

Variance reduction strategies for lattice QCD

Tim Harris*

*Institute for Theoretical Physics,
Department of Physics, ETH Zürich, Switzerland.*

E-mail: harrist@phys.ethz.ch

A significant component of the cost of making predictions from lattice QCD stems from the computation of correlation functions on a given ensemble of gauge fields. This cost depends on the observable of interest and the details of its representation, including any approximation needed to estimate it. Moreover, the variance of such estimators may depend strongly on physical and kinematical parameters such as the lattice spacing, volume or separation, which gives an important insight into the costs of reaching the relevant physical limits. In these proceedings, I review some observables involving quark propagators, including both quark-line connected and disconnected Wick contractions, and discuss variance-reduction schemes based on decompositions of the quark propagators. Such strategies have already proven useful for precision physics observables and in future may help reduce the computational cost of reaching large volumes.

*The 42nd International Symposium on Lattice Field Theory (LATTICE2025)
2-8 November 2025
Tata Institute of Fundamental Research, Mumbai, India*

*Speaker

1. Introduction

Signal-to-noise problems are ubiquitous in Monte Carlo simulations of lattice field theories. In particular, deteriorating signal-to-noise ratios in the various numerical limits required to make physics predictions may significantly hinder the predictive power of lattice computations. Let us immediately consider a concrete example, pure SU(3) gauge theory regulated on a toroidal hypercubic lattice with lattice spacing a and side length L defined by the Wilson action [1]

$$S_g = \frac{2}{g_0^2} a^4 \sum_x s(x), \quad s(x) = \frac{1}{a^4} \sum_{\mu < \nu} \text{Re Tr}\{1 - P_{\mu\nu}(x)\} \quad (1)$$

where $P_{\mu\nu}(x) = U_\mu(x)U_\nu(x + a\hat{\mu})U_\mu^\dagger(x + a\hat{\nu})U_\nu^\dagger(x)$ is the plaquette based at site x oriented in the $\mu\nu$ plane. After the subtraction of its vacuum expectation value, the product of the density $s(x)$ and the (lattice) beta function $dg_0^{-2}/d\ln a$ is a renormalized field [2] which we denote $\phi(x)$, and therefore its connected Euclidean time-correlation function

$$C(x_0 - y_0) = a^3 \sum_{\vec{x}} \left[\langle \phi(x)\phi(y) \rangle - \langle \phi(x) \rangle \langle \phi(y) \rangle \right] \quad (2)$$

has a finite value for non-zero separations, $x_0 - y_0 > 0$, in the continuum and infinite volume limit, and moreover is exponentially suppressed at large separations due to the mass gap m of the theory.

A standard Monte Carlo simulation then yields an estimate for each of the path integrals in eq. (2) via the sample mean, for example

$$\langle \phi(x) \rangle \approx \frac{1}{N} \sum_{i=1}^N \phi^i(x), \quad (3)$$

thanks to the law of large numbers, where the superscript indicates the density is evaluated on the i th field sample. Assuming, then, that the samples are independent and identically distributed, the variance of the plug-in estimator is found to be

$$\frac{\sigma^2}{N} = \frac{1}{N} \left[a^6 \sum_{\vec{x}, \vec{x}'} \delta_{x_0, x_0'} \langle \phi(x)\phi(x') \rangle_c \langle \phi(y)\phi(y) \rangle_c + \mathcal{O}(e^{-m(x_0 - y_0)}) \right] + \dots, \quad (4)$$

where the ellipses denote terms higher order in $1/N$. The utility of this formula is that the statistical uncertainty is completely determined by the physics of the underlying system and therefore much can be learned about the signal-to-noise ratio in the interesting limits.

One immediate consequence of the preceding formula is that, although the correlation function is well-behaved in the limits $a \rightarrow 0$, $L \rightarrow \infty$ and $x_0 - y_0 \rightarrow \infty$, the same is not true for the variance of its Monte Carlo estimator. Indeed, one may expect the following asymptotic behaviour for the signal-to-noise ratio in those same limits

$$\frac{C(x_0 - y_0)}{\sqrt{\sigma^2/N}} \sim a^5 \sqrt{\frac{a^3}{L^3}} e^{-m(x_0 - y_0) \sqrt{N}}, \quad (5)$$

by considering the dominant contributions coming from the points where the fields coincide and these products mix with the vacuum, see Ref. [3] for another worked example. The severity of such problems has long been appreciated [4, 5], and can be (partially) dealt with. In fact the simplest improvement is to average over the volume, and we content ourselves with this modest goal in the rest of these proceedings.

1.1 Translation averaging

As is well-known (e.g. Ref. [6] for an exposition), assuming local translation invariance, an improved estimator for the correlation function is obtained by averaging the product of fields over a region of the lattice, for example on a time-slice

$$C_{\text{vol}}(x_0 - y_0) = \frac{a^3}{L^3} \sum_{\vec{y}} a^3 \sum_{\vec{x}} \left[\langle \phi(x) \phi(y) \rangle - \langle \phi(x) \rangle \langle \phi(y) \rangle \right] \quad (6)$$

which has the beneficial effect of suppressing the variance of the corresponding estimator by a factor $(a/L)^3$ and therefore enhancing the signal-to-noise ratio

$$\frac{C(x_0 - y_0)}{\sqrt{\sigma_{\text{vol}}^2/N}} \sim a^5 e^{-m(x_0 - y_0)} \sqrt{N} \quad (7)$$

Even better, is to avoid integrating the fields over large transverse separations in the first place (see Refs. [7–13] for some examples), and say, cutting off the contribution from the tail after some distance R ,

$$C_{\text{vol},R}(x_0 - y_0) = \frac{a^3}{L^3} \sum_{\vec{y}} a^3 \sum_{\vec{x}} \theta(R^2 - |\vec{x} - \vec{y}|^2) \left[\langle \phi(x) \phi(y) \rangle - \langle \phi(x) \rangle \langle \phi(y) \rangle \right] \quad (8)$$

giving rise to a signal-to-noise ratio which even increases with the lattice size

$$\frac{C(x_0 - y_0)}{\sqrt{\sigma_{\text{vol},R}^2/N}} \sim a^5 \sqrt{\frac{L^3}{R^3}} e^{-m(x_0 - y_0)} \sqrt{N}. \quad (9)$$

That estimator has the attractive feature of improving the statistical and systematic error associated to finite-volume effects in one go, if the regime can be reached for which L/R is sufficiently large that the tail contribution is under control. Of course, the formulae presented so far are applicable only to the asymptotic regions of parameter space: the precise scaling in intermediate regimes may be different and motivate the generalization of the averages to other domains which may or may not be contiguous, as well as the obvious generalization to averaging over the time extent.

Other n -point functions can be analysed in much the same spirit as, almost universally, the problematic features arise from the fact that the variance is not a fully-connected correlation function and disconnected components are practically impossible to avoid. An important exception to this rule will be discussed in the next section. Certain advanced strategies may mitigate the poor scaling in the other variables, by using for example multi-level integration discussed in Sec. 4. However, even implementing the most simple strategy, namely translation averaging, is not always easy.

1.2 Translation averaging in QCD

So far, we have neglected to discuss the cost of the three different estimators presented. Happily, due to the fact that the products of local fields discussed so far factorize, the translation average of the previous section can be evaluated with the help of the convolution theorem and so its cost grows only like $V \ln V$, where $V = (L/a)^3$ in the case above, i.e. not too much more than the original

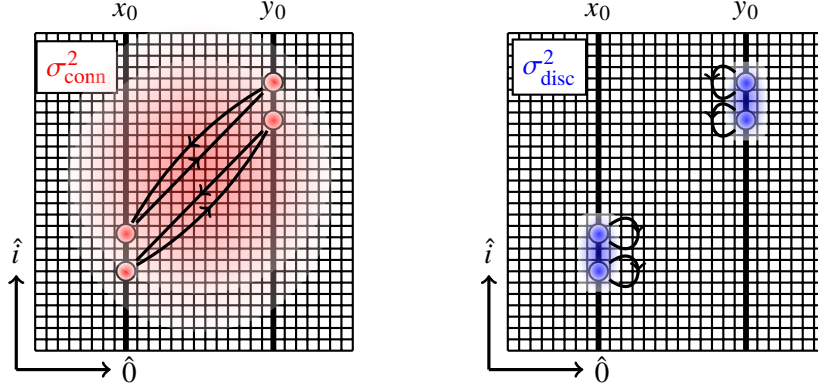


Figure 1: Illustration of the Wick contractions which appear in the variances for a quark-line connected (left) and disconnected (right) primary observable.

estimator. In general, however, the cost to achieve a reduced variance may not be so favourable and one should keep in mind that the relevant metric to be optimized is, for example the computational cost for a fixed precision, $\varepsilon = \sqrt{\sigma^2/N}$, which is

$$\text{total cost} = N \times \text{cost per field} = \varepsilon^{-2} [\sigma^2 \times \text{cost per field}]. \quad (10)$$

A good variance reduction strategy minimizes the product in the brackets, i.e. the effective cost.

In QCD, in the typical representation of the path integral, the local quark fields are integrated out by hand, leaving observables written in terms of products of quark propagators of flavour f , $S_f(x, y) = \langle \psi_f(x) \bar{\psi}_f(y) \rangle_F$, the Greens function for the lattice Dirac operator D_f , and consequently no longer necessarily factorize. To implement translation averaging exactly, the quark propagator should be computed for all x and y which is a problem with $O(V^2)$ complexity and thus the effective cost grows with the volume. In the sense of the above, then, we seek estimators which minimize the effective cost by approximating the translation average, and ideally one for which it scales favourably with the physical volume.

As a prototypical set of observables, we consider the Euclidean-time correlation function of bilinears of fermion fields, and in particular the electromagnetic current, $j_\mu = \sum_f Q_f \bar{\psi}_f \gamma_\mu \psi_f$, where Q_f is the electric charge of the quark flavour f , which is an especially important operator as its hadronic matrix elements parameterize photon-hadron interactions in the Standard Model [14].

The bare spatial current correlator

$$G(x_0, y_0) = \frac{a^3}{L^3} \sum_{\vec{y}} a^3 \sum_{\vec{x}} \sum_{f,g} Q_f Q_g \langle T_{k,f}^{(1)}(x) T_{k,g}^{(1)}(y) - T_{k,fg}^{(2)}(x, y) \delta_{fg} \rangle \quad (11)$$

where the brackets denote the expectation with respect to the effective distribution $\det\{D_f\}^{N_t} e^{-S_g}$, can be written in terms of two traces, the single-propagator trace

$$T_{\mu,f}^{(1)}(x) = \text{tr}\{\gamma_\mu S_f(x, x)\}, \quad (12)$$

which is one factor of the quark-line disconnected graph, and the trace of two quark propagators

$$T_{\mu,fg}^{(2)}(x, y) = \text{tr}\{\gamma_\mu S_f(x, y) \gamma_\mu S_g(y, x)\}, \quad (13)$$

which we refer to as the quark-line connected Wick contraction.

The variances are now defined in terms of the simulated system, and typically cannot be represented in terms of correlation functions of the original fields, so the analysis of the pure gauge theory must be extended to a partially-quenched set-up (e.g. Refs. [15, 16]). Interestingly, the consequences of operating with the non-local Wick contractions as observables is not totally detrimental: the explicit factors of the quark propagator, which is naturally small $O(e^{-m_\pi|x-y|/2})$ independently of the field (its variance is the pion propagator), suppresses the variance of the quark-line connected contractions with physical separations, and even forbids mixing of the effective vertex with the vacuum state, see Fig. 1 for a sketch. That is, the variance, too, is a fully-connected graph and suffers no power-law divergences in the continuum limit while the contributions with large transverse separations are naturally suppressed. Quark-line disconnected diagrams, which arise whenever singlet operators are considered, on the other hand, are presumably afflicted by the issues described in previous section [17, 18], although empirical evidence may suggest the asymptotic regime is not reached [19]. This has important consequences for the method of insertions [20], for example the computation of isospin-breaking effects where disconnected diagrams necessarily arise in the Rome-123 approach [21]. Therefore, it is useful to distinguish the two cases, and in practice it points to the fact that schemes with a good representation of the quark propagator at small distances for the disconnected graphs and at large distances for the connected graphs are essential.

1.3 Stochastic estimates for translation averages

As outlined in the previous section, constructing the translation average explicitly for observables built from quark propagators is prohibitively expensive and even and has an effective cost which increases with the physical volume. That is, it would be more favourable to increase the number of configurations than to increase the volume from the point of view of the statistical precision. Instead, we can opt to estimate the sums over the coordinates stochastically, and in this work we opt to make use of generalizations of the simple Hutchinson estimator for the trace of a matrix [22–26], defined by computing quadratures with a set of dimensionless random fields $\eta_i(x)$ whose independent components have zero mean and unit variance¹ and support on a time-slice x_0

$$a^3 \sum_{\vec{x}} T_{\mu,f}^{(1)}(x) \approx \frac{1}{N_s} \sum_{i=1}^{N_s} a^3 \sum_{\vec{x}} \eta_i^\dagger(x) \gamma_\mu (S_f \eta_i)(x), \quad (14)$$

$$a^6 \sum_{\vec{x}, \vec{y}} T_{\mu,fg}^{(2)}(x, y) \approx \frac{1}{N_s} \sum_{i=1}^{N_s} a^3 \sum_{\vec{y}} (\eta_i^\dagger \gamma_\mu S_f)(y) \gamma_\mu (S_g \eta_i)(y). \quad (15)$$

Indeed, with that choice, the variance of the single-propagator trace can be worked out as [17]

$$\begin{aligned} \sigma^2 = a^6 \sum_{\vec{x}, \vec{x}'} \delta_{x_0, x'_0} & \left[\langle T_{\mu,f}^{(1)}(x) T_{\mu,f}^{(1)*}(x') \rangle - \langle T_{\mu,f}^{(1)}(x) \rangle \langle T_{\mu,f}^{(1)*}(x') \rangle \right. \\ & \left. + \frac{1}{N_s} \langle T_{5,ff}^{(2)}(x, x') \rangle \right]. \end{aligned} \quad (16)$$

¹One may wonder whether the choice of the white noise distribution has an effect on the variance [27], but in fact, it appears to play a minor role in practice, and choosing Gaussian fields simplifies the analysis of the variance.

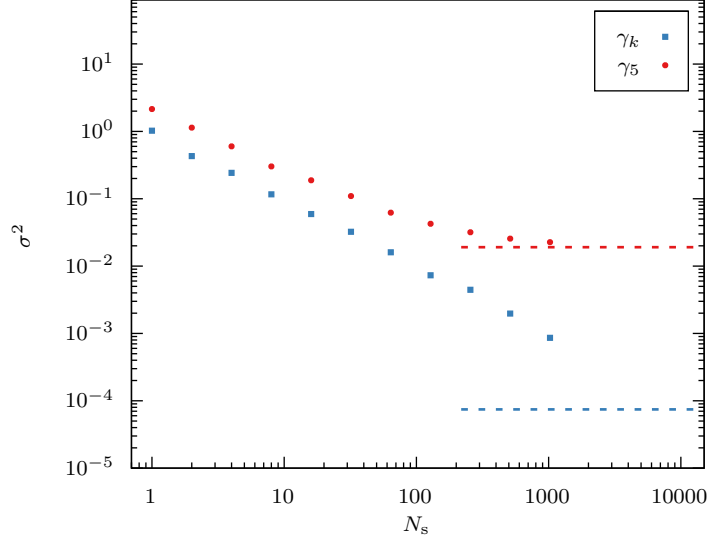


Figure 2: The variance of the Hutchinson estimator for the single-propagator trace $T_{\mu,f}^{(1)}$ for $\mu = k$ (blue squares) and $\mu = 5$ (red circles). The dashed lines indicate the gauge variance which differs by many orders of magnitude for the two cases unlike the stochastic variance which is identical if the reality of observable is ignored.

The first line corresponds to the variance associated with the gauge fields (often referred to as the gauge variance σ_g^2) while the term on the second line is due to the auxiliary random fields and, as expected vanishes in the limit of a large number of fields, $N_s \rightarrow \infty$. This illustrates that, although the estimator provides an unbiased estimate of the translation average, and its cost depends on the number of auxiliary random (source) fields and not explicitly on the volume, the variance receives an extra contribution, which in the case of the vector current is large.

This is visible from simulations performed with $O(a)$ -improved Wilson fermions and a pion mass of $m_\pi \approx 270$ MeV and $m_\pi L \approx 4$ (described in detail in Sec. 3) depicted in Fig. 2 which shows that the additional contribution is enhanced by several orders of magnitude compared to the gauge variance, indicated with the horizontal line. Some indications of this problematic behaviour from the formula of the variance come from the fact that the additional term corresponds to the integrated pion propagator independently of the gamma-matrix in the original estimator, which also has a non-zero value at leading order in perturbation theory, unlike the gauge variance.

A similar exercise reveals an analogous situation for the trace of two propagators for the vector current [6, 28]. That is, in practice such simple stochastic estimators introduce large additional fluctuations and which greatly increase the effective cost even if they solve the volume-scaling problem in principle. The goal then is to find improved estimators based on decompositions of the quark propagators which allow us to efficiently reduce the effective cost by coming up with good and inexpensive approximations to the propagator either at short- or long-distances. If the approximation is cheap, then it can be estimated precisely by sampling frequently, and if the approximation is good then the correction will be suppressed and consequently also have a small error.

2. Multigrid low mode averaging

In order to suppress the variance for the stochastic estimator for the quark-line connected diagram corresponding to the trace of two quark propagators of eq. (13), we seek a good approximate representation of the quark propagator at long distances. To this end, we utilize the concept of deflation of the lattice Dirac operator D , by introducing a subspace spanned by N_v global quark fields $\phi_1(x), \dots, \phi_{N_v}(x)$. In this section, we drop the flavour index f for brevity. Associated with this deflation subspace, one can define a projector R defined through its action on a quark field ψ by $(R\psi)_i = (\phi_i, \psi)$, and an associated little Dirac operator $\hat{Q} = RQR^\dagger$, which acts in the subspace, sometimes known as the Galerkin coarsening, of the Hermitian Dirac operator $Q = \gamma_5 D$.

If the observable is dominated by propagation in this subspace, then one may expect the operator $R^\dagger \hat{Q}^{-1} R$ to furnish us with a good approximation to it. We can form a decomposition by addition and subtraction

$$S(x, y) = \underbrace{\{S(x, y) - (R^\dagger \hat{Q}^{-1} R \gamma_5)(x, y)\}}_{S_0(x, y)} + \underbrace{(R^\dagger \hat{Q}^{-1} R \gamma_5)(x, y)}_{S_1(x, y)} \quad (17)$$

where the term in braces corresponds to the Green's function for the deflated system, and is related to the coarse-grid correction operator as it serves to correct the approximant given by the second term. If the approximation is a good one, one may expect that the first term along with its variance will be highly suppressed. Inserting the decomposition of the quark propagator in the quark-line connected diagram, we arrive at the decomposition

$$T_\mu^{(2)}(x, y) = \text{tr}\{\gamma_\mu \mathcal{S}_0(x, y) \gamma_\mu \mathcal{S}_0(y, x)\} + \text{tr}\{\gamma_\mu \mathcal{S}_0(x, y) \gamma_\mu \mathcal{S}_1(y, x)\} + \text{tr}\{\gamma_\mu \mathcal{S}_1(x, y) \gamma_\mu \mathcal{S}_0(y, x)\} \\ + \text{tr}\{\gamma_\mu \mathcal{S}_1(x, y) \gamma_\mu \mathcal{S}_1(y, x)\} \quad (18)$$

where each term in the first line requires the expensive deflated operator to be computed, while the term on the second line requires only the inversion of the little operator.

If N_v is sufficiently small, then the little operator is clearly cheap to apply, while if it is furthermore well-conditioned, we may expect that it is also inverted without too much effort. In fact, if one takes the deflation subspace to be spanned by a few exact eigenmodes of Q with small in magnitude eigenvalues, which is called low-mode averaging [29–31], then the inversion of the little operator is indeed trivial. In order to suppress the variance in the costly remainder term, then, a small number of low modes needs to provide a good approximation to the quark propagator at long distances. Unfortunately, the required number of low-modes for a constant suppression of the variance grows with the physical volume, which is essentially a consequence of the increasing density described by the Banks–Casher relation [32]. In the following, we circumvent the issue by exploiting the property of the local coherence of the low modes, as suggested already in Ref. [33].

2.1 Local coherence and deflation

The low quark modes exhibit a property known as local coherence (or weak-approximation) which is that the support of such fields on a small domain, of linear size $b = 0.25 - 0.5$ fm, span a small subspace compared with their total number. This suggests using as a basis a few low modes projected to small domains to create a subspace whose size grows with the volume without any extra

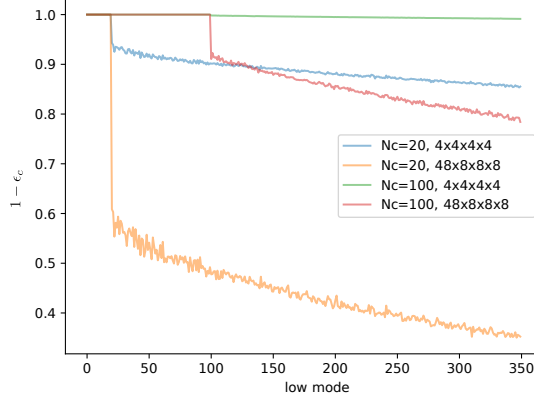


Figure 3: One minus the deficit of the lowest modes in subspaces built from just a few $N_c = 20$ (blue or yellow) or 100 (green or red) low modes. Decreasing the block sizes (indicated in lattice units) increases the subspace size and reduces the deficits further.

cost. The use of such blocked low modes has been instrumental in the definition of efficient deflated or multigrid preconditioned solvers [33–35], and is crucial aspect of forming a local averaging (or aggregation) in gauge theories, and has also been applied to compression of fields [36]. Once the lowest level averaging is defined, the extension to a hierarchical scheme is fairly trivial by simply combining the blocks, which gives a flexible approach to tuning the effective cost [37]. A similar scheme to what follows has been proposed in Ref. [38].

Following Ref. [37], we imagine the $N_v = 2N_c N_b$ block low modes are constructed from N_c exact low modes of the Dirac operator, $\phi_c^{B,\pm}(x) = \theta_B(x)(1 \pm \gamma_5)\phi_c(x)$, where $\theta_B(x)$ is the characteristic function of the block labelled by $B = 1 \dots N_b$. After orthonormalizing, the resulting restricted fields, now called coarse fields because they carry geometric information, have components

$$(R\psi)_c^\pm(x_B) = (\phi_c^{B,\pm}, \psi) \quad (19)$$

where x_B are the coordinates of the block B , and have a structure reminiscent of the original quark field. The alignment of a field ψ with the subspace is large if its deficit

$$\epsilon_\psi = \|\psi - R^\dagger R\psi\| / \|\psi\| \quad (20)$$

is small. Fig. 3 shows that the deficits of first few hundred exact low modes are small for subspaces built with just a small number $N_c = 20, 100$ on an ensemble with $m_\pi L \approx 4$ as described in the next subsection. This indicates that expanding the dimension of the subspace by the blocking procedure should improve the representation of the quark propagator without needing $O(V)$ low modes to begin with. In the following, we investigate empirically both conditions for an efficient deflation variance reduction scheme, namely that the inversion of the little operator is cheap and that the variance on the expensive correction term is suppressed.

2.2 Numerical investigations

In this section, we report the variance of the stochastic estimator for the deflated terms (the first line of eq. (18)) and the approximant (the second line) using $N_f = 2$ $O(a)$ -improved Wilson

$m_\pi L$	estimator	N_s				$\frac{\text{measured cost}}{\text{Hutchinson } N_s = 1}$
		L0	L1	L2	L3	
2.9	Hutchinson	1024	-	-	-	1024
	LMA	16	exact	-	-	16
	MG LMA	1	16	exact	-	1.4
4.3	Hutchinson	2048	-	-	-	2048
	LMA	1024	1024	exact	-	1024
	MG LMA	16	1024	exact	-	65.8
5.8	Hutchinson	4096	-	-	-	4096
	LMA	2048	exact	-	-	2048
	MG LMA	1	16	1024	exact	117

Table 1: Indicative costs to achieve a fixed variance for three different lattice volumes for the undeflated estimator (Hutchinson), low-mode averaging (LMA) and multigrid low-mode averaging (MG LMA). All deflated schemes use $N_c = 50$ exact low modes.

fermions [39] for various lattice sizes $m_\pi L = 2.9, 4.3, 5.8$ with $m_\pi = 270$ MeV and $a = 0.066$ fm. The intermediate volume corresponds to the F7 ensemble generated by the CLS consortium [40] while the others were generated with the openQCD code [41]. The eigensolver from the PRIMME library [42] was used to determine the low modes. This allows us to investigate the hypothesis that a small number of low modes N_c fixed for all lattice sizes is sufficient to suppress the variance as long as the blocked low modes are used in the definition of the deflation subspaces.

For the hierarchical scheme, the propagator is decomposed telescopically into N_ℓ levels where the propagators on levels 0 and $N_\ell - 1$ have a similar form as the two-level case and on the intermediate levels $0 < l < N_\ell - 1$ are given by

$$S_l = \mathcal{R}_l^\dagger Q_l^{-1} \mathcal{R}_l \gamma_5 - \mathcal{R}_{l+1}^\dagger Q_{l+1}^{-1} \mathcal{R}_{l+1} \gamma_5. \quad (21)$$

where the little operator on level l is defined by as $Q_l = \mathcal{R}_l Q \mathcal{R}_l^\dagger$. The restrictors \mathcal{R}_l on coarser levels are just defined by decompositions into larger blocks, which, if they nest, even allow a truly recursive procedure to be constructed.

For the connected correlator, the N_ℓ^2 terms are then estimated in N_ℓ levels

$$G(x_0, y_0) \approx \sum_{l=0}^{N_\ell-1} G_{Ll}(x_0, y_0) \quad (22)$$

where we define each level as the sum of terms

$$G_{Ll}(x_0, y_0) = \frac{1}{N_s} \sum_{i=1}^{N_s} a^3 \sum_{\vec{y}} \sum_{\substack{m,n=0 \\ \min(m,n)=l}}^{N_\ell-1} (\eta_i^\dagger \gamma_k \mathcal{S}_m)(y) \gamma_k (\mathcal{S}_n \eta_i)(y) \quad (23)$$

which collects all the terms which involve at least one inversion of the Dirac operator on level l and coarser levels. Each level requires an inversion of the corresponding coarse or coarser operators

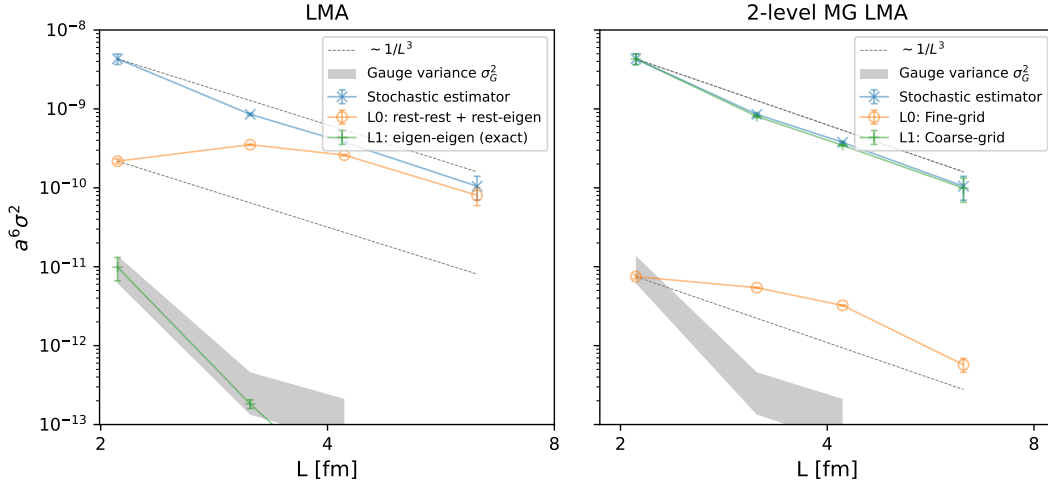


Figure 4: Volume dependence of the variance of the levels of the improved estimators without using block-projection of the low modes (LMA, left) and with block projection (MG LMA, right) for a separation $x_0 - y_0 \approx 1.3$ fm. The variance of the undeflated estimator is also shown with blue points, while the grey band indicates an estimate of the gauge variance. All variances for the stochastic estimators are computed with $N_s = 1$ auxiliary field.

only. Such a decomposition, assuming different sources on each correlator level, has a variance

$$\sigma^2 = \sigma_g^2 + \frac{1}{N_{L0}} \sigma_{L0}^2 + \frac{1}{N_{L1}} \sigma_{L1}^2 + \dots, \quad (24)$$

where the variances due to the auxiliary fields can be written in terms of the deflated propagators on that level. All the parameters which define the estimators which include N_{Lj} , the block sizes and the number of exact low modes N_c can be adjusted to optimize the effective cost.

In Tab. 1, I report the cost of the Hutchinson estimator and deflated schemes without blocking (LMA) and with blocking (MG LMA), either three-levels with block sizes 8^4 (in lattice units) and a four-level scheme with an additional block size of 4^4 . All of the deflated schemes use $N_c = 50$ exact low modes. The cost to reach a fixed precision is in units of a cycle of the plain Hutchinson estimator, which was implemented with spin-diagonal sources [43]. The coarsest level has no blocking and thus can be computed exactly. In all cases we observe that the little operators are at least as well conditioned as the Dirac operator, and could be inverted using a standard GCR solver [44]. The results show that just a few sources are required for the expensive deflated operator which translates into an large reduction in the effective cost, remaining a factor 30 cheaper on the largest volume than without deflation, without requiring a large number of low modes. As expected, the advantage of LMA over the Hutchinson estimator quickly evaporates as the physical volume is increased with a fixed number of low modes, in contrast to MG LMA.

In Fig. 4, the volume dependence of the variance of each level is shown for $N_s = 1$, using again $N_c = 50$ fixed for the deflated estimators, for LMA (left) and MG LMA (right). The grey band indicates the gauge variance. The variance of the single (undeflated) Hutchinson estimator is shown in both cases with the blue points and decreases with L , as expected, roughly like $1/L^3$ (dashed lines). While the variance of the deflated term (orange circles) is suppressed on the smallest volume

in LMA, it quickly saturates the variance of the undeflated estimator, illustrating that the deflation loses efficiency as the volume is increased but the number of modes is kept constant. For MG LMA, when the low modes are blocked, the variance of the deflated term does not saturate the variance of the Hutchinson estimator and decreases with the volume. This illustrates that the increase of the subspace size with the volume using a fixed physical block size keeps this contribution small.

Evidently, such a scheme could be optimized further. The implementation which was used to produce the cost estimates in Tab. 1 relied on a fairly primitive iterative solver. Using a true multi-grid preconditioned solver could improve the coarse-grid solves and accelerate further the computation of the coarse-grid propagators. The advantage of MG LMA is that just a few low modes are sufficient to deflate the variance if they are subsequently blocked. Whether using approximate low modes like those fields used in the set-up stage of the deflated or multigrid solver would provide as good a basis remains to be seen. One interesting observation of this study is that, in order to fulfil the criteria of a good approximation and a well-conditioned little system, both chiralities must be kept in the subspace [44]. Finally, the application to other quark-line connected diagrams with varying number of light-quark propagators, such as static-light meson propagators with one propagator, or baryon propagators with three, would be very interesting, as would the application to fields which have been smoothed with a procedure like distillation [45].

3. Frequency-splitting

As motivated earlier, implementing translation averaging for quark-line disconnected diagrams requires a good and cheap representation of the propagator at short distances. This can be achieved by noting that the quark propagator at short distances is quite independent of the quark mass [17]. Therefore, we can add and subtract the quark propagator with a larger mass

$$S_f(x, x) = \underbrace{\{S_f(x, x) - S_g(x, x)\}}_{S_0(x, x)} + \underbrace{S_g(x, x)}_{S_1(x, x)}, \quad m_f < m_g, \quad (25)$$

where we restrict immediately to the case where $x = y$. When inserted into the single-propagator trace, this leads to a decomposition of the trace into two levels, each of which can be evaluated independently. Exactly like the deflation, then, the variance decomposes into a sum like in eq. (24), and also like that case, it is simple to generalize to a hierarchical scheme by adding and subtracting larger masses. As an identity, the extra propagators are an algorithmic crutch and need not be related to the physical flavour content of the theory. This achieves an additive version quite like a Hasenbusch splitting of the fermion determinant. It has already been noted since a long time, that the variance on the difference is indeed suppressed [46], in fact by $a^2(m_f - m_g)^2$ [17], but some additional improvements can reduce even further the variance of both the levels.

An improved estimator for the trace of S_0 can be written with auxiliary fields with support on the whole lattice volume, using the cyclicity of the trace and the property $D_f - D_g = m_f - m_g$

$$\text{tr}\{\gamma_\mu S_0(x, x)\} \approx (m_g - m_f) \frac{1}{N_s} \sum_i (\eta_i^\dagger S_f)(x) \gamma_\mu (S_g \eta_i)(x) \quad (26)$$

which is very similar to the one-end trick used for twisted-mass fermions [43, 47], where such differences appear naturally. Interestingly, this so-called split-even estimator furnishes us with an

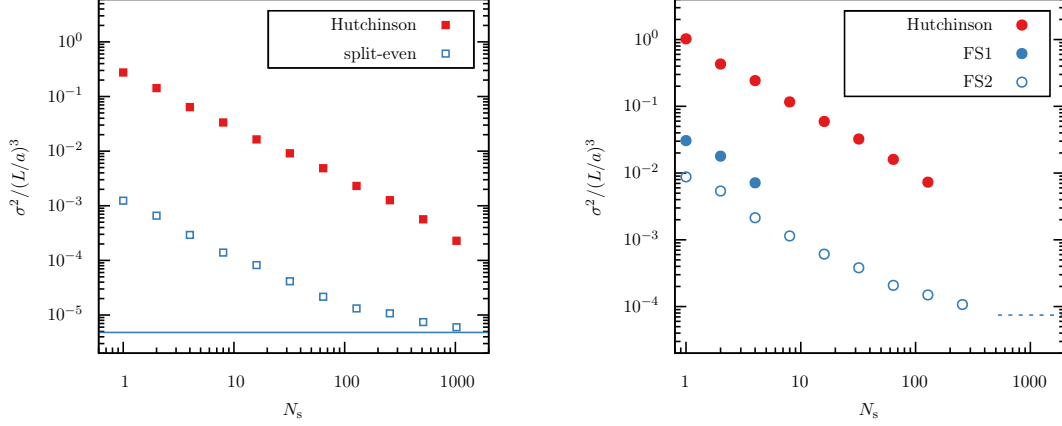


Figure 5: The variance of the estimator for the trace of the difference of propagators \mathcal{S}_0 (left) or the single-propagator trace $T_{k,f}^{(1)}$ (right). In both cases the plain Hutchinson estimator (red points) is compared with the improved estimators, the split-even estimator for the difference (open squares, left) and two variants of the frequency-splitting estimator including the hopping expansion (blue circles, right).

estimator for every x_0 , which can be used to implement translation averaging in time. In addition, the variance is no longer independent of the gamma-matrix, compared with the Hutchinson estimator. We note, due to the charge factors of the electromagnetic current, in a theory with $N_f = 2 + 1$ flavours, the disconnected Wick contractions result exactly in differences of quark propagators of the light and strange quarks, and the split-even estimator can be used by itself, and used for a multitude of related observables [48–51]. Such a representation also works out for domain-wall fermions [52] and has been applied to weak transition amplitudes [53].

For a general flavour content, however, the single-propagator trace is required, and an improved estimator for large quark masses can be constructed by computing the short-distance contribution exactly via the hopping expansion to the k th order [54–56]

$$D_f^{-1} = M_k + D_f^{-1} H^k \quad (27)$$

where $H = -(D_{e0}D_{00}^{-1} + D_{0e}D_{ee}^{-1})$ is the hopping matrix and the sum of the first k terms is given by

$$M_k = (D_{ee} + D_{00})^{-1} \sum_{n=0}^{k-1} H^n. \quad (28)$$

Noting that M_k is a sparse matrix, the trace of those terms can be evaluated exactly by quadratures with so-called probing vectors [57] v^0, \dots, v^{K-1} which satisfy

$$\sum_{n=0}^{K-1} v_{\alpha a}^n(x) v_{\beta b}^n(y) = \delta_{\alpha\beta} \delta_{ab} \delta(x, y) \quad \text{for all } \alpha, \beta, a, b, x, y \text{ where } M_{k, \alpha a \beta b}(x, y) \neq 0. \quad (29)$$

It turns out some fairly efficient schemes can be easily produced for small k , requiring $K = 24(k/2)^4$ vectors, which may be improved upon with schemes such as those put forward in Ref. [58].

$m_\pi L$	estimator	N_s					$\frac{\text{measured cost}}{\text{Hutchinson } N_s = 1}$
		L0	L1	L2	L3	L4	
4	Hutchinson	100	-	-	-	-	100
	FS1	4	16	-	-	-	9.2
	FS2	1	1	2	3	10	5.9

Table 2: Cost to achieve a fixed variance for the single-propagator trace $T_{k,f}^{(1)}$ for the plain Hutchinson estimator and two variants of the frequency-splitting estimator.

3.1 Numerical investigations

In this section we present results for two variants of the frequency-splitting estimator outlined in the previous subsection on the same ensemble with $m_\pi \approx 270$ MeV presented earlier. For a fair comparison, we extend the Hutchinson estimator to one which where the sources have support on all time-slices, so that both estimators can be used to implement translation averaging in time. Firstly, in Fig. 5 (left) we see the variance of the first level \mathcal{S}_0 with the second mass corresponding roughly to the physical strange-quark mass, as a function of the number of auxiliary fields per gauge field. The horizontal line is an estimate of the gauge variance. The Hutchinson estimator of the difference is shown in red solid points while the improved split-even estimator is shown with open blue points. The variance of the split-even estimator is $\mathcal{O}(100)$ suppressed with respect to the Hutchinson estimator for the difference, which itself is only slightly suppressed compared with the single-propagator trace shown in the right-hand panel (solid red circles).

The right-hand panel also shows two variants of the frequency-splitting estimator with either two (FS1) or five (FS2) levels. The FS2 estimator uses a few levels to reach approximately a quark mass corresponding to the physical charm-quark mass, where the hopping expansion with $k = 4$ is already efficient. As well as the FS2 estimator having the minimum variance per cycle, we see that it also has the minimum effective cost when looking at Tab. 2. Optimization schemes like those of Ref. [59], can be useful to further reduce the effective cost, but an order of magnitude speed-up is fairly easy to accomplish, and appears even better at small quark masses [17].

4. Multi-level integration

The techniques discussed so far have concerned reducing the computational cost of implementing translation averaging, but do not provide a solution that deals with the vacuum contributions to the gauge variance, where they arise. One known way to suppress those types of fluctuations is by a local update and averaging procedure, called multi-level integration [60, 61], essentially a generalization of the multi-hit algorithm [62]. If both the distribution and the observable factorize, an improved estimator can be written in which the fields in each factor in the correlation function are averaged independently of one and another

$$\bar{\phi}^i(x) = \frac{1}{n_1} \sum_{j=1}^{n_1} \phi_j^i(x), \quad (30)$$

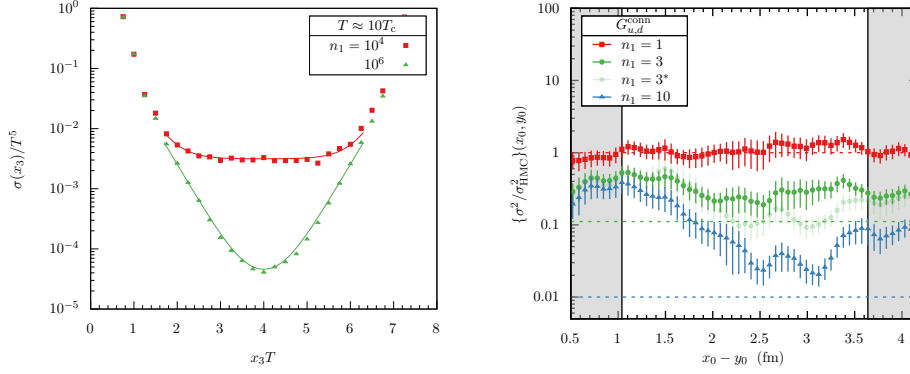


Figure 6: Variance of two-point functions with two-level integration schemes. The left-hand panel shows the (square root of the) variance of the spatial correlator of the action density for a simulation of SU(3) pure gauge theory at high temperature. The right-hand panel shows the variance of the isovector vector correlator using the multi-boson domain decomposed HMC [64], compared with a standard HMC algorithm computed on the F7 ensemble described in the text.

where now the index j labels the samples in a sub-domain keeping the boundary field fixed, which is labelled by i , harking back to the notation of the introduction.

This results in a variance which depends on a non-trivial way on the separation to the boundary d , which, if in the regime where both fields are maximally distant to the boundary $x_0 - y_0 \approx 2d$, results in [63]

$$\sigma^2 = \frac{1}{N} \left[c_2 e^{-2m(x_0 - y_0)} + \frac{c_1}{n_1} e^{-m(x_0 - y_0)} + \frac{c_0}{n_1^2} \right], \quad (31)$$

where we note that the problematic vacuum contributions are now suppressed by n_1^2 , even though the cost of the simulation grows only linearly in n_1 . Choosing $n_1 \sim e^{m(x_0 - y_0)}$ results in a constant signal-to-noise ratio up to $x_0 - y_0$. Such an algorithm costs exponentially less than the standard update algorithm to reach the same signal-to-noise ratio at an equivalent distance. Fig. 6 (left) shows that the formula describes very well the dependence of the measured variance on $x_0 - y_0$ in pure gauge simulations, where it can be applied straightforwardly.

The lack of manifest locality of the usual representation of the fermionic path integral seems to prevent such a scheme working out in QCD. Nevertheless, a proposal has been shown to work out in practice which relies on the observation that an exact factorization is not required to be profitable [65, 66]. If a decomposition exists such that one term (the approximant) is factorizable, while the remainder has a small variance (similar to the decompositions presented for the quark propagator in the previous sections) then a large acceleration can be achieved. The scheme put forward by Cè, Giusti and Schaefer is based on an overlapping parallel Schwarz method, related to the method for solving classical partial differential equations, and the multi-boson representation [67]. A complete exposition of the factorization scheme is beyond the scope of these proceedings, but recent results for such update schemes for the fermion determinant show that indeed the fluctuations in the non-factorized remainder are under control, and that the variance has the expected suppression [64], see Fig. 6 (right). Recent progress was also shown at this conference by Barca et al. [68]. These methods are a promising way to ameliorate the signal-to-noise ratio problems discussed in the

opening section and in particular may unlock ground-state matrix elements needed for precision studies of the Standard Model.

5. Conclusions

In these proceedings, I have reviewed the origins of signal-to-noise ratio problems in lattice QCD, attempting to paint a picture of some obstacles to precision computations of observables. Translation averaging is one of the simplest ways to reduce the variance but it does not come with an acceptable cost for quark propagators if implemented in a naive fashion. I have presented a selection of methods to accelerate the computation of quark propagators for large and small separations, based on decompositions of the propagator and the observable of interest.

For large separations, the quark propagator is well represented in a subspace created from blocked low modes of the Dirac operator, leading to a multigrid low-mode averaging scheme. Furthermore, the little operator is well conditioned if the subspace spans both chiralities of the fields, leading to an efficient deflation scheme which has a favourable scaling with the physical volume. For the quark propagator at short distances, instead, a good approximation is constructed by shifting the quark mass and using the hopping representation for the largest masses. Noting that the sum of the first few hopping terms is a true sparse matrix, at least in the Wilson formulation, enables its trace to be computed exactly with a few probing vectors.

Nevertheless, the vacuum contributions to the variance cause severe signal-to-noise ratio problems that are not solved by translation averaging. A local update procedure such as multi-hit or multi-level integration provides a clear route to improve the computation of n -point functions, and recent innovations have shown that an exact factorization is not necessary for this idea to be useful. Empirical results have shown that the multi-boson domain-decomposed HMC allows one to compute a local average for the contribution which suffers most of the fluctuations, and therefore can speed up the computation exponentially for large separations. To attain the next level of precision in hadronic matrix elements, some combination of the methods presented here should be beneficial, and progress on multiple fronts suggests that significant improvements may be within reach.

Acknowledgments

I extend my sincere thanks to the organizers of Lattice 2025 for a scientifically illuminating program and a wonderful conference. I appreciate all of the valuable discussions with members of the lattice community I have had on topics related to these proceedings, and am especially grateful to Mattia Dalla Brida, Leonardo Giusti, Maxwell T. Hansen, Harvey B. Meyer, Agostino Patella, Mike J. Peardon, my colleagues in the RC* Collaboration and Marina K. Marinković for encouraging work in this direction. I am indebted to Roman Gruber for his persistence in the investigations of multigrid low-mode averaging schemes.

References

- [1] K. G. Wilson. In: *Phys. Rev. D* 10 (1974). Ed. by J. C. Taylor, pp. 2445–2459.
- [2] G. Boyd et al. In: *Nucl. Phys. B* 469 (1996), pp. 419–444. arXiv: [hep-lat/9602007](https://arxiv.org/abs/hep-lat/9602007).

- [3] L. Altenkort et al. In: *Phys. Rev. D* 105.9 (2022), p. 094505. arXiv: [2112.02282 \[hep-lat\]](#).
- [4] G. Parisi. In: *Phys. Rept.* 103 (1984), pp. 203–211.
- [5] G. P. Lepage. In: *Boulder ASI 1989:97-120*. 1989, pp. 97–120.
- [6] M. Lüscher. In: *Les Houches Summer School: Session 93: Modern perspectives in lattice QCD: Quantum field theory and high performance computing*. Feb. 2010, pp. 331–399. arXiv: [1002.4232 \[hep-lat\]](#).
- [7] G. Bali et al. In: *PoS LAT2009* (2009). Ed. by C. Liu and Y. Zhu, p. 149. arXiv: [0911.2407 \[hep-lat\]](#).
- [8] T. Blum et al. In: *Phys. Rev. D* 93.1 (2016), p. 014503. arXiv: [1510.07100 \[hep-lat\]](#).
- [9] K.-F. Liu et al. In: *Phys. Rev. D* 97.3 (2018), p. 034507. arXiv: [1705.06358 \[hep-lat\]](#).
- [10] H. B. Meyer. In: *Eur. Phys. J. C* 77.9 (2017), p. 616. arXiv: [1706.01139 \[hep-lat\]](#).
- [11] M. Lüscher. In: *EPJ Web Conf.* 175 (2018). Ed. by M. Della Morte et al., p. 01002. arXiv: [1707.09758 \[hep-lat\]](#).
- [12] L. Giusti and M. Lüscher. In: *Eur. Phys. J. C* 79.3 (2019), p. 207. arXiv: [1812.02062 \[hep-lat\]](#).
- [13] M. Bruno et al. In: *JHEP* 11 (2023), p. 167. arXiv: [2307.15674 \[hep-lat\]](#).
- [14] D. Bernecker and H. B. Meyer. In: *Eur. Phys. J. A* 47 (2011), p. 148. arXiv: [1107.4388 \[hep-lat\]](#).
- [15] M. Della Morte et al. In: *JHEP* 08 (2005), p. 051. arXiv: [hep-lat/0506008](#).
- [16] M. Della Morte and A. Juttner. In: *JHEP* 11 (2010), p. 154. arXiv: [1009.3783 \[hep-lat\]](#).
- [17] L. Giusti et al. In: *Eur. Phys. J. C* 79.7 (2019), p. 586. arXiv: [1903.10447 \[hep-lat\]](#).
- [18] T. Harris. In: *PoS EuroPLEx2023* (2024), p. 011.
- [19] A. Altherr et al. In: *PoS LATTICE2024* (2025), p. 116. arXiv: [2502.03145 \[hep-lat\]](#).
- [20] A. Altherr et al. In: *JHEP* 10 (2025), p. 158. arXiv: [2506.19770 \[hep-lat\]](#).
- [21] G. M. de Divitiis et al. In: *Phys. Rev. D* 87.11 (2013), p. 114505. arXiv: [1303.4896 \[hep-lat\]](#).
- [22] K. Bitar et al. In: *Nucl. Phys. B* 313 (1989), pp. 348–376.
- [23] M. Hutchinson. In: *Communications in Statistics - Simulation and Computation* 19.2 (1990), pp. 433–450.
- [24] S.-J. Dong and K.-F. Liu. In: *Phys. Lett. B* 328 (1994), pp. 130–136. arXiv: [hep-lat/9308015](#).
- [25] G. M. de Divitiis et al. In: *Phys. Lett. B* 382 (1996), pp. 393–397. arXiv: [hep-lat/9603020](#).
- [26] C. Michael and J. Peisa. In: *Phys. Rev. D* 58 (1998), p. 034506. arXiv: [hep-lat/9802015](#).
- [27] S. Bernardson et al. In: *Computer Physics Communications* 78.3 (1994), pp. 256–264. issn: 0010-4655.

- [28] R. Gruber et al. In: *PoS LATTICE2023* (2024), p. 153. arXiv: [2401.14724 \[hep-lat\]](#).
- [29] L. Giusti et al. In: *JHEP* 04 (2004), p. 013. arXiv: [hep-lat/0402002](#).
- [30] T. A. DeGrand and S. Schaefer. In: *Comput. Phys. Commun.* 159 (2004), pp. 185–191. arXiv: [hep-lat/0401011](#).
- [31] J. Foley et al. In: *Comput. Phys. Commun.* 172 (2005), pp. 145–162. arXiv: [hep-lat/0505023](#).
- [32] T. Banks and A. Casher. In: *Nuclear Physics B* 169.1 (1980), pp. 103–125. issn: 0550-3213.
- [33] M. Lüscher. In: *JHEP* 07 (2007), p. 081. arXiv: [0706.2298 \[hep-lat\]](#).
- [34] R. Babich et al. In: *Phys. Rev. Lett.* 105 (2010), p. 201602. arXiv: [1005.3043 \[hep-lat\]](#).
- [35] A. Frommer et al. In: *SIAM J. Sci. Comput.* 36.4 (2014), A1581–A1608. arXiv: [1303.1377 \[hep-lat\]](#).
- [36] M. A. Clark et al. In: *EPJ Web Conf.* 175 (2018). Ed. by M. Della Morte et al., p. 14023. arXiv: [1710.06884 \[hep-lat\]](#).
- [37] R. Gruber et al. In: *Phys. Rev. D* 111.7 (2025), p. 074508. arXiv: [2412.06347 \[hep-lat\]](#).
- [38] A. Frommer et al. In: (Sept. 2025). arXiv: [2509.11424 \[hep-lat\]](#).
- [39] K. Jansen and R. Sommer. In: *Nucl. Phys. B* 530 (1998). [Erratum: *Nucl.Phys.B* 643, 517–518 (2002)], pp. 185–203. arXiv: [hep-lat/9803017](#).
- [40] P. Fritzsche et al. In: *Nucl. Phys. B* 865 (2012), pp. 397–429. arXiv: [1205.5380 \[hep-lat\]](#).
- [41] M. Lüscher et al. <https://luscher.web.cern.ch/luscher/openQCD>. Accessed: 2024.
- [42] A. Stathopoulos and J. R. McCombs. In: *ACM Transactions on Mathematical Software* 37.2 (2010), 21:1–21:30.
- [43] P. Boucaud et al. In: *Comput. Phys. Commun.* 179 (2008), pp. 695–715. arXiv: [0803.0224 \[hep-lat\]](#).
- [44] R. Gruber. PhD thesis. Zurich, ETH, 2025.
- [45] M. Peardon et al. In: *Phys. Rev. D* 80 (2009), p. 054506. arXiv: [0905.2160 \[hep-lat\]](#).
- [46] V. Gülpers et al. In: *PoS LATTICE2014* (2014), p. 128. arXiv: [1411.7592 \[hep-lat\]](#).
- [47] S. Dinter et al. In: *JHEP* 08 (2012), p. 037. arXiv: [1202.1480 \[hep-lat\]](#).
- [48] M. Cè et al. In: *JHEP* 08 (2022), p. 220. arXiv: [2203.08676 \[hep-lat\]](#).
- [49] A. Boccaletti et al. In: (July 2024). arXiv: [2407.10913 \[hep-lat\]](#).
- [50] D. Djukanovic et al. In: *Phys. Rev. D* 109.9 (2024), p. 094510. arXiv: [2309.06590 \[hep-lat\]](#).
- [51] D. Djukanovic et al. In: *JHEP* 04 (2025), p. 098. arXiv: [2411.07969 \[hep-lat\]](#).
- [52] T. Harris et al. In: *PoS LATTICE2022* (2023), p. 013. arXiv: [2301.03995 \[hep-lat\]](#).
- [53] R. Hodgson et al. In: *PoS LATTICE2024* (2025), p. 258. arXiv: [2501.18358 \[hep-lat\]](#).
- [54] C. Thron et al. In: *Phys. Rev. D* 57 (1998), pp. 1642–1653. arXiv: [hep-lat/9707001](#).

- [55] G. S. Bali et al. In: *Comput. Phys. Commun.* 181 (2010), pp. 1570–1583. arXiv: [0910.3970](#) [[hep-lat](#)].
- [56] V. Gülpers et al. In: *Phys. Rev. D* 89.9 (2014), p. 094503. arXiv: [1309.2104](#) [[hep-lat](#)].
- [57] J. M. Tang and Y. Saad. In: *Numerical Linear Algebra with Applications* 19.3 (2012), pp. 485–501.
- [58] A. Stathopoulos et al. In: *SIAM J. Sci. Comput.* 35.5 (2013), S299–S322. arXiv: [1302.4018](#) [[hep-lat](#)].
- [59] T. Whyte et al. In: *Comput. Phys. Commun.* 294 (2024), p. 108928. arXiv: [2212.04430](#) [[hep-lat](#)].
- [60] M. Lüscher and P. Weisz. In: *JHEP* 09 (2001), p. 010. arXiv: [hep-lat/0108014](#) [[hep-lat](#)].
- [61] H. B. Meyer. In: *JHEP* 01 (2003), p. 048. arXiv: [hep-lat/0209145](#).
- [62] G. Parisi et al. In: *Phys. Lett. B* 128 (1983), pp. 418–420.
- [63] M. García Vera and S. Schaefer. In: *Phys. Rev. D* 93 (2016), p. 074502. arXiv: [1601.07155](#) [[hep-lat](#)].
- [64] M. Dalla Brida et al. In: *Phys. Lett. B* 816 (2021), p. 136191. arXiv: [2007.02973](#) [[hep-lat](#)].
- [65] M. Cè et al. In: *Phys. Rev. D* 93.9 (2016), p. 094507. arXiv: [1601.04587](#) [[hep-lat](#)].
- [66] M. Cè et al. In: *Phys. Rev. D* 95.3 (2017), p. 034503. arXiv: [1609.02419](#) [[hep-lat](#)].
- [67] Lüscher, Martin. In: *Nucl. Phys. B* 418 (1994), pp. 637–648. arXiv: [hep-lat/9311007](#).
- [68] L. Barca et al. In: *Phys. Rev. D* 113.3 (2026), p. 034505. arXiv: [2512.10644](#) [[hep-lat](#)].