EPJ manuscript No.
(will be inserted by the editor)

Locality and multi-level sampling with fermions

Marco Cè

Helmholtz-Institut Mainz, Johannes Gutenberg-Universität Mainz, Germany

Received: date / Revised version: date

Abstract Multi-level Monte Carlo sampling techniques exploit the locality of quantum field theory to provide a solution in purely-bosonic quantum field theories to the signal-to-noise ratio problem that affects the lattice determination of a large class of quantities. However, it is not straightforward to generalize multi-level sampling to lattice theories with fermionic content, such as QCD, due to the loss of manifest locality after the fermion path integral is performed. We discuss how the decrease of the fermion propagator with Euclidean distance induces a systematic approximation of the fermion propagator and the fermion determinant, in which the gauge field dependence of distant spacetime regions is completely factorized. This allows us to apply multi-level sampling to the lattice QCD computation of hadronic observables, such as mesonic and baryonic correlators. In particular, we show an application of this strategy to the disconnected contribution to the correlator of two flavour-singlet pseudoscalar densities, which results in a significant increase of the signal-to-noise ratio when a two-level sampling scheme is used.

PACS. XX.XX.XX No PACS code given

1 Introduction

Numerical Monte Carlo (MC) simulations of Quantum Chromodynamics (QCD) regularized on the lattice have been extremely successful in computing many high-energy physics (HEP) quantities in which the strong interaction plays a rôle. One of the main achievement is the the *ab-initio* determination of the light spectrum of hadrons [1, 2], that can be extracted from the exponential decay at long Euclidean distances of the appropriate correlation functions. This and other achievements have been made possible by the combination of an exponential increase in the available computer power and of significant algorithmic advances in the last couple of decades. This enables state-of-the-art simulations of ensembles with more than 10^8 lattice points and volumes in excess of $(5\,\mathrm{fm})^3$ at physical quark masses [3], that allows full control of systematics effect from the finite lattice spacing, finite volume and non-physical masses. However, the numerical cost to compute hadronic correlation functions to a given target statistical precision is driven by the exponential loss of the signal compared to the sampling noise with increasing Euclidean time separation, known as the signal-to-noise ratio (S/N) problem [4, 5].[1] While the S/N problem has been known for more than three decades, its relevance has been growing in recent years thanks to availability of large lattices, that allows long time separations, and the use of lighter pion masses, that exacerbate the problem. For instance, since the S/N of the nucleon-correlator scales at large separations τ as $\mathrm{e}^{-\mu\tau}$, with $\mu = M_N - 3/2m_\pi \approx 3.7\,\mathrm{fm}^{-1}$ [5], to gain as little as 0.5 fm of additional distance requires a 40-fold increase of the number of samples, and thus an increase by this same factor in the computational cost.

Analogous severe problems afflict the computation of correlators in a large variety of quantum systems, from the harmonic oscillator to Yang-Mills (YM) theory. In some cases, multi-level MC sampling algorithms have been proposed, which lead to an impressive acceleration of the simulations [8–14]. They take advantage of the fact that, when the action and the observables depend locally on the integration variables, the S/N problem can be solved by independent sampling of the local building blocks of the observable.

In the case of theories with fermions, such as QCD, the standard approach to the MC simulation of lattice theory requires to integrate out analytically the fermionic fields. Consider the partition function of QCD on a Euclidean

[1] In spectroscopy studies the S/N is mitigated using complex hadron interpolators, which suppress the excited-state contamination at short distance. However, these methods are not suited for computations that require correlators of bare fields, such as the HVP contribution to $(g-2)_\mu$ [6], or at very long distances, such as nucleon structure studies [7].

lattice with a doublet of mass-degenerate quarks

$$Z = \int \mathcal{D}\big[U, \bar{\psi}, \psi\big]\, \mathrm{e}^{-S_{\mathrm{G}}[U] - \bar{u}Du - \bar{d}Dd} = \int \mathcal{D}[U]\, \det\big\{D^\dagger D\big\} \mathrm{e}^{-S_{\mathrm{G}}[U]}. \tag{1}$$

Performing the Gaussian path integral over ψ and $\bar{\psi}$ results in positive-definite quark determinant factor $\det\{D^\dagger D\}$, that is typically simulated with pseudofermions [15]

$$\det\big\{D^\dagger D\big\} \sim \int \mathcal{D}\big[\phi^\dagger, \phi\big]\, \mathrm{e}^{-\left|D^{-1}\phi\right|^2}. \tag{2}$$

Moreover, using Wick's theorem fermionic observable are expressed in terms of quark propagators.

The quark determinant and the quark propagator are non-local functions of the gauge field, and the manifest locality of the action is lost. This has consequences: since changing a single link requires recomputing the action globally, *local* MC algorithms are not competitive any more.[2] Because of these, state-of-the-art algorithm are variants of the *global* hybrid Monte Carlo (HMC) algorithm [20]. For the same reason, the formulation of multi-level algorithms, based on the locality of the theory, is not straightforward in the presence of fermions.

In addition, the exact fermion path integral results in a partial mitigation of the S/N problem for hadronic correlators that have only connected Wick's contractions. Indeed, numerical evidence shows that the quark propagator on a typical gauge field configuration is still suppressed with Euclidean separation of source and sink when inserted in hadronic quantities. At large distance, it holds

$$\big\|D^{-1}(x,y)\big\| = \mathrm{tr}\big\{D^{-1}(x,y)D^{-1}(x,y)^\dagger\big\}^{1/2} \sim \mathrm{e}^{-\frac{1}{2}M_\pi|x-y|}, \tag{3}$$

where M_π is the mass of the lightest pseudoscalar meson, that is associated to the longest fermionic correlation length in the theory. Therefore, even if the quark propagator and quark determinant are non-local, the dependence of hadronic quantities on gauge links at a physical distance, *e.g.* 0.5 fm, is exponentially suppressed. Building on this observation, in refs [21, 22] we propose a multi-level integration algorithm for lattice QCD that addresses the S/N problem of a large set of hadronic observables.

In our thesis [23], we study a specific hadronic quantity: the mass of the lightest flavour-singlet pseudoscalar meson, *e.g.* the η' meson mass in the three-flavour theory. This mass has a quantum anomaly contribution, and it is connected to the topological susceptibility of the YM theory, that we also computed [24, 25], by the Witten-Veneziano formula [26, 27], an intrinsically non-perturbative relation.[3] However, this mass is extracted from the long distance behaviour of the the flavour-singlet pseudoscalar two-point function, which is affected by a severe S/N problem that significantly limits the statistical precision of standard MC techniques. Therefore, in ref. [23] we propose to apply the multi-level algorithm that we introduced to the flavour-singlet pseudoscalar correlator, focussing in particular on the disconnected Wick's contraction contribution.

In this article we present a self-contained introduction to the method proposed in the thesis, including material from refs [21, 22, 30, 31] and addressing the factorization and multi-level sampling of both the quark propagator and the quark determinant. The article is structured as following: In sect. 2 we introduce the S/N problem in general and in the specific case of fermionic correlators. In sect. 3 we introduce the original multi-level algorithm, explaining the link with locality and the issues introduced by fermionic theories. Sect. 4 is dedicated to the first aspect of the development of multi-level sampling for fermions: the factorization of the quark propagator. Building on it, in sect. 5 we report a numerical study of the multi-level sampling of the disconnected contribution to the pseudoscalar meson propagator, while in sect. 6 the connected contribution case is discussed briefly. The factorization of the quark determinant, as the second aspect of a multi-level algorithm for fermions, is the topic of sect. 7. Finally, in sect. 8 we write a factorized action for lattice QCD and we briefly discuss the implementation of a two-level HMC algorithm.

2 Signal-to-noise ratio of MC observables

As a prototypical manifestation of the S/N problem, let us first study a purely-bosonic theory. Consider the two-point function of a local bosonic field $O(x)$ projected to definite three-momentum $\boldsymbol{p}$

$$G_O(x_0 - y_0, \boldsymbol{p}) = \sum_{\boldsymbol{x}} \mathrm{e}^{\mathrm{i}\boldsymbol{p}\cdot\boldsymbol{x}} \left\langle O(x)O^\dagger(y)\right\rangle = \sum_k \left|\, \langle 0|O|k,\boldsymbol{p}\rangle\right|^2 \mathrm{e}^{-E_{O,k}(\boldsymbol{p})|x_0-y_0|} \sim \mathrm{e}^{-E_{O,0}(\boldsymbol{p})|x_0-y_0|}. \tag{4}$$

[2] Attempts to make proposals including dynamical fermions based on link updates as in refs [16–19] have not been adopted in large scale projects.

[3] See also refs [28, 29] on progress towards the non-perturbative test of the Witten-Veneziano formula.

Assuming that O has non-vacuum quantum numbers and the theory has a mass gap, the exponential decay at large Euclidean time separations $|x_0 - y_0|$ singles out the lightest state compatible with the symmetry transformation properties of the field O. In MC simulations, the path-integral expectation value is estimated on a finite sample of n gauge field configurations U_i, $i = 1, \dots, n$

$$\bar{G}_O(x_0 - y_0, \boldsymbol{p}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{\boldsymbol{x}} \mathrm{e}^{\mathrm{i}\boldsymbol{p}\cdot\boldsymbol{x}} O[U_i](x) O^\dagger[U_i](y). \tag{5}$$

For a large number n of configurations, which we assume to be statistically uncorrelated, $\bar{G}_O$ approximates the field-theoretic expectation value with an error of order $\sqrt{\sigma^2(G_O)/n}$ [4, 5]. The variance $\sigma^2(G_O)$ of G_O has a field-theoretic expression

$$\sigma^2(G_O)(x_0 - y_0, \boldsymbol{p}) = \sum_{\boldsymbol{x},\boldsymbol{x'}} \mathrm{e}^{-\mathrm{i}\boldsymbol{p}\cdot\boldsymbol{x'}} \left\langle O(x) O^\dagger(x + x') O^\dagger(y) O(y) \right\rangle - |G_O(x_0 - y_0, \boldsymbol{p})|^2. \tag{6}$$

Inserting a full set of eigenstates of the Hamiltonian, we observe that $\langle 0|OO^\dagger|0\rangle \neq 0$ since it has an overlap with the vacuum. Therefore, the first term on the r.h.s. is not a connected correlation function in the field-theoretical sense and it has a non-zero limit for $|x_0 - y_0| \to \infty$. Thus, the S/N of the MC correlator decays exponentially with $|x_0 - y_0|$

$$\frac{G_O}{\sqrt{\sigma^2(G_O)/n}} \sim \sqrt{n} \mathrm{e}^{-E_{O,0}(\boldsymbol{p})|x_0 - y_0|}. \tag{7}$$

In many cases, the exponential decay of S/N with separation $|x_0 - y_0|$ is a main limitation to the accuracy with which the spectrum and the matrix elements can be studied. Indeed, most of the systematic effects, such as finite volume and finite lattice spacing, can be solved with a polynomial increase of the simulation cost. On the contrary, one way to deal with excited states systematics is to increase the source-sink separation, but this requires an exponential increase of the number of configurations n, and thus of the simulation cost, to maintain the same statistical accuracy.

2.1 Variance of fermionic correlators

The presence of fermions in the theory modifies the S/N asymptotic described in the previous sections. Consider the correlation function of two quark bilinear fields with Dirac structure Γ and non-singlet flavour

$$G^{ud}_\Gamma(x_0 - y_0, \boldsymbol{p}) = -\sum_{\boldsymbol{x}} \mathrm{e}^{\mathrm{i}\boldsymbol{p}\cdot\boldsymbol{x}} \left\langle [\bar{d}\Gamma u](x) [\bar{u}\Gamma d](y) \right\rangle = -\sum_{\boldsymbol{x}} \mathrm{e}^{\mathrm{i}\boldsymbol{p}\cdot\boldsymbol{x}} \left\langle W^{\mathrm{con}}_{\Gamma,ud}(x, y) \right\rangle \sim \mathrm{e}^{-E^{ud}_{\Gamma,0}(\boldsymbol{p})|x_0 - y_0|}. \tag{8}$$

where $W^{\mathrm{con}}_{\Gamma,ud}$ is the connected Wick's contraction

$$W^{\mathrm{con}}_{\Gamma,ud}(x, y) = [\overbrace{\bar{d}(x)\Gamma u(x)}][\overbrace{\bar{u}(y)\Gamma d(y)}] = -\operatorname{tr}\{\gamma_5 \Gamma D_u^{-1}(x, y) \Gamma \gamma_5 D_d^{-1}(x, y)^\dagger\}, \tag{9}$$

and $E^{ud}_{\Gamma,0}(\boldsymbol{p})$ is the exponential decay rate at large temporal separations. Since the quark path integral is computed exactly, the MC variance is due only to fluctuations of the gauge field

$$\sigma^2(G^{ud}_\Gamma)(x_0 - y_0, \boldsymbol{p}) = \left\langle \left| \sum_{\boldsymbol{x}} \mathrm{e}^{\mathrm{i}\boldsymbol{p}\cdot\boldsymbol{x}} W^{\mathrm{con}}_{\Gamma,ud}(x, y) \right|^2 \right\rangle - \left| G^{ud}_\Gamma(x_0 - y_0, \boldsymbol{p}) \right|^2 \sim \mathrm{e}^{-M_{\pi\pi}|x_0 - y_0|}. \tag{10}$$

While the second term in the r.h.s. is just the square of the two-point function, the first term is obtained from the Wick's contraction of the two-point function of $[\bar{d}\Gamma u][\bar{d}'\Gamma u']$, where the two quark and two antiquark fields have different flavour in order to avoid disconnected Wick's contractions. This term can be shown to behave asymptotically as the lightest state with at least four quark lines, irrespective of Γ and $\boldsymbol{p}$, that is identified with a pair of zero-momentum pions. Therefore, any mesonic correlator with non-singlet flavour has a S/N that at long distances varies according to

$$\frac{G^{ud}_\Gamma}{\sqrt{\sigma^2(G^{ud}_\Gamma)/n}} \sim \sqrt{n} \mathrm{e}^{-\left(E^{ud}_{\Gamma,0}(\boldsymbol{p}) - M_\pi\right)|x_0 - y_0|}, \tag{11}$$

where we set $M_{\pi\pi} = 2M_\pi$ working in a large volume and assuming that there is no bound state. Thus, as a direct consequence of the fact that the quark path integral is computed exactly, the S/N problem is mitigated. The pion plays a special rôle: from Eq (11) is evident that both the mean and the width of the distribution of the pseudoscalar correlator decrease with distance at the same rate, thus there is no exponential S/N degradation in the pion propagator.

Therefore, eq. (11) provides a heuristic argument in support of eq. (3), *i.e.* the numerical evidence that the quark propagator on a typical gauge field configuration $\left\|D^{-1}(x,y)\right\| = -W^{\text{con}}_{\gamma_5,ud}(x,y)$ is suppressed exponentially with the pion mass at a large Euclidean separations between source and sink. This provides a partial mitigation of the S/N that extends to other fermionic correlators. A notable example is the nucleon propagator: its variance contains six quark lines connecting source to sink, and thus its S/N decays asymptotically with the exponential rate $M_N - (3/2)M_\pi$ [5].

The quark-line suppression does not help in the case of flavour-singlet correlators. Indeed, in this case both connected and disconnected Wick's contractions are present

$$G^{uu}_{\Gamma}(x_0 - y_0, \boldsymbol{p}) = -\sum_{\boldsymbol{x}} \mathrm{e}^{\mathrm{i}\boldsymbol{p}\cdot\boldsymbol{x}} \left\langle [\bar{u}\Gamma u](x)[\bar{u}\Gamma u](y)\right\rangle = -\sum_{\boldsymbol{x}} \mathrm{e}^{\mathrm{i}\boldsymbol{p}\cdot\boldsymbol{x}} \left\langle W^{\text{con}}_{\Gamma,uu}(x,y)\right\rangle - \sum_{\boldsymbol{x}} \mathrm{e}^{\mathrm{i}\boldsymbol{p}\cdot\boldsymbol{x}} \left\langle W^{\text{dis}}_{\Gamma,uu}(x,y)\right\rangle, \tag{12}$$

$$W^{\text{dis}}_{\Gamma,uu}(x,y) = [\overbrace{\bar{u}(x)\Gamma u}(x)][\overbrace{\bar{u}(y)\Gamma u}(y)] = \mathrm{tr}\big\{\Gamma D^{-1}_u(x,x)\big\}\,\mathrm{tr}\big\{\Gamma D^{-1}_u(y,y)\big\}. \tag{13}$$

The noise of G^{uu}_{Γ} is then usually dominated by the variance of the disconnected Wick's contraction that is not suppressed with distance. Intuitively, this is because there are no quark propagator lines that connects x with y in the variance of $W^{\text{dis}}_{\Gamma,uu}(x,y)$. Instead, as in the bosonic case, the variance has a non-zero limit for $|x_0 - y_0| \to \infty$.

3 Multi-level methods and locality

A way to improve the MC sampling exploiting the locality of quenched lattice QCD, known as the multi-hit method [8], was originally introduced to achieve smaller statistical errors on Wilson and Polyakov loops. In ref. [9], Lüscher and Weisz further developed this idea and proposed the multi-level algorithm as a full solution to the exponential S/N problem. They achieved a boost of the standard $n^{-1/2}$ scaling of the statical error by splitting the MC estimator in two (or more) levels of sampling. At the lowest level-0, the observable is averaged on n_0 configurations that spans the whole gauge field. At level-1 (and higher), n_1 gauge field configurations are generated for each level-0 configuration. The level-1 updates are characterized by the independent update of different spacetime regions of the lattice, and they are used to independently average the observable in distinct regions. The original multi-level idea has been subsequently extended to different bosonic theories and observables [10–14].

To understand how multi-level methods are based on locality, following ref. [10] consider a partition of the spacetime manifold in mutually disjoint subsets Λ_0, Λ_1 and Λ_b, with the requirement that any continuous path from Λ_0 to Λ_1 necessary passes through Λ_b. We denote with U_{Λ_i} a gauge field configuration supported in Λ_i. Then, the theory is local if the probability density in the path integral can be written as

$$\mathrm{d}P[U_{\Lambda_0\cup\Lambda_1}] = \int_{\Lambda_b} \mathrm{d}P[U_{\Lambda_b}]\mathrm{d}P_0[U_{\Lambda_0}]\mathrm{d}P_1[U_{\Lambda_1}], \tag{14}$$

i.e. there exists P_0 and P_1 such that U_{Λ_0} and U_{Λ_1} influence each other only through U_{Λ_b} [10]. This definition of locality automatically realizes a domain decomposition that factorizes the theory into sectors that are not directly influenced.

The locality condition in eq. (14) is satisfied if the action $S[U]$ can be written as $S_0[U_{\Omega_0}] + S_b[U_{\Lambda_b}] + S_1[U_{\Omega_1}]$, where $\Omega_i = \Lambda_i \cup \Lambda_b$, setting

$$\mathrm{d}P[U_{\Lambda_b}] = \frac{\mathcal{Z}_0\mathcal{Z}_1}{\mathcal{Z}}\,\mathcal{D}[U_{\Lambda_b}]\,\mathrm{e}^{-S_b[U_{\Lambda_b}]}, \qquad \mathrm{d}P_0[U_{\Lambda_0}] = \frac{1}{\mathcal{Z}_0}\,\mathcal{D}[U_{\Lambda_0}]\,\mathrm{e}^{-S_0[U_{\Omega_0}]}, \qquad \mathrm{d}P_1[U_{\Lambda_1}] = \frac{1}{\mathcal{Z}_1}\,\mathcal{D}[U_{\Lambda_1}]\,\mathrm{e}^{-S_1[U_{\Omega_1}]}. \tag{15}$$

Most lattice discretizations of YM theory are local in this sense. For instance, a partition in thick time slices of Wilson's plaquette action is given in ref. [9], with the boundaries Λ_b defined as the set of spatial links on a fixed time slice.

3.1 Factorization of the path integral

Consider two fields O_0 and O_1 localized respectively in Ω_0 and Ω_1. As a direct consequence of locality, their correlator has a factorized path integral expression

$$G_{O_0O_1} = \langle O_0[U]O_1[U]\rangle = \int \mathrm{d}P[U]O_0[U]O_1[U] = \int_{\Lambda_b} \mathrm{d}P[U_{\Lambda_b}]\left\langle\!\!\left\langle O_0[U_{\Omega_0}]\right\rangle\!\!\right\rangle_{\Lambda_0}\left\langle\!\!\left\langle O_1[U_{\Omega_1}]\right\rangle\!\!\right\rangle_{\Lambda_1}, \tag{16}$$

with

$$\left\langle\!\!\left\langle O_0\right\rangle\!\!\right\rangle_{\Lambda_0}[U_{\Lambda_b}] = \int_{\Lambda_0} \mathrm{d}P_0[U_{\Lambda_0}]O[U_{\Omega_0}], \qquad \left\langle\!\!\left\langle O_1\right\rangle\!\!\right\rangle_{\Lambda_1}[U_{\Lambda_b}] = \int_{\Lambda_1} \mathrm{d}P_1[U_{\Lambda_1}]O_1[U_{\Omega_1}]. \tag{17}$$

Thus, the path integral average is factorized into independent averages in disjoint domains Λ_0 and Λ_1, that are functionals of the boundary field U_{Λ_b}, times an average over the "boundary" Λ_b. In eqs (17), we introduced the notation $\langle\!\langle \bullet \rangle\!\rangle_\Lambda$ to denote the field-theoretical *sub-lattice* expectation value in the domain Λ at fixed boundary conditions. Its worth noting that this expectation value is still a functional of the field on the boundary. We can perform a further step on eq. (16): re-expressing $\mathcal{Z}_0$ and $\mathcal{Z}_1$ in path-integral form and using eq. (14), we obtain a standard path integral on the whole lattice

$$G_{O_0O_1} = \int \mathrm{d}P[U] \, \langle\!\langle O_0 \rangle\!\rangle_{\Lambda_0} [U_{\Lambda_b}] \, \langle\!\langle O_1 \rangle\!\rangle_{\Lambda_1} [U_{\Lambda_b}] = \left\langle \langle\!\langle O_0 \rangle\!\rangle_{\Lambda_0} \, \langle\!\langle O_1 \rangle\!\rangle_{\Lambda_1} \right\rangle , \tag{18}$$

where the sub-lattice expectation values are treated as fields depending on the boundary field only.

There are many different ways in which this factorization can be iterated. One possibility is to further subdivide the Λ_0 and Λ_1, to obtain a factorization in more than two regions. A second possibility is to iterate the factorization on $\langle\!\langle O_0 \rangle\!\rangle_{\Lambda_0}$ and $\langle\!\langle O_1 \rangle\!\rangle_{\Lambda_1}$ to obtain a nested expression with three or more levels.

3.2 The multi-level MC algorithm

The two-point function in eq. (18) has a simple realization in terms of a multi-level MC algorithm [9]. In the simple case of a two-domain two-level realization:

level-0: a number n_0 of field configurations $\{U_i\}$, $i = 1, \ldots, n_0$ defined on the whole lattice is generated, using a standard MC;

level-1: starting from every level-0 field configuration, a number n_1 of field configurations $\{U_{i,j}\}$, $j = 1, \ldots, n_1$ is generated updating independently fields in the two regions and keeping fixed fields in the boundary region Λ_b.

This requires computer time comparable to the generation of $n_0 n_1$ configurations with a standard single-level algorithm. The two-point function is estimated as two-level averaging process

$$G^{\mathrm{2lvl}}_{O_0O_1} = \frac{1}{n_0} \sum_{i=1}^{n_0} \left\{ \left[\frac{1}{n_1} \sum_{j=1}^{n_1} O_0[U_{i,j}] \right] \left[\frac{1}{n_1} \sum_{j=1}^{n_1} O_1[U_{i,j}] \right] \right\} . \tag{19}$$

It is important to notice that every level-1 field configuration in the domain Λ_0 combined with every level-1 field configuration in the domain Λ_1 is an independent configuration for both O_0 and O_1 fields. This means that at level-1 we have effectively n_1^2 configurations.

Eq. (19) is an improved estimator of $G_{O_0O_1}$. Suppose that O_0 and O_1 show no dependence at all on the boundary field. Then, each term in the square parenthesis in eq. (19) estimates the field-theoretical expectation value with an error of order $\sqrt{\sigma^2(O_i)/n_1}$. This, combined with the level-0 average, results in an error scaling with $(n_0 n_1^2)^{-1/2}$, instead of the $(n_0 n_1)^{-1/2}$ of the standard case. In general, O_0 and O_1 retain, at least indirectly, a dependence on the quantum fluctuations of the boundary field. This induces in the estimator in eq. (19) a contribution to the error that scales with ${n_0}^{-1/2}$. This scaling will dominate for $n_1 \to \infty$, but how large n_1 can be before it becomes dominant depends on the details of the observable and the domain decomposition. With a sensible setup, the dependency is small and one can take quite large n_1 before the boundary contribution becomes relevant.

3.3 Theories with fermions

The validity of eq. (16) relies on some necessary conditions: first, the theory must be local, *i.e.* it is possible to factorize the action of the theory in domains as described in eq. (14); second, the fields in the expectation value must be localized to a domain, or can be written as a product of localized fields. However, in the numerical simulation of lattice theory with fermions both conditions are not satisfied. Indeed, after integrating out fermions explicitly,

1. the path-integral Boltzmann weight includes the determinant of the Dirac operator, that is a non-local functional of the gauge field and it does not satisfy eq. (14);
2. fermionic observables are given by the inverse of the Dirac operator, that is also a non-local functional of the gauge field and it is not factorizable in contribution localized to a region.

In the rest of this article, we generalize multi-level sampling to theories with fermions, addressing the two issues separately. Taking QCD as the reference theory, we develop approximations to the quark propagator and to the quark determinant that allow to write them as the product of local factors. In both cases, the approximation can be systematically improved and the exact propagator and determinant are recovered in an appropriate limit. It is important to stress that the proposed algorithm in unbiased in spite of the use of an approximation. Even far from

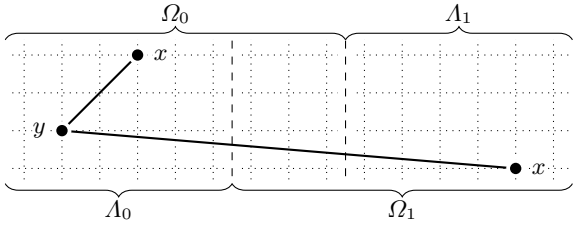

Figure 1. Sketch of the partition of the lattice in Ω_i and Λ_i regions. Above the lattice we label the regions Ω_0 and Λ_1 that leads to the block decomposition in eqs (20) and (21). Below we label the regions Λ_0 and Ω_1 that corresponds to the block decomposition in eq. (22). Starting from a point source $y \in \Lambda_0$, a quark propagator can be evaluated on a sink x localized either in Λ_0 or in Λ_1.

the aforementioned limit, the correction to the approximation is small and we take its contribution into account. As we will show, this is obtained using standard variance reduction ideas [32–36] and reweighting techniques, for the propagator and determinant factorization respectively.

In sect. 4 we first study the factorization of the quark propagator. The resulting approximated quark propagator allows us to write hadronic observables that can be computed using a multi-level estimator such as the one in eq. (19), at least in the quenched approximation of QCD, in which the determinant contribution to the path integral is neglected. The quark propagator approximation is such that the correction term is suppressed with respect to exact non-factorizable observable on every representative gauge field configuration. Estimating this small correction with a standard MC results in a reduced impact on the S/N.

Then, in sect. 7 we study the factorization of the quark determinant and we describe an algorithm to perform multi-level sampling in QCD with dynamical fermions.

4 Factorization of the quark propagator

The quark propagator $D^{-1}(x, y)$, defined as the solution of the Dirac equation on a point-like source in y, has a non-local dependence on all the gauge links of the lattice Ω. However, the locality of the underlying physics suggests to us that the dependence of $D^{-1}(x, y)$ on gauge links should be decreasing for links further away from x and y. In this section, we use this physical intuition to introduce a factorization of the gauge field dependence of the quark propagator in a way that is suitable for multi-level MC sampling.

Let us introduce a region $\Omega_0 \subset \Omega$ of the lattice Ω such that the source point $y \in \Omega_0$, and denote the complementary as $\Lambda_1 = \Omega \setminus \Omega_0$. The Dirac operator then assumes the block form[4]

$$D = \begin{pmatrix} D_{\Omega_0} & D_{\partial\Omega_0} \\ D_{\partial\Lambda_1} & D_{\Lambda_1} \end{pmatrix}, \tag{20}$$

where we are using the notation and conventions of ref. [37], also introduced in appendix A. Using the block decomposition of the quark propagator matrix D^{-1} from eq. (65a) we have

$$D^{-1} = \begin{pmatrix} D^{-1}_{\Omega_0} + D^{-1}_{\Omega_0} D_{\partial\Omega_0} D^{-1} D_{\partial\Lambda_1} D^{-1}_{\Omega_0} & -D^{-1}_{\Omega_0} D_{\partial\Omega_0} D^{-1} P^{\top}_{\Lambda_1} \\ -P_{\Lambda_1} D^{-1} D_{\partial\Lambda_1} D^{-1}_{\Omega_0} & P_{\Lambda_1} D^{-1} P^{\top}_{\Lambda_1} \end{pmatrix}, \tag{21}$$

were we used (66a) to express the Schur complement $[Q/Q_{\Omega_0}]^{-1}$ as $P_{\Lambda_1} Q^{-1} P^{\top}_{\Lambda_1}$. In addition, we define a region $\Lambda_0 \subset \Omega_0$ around the source point $y \in \Lambda_0$ such that Λ_0 and Λ_1 are separated by an intermediate region that we call Λ_b, as depicted graphically in fig. 1. The complementary region is $\Omega_1 = \Omega \setminus \Lambda_0$, and $\Lambda_b = \Omega_0 \cap \Omega_1$. This leads to a second block decomposition formula, based on Λ_0 and Ω_1, that we obtain using eq. (65b)

$$D^{-1} = \begin{pmatrix} P_{\Lambda_0} D^{-1} P^{\top}_{\Lambda_0} & -P_{\Lambda_0} D^{-1} D_{\partial\Lambda_0} D^{-1}_{\Omega_1} \\ -D^{-1}_{\Omega_1} D_{\partial\Omega_1} D^{-1} P^{\top}_{\Lambda_0} & D^{-1}_{\Omega_1} + D^{-1}_{\Omega_1} D_{\partial\Omega_1} D^{-1} D_{\partial\Lambda_0} D^{-1}_{\Omega_1} \end{pmatrix}. \tag{22}$$

As we show in the following, these two formulae are the starting point to rewrite the quark propagator $D^{-1}(x, y)$ in a way such that the gauge field dependence on Λ_0 and Λ_1 is completely factorized. Having chosen a source point $y \in \Lambda_0$, we can either consider sink points x that are close to y, so that we have $x \in \Lambda_0$, or consider y that are at a distance from y, so that $x \in \Lambda_1$. We study the former, denoted as "same region", in sect. 4.1, and the latter, "different regions", in sect. 4.2.

[4] The action of D_{Ω_0} (and D_{Λ_1} respectively) on a quark field in Ω_0 (Λ_1) is equivalent to the action of the full Dirac operator D with Dirichlet boundary conditions on $\partial\Omega_0$ ($\partial\Lambda_1$) [37].

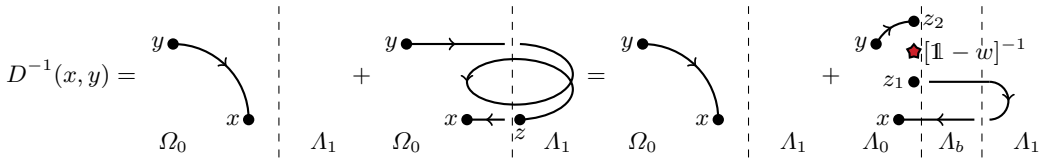

Figure 2. Pictorial representation of the factorization of the quark propagator for $y, x \in \Omega_0$. The first equation is a representation of eq. (23), while the last of eq. (26). ✪ represents the insertion of the Neumann series of w defined in eq. (27).

4.1 Source and sink in the same region

When $x, y \in \Lambda_0$, including the propagator trace case $y = x$, the upper left element from eq.(21) is the relevant one. Writing explicitly all the spacetime indices, we have

$$D^{-1}(x,y) = D^{-1}_{\Omega_0}(x,y) + \sum_{z \in \partial\Omega_0} \left[D^{-1}_{\Omega_0} D_{\partial\Omega_0}\right](x,z) \left[D^{-1} D_{\partial\Lambda_1} D^{-1}_{\Omega_0}\right](z,y), \qquad x, y \in \Omega_0. \tag{23}$$

In particular, comparing this expression with the lower left element of eq. (21), the last factor in eq. (23) is, up to a sign, a quark propagator from $y \in \Omega_0$ to its exterior boundary. Thus, it gets a suppression $\mathcal{O}\big(\mathrm{e}^{-M_\pi d_y/2}\big)$, where d_y is the distance of y from $\partial\Omega_0$. Similarly, the second term in the r.h.s. of eq. (23) is $\mathcal{O}\big(\mathrm{e}^{-M_\pi d_x/2}\big)$, with d_x being the distance of x from the same boundary. Therefore,

$$\left\| D^{-1}(x,y) - D^{-1}_{\Omega_0}(x,y) \right\| \sim \mathrm{e}^{-M_\pi d}, \qquad \text{with } d = (d_x + d_y)/2. \tag{24}$$

$D^{-1}_{\Omega_0}(x,y)$ is pictured in fig. 2 as including all the paths from x to y that are within the region Ω_0. The second term in the r.h.s of eq. (23) is then represented graphically as the last term in the first equation in fig. 2, including paths in Λ_1 that can wind in Ω_0.

This approximation can be improved. Comparing eq. (21) with eq. (22), we observe that $[D^{-1} D_{\partial\Lambda_1} D^{-1}_{\Omega_0}](z,y) = [D^{-1}_{\Omega_1} D_{\partial\Omega_1} D^{-1}](z,y)$ for $z \in \Lambda_1$ and $y \in \Lambda_0$. Performing this replacement in eq. (23), we have

$$D^{-1}(x,y) = D^{-1}_{\Omega_0}(x,y) + \sum_{z \in \partial\Omega_1} \tilde{w}(x,z) D^{-1}(z,y), \qquad x \in \Lambda_0,\ y \in \Omega_0, \tag{25}$$

where $\tilde{w}(x,z) = [D^{-1}_{\Omega_0} D_{\partial\Omega_0} D^{-1}_{\Omega_1} D_{\partial\Omega_1}](x,z)$ for $z \in \partial\Omega_1$ and $x \in \Lambda_0$. Formally solving for D^{-1}, we arrive at

$$D^{-1}(x,y) = D^{-1}_{\Omega_0}(x,y) + \sum_{z_1, z_2 \in \partial\Omega_1} \tilde{w}(x,z_1) \left[\mathbb{1} - w\right]^{-1}(z_1,z_2) D^{-1}_{\Omega_0}(z_2,y), \tag{26}$$

where we used the properties of $D_{\partial\Omega_1}$ acting form the left to introduce [38],

$$w(z_1,z_2) = \left[P_{\partial\Omega_1} D^{-1}_{\Omega_0} D_{\partial\Omega_0} D^{-1}_{\Omega_1} D_{\partial\Omega_1}\right](z_1,z_2), \qquad z_i \in \partial\Omega_1, \tag{27}$$

that has the crucial property of being an operator living on the $\partial\Omega_1$ boundary only. The $[\mathbb{1} - w]^{-1}$ term in Eq (26) is the sum of the Neumann series of w. The second term in the r.h.s. of Eq (26) is then an infinite sum of products of $D^{-1}_{\Omega_0}$ and $D^{-1}_{\Omega_1}$ propagator, thus any truncation of the sum leads to a factorization of the gauge field dependence in Λ_0 and Λ_1. The series converges fast thanks to the properties of w, that is studied in details in sect. 7.4 in the context of the determinant factorization. Intuitively, since the expression in eq. (27) entails a propagator from $\partial\Omega_1$ to $\partial\Omega_0$ and back, we expect $\|w\| \sim \mathrm{e}^{-M_\pi \Delta}$, where Δ is the "thickness" of Λ_b. The last equation in fig. 2 provides a pictorial representation of eq. (26), where the summed series of w insertions is denoted by a ✪.

4.2 Source and sink in different regions

A second possibility is that x is at a physical distance from y, such that $x \in \Lambda_1$. Then, taking the lower left element from eq. (22) we have

$$D^{-1}(x,y) = -\sum_{z \in \partial\Omega_1} \left[D^{-1}_{\Omega_1} D_{\partial\Omega_1}\right](x,z) D^{-1}(z,y), \qquad x \in \Omega_1,\ y \in \Lambda_0. \tag{28}$$

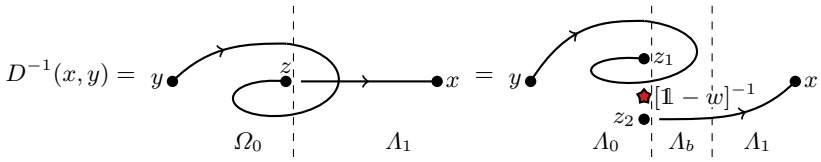

Figure 3. Pictorial representation of the factorization of the quark propagator for $y \in \Lambda_0$ and $x \in \Omega_1$. The first equation is a representation of eq. (28), while the last of eq. (29). ✸ represents the insertion of the Neumann series of w defined in eq. (27).

The r.h.s. of eq. (28) contains a D^{-1} factor that depends on the gauge field of the whole lattice, thus the gauge field dependence is not factorized yet. However, since $D^{-1}(z,y)$ acts on $y \in \Lambda_0$ and it is evaluated in $z \in \partial\Omega_1$, with both well into Ω_0, we can use the factorization derived sect. 4.1, specifically eq. (26), to write

$$D^{-1}(x,y) = -\sum_{z_1,z_2 \in \partial\Omega_1} \left[D^{-1}_{\Omega_1} D_{\partial\Omega_1}\right](x,z_2)\,[\mathbb{1}-w]^{-1}(z_2,z_1) D^{-1}_{\Omega_0}(z_1,y), \qquad x \in \Omega_1,\, y \in \Lambda_0. \tag{29}$$

where w is defined in eq. (27). As in the same-region case in eq. (26), any truncation of the Neumann series of w in eq. (29) leads to a factorization of the gauge field dependence in Λ_0 and Λ_1, that is suitable for the application of multi-level MC sampling. Fig. 3 provides a pictorial representation of eqs (28) and (29). The fast convergence of the series means that the "zeroth-order" approximation

$$D^{-1}(x,y) \simeq -\left[D^{-1}_{\Omega_1} D_{\partial\Omega_1} D^{-1}_{\Omega_0}\right](x,y), \qquad x \in \Omega_1,\, y \in \Lambda_0, \tag{30}$$

is most of the time adequate, with $\|D^{-1} - D^{-1}_{\Omega_1} D_{\partial\Omega_1} D^{-1}_{\Omega_0}\| \sim \mathrm{e}^{-M_\pi \Delta}$, where Δ is the "thickness" of Λ_b.

In a multi-region setup, such as the partition in thick time slices that is introduced in fig. 8, it is possible to iterate the steps that lead to eq. (30) to obtain a fully-factorized quark propagator spanning more than two regions

$$D^{-1}_{\mathrm{f}}(x,y) = (-)^{m-1} \sum_{z_i \in \partial\Omega_i} \left[D^{-1}_{\Omega_m} D_{\partial\Omega_m}\right](x,z_m) \prod_{i=m-1}^{1} \left[D^{-1}_{\Omega_i} D_{\partial\Omega_i}\right](z_{i+1},z_i) D^{-1}_{\Omega_0}(z_1,y), \qquad x \in \Omega_m,\, y \in \Lambda_0, \tag{31}$$

that is pictured in fig. 4. Since in each factor the inverse Dirac operators are such that source and sink are always at a distance Δ from the Dirichlet boundary conditions, $\|D^{-1} - D^{-1}_{\mathrm{f}}(x,y)\| \sim \mathrm{e}^{-M_\pi \Delta}$ still holds.

5 Multi-level integration of the disconnected meson propagator

The decomposition obtained in sect. 4.1 calls for a multi-level integration of disconnected contributions to correlation functions. Consider the disconnected Wick's contraction of the mesonic two-point function in eq. (12), specialized to the pseudoscalar case,

$$W^{\mathrm{dis}}_{\gamma_5}(x,y) = \mathrm{tr}\{\gamma_5 D^{-1}(x,x)\}\,\mathrm{tr}\{\gamma_5 D^{-1}(y,y)\}, \tag{32}$$

and assume that $y \in \Lambda_0$ and $x \in \Lambda_1$. Applying eq. (23) to both quark traces, the Wick's contraction is decomposed as

$$W^{\mathrm{dis}}_{\gamma_5}(x,y) = W^{\mathrm{dis,f}}_{\gamma_5}(x,y) + W^{\mathrm{dis,r}}_{\gamma_5}(x,y) \tag{33}$$

where the factorized contribution is

$$W^{\mathrm{dis,f}}_{\gamma_5}(x,y) = \mathrm{tr}\{\gamma_5 D^{-1}_{\Omega_1}(x,x)\}\,\mathrm{tr}\{\gamma_5 D^{-1}_{\Omega_0}(y,y)\}, \tag{34}$$

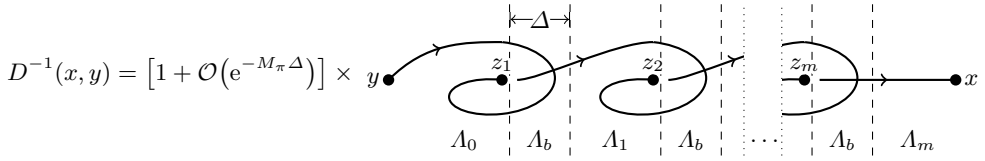

Figure 4. Pictorial representation of the approximated factorized quark propagator in a multi-region setup, from a source $y \in \Lambda_0$ to a sink $x \in \Omega_m$.

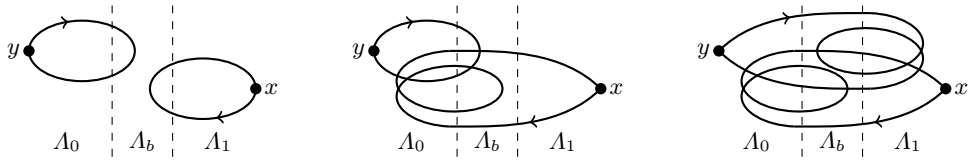

Figure 5. Pictorial representation of the factorization of the disconnected Wick's contraction of the two-point function of quark bilinears. The leftmost picture depicts the factorized contribution $W_{\gamma_5}^{\mathrm{dis,f}}$ defined in eq. (34). The central picture is a representation of the residual contribution $W_{\gamma_5}^{\mathrm{dis,r_1}}$, with $W_{\gamma_5}^{\mathrm{dis,r_2}}$ being similar but with the rôle of x and y exchanged. The rightmost picture depicts the residual contribution $W_{\gamma_5}^{\mathrm{dis,r_3}}$.

while the rest of the contraction is further decomposed in

$$W_{\gamma_5}^{\mathrm{dis,r}}(x,y) = W_{\gamma_5}^{\mathrm{dis,r_1}}(x,y) + W_{\gamma_5}^{\mathrm{dis,r_2}}(x,y) + W_{\gamma_5}^{\mathrm{dis,r_3}}(x,y), \tag{35}$$

where

$$W_{\gamma_5}^{\mathrm{dis,r_1}}(x,y) = \mathrm{tr}\big\{\gamma_5 \left[D_{\Omega_1}^{-1} D_{\partial\Omega_1} D^{-1} D_{\partial\Lambda_0} D_{\Omega_1}^{-1}\right](x,x)\big\}\,\mathrm{tr}\big\{\gamma_5 D_{\Omega_0}^{-1}(y,y)\big\}, \tag{36a}$$

$$W_{\gamma_5}^{\mathrm{dis,r_2}}(x,y) = \mathrm{tr}\big\{\gamma_5 D_{\Omega_1}^{-1}(x,x)\big\}\,\mathrm{tr}\big\{\gamma_5 \left[D_{\Omega_0}^{-1} D_{\partial\Omega_0} D^{-1} D_{\partial\Lambda_1} D_{\Omega_0}^{-1}\right](y,y)\big\}, \tag{36b}$$

$$W_{\gamma_5}^{\mathrm{dis,r_3}}(x,y) = \mathrm{tr}\big\{\gamma_5 \left[D_{\Omega_1}^{-1} D_{\partial\Omega_1} D^{-1} D_{\partial\Lambda_0} D_{\Omega_1}^{-1}\right](x,x)\big\}\,\mathrm{tr}\big\{\gamma_5 \left[D_{\Omega_0}^{-1} D_{\partial\Omega_0} D^{-1} D_{\partial\Lambda_1} D_{\Omega_0}^{-1}\right](y,y)\big\}. \tag{36c}$$

The contribution in eq. (34) has the property of being the product of two factors that are local to Ω_0 and Ω_1. Therefore, it is possible to apply multi-level sampling to it

$$\left\langle W_{\gamma_5}^{\mathrm{dis,f}}(x,y)\right\rangle_{\mathrm{2lvl}} = \left\langle \left\langle\!\!\left\langle \mathrm{tr}\big\{\gamma_5 D_{\Omega_1}^{-1}(x,x)\big\}\right\rangle\!\!\right\rangle \left\langle\!\!\left\langle \mathrm{tr}\big\{\gamma_5 D_{\Omega_0}^{-1}(y,y)\big\}\right\rangle\!\!\right\rangle \right\rangle. \tag{37}$$

Of course, this assumes that is possible generated independent gauge field configuration in different regions, as it is the case when the disconnected Wick's contribution is studied in the quenched theory, or using the corresponding factorization of the Dirac determinant introduced in sect. 7. Conversely, the remaining contribution $W_{\gamma_5}^{\mathrm{dis,r}}$ is not factorized and it cannot be estimated with multi-level techniques. However, the contribution from $W_{\gamma_5}^{\mathrm{dis,r_1}}$ is expected to be suppresses, for a typical configuration, by a factor $\mathrm{e}^{-M_\pi d_x}$ at large temporal separations, with d_x being the distance of x from $\partial\Omega_1$. Similarly, $W_{\gamma_5}^{\mathrm{dis,r_2}}$ is suppressed by a factor $\mathrm{e}^{-M_\pi d_y}$ with d_y being the distance of y from $\partial\Omega_0$. Thus, the noise from the non-factorized contribution is suppressed with distance accordingly. Choosing x and y to be roughly equidistant from Λ_b, the noise decays with an exponential rate that is half of the one of the expected signal of the disconnected contribution, *i.e.* a halving of the exponential S/N problem of the standard MC estimator. The last contribution, $W_{\gamma_5}^{\mathrm{dis,r_3}}$, is expected to be already proportional to $\mathrm{e}^{-M_\pi(d_x+d_y)} \sim \mathrm{e}^{-M_\pi|x_0-y_0|}$. This is of the same order of the expected signal, and therefore the standard level-0 average is adequate.

The factorization can be iterated using eq. (26). With the next iteration, the $W_{\gamma_5}^{\mathrm{dis,r}}$ contribution splits into a term that is suitable for multi-level sampling and a residual term that is further suppressed, at least as $\mathrm{e}^{-M_\pi(d+\Delta)}$, where d is either d_x or d_y and Δ is the thickness of Λ_b. However, this involves the multi-level evaluation of quark lines crossing the boundary that, as discussed in sect. 6, is technically more involved.

5.1 Numerical tests

In ref. [21] we tested the factorization of the disconnected Wick's contraction in a two-level setup with two active region. To disentangle from the complexity of the factorization of the quark determinant, that is described in sect. 7, we performed the test in the quenched approximation of QCD. We employed Wilson's plaquette action discretization with parameters $\beta = 6.0$, $T \times L^3 = 64a \times 24^3 a^3$ and open boundary conditions (OBCs) in the temporal direction, that correspond to a lattice spacing of $a = 0.093\,\mathrm{fm}$ fixed using the value $0.5\,\mathrm{fm}$ for the Sommer scale $r_0/a = 5.368$ [39]. We thus generated $n_0 = 200$ level-0 independent gauge field configurations spaced by 400 molecular-dynamics units (MDUs) using openQCD [40, 41] implementation of the HMC algorithm. For each level-0 configuration, we generated $n_1 = 100$ level-1 configurations updating independently two thick time slices of equal size, $\Lambda_0 = \{x | x_0 < 32a\}$ and $\Lambda_1 = \{x : x_0 > 32a\}$, while keeping the spatial links at $x_0 = 32a$ fixed, also spacing the configurations by 400 MDUs. On these configurations, we studied a doublet of mass-degenerate valence light quark, discretized with the unimproved Wilson–Dirac operator with $\kappa = 0.1560$ that corresponds to $M_\pi = 455\,\mathrm{MeV}$ [42].

The disconnected Wick's contraction is decomposed according to eqs (33) and (35), and the factorized contribution is computed using the two-level estimator in Eq (19)

$$G_{\gamma_5,\mathrm{2lvl}}^{\mathrm{dis,f}}(x_0,y_0) = \frac{1}{n_0}\sum_{i=1}^{n_0}\left[\frac{1}{n_1}\sum_{j=1}^{n_1}\sum_{\boldsymbol{x}} \mathrm{tr}\big\{\gamma_5 D_{\Omega_1}^{-1}[U_{i,j}](x,x)\big\}\right]\left[\frac{1}{n_1}\sum_{j=1}^{n_1}\frac{1}{L^3}\sum_{\boldsymbol{y}} \mathrm{tr}\big\{\gamma_5 D_{\Omega_0}^{-1}[U_{i,j}](y,y)\big\}\right]. \tag{38}$$

The terms in eq. (36) are instead computed using a standard estimator, but still making use of the 100 level-1 configurations as global configurations,[5]

$$G^{\mathrm{dis,r_{12}}}_{\gamma_5}(x_0,y_0) = \frac{1}{n_0 n_1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1}\frac{1}{L^3}\sum_{\boldsymbol{x},\boldsymbol{y}}\left\{W^{\mathrm{dis,r_3}}_{\gamma_5}[U_{i,j}](x,y) + W^{\mathrm{dis,r_1}}_{\gamma_5}[U_{i,j}](x,y)\right\}, \tag{39a}$$

$$G^{\mathrm{dis,r_3}}_{\gamma_5}(x_0,y_0) = \frac{1}{n_0 n_1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1}\frac{1}{L^3}\sum_{\boldsymbol{x},\boldsymbol{y}}W^{\mathrm{dis,r_3}}_{\gamma_5}[U_{i,j}](x,y), \tag{39b}$$

where we summed for convenience the r_1 and r_2 contributions. For both exact propagators D^{-1} and factorized ones $D^{-1}_{\Omega_0}$ and $D^{-1}_{\Omega_1}$, we achieved a significant noise reduction using the hopping parameter expansion identity $\mathrm{tr}\{\gamma_5 D^{-1}\} = \kappa^p\,\mathrm{tr}\{\gamma_5 D^p_{\mathrm{hop}} D^{-1}\}$, where $2\kappa D = \mathbb{1} - \kappa D_{\mathrm{hop}}$ and $p \leq 8$ [43, 44, 35], and estimating the trace stochastically

$$\frac{1}{n_s}\sum_{1}^{n_s}\sum_{\boldsymbol{x}}\eta^\dagger_i(x)\left[\kappa^8 D^8_{\mathrm{hop}} D^{-1}\gamma_5\eta\right](x) \xrightarrow{n_s\to\infty} \sum_{\boldsymbol{x}}\mathrm{tr}\left\{\gamma_5 D^{-1}\right\} \tag{40}$$

on $n_s = 100$ sources η_i of Gaussian noise [45, 46], defined on the whole spacetime volume. No attempt was made at further optimising the stochastic estimator and the number of sources. For recent results combining multi-level sampling and state-of-the-art stochastic estimation of propagator traces, see ref. [47].

The numerical results for $G^{\mathrm{dis,f}}_{\gamma_5,\mathrm{2lvl}}$, $G^{\mathrm{dis,r_{12}}}_{\gamma_5}$ and $G^{\mathrm{dis,r_3}}_{\gamma_5}$ are plotted in fig. 6 as a function of the temporal separation of the pseudoscalar densities. The central values and their errors are shown in the plots on the left and right columns respectively. The sum of the three contributions, $G^{\mathrm{dis}}_{\gamma_5,\mathrm{2lvl}}$, is also shown in each plot on the left for comparison. In all cases $x_0 \in \Omega_1$ and $y_0 \in \Omega_0$ and they are chosen to be as much as possible equidistant from the boundary at $x_0 = 32a$.

The statistical error on $G^{\mathrm{dis,f}}_{\gamma_5,\mathrm{2lvl}}$ in the top-right plot of fig. 6 is a flat function of the temporal separation with sizeable deviations near the boundaries of the domains. Error bars are smaller than the symbols. Up to the largest value that we have, $n_1 = 100$, the error decreases as n_1^{-1}, *i.e.* the two-level MC works at full capacity. The mean value of $G^{\mathrm{dis,f}}_{\gamma_5,\mathrm{2lvl}}$ in the top-left plot is compatible with zero. The correlation between $G^{\mathrm{dis}}_{\gamma_5}$ and $G^{\mathrm{dis,f}}_{\gamma_5,\mathrm{2lvl}}$ goes from 0.9 to 1.0 for temporal separations from $15a$ to $50a$, a value that collapses towards zero when the multi-level is switched on. The statistical error on $G^{\mathrm{dis,r_{12}}}_{\gamma_5}$ in the middle-right plot of fig. 6 shows a strong dependence on the temporal separation. It is compatible with an exponential behaviour of the form $\mathrm{e}^{-M|x_0-y_0|/2}$ as expected from eqs (36a) and (36b), but with an effective mass $aM = 0.14$, lighter than expected and $\approx 67\,\%$ of the pion mass. It decreases as $n_1^{-1/2}$ up to $n_1 = 100$ and, at fixed temporal distance, it becomes the dominant contribution to the error of $G^{\mathrm{dis}}_{\gamma_5,\mathrm{2lvl}}$ once n_1 is large enough. The mean value of $G^{\mathrm{dis,r_{12}}}_{\gamma_5}$ is $\approx 67\,\%$ of the full correlator at $|x_0 - y_0| = 15a$, and it becomes the dominant contribution at larger distances. The statistical errors on $G^{\mathrm{dis,r_3}}_{\gamma_5}$ in the bottom-right plot of fig. 6 decreases exponentially as $\mathrm{e}^{-M|x_0-y_0|/2}$ as expected from eq. (36c), and it scales as $n_1^{-1/2}$.

Our best estimate of the disconnected contribution is shown in fig. 7, where also the result without multi-level is plotted for comparison. A clear picture emerges. Without multi-level, at large temporal separations, the statistical error on the standard estimate of the disconnected Wick's contraction contribution to the pseudoscalar propagator is dominated by the one on $G^{\mathrm{dis,f}}_{\gamma_5}$. The second largest contribution is the statistical error on $G^{\mathrm{dis,r_{12}}}_{\gamma_5}$ which, however, is exponentially suppressed as $\mathrm{e}^{-M|x_0-y_0|/2}$. Once the two-level integration is switched on, the error on $G^{\mathrm{dis,r_{12}}}_{\gamma_5}$ decreases as n_1^{-1}, while the one on $G^{\mathrm{dis,r_{12}}}_{\gamma_5}$ continues to scale as $n_1^{-1/2}$. The right plot in fig. 7 shows that, for $n_1 = 100$, the error on the multi-level estimator decreases with $\mathrm{e}^{-M|x_0-y_0|/2}$ up to a temporal separation of $\approx 30a$, and it is dominated by the flat error on $G^{\mathrm{dis}}_{\gamma_5,\mathrm{2lvl}}$ for temporal separations $\gtrsim 35a$. As it is shown in the left plot, this results in a S/N larger than one for ten additional time slices. The parameter n_1 can thus be tuned, up to a prefactor of $\mathcal{O}(1)$, so that $n_1 \sim \mathrm{e}^{Md}$ with d being the maximum temporal separation in which one is interested in. In this way, the error on the factorized contribution is reduced to the level of (or below) the uncertainty on $G^{\mathrm{dis,r_{12}}}_{\gamma_5}$ at the same cost of generating $n_0 n_1$ global configurations. The net computational gain is therefore proportional to n_1, and a good statistical precision is reached with a total number of updates $n_0 n_1$ proportional to $\mathrm{e}^{M_\pi|x_0-y_0|}$. Notice that the factor at the exponent is halved with respect to the standard MC.

6 Multi-level integration of connected meson and baryon propagators

In parallel to the disconnected contributions discussed in the previous section, using the factorization obtained in sect. 4.2 it is possible to apply multi-level sampling to *connected* Wick's contraction contributions to hadron propa-

[5] On any given configuration, the quark traces in eq. (36) are easy computed as the difference between exact the quark propagators and factorized ones, *e.g.* $\mathrm{tr}\left\{\gamma_5\left[D^{-1}_{\Omega_1} D_{\partial\Omega_1} D^{-1} D_{\partial\Lambda_0} D^{-1}_{\Omega_1}\right]\right\} = \mathrm{tr}\left\{\gamma_5 D^{-1}\right\} - \mathrm{tr}\left\{\gamma_5 D^{-1}_{\Omega_1}\right\}$.

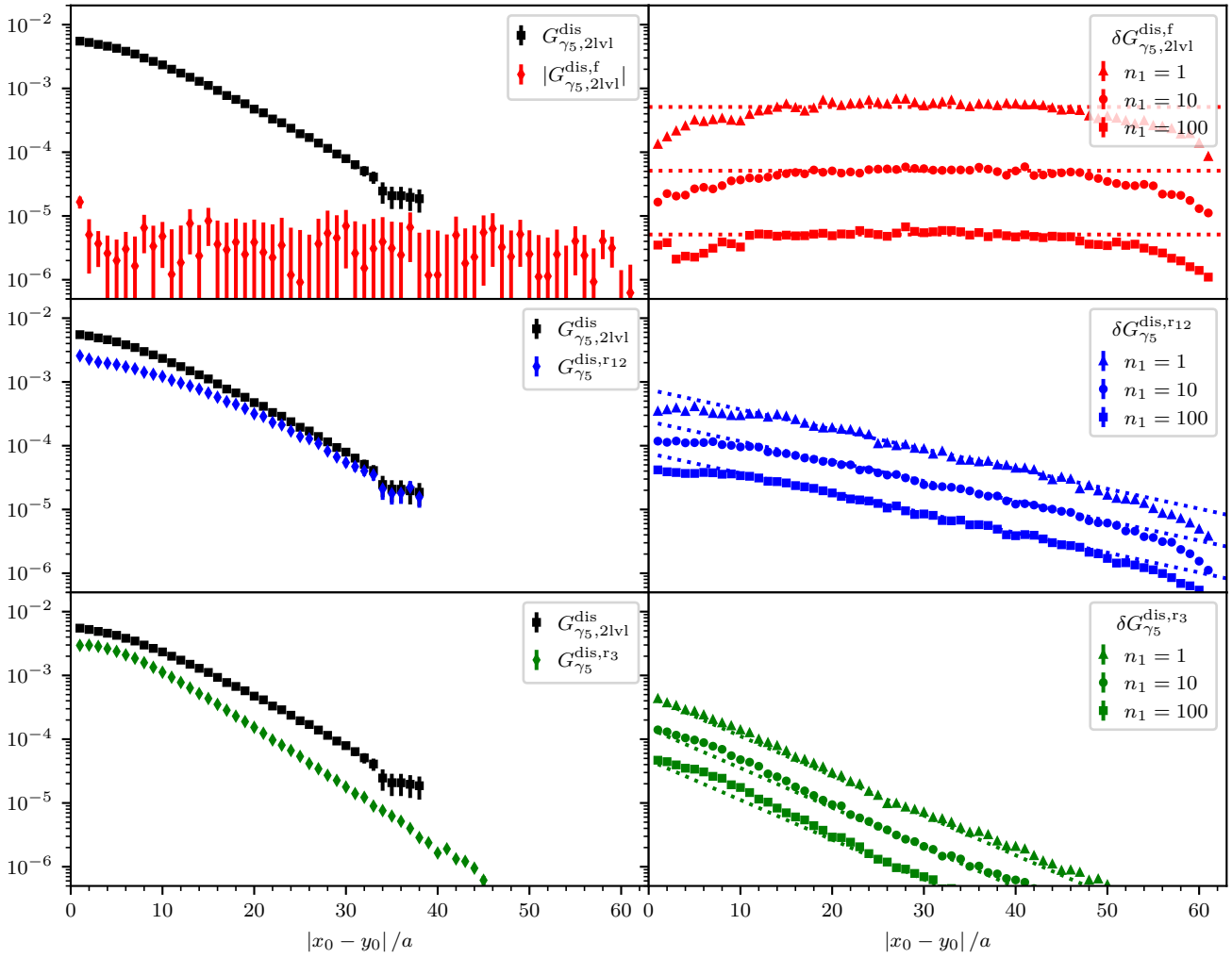

Figure 6. The plots in the left column show, from top to bottom, the contributions $G^{\mathrm{dis}}_{\gamma_5,\mathrm{2lvl}}$, $G^{\mathrm{dis,r_{12}}}_{\gamma_5}$ and $G^{\mathrm{dis,r_3}}_{\gamma_5}$ as function of temporal separation $|x_0 - y_0|$, together with the best estimate of the full correlator given by the sum o the three. The plots in the right column show, for each of the three contributions, the corresponding statistical error as function of temporal separation and for various values of level-1 samples n_1.

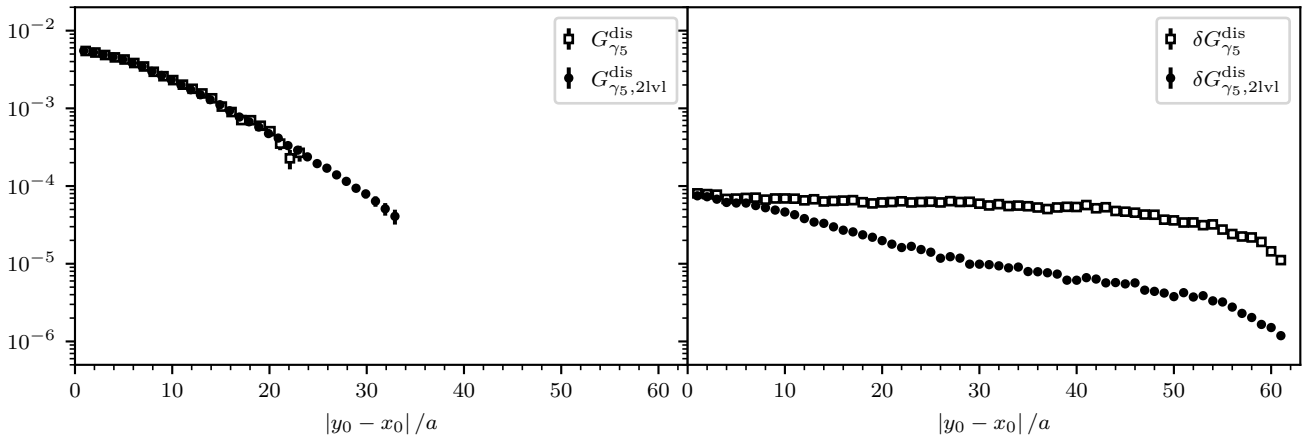

Figure 7. The best estimate of $G^{\mathrm{dis}}_{\gamma_5}$ is shown in the left plot, together with its statistical error in the right plot, as a function of temporal separation $|x_0 - y_0|$, both with and without two-level sampling. In the latter case, the $n_1 = 100$ level-1 configurations are treated as if they were correlated level-0 ones.

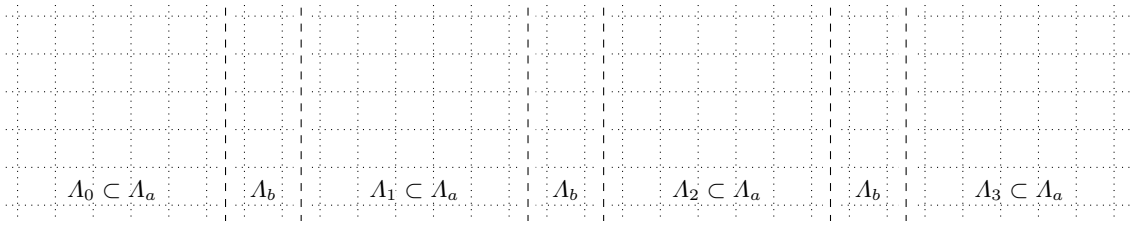

Figure 8. Sketch of a possible decomposition of the lattice in Λ_a and Λ_b thick-time-slice regions. Λ_a is the union of the thick time slices, labelled by Λ_i, $i = 0, \ldots, 3$, whose links are active at level-1 of two-level integration. Λ_b is the union of the thick time slices that act a buffer between Λ_i regions and whose links are not updated at level-1.

gators. The factorization and multi-level integration of two-point functions, such as and the nucleon propagator and the pion propagator, which however is not affected by the S/N problem, has been covered in ref. [21] in the zero-momentum case. In ref. [31], the study of connected Wick's contraction has been extended to the vector correlator and the pseudoscalar one at non-zero momentum. They are both interesting observables, with the former playing a prominent rôle in the computation of the hadronic vacuum polarization (HVP) contribution to $(g-2)_\mu$ and the latter being a building block of transition matrix elements such as heavy meson decay form factors.

In all these cases, numerical tests in the quenched approximation show that multi-level sampling is effective in addressing the S/N problem. Here we choose to discuss only a technical detail of the implementation: The evaluation of the matrix product of quark propagator factors such as those in eqs (26) and (29) requires to handle and average independently quark propagator matrices with open indices on a region boundary. While this is avoided in the simplest implementation of the disconnected Wick's contraction case discussed in sect. 5, the lowest-order factorized approximation of a connected Wick's contraction contains a number of factorized quark line as in eq. (30), thus the matrix product of $D^{-1}_{\Omega_1} D_{\partial\Omega_1}$ and $D^{-1}_{\Omega_0}$. In the case of a decomposition in two thick-time-slice regions, the product of matrix factors local to Ω_0 and Ω_1 amounts to a contraction of size $|\partial\Omega_1|^\ell = (L/a)^{3\ell}$, where L/a is the number of lattice sites in any spatial direction and ℓ is the number of quark lines in the Wick's contraction, *e.g.* $\ell = 2$ for mesons and $\ell = 3$ for baryons. This is numerically not feasible.

A solution is to transform the matrix product in a reduced number of ordinary products, introducing a projection $P_L = \sum_{i=1}^{N_m} \phi_i \phi_i^\dagger$, where ϕ_i are N_m orthonormal vectors supported on $\partial\Omega_1$. The projection is then used to define

$$D^{-1}_{\mathrm{f},P_L}(x,y) = -\sum_{i=1}^{N_m} \left[D^{-1}_{\Omega_1} D_{\partial\Omega_1} \phi_i\right](x) \left[\phi_i^\dagger D^{-1}_{\Omega_0}\right](y), \qquad x \in \Omega_1,\, y \in \Lambda_0. \tag{41}$$

Using this quark propagator, connected Wick's contractions reduce to a sum of N_m^ℓ products. However, eq. (41) introduces a further approximation on top of eq. (30). Therefore, cutting the quark lines with the projection is a solution only if it is possible to obtain a good approximation, at the level of eq. (30) or better, with a reasonable number N_m of vectors.

In ref. [21] we showed how this is obtained for both mesonic and baryonic two-point functions with two choices of set of vectors ϕ_i: those which span the deflation subspace as defined in ref. [48], and N_m orthonormal vectors constructed by applying 10 inverse iterations of the Wilson-Dirac operator defined on a thick-time-slice region surrounding $\partial\Omega_1$.

An alternative strategy that does not involve projection on ϕ_i has recently been proposed in ref. [47] and has been successfully applied to meson two-point functions, with the extension to baryons currently under development.

7 Factorization of the quark determinant

The aim of this section is to rewrite the quark determinant in a way that makes the gauge-field dependence local to a region of the lattice. To this purpose, we first introduce a block decomposition of the Dirac operator the factorize the bulk contribution of a region to the determinant, from a small non-local term that takes into account contributions between regions. We then approach the latter term with a polynomial approximation of its inverse, which can be written in terms of multiboson fields with a gauge-field local action.

7.1 Block decomposition of the determinant

To the purpose of the derivation, we partition the lattice in two regions, Λ_a and Λ_b, that can have disconnected subregions. This corresponds to a LDU-decomposition of the Hermitian Dirac operator $Q = \gamma_5 D$ as in eq. (64a), from

which an equation between determinants follows

$$\det Q = \det Q_{\Lambda_b} \cdot \det Q/Q_{\Lambda_b}. \tag{42}$$

Λ_a is the union of the active regions of the lattice, to be updated with level-1 updates, as shown in fig. 8. Similarly, Λ_b is the union of inactive *buffer* regions, that act as a boundary between active regions and whose links are not updated by level-1 updates. Partitioning Λ_a in even and odd active regions, denoted respectively by Λ_e and Λ_o, results in a further block-decomposition of the Schur complement Q/Q_{Λ_b}

$$\begin{aligned} Q/Q_{\Lambda_b} = Q_{\Lambda_a} - Q_{\partial\Lambda_a} Q_{\Lambda_b}^{-1} Q_{\partial\Lambda_b} &= \begin{pmatrix} Q_{\Lambda_e} - Q_{\partial\Lambda_e} Q_{\Lambda_b}^{-1} Q_{\partial\Lambda_b} P_{\Lambda_e}^\top & -Q_{\partial\Lambda_e} Q_{\Lambda_b}^{-1} Q_{\partial\Lambda_b} P_{\Lambda_o}^\top \\ -Q_{\partial\Lambda_o} Q_{\Lambda_b}^{-1} Q_{\partial\Lambda_b} P_{\Lambda_e}^\top & Q_{\Lambda_o} - Q_{\partial\Lambda_o} Q_{\Lambda_b}^{-1} Q_{\partial\Lambda_b} P_{\Lambda_o}^\top \end{pmatrix} \\ &= \begin{pmatrix} Q_{\Omega_e}/Q_{\Lambda_b} & -Q_{\partial\Lambda_e} Q_{\Lambda_b}^{-1} Q_{\partial\Lambda_b} P_{\Lambda_o}^\top \\ -Q_{\partial\Lambda_o} Q_{\Lambda_b}^{-1} Q_{\partial\Lambda_b} P_{\Lambda_e}^\top & Q_{\Omega_o}/Q_{\Lambda_b} \end{pmatrix}, \end{aligned} \tag{43}$$

where in the last equation we introduced the region $\Omega_e = \Lambda_e \cup \Lambda_b$ ($\Omega_o = \Lambda_o \cup \Lambda_b$ respectively) as the union of Λ_e (Λ_o) and Λ_b, and we identified the diagonal blocks with the Schur complement of the block Q_{Λ_e} (Q_{Λ_o}) of the Dirac operator Q_{Ω_e} (Q_{Ω_o}). The off-diagonal blocks of Q/Q_{Λ_b} are suppressed w.r.t. the diagonal ones, because they can be represented by a quark line that propagates from Λ_e to Λ_o, or vice-versa, over the thickness of Λ_b. This result in a suppression according to eq. (3). To make this argument sound, we proceed by pre-conditioning the Q/Q_{Λ_b} with the inverse of its diagonal blocks, $\mathrm{diag}([Q_{\Omega_e}/Q_{\Lambda_b}]^{-1}, [Q_{\Omega_o}/Q_{\Lambda_b}]^{-1})$

$$\begin{aligned} Q/Q_{\Lambda_b} &= \begin{pmatrix} [Q_{\Omega_e}/Q_{\Lambda_b}]^{-1} & \\ & [Q_{\Omega_e}/Q_{\Lambda_b}]^{-1} \end{pmatrix}^{-1} \begin{pmatrix} \mathbb{1} & -[Q_{\Omega_e}/Q_{\Lambda_b}]^{-1} Q_{\partial\Lambda_e} Q_{\Lambda_b}^{-1} Q_{\partial\Lambda_b} P_{\Lambda_o}^\top \\ -[Q_{\Omega_e}/Q_{\Lambda_b}]^{-1} Q_{\partial\Lambda_o} Q_{\Lambda_b}^{-1} Q_{\partial\Lambda_b} P_{\Lambda_e}^\top & \mathbb{1} \end{pmatrix} \\ &= \begin{pmatrix} P_{\Lambda_e} Q_{\Omega_e}^{-1} P_{\Lambda_e}^\top & \\ & P_{\Lambda_o} Q_{\Omega_o}^{-1} P_{\Lambda_o}^\top \end{pmatrix}^{-1} \tilde{W}, \qquad \tilde{W} = \begin{pmatrix} \mathbb{1} & P_{\Lambda_e} Q_{\Omega_e}^{-1} Q_{\partial\Lambda_b} P_{\Lambda_o}^\top \\ P_{\Lambda_o} Q_{\Omega_o}^{-1} Q_{\partial\Lambda_b} P_{\Lambda_e}^\top & \mathbb{1} \end{pmatrix}, \end{aligned} \tag{44}$$

where in the last equation we used the properties of the Schur complement in eqs (66) and introduced $\tilde{W}$. We have

$$\det Q = \frac{\det \tilde{W}}{\det Q_{\Lambda_b}^{-1} \cdot \det\{P_{\Lambda_e} Q_{\Omega_e}^{-1} P_{\Lambda_e}^\top\} \cdot \det\{P_{\Lambda_o} Q_{\Omega_o}^{-1} P_{\Lambda_o}^\top\}}. \tag{45}$$

The factor $\det \tilde{W}$ can be further simplified as $\det \tilde{W}/\tilde{W}_{\Lambda_o} \cdot \det \tilde{W}_{\Lambda_o}$, where the second factor evaluates to unity and the first factor is the Schur complement

$$\tilde{W}/\tilde{W}_{\Lambda_o} = \tilde{W}_{\Lambda_e} - \tilde{W}_{\Lambda_e,o} \tilde{W}_{\Lambda_o}^{-1} \tilde{W}_{\Lambda_o,e} = \mathbb{1} - P_{\Lambda_e} Q_{\Omega_e}^{-1} Q_{\partial\Lambda_b} P_{\Lambda_o}^\top P_{\Lambda_o} Q_{\Omega_o}^{-1} Q_{\partial\Lambda_b} P_{\Lambda_e}^\top. \tag{46}$$

In the last equation, because of the action of $Q_{\partial\Lambda_b}$, the projection P_{Λ_o} is equivalent to the projection $P_{\partial\Lambda_o}$ on the boundary of Λ_o, as defined in appendix A. Similarly, using the Schur complement of the block Λ_e, the determinant is not modified by the substitution of P_{Λ_e} with $P_{\partial\Lambda_e}$, which leads to the equation

$$\det \tilde{W} = \det\{\mathbb{1} - w\}, \qquad w = P_{\partial\Lambda_e} Q_{\Omega_e}^{-1} Q_{\partial\Lambda_b} P_{\partial\Lambda_o} Q_{\Omega_o}^{-1} Q_{\partial\Lambda_b}, P_{\partial\Lambda_e}^\top \tag{47}$$

where we introduced the operator w that is supported on the internal boundaries of Λ_e only. The determinant of the Hermitian Dirac operator can thus be rewritten as

$$\det Q = \frac{\det\{\mathbb{1} - w\}}{\det Q_{\Lambda_b}^{-1} \cdot \det\{P_{\Lambda_e} Q_{\Omega_e}^{-1} P_{\Lambda_e}^\top\} \cdot \det\{P_{\Lambda_o} Q_{\Omega_o}^{-1} P_{\Lambda_o}^\top\}}. \tag{48}$$

Neither $\tilde{W}$ nor w are Hermitian operators. However, all the other factors in eq. (48) are real because they are determinants of Hermitian operators, thus also $\tilde{W}$ and $(\mathbb{1} - w)$ have a real determinant.

7.2 Comparison with DD-HMC algorithm

The factorization of $\det Q$ obtained in the previous section bears similarities with the factorization introduced by the Schwarz-preconditioned HMC algorithm, also known as DD-HMC algorithm [49]. This is evident rewriting eq. (48) as

$$\det Q = \det Q_{\Lambda_b} \cdot \det Q_{\Omega_e}/Q_{\Lambda_b} \cdot \det Q_{\Omega_o}/Q_{\Lambda_b} \cdot \det\{\mathbb{1} - P_{\partial\Lambda_e} Q_{\Omega_e}^{-1} Q_{\partial\Lambda_b} P_{\partial\Lambda_o} Q_{\Omega_o}^{-1} Q_{\partial\Lambda_b} P_{\partial\Lambda_e}^\top\}, \tag{49}$$

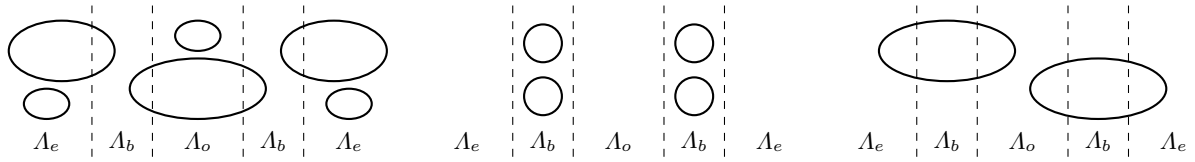

Figure 9. Pictorial representation of the contribution to the factorization of the quark determinant in eq. (48). The leftmost picture represent quark loops that contributes to either $\det\{P_{\Lambda_e} Q^{-1}_{\Omega_e} P^\top_{\Lambda_e}\}$ or $\det\{P_{\Lambda_o} Q^{-1}_{\Omega_o} P^\top_{\Lambda_o}\}$, *i.e.* quark loops that span Ω_e or Ω_o. The central picture shows quark loops limited to the buffer region Λ_b only, that contribute to $\det Q^{-1}_{\Omega_b}$. Finally, the rightmost picture depicts quark loops that extends both on Λ_e and Λ_o across Λ_b, and are the residual contribution to the factorized determinant encoded in $\det\{\mathbb{1} - w\}$.

and comparing it with eq. (2.2) of ref. [49]

$$\det Q = \det Q_{\Lambda_e} \cdot \det Q_{\Lambda_o} \cdot \det\left\{\mathbb{1} - P_{\partial\Lambda_e} Q^{-1}_{\Lambda_e} Q_{\partial\Lambda_e} P_{\partial\Lambda_o} Q^{-1}_{\Lambda_o} Q_{\partial\Lambda_o}\right\}. \tag{50}$$

Namely, the latter expression is obtained from the former in the limit of an empty buffer region Λ_b. The crucial difference between the two derivations is that the one in sect. 7.1 keeps an explicit dependence on the separation between Λ_e and Λ_o provided by Λ_b. As we show in the following, the size or "thickness" of Λ_b can be used to control the smallness of the last factor in eq. (48). Conversely, there is no separation between Λ_e and Λ_o in the DD-HMC algorithm and thus no reason to expect the last factor in eq. (50) to be small. In addition, the factorization in eq. (48) can be shown to be equivalent to the one resulting from the overlapping Schwarz alternating procedure (SAP) preconditioning of the Dirac operator [50, 38], instead of the non-overlapping SAP preconditioning of ref. [49].

7.3 Locality of the factorized determinant

The expression in eq. (48) is a product of factors that, with the exclusion of $\det\{\mathbb{1} - w\}$, have a localized gauge field dependence. Specifically, $\det\{P_{\Lambda_e} Q^{-1}_{\Omega_e} P^\top_{\Lambda_e}\}$ depends only on the gauge field in Ω_o, while $\det\{P_{\Lambda_o} Q^{-1}_{\Omega_o} P^\top_{\Lambda_o}\}$ depends only on the gauge field in Ω_o, and $\det Q^{-1}_{\Omega_b}$ depends only on the gauge field in Λ_b. This suggests that eq. (48) is a step towards an independent multi-level sampling of the gauge field in Λ_e and Λ_o. Indeed, a change in the gauge field in Λ_e contributes to a change of the factor $\det\{P_{\Lambda_e} Q^{-1}_{\Omega_e} P^\top_{\Lambda_e}\}$ and $\det\{\mathbb{1} - w\}$ only, with the former including the bulk contribution from within region Ω_e, and the latter including the long-range contribution that extends to Λ_o. An equivalent statement holds for a change in the gauge field in Λ_o.

Choosing a decomposition of the lattice in thick time slices regions as in fig. 8, the different contributions to eq. (48) are represented pictorially in fig. 9 in terms of different quark-loop contributions to the determinant.

It is clear that, if the long-range contribution from $\det\{\mathbb{1} - w\}$ is ignored, the gauge-field dependence is completely factorized. Thus, we further study the operator w and its contribution to the determinant.

7.4 Properties of w

The form of the operator w defined in eq. (47) allows to estimate the contribution of the determinant factor $\det\{\mathbb{1} - w\}$. First, we rewrite the two factors in its definition in the equivalent form

$$P_{\partial\Lambda_e} Q^{-1}_{\Omega_e} Q_{\partial\Lambda_b} P^\top_{\partial\Lambda_o} = P_{\partial\Lambda_e} Q^{-1} \left[Q_{\partial\Lambda_b} + Q_{\partial\Lambda_o} Q^{-1}_{\Omega_e} Q_{\partial\Lambda_b}\right] P^\top_{\partial\Lambda_o}, \tag{51a}$$

$$P_{\partial\Lambda_o} Q^{-1}_{\Omega_o} Q_{\partial\Lambda_b} P^\top_{\partial\Lambda_e} = P_{\partial\Lambda_o} Q^{-1} \left[Q_{\partial\Lambda_b} + Q_{\partial\Lambda_e} Q^{-1}_{\Omega_o} Q_{\partial\Lambda_b}\right] P^\top_{\partial\Lambda_e}. \tag{51b}$$

Therefore, w contains two factors of the propagator Q^{-1} between the boundaries of the Λ_e and Λ_o regions. According to eq. (3), on every representative gauge configuration the operator gets a suppression $[\|Q^{-1}\|\,(\Delta)]^2$, where Δ is the "thickness" of region Λ_b. If Δ is large enough, the suppression approaches the asymptotic exponential rate $\bar{\delta} = \mathrm{e}^{-M_\pi \Delta}$. Moreover, the matrix w can be written as the product of two Hermitian matrices,

$$w = \left[P_{\partial\Lambda_e} Q^{-1}_{\Omega_e} P^\top_{\partial\Lambda_e}\right] \left[Q_{\partial\Lambda_e} Q^{-1}_{\Lambda_b} Q_{\partial\Lambda_b} Q^{-1}_{\Omega_o} Q_{\partial\Lambda_o} Q^{-1}_{\Lambda_b} Q_{\partial\Lambda_b} P^\top_{\partial\Lambda_e}\right], \tag{52}$$

acting on the boundary of region Λ_e. In turns, this implies that w is similar to $w^\dagger$ [51], thus the characteristic polynomial of w has real coefficients, the complex eigenvalues δ_i come in conjugate pairs and the spectrum of w is symmetric with respect to the real axis. This confirms that $\det\{\mathbb{1} - w\}$ is real, as anticipated.

In ref. [22], we tested these properties on a set of 200 configurations generated with Wilson's plaquette action and two flavour of non-perturbatively $\mathcal{O}(a)$-improved Wilson quarks using the openQCD [40, 41] package. The parameters

Table 1. Properties of the spectrum of w studied on 200 configurations for different values of Δ. f_{Re} is the fraction of configurations for which the eigenvalue with the largest magnitude is real.

Δ/a	$\bar{\delta}$	$\langle \max_i \lvert\delta_i\rvert \rangle$	$\sigma(\max_i \lvert\delta_i\rvert)$	$\max \max_i \lvert\delta_i\rvert$	$f_{\mathrm{Re}}(\%)$
8	0.3273	0.2886	0.0616	0.5130	48.5
12	0.1710	0.1692	0.0453	0.3193	46.5
16	0.1072	0.0951	0.0284	0.1977	45.5

are $\beta = 5.3$, $c_{\mathrm{SW}} = 1.909\,52$, $c_{\mathrm{F}} = c'_{\mathrm{F}} = 1$, $\kappa = 0.136\,25$ $T \times L^3 = 64a \times 32^3 a^3$ and OBCs in time, that correspond to a lattice spacing of $a = 0.0652(6)$ fm and a pion mass of $M_\pi = 1454(5) = 440(5)$ MeV [52]. We setup a partition in three thick time slices Λ_e, Λ_o and Λ_b and we compute with the Arnoldi iteration the 60 approximate eigenvalues δ_i of w with the largest magnitude, for three different choices $\Delta = 8a$, $12a$ and $16a$ of the thickness of Λ_b. In the central plot in fig. 10 we show the computed eigenvalues for the $\Delta = 8a$ case. As expected, the eigenvalues are either real or come in complex conjugate pairs. The eigenvalue with largest magnitude for each configuration is highlighted with a grey dot and it can be either real or a pair with opposite imaginary part, with both possibilities being common. The right plot in fig. 10 shows the distribution of the absolute values of the eigenvalues, with also the eigenvalue with the largest magnitude for each configuration in grey. In both plots the blue circle or line indicates the absolute value $\bar{\delta}$. Results for also $\Delta = 12a$ and $16a$ is reported in table 1.

A clear message emerges form these data: the largest eigenvalue of w deceases proportionally to $\bar{\delta}$, with a $\mathcal{O}(1)$ prefactor. This implies that $(\mathbb{1} - w)$ as a large spectral gap if Δ is properly tuned.

7.5 Polynomial approximation

In this section, we estimate the residual global factor $\det\{\mathbb{1} - w\}$ in Eq (48) using a variant of the multiboson algorithm [53–56] to $\det\{\mathbb{1} - w\}$, *i.e.* we use a polynomial approximation of the inverse of $(\mathbb{1} - w)$ to obtain a local formulation of the determinant. In Lüscher's original multiboson proposal [53] the polynomial approximation of the inverse is applied to the Dirac operator squared Q^2, whose condition number roughly proportional to a^{-2} and generally rather large. This drives up significantly the order of the polynomial to obtain a faithful approximation of the inverse over the whole spectral range of Q^2, especially on fine lattices. In turn, a large number of multiboson fields has been shown to induce large MC autocorrelation times that limit the applicability of the algorithm [57].

Conversely, the condition number of $(\mathbb{1} - w)$ is $\epsilon \sim (1 + \mathrm{e}^{-M_\pi \Delta})/(1 - \mathrm{e}^{-M_\pi \Delta})$, which is $\mathcal{O}(1)$. This suggests that a low-order polynomial might be sufficient to cover the spectral range. However, w has complex eigenvalue and the polynomial must be chosen to faithfully represent those too [54]. Formally applying the polynomial approximation $P_N(z)$ in eq. (67) from appendix B to the $(\mathbb{1} - w)$ matrix, it results

$$\frac{\det\{\mathbb{1} - R_{N+1}(\mathbb{1} - w)\}}{\det\{\mathbb{1} - w\}} = \det\{P_N(\mathbb{1} - w)\} = c_N \prod_{k=1}^{N} \det\{\mathbb{1} - z_k - w\}, \tag{53}$$

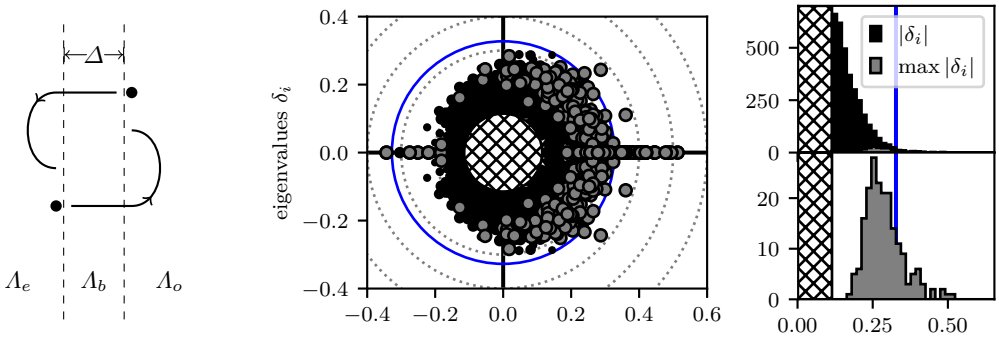

Figure 10. The picture on the left is a pictorial representation of the two factor contribution to the operator w in eq. (47). The central plot shows on the complex plane the eigenvalues δ_i of w for 200 configuration, while the distribution of their absolute values in plotted on the right. In both plots, $\Delta = 8a$, the blue circle and line denote $\bar{\delta}$, the eigenvalue (or the complex conjugate pair) with the largest magnitude for each configuration is plotted in grey, and the bulk of eigenvalues with magnitude $|\delta_i| \leq 0.35\bar{\delta}$ is hidden under the hatched region.

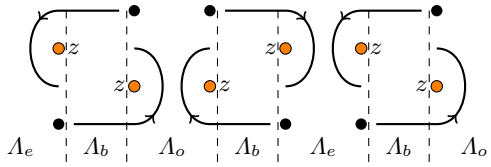

Figure 11. Pictorial representation of the terms contributing to the multiboson action $|W_{\sqrt{1-z_k}}\chi_k|^2$ defined in eq. (61).

where z_k are the roots of $P_N(z)$ in eq. (68) and c_N is a known constant. Requiring N to be even and the roots to be ordered such that $z_{N-k} = \bar{z}_k$, we combine the complex-conjugate roots in manifestly-positive terms

$$\det\{\mathbb{1} - z_{N-i} - w\}\det\{\mathbb{1} - z_k - w\} = \det\big\{(\mathbb{1} - z_k - w)^\dagger(\mathbb{1} - z_k - w)\big\} = \det\Big\{W^\dagger_{\sqrt{1-z_k}}W_{\sqrt{1-z_k}}\Big\}, \tag{54}$$

where we used that w is similar to $w^\dagger$ and, applying the same reasoning used to derive eq. (47), we introduced

$$W_y = \begin{pmatrix} y\mathbb{1} & P_{\partial\Lambda_e}Q^{-1}_{\Omega_e}Q_{\partial\Lambda_b}P^\top_{\partial\Lambda_o} \\ P_{\partial\Lambda_o}Q^{-1}_{\Omega_o}Q_{\partial\Lambda_b}P^\top_{\partial\Lambda_e} & y\mathbb{1} \end{pmatrix}, \tag{55}$$

that, differently from $\tilde{W}$, lives only on the internal boundaries of regions Λ_e and Λ_o. Combining eqs (48), (53) and (54) one arrives at the final expression of the factorized quark determinant

$$\frac{\det Q}{\det\{\mathbb{1} - R_{N+1}(\mathbb{1} - w)\}} = \frac{c_N^{-1}\prod_{k=1}^{N/2}\det\Big\{W^\dagger_{\sqrt{1-z_k}}W_{\sqrt{1-z_k}}\Big\}^{-1}}{\det Q^{-1}_{\Lambda_b}\cdot\det\big\{P_{\Lambda_e}Q^{-1}_{\Omega_e}P^\top_{\Lambda_e}\big\}\cdot\det\big\{P_{\Lambda_o}Q^{-1}_{\Omega_o}P^\top_{\Lambda_o}\big\}}. \tag{56}$$

Given a precision requirement, the order of the polynomial is estimated using eq. (69) from appendix B. It holds

$$\|\mathbb{1} - (\mathbb{1} - w)P_N(\mathbb{1} - w)\| = \|R_{N+1}(\mathbb{1} - w)\| \leq \max_i |\delta_i|^{N+1} = |\delta|^{N+1}_{\max}, \tag{57}$$

where the matrix norm is $\|A\| = \sup\{\|Av\| : \|v\| = 1\}$. Then, eq. (53) implies

$$\det\{\mathbb{1} - w\}\det\{P_N(\mathbb{1} - w)\} = \det\{\mathbb{1} - R_{N+1}(\mathbb{1} - w)\} = 1 + \mathrm{tr}\{R_{N+1}(\mathbb{1} - w)\} + \ldots, \tag{58}$$

assuming that $|\delta|^{N+1}_{\max} \ll 1$ and $\mathrm{tr}\{R_{N+1}(\mathbb{1} - w)\} \ll 1$. At first order in the expansion, the relative error on the determinant is therefore

$$|\mathrm{tr}\{R_{N+1}(\mathbb{1} - w)\}| \leq \sum_i |\delta_i|^{N+1} \leq \sum_{i=1}^{N_{\mathrm{ev}}} |\delta_i|^{N+1} + (6L^3/a^3 - N_{\mathrm{ev}})|\delta_{N_{\mathrm{ev}}+1}|^{N+1}, \tag{59}$$

where in the last equality the contribution from the N_{ev} eigenvalues with the largest magnitude, *e.g.* the ones computed in sect. 7.4, has been treated separately form the bulk of the modes, which satisfy $|\delta_i| \ll \bar{\delta}$.

Given the distribution of eigenvalues of w that emerges from the numerical study in sect. 7.4, the circle centred in one with radius one is a natural choice for the polynomial approximating $(\mathbb{1} - w)^{-1}$. However, the approximation could be optimized using an ellipse also centred in one and tuning the focal distance.

8 The multiboson action

The determinant factors in eq. (56) are represented by scalar fields [15]. Working with $N_{\mathrm{f}} = 2$ flavours of mass-degenerate quarks, we represent $\det Q^2$ as

$$\begin{aligned}\frac{\det Q^2}{\det\{\mathbb{1} - R_{N+1}(\mathbb{1} - w)\}^2} = C\int\mathcal{D}[\phi_{\Lambda_e},\phi^\dagger_{\Lambda_e}]\,\mathrm{e}^{-\left|P_{\Lambda_e}Q^{-1}_{\Omega_e}P^\top_{\Lambda_e}\phi_{\Lambda_e}\right|^2}\cdot\int\mathcal{D}[\phi_{\Lambda_o},\phi^\dagger_{\Lambda_o}]\,\mathrm{e}^{-\left|P_{\Lambda_o}Q^{-1}_{\Omega_o}P^\top_{\Lambda_o}\phi_{\Lambda_o}\right|^2} \\ \cdot\int\mathcal{D}[\phi_{\Lambda_b},\phi^\dagger_{\Lambda_b}]\,\mathrm{e}^{-\left|Q^{-1}_{\Lambda_b}\phi_{\Lambda_b}\right|^2}\cdot\prod_{k=1}^{N}\int\mathcal{D}[\chi_k,\chi^\dagger_k]\,\mathrm{e}^{-\left|W_{\sqrt{1-z_k}}\chi_k\right|^2},\end{aligned} \tag{60}$$

where C is an irrelevant numerical constant. The three bulk contributions to the determinant are estimated with standard techniques using pseudofermion scalar fields ϕ_{Λ_i} that are supported on Λ_i. The long-range contribution encoded in the $N_{\rm f} \cdot N/2$ terms $\det\{W_y^\dagger W_y\}$ is instead represented by N *multiboson* scalar fields χ_k that live on $\partial\Lambda_b$. The action for multiboson field is

$$\left|W_{\sqrt{1-z_k}}\chi_k\right|^2 = \sum_{i=\{e,o\},j\neq i} \left|z\chi_{\Lambda_i,k} + P_{\partial\Lambda_i} Q_{\Omega_i}^{-1} Q_{\Lambda_b}\chi_{\Lambda_j,k}\right|^2, \qquad \chi_{\Lambda,k} = P_{\partial\Lambda}\chi_k. \tag{61}$$

The dependence of this action from the gauge field in Λ_e and Λ_o is thus factorized, and the HMC forces of links in different regions are independent. Moreover, the terms that contribute to the force of Λ_e always start on the inner boundary of Λ_o, and *vice versa*. Thus, the forces of the multiboson action are also suppressed exponentially in the thickness Δ of Λ_b. In fig. 11 we represent graphically the terms of the sum in eq. (61) in the case of a partition of the lattice in thick time slices.

We point out that, while we choose for simplicity to represent the determinant of a quark doublet, the multiboson representation of $\det\{\mathbb{1} - w\}$ has been derived for a single quark flavour. Therefore, the factorization in eq. (56) is suitable of the simulation of an arbitrary number of quark flavours with non-degenerate masses, employing $N_{\rm f} \cdot N/2$ multiboson fields and the standard RHMC technique for the bulk determinant contributions. Similarly, as an improvement over using a single pseudofermion for each active region as in eq. (61), techniques such as the Hasenbusch mass splitting [58, 59] and multiple time-step integration [60] can be used.

8.1 Multi-level sampling and reweighting

The expectation value of a generic field O can be written as

$$\langle O\rangle = \frac{\langle O\mathcal{W}_N\rangle_N}{\langle\mathcal{W}_N\rangle_N} = \frac{\langle O^{\rm f}\rangle_N}{\langle\mathcal{W}_N\rangle_N} + \frac{\langle O^{\rm f} - O\mathcal{W}_N\rangle_N}{\langle\mathcal{W}_N\rangle_N}, \qquad \mathcal{W}_N = \det\{\mathbb{1} - R_{N+1}(\mathbb{1} - w)\}^{N_{\rm f}} \tag{62}$$

where $O^{\rm f}$ is a factorized approximation of O, such as the factorized disconnected Wick's contraction introduced in sect. 5, and $\langle\bullet\rangle_N$ denotes the expectation value in the theory defined by the multiboson action at finite N. Since both the action and the observable are factorized, the expectation value $\langle O^{\rm f}\rangle_N$ can be computed with a multi-level algorithm by generating gauge field configurations with the multiboson action at finite N. The reweighting factor $\mathcal{W}_N$ is easily evaluated with, in the $N_{\rm f} = 2$ case,

$$\mathcal{W}_N = \frac{\int \mathcal{D}[\eta,\eta^\dagger]\, \mathrm{e}^{-\left|[\mathbb{1}-R_{N+1}(\mathbb{1}-w)]^{-1}\eta\right|^2}}{\int \mathcal{D}[\eta,\eta^\dagger]\, \mathrm{e}^{-|\eta|^2}}. \tag{63}$$

8.2 Implementation and numerical tests

The HMC of a two-active-region version of eq. (60) has been implemented in ref. [22]. We refer to that publication for the technical details of the implementation, based on an heavily modified version of the openQCD [40, 41] package, such as how to perform the heatbath of the multiboson fields at the beginning of the the level-1 MC chain. A set of level-1 configurations has been generated using the parameters in sect. 7.4 and updating independently two thick time slices separated by Λ_b of thickness $\Delta = 12a \approx 0.8\,\mathrm{fm}$, corresponding to $M_\pi\Delta \approx 1.7$. We simulated eq. (60) using 5 mass-preconditioned pseudofermions for each bulk term and 12 multiboson fields. The latter are integrated on the outermost of three time-step integration levels and result in a small overhead on the computational cost of a standard HMC simulating the bulk action only. Numerical tests have been performed for correlators of the gluonic observables such as the YM energy and topological charge densities, and for the factorized disconnected Wick's contraction studied in sect. 5, although with limited statistics. In all cases, the results are in line with expectations: When two local fields are in different active regions and a two-level estimator is used, its variance scales with n_1^{-2}. Thus, the two-level MC works at full capacity, resulting in a net gain in the S/N. Moreover, no particular freezing of the links is observed.

9 Conclusions

We described all the steps necessary to implement a multi-level MC integration scheme in lattice theories with fermionic content. The proposed method relies on the locality of the lattice-discretized Dirac operator, and on the configuration by configuration exponential decrease of its inverse with the distance between the sink and the source. It works by

decomposing both the fermion propagator and the fermion determinant in factors that have a gauge field dependence local to distinct spacetime regions. The propagator factorization is applied to both connected and disconnected Wick's contractions of correlators. The determinant factorization is a combination of a decomposition in overlapping domains, and the multiboson representation of the small interaction among the gauge fields of distant regions. The resulting action is local in the gauge field and can be simulated by variant of the standard HMC algorithm.

We reported the application of the method to the computation of the disconnected contribution to the correlator of two flavour-singlet pseudoscalar densities in quenched QCD. This numerical test implements a very simple two-region two-level setup and shows a clear gain in the S/N, with the number of configuration needed to reach a given statical precision being proportional to the square root of these required in the standard case.

The generalization of multi-level sampling to lattice QCD opens new perspective. Many computations that are affected by the exponential S/N degradation are expected to profit from these improvement, with prime examples being correlators with disconnected contributions and baryonic multi-point functions.

Acknowledgements

The results presented in this article have been obtained in collaboration with Leonardo Giusti and Stefan Schaefer, and have originally been published in refs [21, 22, 31]. I am especially indebted to Leonardo Giusti for his mentoring throughout my PhD. I thank Tim Harris for a critical reading of the manuscript.

A Two-region block decomposition

In this section we collect elementary formulae about the block-decomposition of lattice-sized matrices, such as the lattice Dirac operator D, which includes the mass term, or its Hermitian version Q. Consider a region of the lattice $\Lambda_0 \in \Omega$ and its complementary $\Lambda_1 = \Omega \setminus \Lambda_0$. Following the notation of ref. [37], we define the *exterior boundary* $\partial\Lambda_i$ of Λ_i. Properly ordering the lattice sites so that all those in Λ_0 come before the ones in Λ_1, we write the matrix Q in its block form and perform a LDU decomposition of the blocks

$$Q = \begin{pmatrix} Q_{\Lambda_0} & Q_{\partial\Lambda_0} \\ Q_{\partial\Lambda_1} & Q_{\Lambda_1} \end{pmatrix} = \begin{pmatrix} \mathbb{1} & \\ Q_{\partial\Lambda_1} Q_{\Lambda_0}^{-1} & \mathbb{1} \end{pmatrix} \begin{pmatrix} Q_{\Lambda_0} & \\ & Q/Q_{\Lambda_0} \end{pmatrix} \begin{pmatrix} \mathbb{1} & Q_{\Lambda_0}^{-1} Q_{\partial\Lambda_0} \\ & \mathbb{1} \end{pmatrix}, \tag{64a}$$

where $Q/Q_{\Lambda_0} = Q_{\Lambda_1} - Q_{\partial\Lambda_1} Q_{\Lambda_0}^{-1} Q_{\partial\Lambda_0}$ is the Schur complement of the block Q_{Λ_0} of Q. This leads to the following block decomposition for the inverse of Q

$$Q^{-1} = \begin{pmatrix} Q_{\Lambda_0}^{-1} + Q_{\Lambda_0}^{-1} Q_{\partial\Lambda_0} [Q/Q_{\Lambda_0}]^{-1} Q_{\partial\Lambda_1} Q_{\Lambda_0}^{-1} & -Q_{\Lambda_0}^{-1} Q_{\partial\Lambda_0} [Q/Q_{\Lambda_0}]^{-1} \\ -[Q/Q_{\Lambda_0}]^{-1} Q_{\partial\Lambda_1} Q_{\Lambda_0}^{-1} & [Q/Q_{\Lambda_0}]^{-1} \end{pmatrix}. \tag{65a}$$

Equivalently, swapping the rôle of Λ_0 and Λ_1,

$$Q = \begin{pmatrix} \mathbb{1} & Q_{\partial\Lambda_0} Q_{\Lambda_1}^{-1} \\ & \mathbb{1} \end{pmatrix} \begin{pmatrix} Q/Q_{\Lambda_1} & \\ & Q_{\Lambda_1} \end{pmatrix} \begin{pmatrix} \mathbb{1} & \\ Q_{\Lambda_1}^{-1} Q_{\partial\Lambda_1} & \mathbb{1} \end{pmatrix}, \tag{64b}$$

where $Q/Q_{\Lambda_1} = Q_{\Lambda_0} - Q_{\partial\Lambda_0} Q_{\Lambda_1}^{-1} Q_{\partial\Lambda_1}$ and the same block decomposition of the inverse of Q can be written as

$$Q^{-1} = \begin{pmatrix} [Q/Q_{\Lambda_1}]^{-1} & -[Q/Q_{\Lambda_1}]^{-1} Q_{\partial\Lambda_0} Q_{\Lambda_1}^{-1} \\ -Q_{\Lambda_1}^{-1} Q_{\partial\Lambda_1} [Q/Q_{\Lambda_1}]^{-1} & Q_{\Lambda_1}^{-1} + Q_{\Lambda_1}^{-1} Q_{\partial\Lambda_1} [Q/Q_{\Lambda_1}]^{-1} Q_{\partial\Lambda_0} Q_{\Lambda_1}^{-1} \end{pmatrix}. \tag{65b}$$

The inverse of the Schur complement appears in various blocks of the inverse of Q. In particular,

$$[Q/Q_{\Lambda_0}]^{-1} = P_{\Lambda_1} Q^{-1} P_{\Lambda_1}^{\top}, \tag{66a}$$

$$[Q/Q_{\Lambda_1}]^{-1} = P_{\Lambda_0} Q^{-1} P_{\Lambda_0}^{\top}, \tag{66b}$$

$$-Q_{\Lambda_0}^{-1} Q_{\partial\Lambda_0} [Q/Q_{\Lambda_0}]^{-1} = -[Q/Q_{\Lambda_1}]^{-1} Q_{\partial\Lambda_0} Q_{\Lambda_1}^{-1} = P_{\Lambda_0} Q^{-1} P_{\Lambda_1}^{\top}, \tag{66c}$$

$$-[Q/Q_{\Lambda_0}]^{-1} Q_{\partial\Lambda_1} Q_{\Lambda_0}^{-1} = -Q_{\Lambda_1}^{-1} Q_{\partial\Lambda_1} [Q/Q_{\Lambda_1}]^{-1} = P_{\Lambda_1} Q^{-1} P_{\Lambda_0}^{\top}, \tag{66d}$$

where P_{Λ_i} are projection matrices of the right shape so that $[P_{\Lambda_i}\psi](x)$ is defined to be equal to $\psi(x)$ only for $x \in \Lambda_i$.

A.1 Boundaries

Discretizations of the Dirac operator such as the (improved) Wilson-Dirac operator are *ultra-local*, *i.e.* D is a sparse matrix with elements connecting only nearest-neighbour sites, or possibly a finite stencil of near sites. We use this property to define of the *exterior boundary* $\partial\Lambda_i$ of Λ_i as the set of lattice sites in the complementary of Λ_i that are in the image of D acting on Λ_i. With this definition, the off-diagonal blocks $Q_{\partial\Lambda_i}$ in the decomposition in eq. (64a) acts on $\partial\Lambda_i$ and their image is in Λ_i. Projection matrices $P_{\partial\Lambda_i}$ are defined accordingly. For instance, in the case of the (improved) Wilson-Dirac operator that connects only nearest-neighbour sites, the exterior boundary $\partial\Lambda_i$ is aptly chosen as the set of points that are at distance 1 from Λ_i.

B Polynomial approximation of the inverse

The Chebyshev polynomials offer an asymptotically-optimal polynomial approximation of the multiplicative inverse $1/z$ when $z \in \mathbb{C} \setminus \{0\}$ is within an ellipse that does not contain the origin [61, 62]. When z is contained in an ellipse centred at a distance d from the origin on the positive real axis, with a major and minor radii a and b respectively and with focal distance $c = \sqrt{a^2 - b^2}$, the polynomial approximation of $1/z$ of order N is

$$P_N(z) = \frac{1 - R_{N+1}(z)}{z} = c_N \prod_{k=1}^{N} (z - z_k), \qquad R_{N+1}(z) = \frac{T_{N+1}\left(\frac{d-z}{c}\right)}{T_{N+1}\left(\frac{d}{c}\right)}, \tag{67}$$

where $R_{N+1}(z)$ is a polynomial of degree $N+1$ and $T_k(z)$ are the Chebyshev polynomial of the first kind of degree k. The N roots z_k of $P_N(z)$ are obtained by requiring that $R_{N+1}(z_k) = 1$ and $z_k \neq 0$, and they are given by

$$z_k = d\left(1 - \cos\frac{2\pi k}{N+1}\right) - \mathrm{i}\sqrt{d^2 - c^2}\sin\frac{2\pi k}{N+1}, \qquad k = 1, \dots, N. \tag{68}$$

They lie on the ellipse in the complex plane with centre d, foci $d \pm c$, and which passes through zero. By using the definition of the Chebyshev polynomials, a uniform error bound on the approximation is given by

$$|1 - zP_N(z)| = |R_{N+1}(z)| \leq \left(\frac{a + \sqrt{a^2 - c^2}}{d + \sqrt{d^2 - c^2}}\right)^{N+1} \left\{1 + \left[\frac{a}{c} + \sqrt{\frac{a^2}{c^2} - 1}\right]^{-2N-2}\right\}. \tag{69}$$

In the limit $c \to 0$ the ellipse becomes a circle centred in d with radius $a = b$, it holds

$$R_{N+1}(z) = \left(\frac{d - z}{d}\right)^{N+1}, \qquad |1 - zP_N(z)| = |R_{N+1}(z)| \leq \left(\frac{a}{d}\right)^{N+1}. \tag{70}$$

Zarantonello's lemma [62, Lemma 6.26] guarantees that the polynomial is optimal in this case. The roots of $P_N(z)$ are given by eq. (68) and they lie on a circle centred in d of radius d.

References

1. S. Durr et al., Science **322**, 1224–1227 (2008).
2. S. Borsanyi et al., Science **347**, 1452–1455 (2015).
3. D. Mohler, S. Schaefer, and J. Simeth, EPJ Web Conf. **175**, 02010 (2018).
4. G. Parisi, Phys. Rep. **103**, 203–211 (1984).
5. G. P. Lepage, in Boulder ASI 1989 (1989), pp. 97–120.
6. H. B. Meyer and H. Wittig, Prog. Part. Nucl. Phys. **104**, 46–96 (2019).
7. O. Bär, Phys. Rev. **94**, 054505 (2016).
8. G. Parisi, R. Petronzio, and F. Rapuano, Phys. Lett. B **128**, 418–420 (1983).
9. M. Lüscher and P. Weisz, JHEP **0109**, 010 (2001).
10. H. B. Meyer, JHEP **0301**, 048–048 (2003).
11. M. Della Morte and L. Giusti, Comput. Phys. Commun. **180**, 813–818 (2009).
12. M. Della Morte and L. Giusti, Comput. Phys. Commun. **180**, 819–826 (2009).
13. M. Della Morte and L. Giusti, JHEP **1105**, 056 (2011).
14. M. García Vera and S. Schaefer, Phys. Rev. D **93**, 074502 (2016).

15. D. H. Weingarten and D. N. Petcher, Phys. Lett. B **99**, 333–338 (1981).
16. M. Hasenbusch, Phys. Rev. D **59**, 054505 (1999).
17. F. Knechtli and U. Wolff, Nucl. Phys. B **663**, 3–32 (2003).
18. A. Hasenfratz, P. Hasenfratz, and F. Niedermayer, Phys. Rev. D **72**, 114508 (2005).
19. J. Finkenrath, F. Knechtli, and B. Leder, Comput. Phys. Commun. **184**, 1522–1534 (2013).
20. S. Duane, A. Kennedy, B. J. Pendleton, and D. Roweth, Phys. Lett. B **195**, 216–222 (1987).
21. M. Cè, L. Giusti, and S. Schaefer, Phys. Rev. D **93**, 094507 (2016).
22. M. Cè, L. Giusti, and S. Schaefer, Phys. Rev. D **95**, 034503 (2017).
23. M. Cè, "Solving the $U_A(1)$ problem of QCD: new computational strategies and results", PhD thesis (Scuola Normale Superiore di Pisa, 2017).
24. M. Cè, C. Consonni, G. P. Engel, and L. Giusti, Phys. Rev. D **92**, 074502 (2015).
25. M. Cè, M. García Vera, L. Giusti, and S. Schaefer, Phys. Lett. B **762**, 232–236 (2016).
26. E. Witten, Nucl. Phys. B **156**, 269–283 (1979).
27. G. Veneziano, Nucl. Phys. B **159**, 213–224 (1979).
28. K. Cichy, E. Garcia-Ramos, K. Jansen, K. Ottnad, and C. Urbach, JHEP **1509**, 020 (2015).
29. P. Dimopoulos et al., (2018), arXiv:`1812.08787 [hep-lat]`.
30. M. Cè, L. Giusti, and S. Schaefer, PoS **LATTICE2016**, 263 (2017).
31. M. Cè, L. Giusti, and S. Schaefer, EPJ Web Conf. **175**, 11005 (2018).
32. L. Giusti, P. Hernandez, M. Laine, P. Weisz, and H. Wittig, JHEP **0404**, 013 (2004).
33. T. DeGrand and S. Schaefer, Comput. Phys. Commun. **159**, 185–191 (2004).
34. S. Collins, G. Bali, and A. Schäfer, PoS **LATTICE2007**, 141 (2007).
35. G. S. Bali, S. Collins, and A. Schäfer, Comput. Phys. Commun. **181**, 1570–1583 (2010).
36. T. Blum, T. Izubuchi, and E. Shintani, Phys. Rev. D **88**, 094503 (2013).
37. M. Lüscher, Comput. Phys. Commun. **156**, 209–220 (2004).
38. L. Giusti, M. Cè, and S. Schaefer, EPJ Web Conf. **175**, 01003 (2018).
39. M. Guagnelli, R. Sommer, and H. Wittig, Nucl. Phys. B **535**, 389–402 (1998).
40. M. Lüscher and S. Schaefer, `https://luscher.web.cern.ch/luscher/openQCD/`.
41. M. Lüscher and S. Schaefer, Comput. Phys. Commun. **184**, 519–528 (2013).
42. C. R. Allton, V. Giménez, L. Giusti, and F. Rapuano, Nucl. Phys. B **489**, 427–452 (1997).
43. C. Thron, S. J. Dong, K. F. Liu, and H. P. Ying, Phys. Rev. D **57**, 1642–1653 (1998).
44. C. McNeile and C. Michael, Phys. Rev. D **63**, 114503 (2001).
45. R. Sommer, Nucl. Phys. B Proc. Suppl. **42**, 186–193 (1995).
46. M. Foster and C. Michael, Phys. Rev. D **59**, 074503 (1999).
47. L. Giusti, T. Harris, A. Nada, and S. Schaefer, in 36th International Symposium on Lattice Field Theory (Lattice 2018) East Lansing, MI, United States, July 22-28, 2018, Vol. LATTICE2018 (2018), p. 028.
48. M. Lüscher, JHEP **0707**, 081–081 (2007).
49. M. Lüscher, Comput. Phys. Commun. **165**, 199–220 (2005).
50. M. Lüscher, "Schwarz factorization of the quark determinant", unpublished notes.
51. H. Radjavi and J. P. Williams, Michigan Math. J **16**, 177–185 (1969).
52. P. Fritzsch et al., Nucl. Phys. B **865**, 397–429 (2012).
53. M. Lüscher, Nucl. Phys. B **418**, 637–648 (1994).
54. A. Boriçi and Ph. de Forcrand, Nucl. Phys. B **454**, 645–660 (1995).
55. A. Boriçi and Ph. de Forcrand, Nucl. Phys. Proc. Suppl. **47**, 800–803 (1996).
56. B. Jegerlehner, Nucl. Phys. B **465**, 487–504 (1996).
57. B. Jegerlehner, Nucl. Phys. B Proc. Suppl. **42**, 879–881 (1995).
58. M. Hasenbusch, Phys. Lett. B **519**, 177–182 (2001).
59. M. Hasenbusch and K. Jansen, Nucl. Phys. Proc. Suppl. **106-107**, 1076–1078 (2002).
60. J. C. Sexton and D. H. Weingarten, Nucl. Phys. B **380**, 665–677 (1992).
61. T. A. Manteuffel, Numer. Math. **28**, 307–327 (1977).
62. Y. Saad, *Iterative methods for sparse linear systems*, second edition (SIAM, 2003).