\dagger = R^{-1} \tag{2.102}$$

(cf. Section 2.1.4). For C one can take the rational function

$$C = A^{-1/2}\frac{(Q_\mathrm{s}+i\mu_2)\dots(Q_\mathrm{s}+i\mu_{2n})}{(Q_\mathrm{s}+i\mu_1)\dots(Q_\mathrm{s}+i\mu_{2n-1})}, \tag{2.103}$$

but there is no similarly simple choice for the other operator. However, since

$$Z = \frac{Q_\mathrm{s}^2R^2}{M^2} - 1 \tag{2.104}$$

is of order δ, the series

$$B = \sqrt{M}(1+Z)^{1/4} = \sqrt{M}\left\{1 + \tfrac{1}{4}Z - \tfrac{3}{32}Z^2 + \dots\right\} \tag{2.105}$$

converges rapidly and may be truncated after the first few terms. The strange-quark pseudo-fermion fields can thus be generated with a computational effort equivalent to the one required for a few applications of the operator R to a given quark field.

(b) Strange-quark force. As already mentioned, only the second term of the pseudo-fermion action (2.96) is included in the molecular-dynamics Hamilton function. The computation of the associated force,

$$F_\mathrm{s}^a(x,\mu) = -\frac{2A}{M^2}\sum_{k=1}^{n} r_{2k}\mathrm{Re}\left(\psi_k, Q_\mathrm{s}\gamma_5\delta^a_{x,\mu}D_\mathrm{s}\psi_k\right), \tag{2.106}$$

requires the n linear systems (2.101) to be solved. The computer time needed for this calculation is therefore essentially the same as for one application of the operator R.

(c) Acceptance step. The acceptance probability is calculated as usual except for the fact that the change of the first term of the pseudo-fermion action (2.96),

$$(\phi_1, (|Q_\mathrm{s}|R)^{-1}\phi_1)\big|_{U=U_\tau} - (\phi_1, (|Q_\mathrm{s}|R)^{-1}\phi_1)\big|_{U=U_0}, \tag{2.107}$$

must be added to the difference (2.72) of the molecular-dynamics Hamilton function. Note that the action difference (2.107) can be computed by expanding $(|Q_\mathrm{s}|R)^{-1}\phi_1$ in powers of Z.

2.6.5 Appendix: Coefficients of the optimal rational approximation

The analytic expressions for the coefficients of the rational function (2.97) that minimizes the approximation error (2.98) involve the Jacobi elliptic functions $\mathrm{sn}(u,k)$,

$\mathrm{cn}(u,k)$ and the complete elliptic integral $K(k)$ (see Abramowitz and Stegun (1972), for example, for the definition of these functions). Explicitly, they are given by

$$a_r = \frac{\mathrm{cn}^2(rv,k)}{\mathrm{sn}^2(rv,k)}, \quad r = 1,2,\ldots,2n, \tag{2.108}$$

where

$$k = \sqrt{1-\epsilon}, \qquad v = \frac{K(k)}{2n+1}. \tag{2.109}$$

The formulae for the amplitude A and the error δ,

$$A = \frac{2}{1+\sqrt{1-d^2}} \frac{c_1 c_3 \ldots c_{2n-1}}{c_2 c_4 \ldots c_{2n}}, \tag{2.110}$$

$$\delta = \frac{d^2}{\left(1+\sqrt{1-d^2}\right)^2}, \tag{2.111}$$

involve the coefficients

$$c_r = \mathrm{sn}^2(rv,k), \quad r = 1,2,\ldots,2n, \tag{2.112}$$

$$d = k^{2n+1} \left(c_1 c_3 \ldots c_{2n-1}\right)^2. \tag{2.113}$$

All these expressions are free of singularities and can be programmed straightforwardly, using the well-known methods for the numerical evaluation of the Jacobi elliptic functions. An ISO C program that calculates the coefficients A, $a_1,\ldots,a_{2n}$ and the error δ to machine precision can be downloaded from `http://cern.ch/luscher/`.

3
Variance reduction methods

Lattice QCD simulations produce ensembles $\{U_1, \ldots, U_N\}$ of gauge fields, which are representative of the functional integral at the specified gauge coupling and sea-quark masses. In this chapter it is taken for granted that autocorrelation effects can be safely neglected, i.e. that the separation in simulation time of subsequent field configurations is sufficiently large for this to be the case. As discussed in Section 2.2.4, the expectation value of any (real or complex) observable $\mathcal{O}(U)$ may then be calculated through

$$\langle \mathcal{O} \rangle = \frac{1}{N} \sum_{k=1}^{N} \mathcal{O}(U_k) + \mathrm{O}(N^{-1/2}), \tag{3.1}$$

where the statistical error is, to leading order in $1/N$, given by

$$\sigma(\mathcal{O}) = \frac{\sigma_0(\mathcal{O})}{N^{1/2}}, \qquad \sigma_0(\mathcal{O}) = \langle |\mathcal{O} - \langle \mathcal{O} \rangle|^2 \rangle^{1/2}. \tag{3.2}$$

In practice, the error is estimated from the variance of the "measured" values $\mathcal{O}(U_k)$, which is a correct procedure up to subleading terms.

For a given observable $\mathcal{O}$, another observable $\mathcal{O}'$ satisfying

$$\langle \mathcal{O}' \rangle = \langle \mathcal{O} \rangle, \qquad \sigma_0(\mathcal{O}') \ll \sigma_0(\mathcal{O}), \tag{3.3}$$

can sometimes be found. The desired expectation value is then obtained with a much smaller statistical error if $\mathcal{O}$ is replaced by $\mathcal{O}'$. Most variance-reduction methods are based on this simple observation. The construction of effective alternative observables is non-trivial, however, and may involve auxiliary stochastic variables and transformations of the functional integral.

The discussion in this chapter is often of a general nature and applies to most forms of lattice QCD, but if not specified otherwise, the Wilson formulation will be assumed, with or without $\mathrm{O}(a)$-improvement and with two or more flavours of sea quarks.

3.1 Hadron propagators

The calculation of the properties of the light mesons and baryons is a central goal in lattice QCD. Finding good variance-reduction methods proves to be difficult in this field, but some important progress has nevertheless been made. In this section, the aim is to shed some light on the problem by discussing the statistical variance of hadron propagators.

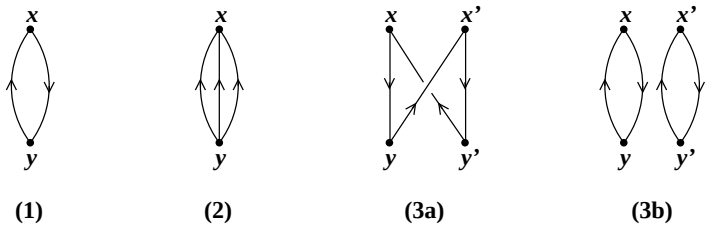


Fig. 3.1 Quark-line diagrams contributing to the pion propagator (1), the nucleon propagator (2) and the pion four-point function (3a and 3b). Directed lines from y to x stand for the light-quark propagator $S(x, y)$. At the vertices, the spinor indices are contracted with the appropriate colour tensor and Dirac matrix (γ_5 in the case of the pion-field vertices).

3.1.1 The pion propagator

Once the sea-quark fields are integrated out, the correlation functions of local fields like the isospin pseudo-scalar density

$$P^a = \overline{\psi}\tfrac{1}{2}\tau^a\gamma_5\psi, \qquad \psi = \begin{pmatrix} u \\ d \end{pmatrix}, \qquad \tau^a\text{: Pauli matrices,} \tag{3.4}$$

become expectation values of quark-line diagrams. The diagrams are products of quark propagators, Dirac matrices and invariant colour tensors, with all Dirac and colour indices properly contracted (see Fig. 3.1).

For any given gauge field, the light-quark propagator is determined through the field equation

$$DS(x, y) = \delta_{xy} \tag{3.5}$$

and the chosen boundary conditions. The spinor indices have been suppressed in this equation, but one should keep in mind that $S(x, y)$ is, for fixed x and y, a complex 12×12 matrix in spinor space. In view of the γ_5-hermiticity of the Dirac operator, forward and backward propagators are related by

$$\gamma_5 S(y, x)\gamma_5 = S(x, y)^\dagger, \tag{3.6}$$

where the hermitian conjugation refers to spinor space only.

In lattice QCD, the pion mass m_π is usually determined from the exponential decay of the "pion propagator"

$$\langle P^a(x)P^b(y)\rangle = -\tfrac{1}{2}\delta^{ab}\langle \mathrm{tr}\{S(x, y)S(x, y)^\dagger\}\rangle \tag{3.7}$$

at large distances $|x - y|$. It is advantageous for this calculation to pass to the zero-momentum component of the propagator,

$$g_\pi(x_0 - y_0) = \langle \mathcal{O}_\pi(x_0, y)\rangle, \tag{3.8}$$

$$\mathcal{O}_\pi(x_0, y) = \sum_{\vec{x}} \mathrm{tr}\{S(x, y)S(x, y)^\dagger\}, \tag{3.9}$$

whose asymptotic form at large times,

$$g_\pi(t) \underset{t\to\infty}{\propto} \mathrm{e}^{-m_\pi t} + \mathrm{O}(\mathrm{e}^{-3m_\pi t}), \tag{3.10}$$

is dominated by a single exponential with exponent equal to the pion mass.

3.1.2 Statistical error estimation

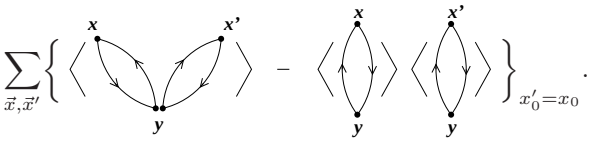
For a given gauge field and source point y, the calculation of the propagator $S(x,y)$ allows the observable $\mathcal{O}_\pi(x_0,y)$ to be evaluated at all times x_0. The ensemble average of $\mathcal{O}_\pi(x_0,y)$ then provides a stochastic estimate of the zero-momentum pion propagator, from which one may be able to extract the pion mass.

The statistical error of the average of $\mathcal{O}_\pi(x_0,y)$ is proportional to the square root of the a priori variance

$$\sigma_0(\mathcal{O}_\pi)^2 = \langle \mathcal{O}_\pi(x_0,y)^2\rangle - \langle \mathcal{O}_\pi(x_0,y)\rangle^2, \tag{3.11}$$

which may symbolically be written in the form

$$\sigma_0(\mathcal{O}_\pi)^2 = \sum_{\vec{x},\vec{x}'} \Big\{ \Big\langle \quad \Big\rangle - \Big\langle \quad \Big\rangle \Big\langle \quad \Big\rangle \Big\}_{x_0'=x_0} . \tag{3.12}$$



An important point to note here is that the diagram in the first term also contributes to the $\pi\pi$ propagator $\langle P^a(x)P^b(x')P^a(y)P^b(y)\rangle$ (see Fig. 3.1). Moreover, the second term is just the square of the pion propagator. The variance is therefore expected to decay like $\mathrm{e}^{-2m_\pi|x_0-y_0|}$ at large time separations, which implies that the pion propagator is obtained with a nearly time-independent relative statistical error.

Lattice QCD simulations tend to confirm this and they also show that the quark propagator typically falls off like

$$\mathrm{tr}\{S(x,y)S(x,y)^\dagger\}^{1/2} \propto \mathrm{e}^{-\frac{1}{2}m_\pi|x-y|} \tag{3.13}$$

at large distances, for every gauge field in a representative ensemble of fields. There is actually little room for a different behaviour of the propagator, since both the mean and width of the distribution of $\mathrm{tr}\{S(x,y)S(x,y)^\dagger\}$ decay exponentially with the same exponent.

3.1.3 Baryons and the exponential SNR problem

The diagrams contributing to the nucleon two-point function involve 3 quark propagators. At zero spatial momentum and for every gauge-field configuration, the associated observable $\mathcal{O}_N(x_0,y)$ thus falls off roughly like $\mathrm{e}^{-\frac{3}{2}m_\pi|x_0-y_0|}$ at large time separations. The nucleon propagator, however, decays much more rapidly, since the exponent (the nucleon mass m_N) is significantly larger than $\frac{3}{2}m_\pi$.

In the sum over all gauge fields, there must therefore be important cancellations among the measured values of $\mathcal{O}_N(x_0,y)$. More precisely, after averaging over N configurations, the nucleon propagator is obtained with a signal-to-noise ratio (SNR)

proportional to $\sqrt{N}\mathrm{e}^{-(m_N-\frac{3}{2}m_\pi)|x_0-y_0|}$. The number of measurements required for a specified statistical accuracy at a given time separation thus scales like

$$N \propto \mathrm{e}^{(2m_N-3m_\pi)|x_0-y_0|}. \tag{3.14}$$

Note that the exponent $2m_N - 3m_\pi$ is, in all cases of interest, not small and as large as $7\,\mathrm{fm}^{-1}$ at the physical point. The calculated values of the nucleon propagator therefore tend to rapidly disappear in the statistical noise at large time separations $|x_0 - y_0|$.

Computations of other hadron propagators are similarly affected by an exponential loss of significance, the only exception being the propagators of the stable pseudo-scalar mesons. The masses and decay constants of the latter can thus be determined far more easily than those of the vector mesons and the baryons.

3.2 Using random sources

As already noted in Section 2.1.3, the link variables in distant regions of a large lattice are practically decoupled. Hadron propagators calculated at widely separated source points therefore tend to be sampled independently and their average consequently has smaller statistical fluctuations than the propagator at a single source point. Averaging over as many source points as possible is thus desirable, but tends to be expensive in terms of computer time, because the quark propagators must be recomputed at each point.

The random source method performs the sum over all or a selected set of source points stochastically (Michael and Peisa, 1998). In many cases, a significant acceleration of the computation can be achieved in this way. The idea of introducing random sources is also quite interesting from a purely theoretical point of view, because it extends the notion of an observable to stochastic observables (i.e. functions of the gauge field that depend on random auxiliary variables).

3.2.1 Gaussian random fields

Random sources may be thought of as a set of additional fields that are decoupled from the dynamical fields and therefore do not change the physics content of the theory. In the case considered here, the added fields are a multiplet

$$\eta_i(\vec{x}), \quad i = 1, \ldots, N_{\mathrm{src}}, \tag{3.15}$$

of pseudo-fermion fields on the fixed-time spatial lattice. Their action is taken to be

$$S_{\mathrm{src}}(\eta) = \sum_{i=1}^{N_{\mathrm{src}}} (\eta_i, \eta_i), \tag{3.16}$$

where the scalar product is the obvious one for such fields. For each gauge-field configuration in a representative ensemble of fields, the source fields are chosen randomly with probability density proportional to $\mathrm{e}^{-S_{\mathrm{src}}(\eta)}$. Evidently, the set of fields obtained in this way is a representative ensemble for the joint probability density of the gauge field and the source fields.

One may now consider observables $\mathcal{O}(U,\eta)$ that depend on both the gauge field and the sources. When the latter are integrated out, such observables reduce to ordinary (non-stochastic) observables. Noting

$$\langle \eta_i(\vec{x})\eta_j(\vec{y})^\dagger \rangle_{\rm src} = \delta_{ij}\delta_{\vec{x}\vec{y}}, \tag{3.17}$$

the integral over the source fields is easily worked out, using Wick's theorem, if $\mathcal{O}(U,\eta)$ is a polynomial in the source fields. In the case of the observable

$$\mathcal{O} = \frac{1}{N_{\rm src}} \sum_{i=1}^{N_{\rm src}} \sum_{\vec{x},\vec{y}} \eta_i(\vec{x})^\dagger S(x,y)|_{x_0=y_0} \eta_i(\vec{y}), \tag{3.18}$$

for example, the calculation yields

$$\langle \mathcal{O} \rangle_{\rm src} = \sum_{\vec{x}} {\rm tr}\{S(x,x)\}. \tag{3.19}$$

The random sources thus allow the trace (3.19) to be estimated stochastically.

3.2.2 The pion propagator revisited

For any fixed time y_0, the spinor fields

$$\phi_i(x,y_0) = \sum_{\vec{y}} S(x,y)\eta_i(\vec{y}), \qquad i = 1,\ldots,N_{\rm src}, \tag{3.20}$$

can be computed by solving the Dirac equations $D\phi_i(x,y_0) = \delta_{x_0 y_0}\eta_i(\vec{x})$. It is then straightforward to show that the observable

$$\hat{\mathcal{O}}_\pi(x_0,y_0) = \frac{1}{N_{\rm src}V_3} \sum_{i,\vec{x}} |\phi_i(x,y_0)|^2\,, \qquad V_3\text{: spatial lattice volume}, \tag{3.21}$$

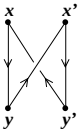
has the same expectation value as $\mathcal{O}_\pi(x_0,y)$ and may therefore be used in place of the latter in a calculation of the pion propagator. Note that through the average over the source fields,

$$\langle \hat{\mathcal{O}}_\pi(x_0,y_0) \rangle_{\rm src} = \frac{1}{V_3} \sum_{\vec{y}} \mathcal{O}_\pi(x_0,y), \tag{3.22}$$

one effectively sums over all source points at time y_0.

Whether the new observable is any better than the old one depends on whether it has smaller statistical fluctuations or not. A short calculation, using Wick's theorem, shows that the associated a priori variance is given by

$$\sigma_0(\hat{\mathcal{O}}_\pi)^2 = \sigma_0(\overline{\mathcal{O}}_\pi)^2 + \frac{1}{N_{\rm src}V_3^2} \sum_{\vec{x},\vec{x}'} \sum_{\vec{y},\vec{y}'} \Big\{ \Big\langle$$



$$\Big\rangle \Big\}_{x_0'=x_0, y_0'=y_0}, \tag{3.23}$$

where $\overline{\mathcal{O}}_\pi(x_0,y_0)$ denotes the volume-averaged observable (3.22). Recalling the exponential decay (3.13) of the quark propagator, the second term is readily estimated to

be of order $(N_{\mathrm{src}} m_\pi^3 V_3)^{-1}$ at large volumes. Since the first term scales like $(m_\pi^3 V_3)^{-1}$, it tends to be the dominant contribution to the variance already for moderately large numbers of source fields (setting $N_{\mathrm{src}} = 12$, for example, is often sufficient).

With respect to a calculation of the pion propagator at a single source point, the random source method described here thus achieves a reduction of the statistical error by a factor proportional to $(m_\pi^3 V_3)^{-1/2}$ for approximately the same computational cost. The use of random sources is therefore recommended in this case, particularly so on large lattices.

3.2.3 Further applications

Random sources are an interesting and useful tool. Here only one particular application and variant of the method was discussed. Source fields on different subsets of points can be considered as well as random fields taking values in a group or, more generally, fields with any non-gaussian distribution. The method combines well with low-mode averaging techniques (Neff *et al.*, 2001; Giusti *et al.*, 2004; DeGrand and Schaefer, 2004; Bali *et al.*, 2005; Foley *et al.*, 2005) and, to mention just one further example, it also played a central rôle in a recent calculation of the spectral density of the hermitian Wilson–Dirac operator (Giusti and Lüscher, 2009).

In the case of the vector-meson and baryon propagators, the application of random source methods is complicated by the fact that the index contractions at the vertices of the quark-line diagrams couple different components of the quark propagators. Different kinds of random sources (one for each component, for example) then need to be introduced to be able to write down a correct random-source representation of the hadron propagator. The variance of such observables typically involves many diagrams, among them often also disconnected ones. As a consequence, the use of random sources for these propagators is not obviously profitable, at least as long as the spatial extent of the lattices considered is not significantly larger than 2 or 3 fm.

Random source methods should preferably be applied only after a careful analysis of the variance of the proposed stochastic observables. Such an analysis can be very helpful in deciding which kind of random fields to choose and how exactly the stochastic observable must be constructed in order to achieve a good scaling of the statistical error with the lattice volume.

3.3 Multilevel simulations

Random source methods can lead to an important reduction of statistical errors, but are unable to overcome the exponential SNR problem encountered in the case of the nucleon propagator, for example. In a multilevel simulation, the improved observables are constructed through a stochastic process, i.e. through a secondary or nested simulation. Exploiting the locality of the theory, an exponential reduction of the statistical error can then be achieved in certain cases.

So far multilevel simulations have been limited to bosonic field theories, essentially because manifest locality is lost when the fermion fields are integrated out. Whether this limitation is a transient one is unclear at present, but it is certainly worth explaining the idea here and to show its impressive potential.

3.3.1 Statistical fluctuations of Wilson loops

In the following, a multilevel algorithm for the computation of the expectation values of Wilson loops in the pure SU(3) lattice gauge theory will be described. The lattice action is assumed to be the Wilson plaquette action.

Similarly to the hadron propagators, Wilson loop expectation values suffer from an exponential SNR problem. Let $\mathcal{C}$ be a $T \times R$ rectangular loop in the (x_0, x_1)-plane, $U(\mathcal{C})$ the ordered product of the link variables around the loop and $\mathcal{W} = \mathrm{tr}\{U(\mathcal{C})\}$ its trace. For large loops, the expectation value of $\mathcal{W}$ satisfies the area law

$$\langle \mathcal{W} \rangle \sim \mathrm{e}^{-\sigma A}, \qquad A = TR, \tag{3.24}$$

where $\sigma \simeq 1\,\mathrm{GeV/fm}$. Since $\mathcal{W}$ is a number of order 1 for every gauge-field configuration, the a priori variance $\sigma_0(\mathcal{W})$ is practically equal to $\langle |\mathcal{W}|^2 \rangle$ and thus of order 1 too. One therefore needs to generate ensembles of at least

$$N = \langle \mathcal{W} \rangle^{-2} \sim \mathrm{e}^{2\sigma A} \tag{3.25}$$

configurations to be able to calculate the expectation value $\langle \mathcal{W} \rangle$ to a useful precision.

For loop areas A equal to 1, 2 and 4 fm^2, for example, the minimal ensemble sizes are thus estimated to be 2×10^4, 6×10^8 and 4×10^{17}, respectively. These figures may not be exactly right, because the Wilson loop is averaged over all possible translations in practice and since there are important subleading corrections to the area law (3.24). However, for large areas A, the minimal ensemble size is essentially determined by the rapidly growing exponential factor (3.25).

3.3.2 Factorization and sublattice expectation values

A multilevel algorithm invented many years ago is the "multihit method" (Parisi *et al.*, 1983). In this case, a reduction of the statistical error by an exponential factor with exponent proportional to T is achieved by replacing the time-like link variables along the Wilson loop through stochastic estimates of their average values in presence of the other link variables. The method thus exploits the fact that the Wilson loop factorizes into a product of link variables.

If the loop is instead factorized into a product of two-link operators, an exponential error reduction is obtained with an exponent proportional to the area A (Lüscher and Weisz, 2001). A two-link operator is an object with four SU(3) indices,

$$\mathbb{T}(x_0)_{\alpha\beta\gamma\delta} = U(x,0)^*_{\alpha\beta} U(x + R\hat{1}, 0)_{\gamma\delta}, \tag{3.26}$$

residing at time x_0 and $\vec{x} = 0$ (see Fig. 3.2). Two-link operators at adjacent times can be multiplied and their product from time 0 to $T - a$ yields the tensor product of the time-like lines of the Wilson loop. The latter is then given by

$$\mathcal{W} = \mathrm{tr}\{\mathbb{L}(0)\mathbb{T}(0)\mathbb{T}(a) \ldots \mathbb{T}(T-a)\mathbb{L}(T)\}, \tag{3.27}$$

where the product

$$\mathbb{L}(x_0)_{\alpha\beta} = \big\{U(x,1)U(x + a\hat{1}, 1) \ldots U(x + (R-a)\hat{1}, 1)\big\}_{\alpha\beta} \tag{3.28}$$

denotes the spatial Wilson line at time x_0.

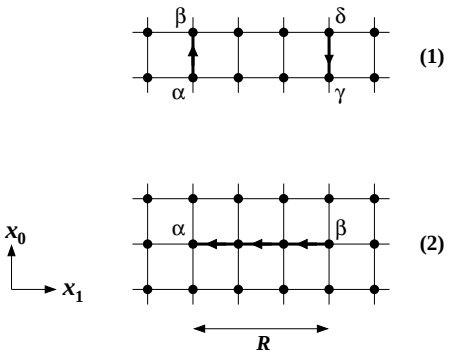


Fig. 3.2 A $T \times R$ Wilson loop in the (x_0, x_1)-plane with lower-left corner at $x = 0$ can be factorized into the product (3.27) of two-link and line operators. The SU(3) indices of the two-link operator (3.26) are assigned as shown in figure (1). The line operators (3.28) coincide with the spatial lines of the Wilson loop and have only two indices, as indicated in figure (2).

Consider now the sublattice bounded by the equal-time hyperplanes at some time y_0 and some later time z_0 (see Fig. 3.3). The Wilson action couples the link variables inside the sublattice (the "interior" link variables) to themselves and the spatial fields on the boundaries, but there is no interaction with the field variables elsewhere on the lattice. If $\mathcal{O}$ is any function of the interior link variables, its sublattice expectation value is defined by

$$[\mathcal{O}] = \frac{1}{\mathcal{Z}_{\rm int}} \int \mathrm{D}[U]_{\rm int}\, \mathcal{O}(U)\, \mathrm{e}^{-S(U)}, \tag{3.29}$$

where one integrates over the interior variables. Note that the part of the action that does not depend on the latter drops out in the expectation value, which is thus a function of the spatial link variables at time y_0 and z_0 only.

If the lattice is decomposed into non-overlapping sublattices of this kind, the functional integral divides into an integral over the interior variables and an integral over the boundary fields. For even T/a, for example, the Wilson loop expectation value may be rewritten in the form

$$\mathcal{W} = \langle \mathrm{tr}\{\mathbb{L}(0)[\mathbb{T}(0)\mathbb{T}(a)][\mathbb{T}(2a)\mathbb{T}(3a)] \ldots \mathbb{L}(T)\}\rangle. \tag{3.30}$$

The outer expectation value in this expression is the usual one involving an integration over all field variables. However, since the observable depends only on the spatial fields at time $x_0 = 0, 2a, 4a, \ldots, T$, the integral over the interior field variables yields the product of the sublattice partition functions. This factor exactly cancels the product of the normalization factors of the sublattice expectation values.

The sublattice expectation value of a product of two-link operators is a correlation function of two segments of Wilson lines in presence of the boundary fields. A single segment transforms non-trivially under the center symmetry of the theory, where all time-like link variables at a given time are multiplied by an element of the center of

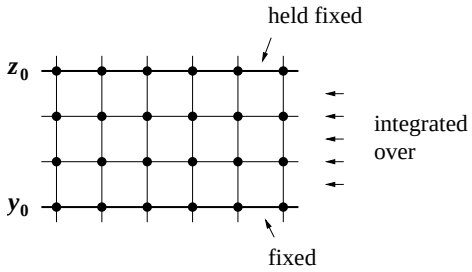


Fig. 3.3 Factorized representations of the Wilson loop expectation value such as (3.30) are based on a division of the lattice into sublattices separated by equal-time hyperplanes. The sublattice expectation values $[\ldots]$ in these expressions involve an integration over the field variables residing on the links between the hyperplanes, all other link variables being held fixed.

SU(3). Barring spontaneous symmetry breaking, its sublattice expectation value therefore vanishes and the correlation function of two segments is consequently expected to go to zero at large separations R. Experience actually suggests that

$$[\mathbb{T}(y_0)\mathbb{T}(y_0+a)\ldots\mathbb{T}(z_0-a)] \sim \mathrm{e}^{-\sigma(z_0-y_0)R} \tag{3.31}$$

if the time difference $z_0 - y_0$ is larger than, say, 0.5 fm or so. Recalling eqn (3.30), the rapid decay of the Wilson loop expectation value at large times T (and the area law if eqn (3.31) holds) is thus seen to arise from a product of small factors.

3.3.3 Multilevel update scheme

The foregoing suggests that the expectation value of the Wilson loop may be accurately calculated, at any value of T, starting from a factorized representation like (3.30). An algorithm that implements the idea for this particular factorization proceeds in cycles consisting of the following steps:

(a) *Update the gauge field N_0 times using a combination of the heatbath and microcanonical link-update algorithms.*

(b) *Estimate the expectation values $[\mathbb{T}(x_0)\mathbb{T}(x_0+a)]$ by updating the field inside the associated sublattices N_1 times and by averaging the product of the two-link operators over the generated configurations.*

(c) *Compute the trace* $\mathrm{tr}\{\mathbb{L}(0)[\mathbb{T}(0)\mathbb{T}(a)][\mathbb{T}(2a)\mathbb{T}(3a)]\ldots\mathbb{L}(T)\}$ *using the estimates of the sublattice expectation values obtained in step (b).*

Each cycle thus yields an estimate of the trace $\mathrm{tr}\{\mathbb{L}(0)[\mathbb{T}(0)\mathbb{T}(a)]\ldots\mathbb{L}(T)\}$. A moment of thought shows that these estimates are averages of the Wilson loop over a particular set field configurations generated through a valid simulation algorithm. Their average over many cycles therefore coincides with the expectation value of the Wilson loop up to statistical errors.

As explained in Section 3.3.2, the trace $\mathrm{tr}\{\mathbb{L}(0)[\mathbb{T}(0)\mathbb{T}(a)]\ldots\mathbb{L}(T)\}$ tends to decay exponentially, for every field configuration in a representative ensemble of fields. The

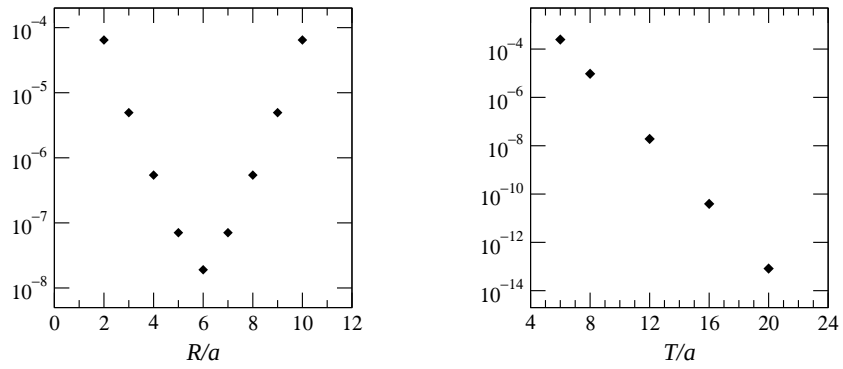


Fig. 3.4 Values of the correlation function of the Polyakov loop on a $(T/a) \times 12^3$ lattice with spacing $a \simeq 0.17$ fm and periodic boundary conditions, plotted as a function of the distance R at $T = 12a$ (left figure) and as a function of T at $R = 6a$ (right figure).

statistical error of the stochastic estimates of the trace produced by the multilevel algorithm (a)–(c) is therefore guaranteed to fall off exponentially as well. With respect to a straightforward computation of the Wilson loop expectation value, the sublattice averaging thus achieves an exponential error reduction.

For illustration, some results obtained in the course of an early application of the multilevel algorithm are shown in Fig. 3.4. In order to simplify the situation a little bit, the correlation function of two Polyakov loops (i.e. Wilson lines that wrap around the lattice in the time direction) was considered in this study. The line operators $\mathbb{L}(x_0)$ are then not needed and the observable coincides with the trace of a product of T/a two-link operators, T being the time-like extent of the lattice. As is evident from the data plotted in the figure, the multilevel algorithm allows the exponential decay of the correlation function to be followed over many orders of magnitude, which would not be possible (or only with an astronomical computer budget) using the standard simulation techniques.

3.3.4 Remarks and further developments

Multilevel algorithms of the kind described here have been employed in studies of the string behaviour (Lüscher and Weisz, 2002; Kratochvila and de Forcrand, 2003; Pepe and Wiese, 2009), the glueball spectrum (Meyer, 2003, 2004) and the bulk viscosity (Meyer, 2008) in pure gauge theories. Another, closely related multilevel algorithm has recently been proposed for the calculation of the energies of the lightest states with a specified (non-trivial) transformation behaviour under the exact symmetries of the theory (Della Morte and Giusti, 2009). In all these cases, an important and sometimes impressive acceleration of the simulation was achieved.

The exponential SNR problem is, however, often not completely eliminated. Computations of Wilson loop expectation values, for example, require sublattice expectation values of products of two-link operators to be determined to some accuracy. In view of eqn (3.31), the number of sublattice updates that need to be performed in this

part of the calculation is therefore expected to scale roughly like $e^{2\sigma(z_0-y_0)R}$. With respect to a one-level simulation, the exponential growth of the required computational effort is nevertheless dramatically reduced, because the exponent now increases only proportionally to the distance R rather than the area A.

4
Statistical error analysis

The estimation of the statistical errors in numerical lattice QCD appears to be an easy topic. However, the physical quantities often need to be extracted from the simulation data through some non-linear procedure, which may involve complicated fits and extrapolations. The correct propagation of the errors becomes a non-trivial task under these conditions. Resampling techniques such as the jackknife and bootstrap methods (Efron and Tibshirani, 1993) allow the errors of the calculated physical quantities to be estimated with minimal effort, but are a bit of magic and are actually known to be incorrect in certain special cases (the jackknife method, for example, in general gives wrong results for the statistical error of the median of an observable).

The aim in this chapter is to build up a conceptually solid framework in which the error propagation is made transparent. In particular, the correctness of the jackknife method can then be established for a class of quantities, which includes most cases of interest.

4.1 Primary observables

Wilson loops, quark-line diagrams and any other function of the gauge field, including the stochastic ones considered in Chapter 3, are the primary observables in lattice QCD. Their expectation values are the first quantities calculated in a simulation project and all physical quantities are eventually obtained from these.

4.1.1 Correlation functions and data series

Let A_r be some real-valued primary observables labeled by an index r. As before, the lattice QCD expectation value of an observable $\mathcal{O}$ is denoted by $\langle\mathcal{O}\rangle$. The quantities of interest are then the n-point correlation functions $\langle A_{r_1} \dots A_{r_n}\rangle$. Independently of any locality properties, their connected parts $\langle A_{r_1} \dots A_{r_n}\rangle_{\mathrm{c}}$ may be defined in the familiar way. In particular,

$$\langle A_r\rangle_{\mathrm{c}} = \langle A_r\rangle, \tag{4.1}$$

$$\langle A_r A_s\rangle_{\mathrm{c}} = \langle A_r A_s\rangle - \langle A_r\rangle\langle A_s\rangle, \tag{4.2}$$

$$\begin{aligned}\langle A_r A_s A_t\rangle_{\mathrm{c}} = {} & \langle A_r A_s A_t\rangle - \langle A_r\rangle\langle A_s A_t\rangle - \langle A_s\rangle\langle A_t A_r\rangle - \langle A_t\rangle\langle A_r A_s\rangle \\ & + 2\langle A_r\rangle\langle A_s\rangle\langle A_t\rangle.\end{aligned} \tag{4.3}$$

For an arbitrary number n of observables, the relation between the connected and the full correlation functions is given by the moment-cumulant transformation (see Section 4.1.4).

Suppose now that a representative ensemble of N statistically independent gauge-field configuration has been generated in the course of a simulation. The evaluation of the primary observable A_r on these configurations yields a series

$$a_{r,1}, a_{r,2}, \ldots, a_{r,N} \tag{4.4}$$

of "measured" values of this observable, whose average

$$\bar{a}_r = \frac{1}{N} \sum_{i=1}^{N} a_{r,i} \tag{4.5}$$

provides a stochastic estimate of its expectation value $\langle A_r \rangle$.

If $a_{r,i}$ and $a_{s,i}$, $i = 1, \ldots, N$, are the data series obtained for the observables A_r and A_s, the data series for the product $A_r A_s$ is $a_{r,i} a_{s,i}$. The product is actually just another primary observable, even in the case of the stochastic observables considered in Section 3.2 if the random sources are included as additional fields in the ensemble of fields generated by the simulation. A stochastic estimate of the n-point correlation function $\langle A_{r_1} \ldots A_{r_n} \rangle$ is thus obtained by calculating the average of $a_{r_1,i} \ldots a_{r_n,i}$.

4.1.2 Simulation statistics

As previously noted in Section 2.2.4, the question by how much $\bar{a}_r$ deviates from $\langle A_r \rangle$ can only be answered statistically when the simulation is repeated many times. In the following, the average over infinitely many simulations of any function ϕ of the measured values of the primary observables is denoted by $\langle\langle \phi \rangle\rangle$. The standard deviation of $\bar{a}_r$ from $\langle A_r \rangle$, for example, is given by $\langle\langle (\bar{a}_r - \langle A_r \rangle)^2 \rangle\rangle^{1/2}$.

Simulation statistics becomes a useful tool when the following two assumptions are made. First it must be guaranteed that the measurements of the primary observables are unbiased, i.e. that the equality

$$\langle\langle a_{r,i} \rangle\rangle = \langle A_r \rangle \tag{4.6}$$

holds for all i and all primary observables A_r. The second assumption is that the fields generated by the simulations are statistically independent. In particular,

$$\langle\langle a_{r,i} a_{s,j} \rangle\rangle = \langle\langle a_{r,i} \rangle\rangle \langle\langle a_{s,j} \rangle\rangle \quad \text{if} \quad i \neq j. \tag{4.7}$$

More generally, the average $\langle\langle a_{r_1,i_1} \ldots a_{r_n,i_n} \rangle\rangle$ factorizes into a product of averages, one for each subset of the factors $a_{r,i}$ with a given value of the configuration index i. As discussed in Section 2.2.4, both conditions are fulfilled if the residual autocorrelations among the fields in the representative ensembles generated by the simulation algorithm are negligible.

In practice the statistical independence of the measured values should be carefully checked by computing the associated autocorrelation functions. However, since the latter is estimated from the data and is therefore subject to statistical fluctuations, a

reliable determination of the autocorrelation times can sometimes be difficult, particularly so if the available data series is not very long. The issue of "calculating the error of the error" was addressed by Madras and Sokal (1988) and in greater detail again by Wolff (2004) (see also Lüscher (2005), Appendix E, for some additional material).

4.1.3 Distribution of the mean values

The mean values $\bar{a}_r, \bar{a}_s, \ldots$ of the measured values of the primary observables are correlated to some extent, because the underlying ensemble of gauge-field configurations is the same. It is possible to work out the correlations

$$\langle\!\langle \bar{a}_{r_1} \ldots \bar{a}_{r_k} \rangle\!\rangle = \frac{1}{N^k} \sum_{i_1=1}^{N} \cdots \sum_{i_k=1}^{N} \langle\!\langle a_{r_1,i_1} \ldots a_{r_k,i_k} \rangle\!\rangle \tag{4.8}$$

in terms of the correlation functions $\langle A_{r_1} \ldots A_{r_n} \rangle$. In the case of the two-point correlation functions, for example, one obtains

$$\langle\!\langle \bar{a}_r \bar{a}_s \rangle\!\rangle = \frac{1}{N^2} \sum_{i,j=1}^{N} \langle\!\langle a_{r,i} a_{s,j} \rangle\!\rangle = \langle A_r \rangle \langle A_s \rangle + \frac{1}{N} \langle A_r A_s \rangle_{\mathrm{c}}, \tag{4.9}$$

where use was made of eqns (4.6) and (4.7).

For any $k \geq 1$, the analogous expression for the correlation function (4.8) reads

$$\langle\!\langle \bar{a}_{r_1} \ldots \bar{a}_{r_k} \rangle\!\rangle = \sum_{l=1}^{k} \frac{1}{N^{k-l} l!} \sum_{P \in \mathcal{P}_{k,l}} \langle A_{P_1} \rangle_{\mathrm{c}} \ldots \langle A_{P_l} \rangle_{\mathrm{c}}. \tag{4.10}$$

The second sum in this formula runs over the set $\mathcal{P}_{k,l}$ of all partitions $P = (P_1, \ldots, P_l)$ of the set $\{1, \ldots, k\}$ into l non-empty subsets. Furthermore,

$$A_{P_i} = \prod_{j \in P_i} A_{r_j}. \tag{4.11}$$

Note that the subsets $P_1, \ldots, P_l$ are ordered and therefore distinguished. In particular, at large N the dominant term is

$$\langle\!\langle \bar{a}_{r_1} \ldots \bar{a}_{r_k} \rangle\!\rangle = \langle A_{r_1} \rangle \ldots \langle A_{r_k} \rangle + \mathrm{O}(N^{-1}). \tag{4.12}$$

The proof of eqn (4.10) is a bit technical and is deferred to Section 4.1.4.

The statistical properties of the deviations

$$\delta \bar{a}_r = \bar{a}_r - \langle A_r \rangle \tag{4.13}$$

can now be determined as follows. First note that eqn (4.9) may be rewritten in the form

$$\langle\!\langle \delta \bar{a}_r \delta \bar{a}_s \rangle\!\rangle = \frac{1}{N} \langle A_r A_s \rangle_{\mathrm{c}}. \tag{4.14}$$

On average, the magnitude of the deviations $\delta \bar{a}_r$ is thus proportional to $N^{-1/2}$.

A more detailed characterization of the distribution of the deviations is obtained by working out their higher-order correlations. Starting from eqn (4.10), it is possible to show (Section 4.1.4) that

$$\langle\!\langle \delta\bar{a}_{r_1} \dots \delta\bar{a}_{r_k} \rangle\!\rangle = \sum_{l=1}^{k} \frac{1}{N^{k-l} l!} \sum_{P \in \tilde{\mathcal{P}}_{k,l}} \langle A_{P_1} \rangle_{\mathrm{c}} \dots \langle A_{P_l} \rangle_{\mathrm{c}}, \tag{4.15}$$

where $\tilde{\mathcal{P}}_{k,l} \subset \mathcal{P}_{k,l}$ denotes the set of partitions of $\{1, \dots, k\}$ into l subsets with two or more elements. In particular, for all even k

$$\langle\!\langle \delta\bar{a}_{r_1} \dots \delta\bar{a}_{r_k} \rangle\!\rangle = \frac{1}{N^{k/2}} \left\{ \langle A_{r_1} A_{r_2} \rangle_{\mathrm{c}} \dots \langle A_{r_{k-1}} A_{r_k} \rangle_{\mathrm{c}} + \text{permutations} \right\} + \dots, \tag{4.16}$$

while for all odd k the leading terms are of order $N^{-(k+1)/2}$, because the admissible partitions contain at least one subset with 3 or more elements. Taken together, these results show that the joint probability distribution of the scaled deviations $\sqrt{N}\delta\bar{a}_r$ is gaussian to leading order in $1/N$, with mean zero and variance $\langle A_r A_s \rangle_{\mathrm{c}}$.

4.1.4 Appendix: Proof of eqns (4.10) and (4.15)

Let J_r be real-valued sources for the selected primary observables A_r. The generating function of the correlation functions $\langle A_{r_1} \dots A_{r_n} \rangle$ is a formal power series

$$\mathcal{Z}(J) = 1 + \sum_{n=1}^{\infty} \frac{1}{n!} \sum_{r_1} \dots \sum_{r_n} \langle A_{r_1} \dots A_{r_n} \rangle J_{r_1} \dots J_{r_n} \tag{4.17}$$

in these sources. Similarly, the generating function of the connected parts of the correlation functions is given by

$$\mathcal{W}(J) = \sum_{n=1}^{\infty} \frac{1}{n!} \sum_{r_1} \dots \sum_{r_n} \langle A_{r_1} \dots A_{r_n} \rangle_{\mathrm{c}} J_{r_1} \dots J_{r_n}. \tag{4.18}$$

The moment-cumulant transformation is then summarized by the identity

$$\mathcal{Z}(J) = \mathrm{e}^{\mathcal{W}(J)} \tag{4.19}$$

among formal power series.

The left-hand side of eqn (4.10) is related to the generating function $\mathcal{Z}(J)$ through

$$\begin{aligned} \langle\!\langle \bar{a}_{r_1} \dots \bar{a}_{r_k} \rangle\!\rangle &= \frac{1}{N^k} \left. \frac{\partial^k \langle\!\langle \exp\{\sum_{r,i} a_{r,i} J_r\} \rangle\!\rangle}{\partial J_{r_1} \dots \partial J_{r_k}} \right|_{J=0} \\ &= \frac{1}{N^k} \left. \frac{\partial^k \mathcal{Z}(J)^N}{\partial J_{r_1} \dots \partial J_{r_k}} \right|_{J=0}. \end{aligned} \tag{4.20}$$

Use has here been made of the statistical independence of the data $a_{r,i}$ at different values of the index i and of the fact that their products at fixed i are unbiased estimators

of the correlation functions of the primary observables. The insertion of eqn (4.19) and the subsequent expansion of the exponential function $\mathrm{e}^{\mathcal{W}(J)}$ now leads to the formula

$$\langle\!\langle \bar{a}_{r_1} \dots \bar{a}_{r_k} \rangle\!\rangle = \sum_{l=1}^{k} \frac{1}{N^{k-l} l!} \left. \frac{\partial^k \mathcal{W}(J)^l}{\partial J_{r_1} \dots \partial J_{r_k}} \right|_{J=0} . \tag{4.21}$$

Each derivative in this expression acts on the l factors $\mathcal{W}(J) \dots \mathcal{W}(J)$ one by one. It is then not difficult to convince oneself that the possible distributions of the derivatives to the factors match the possible partitions $P \in \mathcal{P}_{k,l}$, thus proving eqn (4.10).

The proof of eqn (4.15) proceeds in the same way. In order to obtain the correlation functions of the deviations $\delta\bar{a}_r$, it suffices to substitute $a_{r,i} \to a_{r,i} - \langle A_r \rangle$ on the right of eqn (4.20). The generating function in eqn (4.21) then gets replaced by

$$\tilde{\mathcal{W}}(J) = \mathcal{W}(J) - \sum_r \langle A_r \rangle J_r . \tag{4.22}$$

Since at least two derivatives must act on each factor $\tilde{\mathcal{W}}(J)$, the possible distributions of the derivatives to the factors now match the partitions $P \in \tilde{\mathcal{P}}_{k,l}$.

4.2 Physical quantities

The correlation functions of the primary observables only rarely have an immediate physical meaning. In the present context, any well-defined function of the expectation values $\langle A_r \rangle, \langle A_r A_s \rangle, \dots$ is referred to as a physical quantity. Ratios of the expectation values of Wilson loops, for example, are considered to be physical quantities as well as the heavy-quark potential, which is a limit of such ratios.

4.2.1 From the primary observables to the physical quantities

In practice, the correlation functions required for the calculation of a physical quantity are approximated by the averages of the measured values of the appropriate primary observables. A prototype computation of the pion mass m_π, for example, starts from the data series for the observables $\mathcal{O}_\pi(x_0, y)$ at all times x_0 and, say, a single source point y (cf. Section 3.1.1). The so-called effective mass

$$m_{\mathrm{eff}}(x_0) = -a^{-1} \ln \frac{\overline{\mathcal{O}}_\pi(x_0 + a, y)}{\overline{\mathcal{O}}_\pi(x_0, y)} \tag{4.23}$$

is then computed and fitted by a constant in a sensible range of x_0, where the contributions of the higher-energy states to the pion propagator are negligible with respect to the statistical errors. As long as this condition is satisfied, the fitted constant provides a stochastic estimate of the pion mass on the given lattice.

Once a definite fit procedure is adopted, the so calculated values of the pion mass become a complicated but unambiguously defined function of the data series for the observables $\mathcal{O}_\pi(x_0, y)$. It should be noted, however, that this function is not simply a function of the average values of these observables. The fit also requires an estimate of the covariance matrix of the latter as input, which one obtains from the available

data through the jackknife method or in some other way. Evidently, when fits of fitted quantities are considered, as may be the case in studies of the quark-mass dependence of the pion mass, the functional dependence of the calculated quantities on the primary data series is further obscured.

4.2.2 Stochastic estimators

A stochastic estimator of a physical quantity Q is any function ϕ of N and the measured values of the primary observables A_r such that

$$Q = \lim_{N\to\infty} \phi \quad \text{with probability 1.} \tag{4.24}$$

In the following, a class of stochastic estimators with some additional properties will be considered. More precisely, it will be assumed that the asymptotic expansion

$$\phi \underset{N\to\infty}{\sim} \sum_{k=0}^{\infty} N^{-k} \phi^{(k)}(\bar{a}_{r_1}, \bar{a}_{r_2}, \ldots) \tag{4.25}$$

holds, where the coefficients $\phi^{(k)}$ are smooth functions of their arguments. The number of arguments may grow with k but is required to be finite for all k.

It may not be obvious at this point why one needs to consider stochastic estimators with an explicit dependence on N. The leading coefficient in the expansion (4.25) is in fact a valid stochastic estimator for Q which depends only on the averages $\bar{a}_r$ of the primary data series. However, when a stochastic estimator is implicitly defined through a fit procedure, for example, its leading coefficient at large N may be inaccessible in practice. In these cases one does not really have the choice and is forced to deal with the operationally well-defined but N-dependent estimators.

A simple example of an admissible stochastic estimator is provided by the effective mass (4.23). The rôle of the physical quantity Q is here played by the effective mass calculated from the exact pion propagator $g_\pi(x_0 - y_0)$. More complicated, N-dependent stochastic estimators will be discussed in Section 4.3.

4.2.3 Bias and covariance matrix

Let Q_α be some physical quantities, labeled by an index α, and ϕ_α stochastic estimators of these. On average, the values of the estimators obtained in a simulation approximate the physical quantities up to a deviation given by

$$B_\alpha = \langle\langle \delta\phi_\alpha \rangle\rangle, \qquad \delta\phi_\alpha = \phi_\alpha - Q_\alpha. \tag{4.26}$$

B_α is referred to as the bias of the chosen estimators. The statistical fluctuations of the measured values are described by the covariance matrix

$$C_{\alpha\beta} = \langle\langle \delta\phi_\alpha \delta\phi_\beta \rangle\rangle \tag{4.27}$$

and are usually significantly larger than the bias.

For $N \to \infty$, the bias and the covariance matrix can be expanded in a series in inverse powers of N with coefficients depending on the correlation functions $\langle A_{r_1} \dots A_{r_n} \rangle$ of the primary observables. To show this, first recall that

$$\bar{a}_r = \langle A_r \rangle + \delta\bar{a}_r, \qquad \delta\bar{a}_r = \mathrm{O}(N^{-1/2}). \tag{4.28}$$

The Taylor expansion of the coefficient functions in eqn (4.25) in powers of the deviations $\delta\bar{a}_r$ then leads to the expression

$$\begin{aligned} \phi_\alpha = \hat{\phi}_\alpha^{(0)} &+ \sum_r \partial_r \hat{\phi}_\alpha^{(0)} \delta\bar{a}_r \\ &+ \frac{1}{N}\hat{\phi}_\alpha^{(1)} + \tfrac{1}{2}\sum_{r,s} \partial_r \partial_s \hat{\phi}_\alpha^{(0)} \delta\bar{a}_r \delta\bar{a}_s + \mathrm{O}(N^{-3/2}) \end{aligned} \tag{4.29}$$

in which

$$\hat{\phi}_\alpha^{(k)} = \phi_\alpha^{(k)}(\langle A_{r_1}\rangle, \langle A_{r_2}\rangle, \dots), \qquad \partial_r = \frac{\partial}{\partial\langle A_r\rangle}. \tag{4.30}$$

Evidently, since the leading term $\hat{\phi}_\alpha^{(0)}$ coincides with Q_α, the deviation $\delta\phi_\alpha$ is given by the sum of all other terms on the right of eqn (4.29).

It is important to realize that the only stochastic variables in the expansion (4.29) are the deviations $\delta\bar{a}_r$. The averages (4.26) and (4.27) over repeated simulations can therefore be computed straightforwardly using the results obtained in Section 4.1. In a few lines one then obtains the result

$$B_\alpha = \frac{1}{N}\Big\{\hat{\phi}_\alpha^{(1)} + \tfrac{1}{2}\sum_{r,s} \partial_r \partial_s \hat{\phi}_\alpha^{(0)} \langle A_r A_s \rangle_{\mathrm{c}}\Big\} + \mathrm{O}(N^{-2}), \tag{4.31}$$

$$C_{\alpha\beta} = \frac{1}{N}\sum_{r,s} \partial_r \hat{\phi}_\alpha^{(0)} \partial_s \hat{\phi}_\beta^{(0)} \langle A_r A_s \rangle_{\mathrm{c}} + \mathrm{O}(N^{-2}). \tag{4.32}$$

In particular, the bias of the stochastic estimators is of order N^{-1} while their statistical fluctuations are of order $N^{-1/2}$.

4.3 Jackknife error estimation

The bias B_α and the covariance matrix $C_{\alpha\beta}$ usually need to be estimated from the data. Such estimates can in principle be obtained via eqns (4.31) and (4.32) by calculating the expectation values and two-point functions of the relevant primary observables. The coefficient functions $\phi^{(0)}$ and $\phi^{(1)}$ must be known explicitly in this case.

The jackknife method allows the bias and covariance matrix to be estimated even if the coefficient functions are inaccessible or too complicated to be used directly. More precisely, the method constructs stochastic estimators for the large-N limits of NB_α and $NC_{\alpha\beta}$ (which are functions of the expectation values of the primary observables and therefore physical quantities).

4.3.1 Jackknife samples

A jackknife sample of the measured values $a_{r,1}, \ldots, a_{r,N}$ of a primary observable A_r is obtained by omitting one measurement from the full series. If, say, the i'th measurement is omitted, the corresponding jackknife sample consists of the measurements

$$a_{r,1}, \ldots, a_{r,i-1}, a_{r,i+1}, \ldots a_{r,N}. \tag{4.33}$$

Evidently, there are N distinct jackknife samples of $N-1$ measurements, labeled by the number i of the omitted measurement.

The average of the measurements included in the jackknife sample number i is denoted by

$$\bar{a}^J_{r,i} = \frac{1}{N-1} \sum_{j=1, j\neq i}^{N} a_{r,j}. \tag{4.34}$$

More generally, a stochastic estimator ϕ assumes some value ϕ^J_i if the i'th measurement of the primary observables is discarded. Note that the jackknife samples are treated like any other measurement series of length $N-1$ in this context. From the expansion (4.25) one then infers that

$$\phi^J_i \underset{N\to\infty}{\sim} \sum_{k=0}^{\infty} (N-1)^{-k} \phi^{(k)}(\bar{a}^J_{r_1,i}, \bar{a}^J_{r_2,i}, \ldots), \tag{4.35}$$

where the coefficient functions $\phi^{(k)}$ are the same as before.

4.3.2 Estimators for the bias and the covariance matrix

The jackknife estimators for the bias and the covariance matrix are now given by the elegant formulae

$$B^J_\alpha = \sum_{i=1}^{N} (\phi^J_{\alpha,i} - \phi_\alpha), \tag{4.36}$$

$$C^J_{\alpha\beta} = \sum_{i=1}^{N} (\phi^J_{\alpha,i} - \phi_\alpha)(\phi^J_{\beta,i} - \phi_\beta). \tag{4.37}$$

An important result of the discussion below is going to be that the scaled expressions NB^J_α and $NC^J_{\alpha\beta}$ are stochastic estimators in the sense of Section 4.2.2, the associated physical quantities being

$$\lim_{N\to\infty} NB_\alpha = \hat{\phi}^{(1)}_\alpha + \tfrac{1}{2} \sum_{r,s} \partial_r \partial_s \hat{\phi}^{(0)}_\alpha \langle A_r A_s \rangle_{\mathrm{c}}, \tag{4.38}$$

$$\lim_{N\to\infty} NC_{\alpha\beta} = \sum_{r,s} \partial_r \hat{\phi}^{(0)}_\alpha \partial_s \hat{\phi}^{(0)}_\beta \langle A_r A_s \rangle_{\mathrm{c}}. \tag{4.39}$$

The bias B_α and the covariance matrix $C_{\alpha\beta}$ are therefore approximated by the estimators B^J_α and $C^J_{\alpha\beta}$ up to statistical fluctuations of order $N^{-3/2}$ and further terms of order N^{-2}.

The expansion of the jackknife estimators (4.36) and (4.37) for $N \to \infty$ is obtained straightforwardly by substituting

$$\bar{a}^J_{r,i} = \bar{a}_r + \frac{1}{N-1}(\bar{a}_r - a_{r,i}) \tag{4.40}$$

in eqn (4.35) and by systematically expanding all terms in powers of N^{-1}. In the case of the covariance matrix, for example, this leads to the expression

$$C^J_{\alpha\beta} = \frac{1}{N}\sum_{r,s} \bar{\partial}_r \phi^{(0)}_\alpha(\bar{a}_{r_1}, \ldots)\bar{\partial}_s \phi^{(0)}_\beta(\bar{a}_{r_1}, \ldots)(\bar{a}_{rs} - \bar{a}_r\bar{a}_s) + \mathrm{O}(N^{-2}), \tag{4.41}$$

where the notation has been simplified by setting

$$\bar{a}_{rs} = \frac{1}{N}\sum_{i=1}^{N} a_{r,i}a_{s,i}, \qquad \bar{\partial}_r = \frac{\partial}{\partial \bar{a}_r}. \tag{4.42}$$

The higher-order terms in eqn (4.41) involve averages of increasingly longer products of the measured values $a_{r,i}$, but are otherwise of the same structure as the leading one. Since the corresponding products of the primary observables are primary observables too, the expansion shows that $NC^J_{\alpha\beta}$ is a stochastic estimator of the kind specified in Section 4.2.2. Moreover, proceeding as in Section 4.2.3, the physical quantity approximated by $NC^J_{\alpha\beta}$ is seen to coincide with the expression on the right of eqn (4.39), as asserted above.

An interesting feature of the jackknife method is the fact that it does not require the derivatives of the coefficient function $\phi^{(0)}$ in the asymptotic expressions (4.38) and (4.39) to be calculated. As is evident from the derivation of eqn (4.41), the method effectively performs a numerical differentiation by probing the stochastic estimators ϕ_α through the jackknife samples. The estimators are thus implicitly assumed to vary smoothly on the scale of the deviations $\bar{a}^J_{r,i} - \bar{a}_r = \mathrm{O}(N^{-1})$.

4.3.3 Error propagation

The jackknife method is easy to apply, because it requires no more than the evaluation of the stochastic estimators of interest for the full sample and the jackknife samples of the measured values of the primary observables. Note that $\psi^J_i = \psi(\phi^J_{1,i}, \ldots, \phi^J_{m,i})$ if ψ is a function of the stochastic estimators $\phi_1, \ldots, \phi_m$. The calculation can therefore often be organized in a hierarchical manner, where one proceeds from the primary observables to more and more complicated stochastic estimators.

In the case of the computation of the pion mass sketched in Section 4.2.1, for example, the estimation of the statistical errors starts from the jackknife averages $\overline{\mathcal{O}}_\pi(x_0, y)^J_i$ of the primary observables $\mathcal{O}_\pi(x_0, y)$. The bias and covariance matrix of the effective mass are then obtained from the jackknife sample values

$$m_{\mathrm{eff}}(x_0)^J_i = -a^{-1} \ln \frac{\overline{\mathcal{O}}_\pi(x_0 + a, y)^J_i}{\overline{\mathcal{O}}_\pi(x_0, y)^J_i} \tag{4.43}$$

using eqns (4.36) and (4.37). In particular, the jackknife estimator of the covariance matrix is given by

$$C^J_{x_0x'_0} = \sum_{i=1}^{N} \{m_{\mathrm{eff}}(x_0)^J_i - m_{\mathrm{eff}}(x_0)\}\{m_{\mathrm{eff}}(x'_0)^J_i - m_{\mathrm{eff}}(x'_0)\}. \tag{4.44}$$

The statistical error of $m_{\mathrm{eff}}(x_0)$, for example, is estimated to be $(C^J_{x_0x_0})^{1/2}$.

In the next step, the pion mass m_π is computed by minimizing the χ^2-statistic

$$\chi^2 = \sum_{x_0=t_0}^{t_1} \sum_{x'_0=t_0}^{t_1} \{m_\pi - m_{\mathrm{eff}}(x_0)\}[(C^J)^{-1}]_{x_0x'_0}\{m_\pi - m_{\mathrm{eff}}(x'_0)\} \tag{4.45}$$

in some range $[t_0, t_1]$ of time. The pion mass

$$m_\pi = \frac{\sum_{x_0,x'_0=t_0}^{t_1}[(C^J)^{-1}]_{x_0x'_0} m_{\mathrm{eff}}(x'_0)}{\sum_{x_0,x'_0=t_0}^{t_1}[(C^J)^{-1}]_{x_0x'_0}} \tag{4.46}$$

thus calculated is algebraically expressed through the effective mass and the scaled jackknife estimator of the associated covariance matrix. In particular, since all of these are stochastic estimators of the kind specified in Section 4.2.2, the same is the case for the pion mass determined in this way. Its statistical error can therefore be computed using the jackknife method.

Following the general rules, the application of the jackknife method requires the pion mass to be recomputed for each jackknife sample of the primary data. In particular, the covariance matrix of the effective mass must be recomputed, which requires the jackknife samples of the jackknife samples to be considered, i.e. samples where a second measurement is discarded from the full data series. The computational effort thus grows like N^2, but the calculation is otherwise entirely straightforward.

In practice a simplified procedure is often adopted, where the statistical error of the covariance matrix is ignored. The matrix is calculated for the full sample in this case and is held fixed when the errors of the effective mass are propagated to those of the pion mass. An advantage of this procedure is that the computational effort increases like N rather than N^2, but the correctness of the error estimation can only be shown for asymptotically large values of N and if the systematic deviation of the effective mass from being constant in time is much smaller than the statistical errors. Moreover, the bias of the pion mass is no longer correctly given by the jackknife formula (4.36). It is therefore advisable to pass to the simplified procedure only after having checked its consistency with the results of the full procedure.

References

Abramowitz, M. and Stegun, I. A. (1972). *Handbook of Mathematical Functions*. Dover Publications, New York.

Achiezer, N. I. (1992). *Theory of Approximation*. Dover Publications, New York.

Aoki, S. et al. (2002). *Phys. Rev.*, **D65**, 094507.

Bali, G. S. et al. (2005). *Phys. Rev.*, **D71**, 114513.

Banks, T. and Casher, A. (1980). *Nucl. Phys.*, **B169**, 103.

Brannick, J. et al. (2008). *Phys. Rev. Lett.*, **100**, 041601.

Brower, R. C., Ivanenko, T., Levi, A. R., and Orginos, K. N. (1997). *Nucl. Phys.*, **B484**, 353.

Brown, F. R. and Woch, T. J. (1987). *Phys. Rev. Lett.*, **58**, 2394.

Cabibbo, N. and Marinari, E. (1982). *Phys. Lett.*, **119B**, 387.

Clark, M. A. et al. (2008*a*). *PoS(LATTICE 2008)*, 035.

Clark, M. A. and Kennedy, A. D. (2007). *Phys. Rev. Lett.*, **98**, 051601.

Clark, M. A., Kennedy, A. D., and Silva, P. J. (2008*b*). *PoS(LATTICE 2008)*, 041.

Creutz, M. (1980). *Phys. Rev.*, **D21**, 2308.

Creutz, M. (1987). *Phys. Rev.*, **D36**, 515.

DeGrand, T. A. and Schaefer, S. (2004). *Comput. Phys. Commun.*, 185.

Del Debbio, L. et al. (2006). *J. High Energy Phys.*, **0602**, 011.

Del Debbio, L. et al. (2007). *J. High Energy Phys.*, **0702**, 056.

Della Morte, M. and Giusti, L. (2009). *Comput. Phys. Commun.*, **180**, 819.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). *Phys. Lett.*, **B195**, 216.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, CRC Press, Boca Raton.

Fabricius, K. and Haan, O. (1984). *Phys. Lett.*, **143B**, 459.

Foley, J. et al. (2005). *Comput. Phys. Commun.*, **172**, 145.

Frezzotti, R. and Jansen, K. (1997). *Phys. Lett.*, **B402**, 328.

Frezzotti, R. and Jansen, K. (1999). *Nucl. Phys.*, **B555**, 395 and 432.

Frommer, A. et al. (1994). *Int. J. Mod. Phys.*, **C5**, 1073.

Giusti, L. et al. (2004). *J. High Energy Phys.*, **0404**, 013.

Giusti, L., Hoelbling, C., Lüscher, M., and Wittig, H. (2003). *Comput. Phys. Commun.*, **153**, 31.

Giusti, L. and Lüscher, M. (2009). *J. High Energy Phys.*, **0903**, 013.

Golub, G. H. and van Loan, C. F. (1989). *Matrix Computations* (2nd edn). The Johns Hopkins University Press, Baltimore.

Hairer, E., Lubich, C., and Wanner, G. (2006). *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations* (2nd edn). Springer, Berlin.

Hasenbusch, M. (2001). *Phys. Lett.*, **B519**, 177.
Hasenbusch, M. and Jansen, K. (2003). *Nucl. Phys.*, **B659**, 299.
Horvath, I., Kennedy, A. D., and Sint, S. (1999). *Nucl. Phys. (Proc.Suppl.)*, **73**, 834.
James, F. (1994). *Comput. Phys. Commun.*, **79**, 111. [E: *ibid.* **97** (1996) 357].
Jegerlehner, B. (1996). arXiv:hep-lat/9612014 (unpublished).
Kennedy, A. D. and Pendleton, B. J. (1985). *Phys. Lett.*, **156B**, 393.
Knuth, D. E. (1997). *The Art of Computer Programming* (3rd edn). Volume 2. Addison–Wesley, Reading MA.
Kratochvila, S. and de Forcrand, Ph. (2003). *Nucl. Phys.*, **B671**, 103.
Leimkuhler, B. and Reich, S. (2004). *Simulating Hamiltonian Dynamics*. Cambridge University Press, Cambridge.
Lüscher, M. (1994). *Comput. Phys. Commun.*, **79**, 100.
Lüscher, M. (2004). *Comput. Phys. Commun.*, **156**, 209.
Lüscher, M. (2005). *Comput. Phys. Commun.*, **165**, 199.
Lüscher, M. (2007*a*). *J. High Energy Phys.*, **0707**, 081.
Lüscher, M. (2007*b*). *J. High Energy Phys.*, **0712**, 011.
Lüscher, M. and Weisz, P. (2001). *J. High Energy Phys.*, **09**, 010.
Lüscher, M. and Weisz, P. (2002). *J. High Energy Phys.*, **07**, 049.
Madras, N. and Sokal, A. D. (1988). *J. Stat. Phys.*, **50**, 109.
Metropolis, N. et al. (1953). *J. Chem. Phys.*, **21**, 1087.
Meyer, H. B. (2003). *J. High Energy Phys.*, **0301**, 048.
Meyer, H. B. (2004). *J. High Energy Phys.*, **0401**, 030.
Meyer, H. B. (2008). *Phys. Rev. Lett.*, **100**, 162001.
Michael, C. and Peisa, J. (1998). *Phys. Rev.*, **D58**, 034506.
Neff, H. et al. (2001). *Phys. Rev.*, **D64**, 114509.
Omelyan, I. P., Mryglod, I. M., and Folk, R. (2002). *Phys. Rev.*, **E65**, 056706.
Omelyan, I. P., Mryglod, I. M., and Folk, R. (2003). *Comput. Phys. Commun.*, **151**, 272.
Parisi, G., Petronzio, R., and Rapuano, F. (1983). *Phys. Lett.*, **B128**, 418.
Pepe, M. and Wiese, U.-J. (2009). *Phys. Rev. Lett.*, **102**, 191601.
Saad, Y. (2003). *Iterative Methods for Sparse Linear Systems* (2nd edn). SIAM, Philadelphia. See also `http://www-users.cs.umn.edu/~saad/`.
Sexton, J. C. and Weingarten, D. H. (1992). *Nucl. Phys.*, **B380**, 665.
Takaishi, T. and de Forcrand, Ph. (2002). *Int. J. Mod. Phys.*, **C13**, 343.
Takaishi, T. and de Forcrand, Ph. (2006). *Phys. Rev.*, **E73**, 036706.
Tierney, L. (1994). *Ann. Statist.*, **22**, 1701.
Ukawa, A. (2002). *Nucl. Phys. B (Proc. Suppl.)*, **106–107**, 195.
Urbach, C., Jansen, K., Shindler, A., and Wenger, U. (2006). *Comput. Phys. Commun.*, **174**, 87.
van der Vorst, H. A. (1992). *SIAM J. Sci. Stat. Comput.*, **13**, 631.
Wilcox, W. (2007). *PoS(LATTICE 2007)*, 025.
Wolff, U. (2004). *Comput. Phys. Commun.*, **156**, 143. [E: *ibid.* **176** (2007) 383].