

Optimizing stochastic algorithms for hadron correlation function computations in lattice QCD using a localized distillation basis

Nicolas Lang^{1,*}, Robert G. Edwards^{2,†} and Michael J. Peardon^{3,‡}

(for the Hadron Spectrum Collaboration)

¹*Instituto de Física Corpuscular (IFIC), Universitat de Valencia, 46071, Valencia, Spain*

²*Thomas Jefferson National Accelerator Facility, 12000 Jefferson Avenue, Newport News, Virginia 23606, USA*

³*School of Mathematics, Trinity College, Dublin 2, Ireland*



(Received 14 February 2025; accepted 16 May 2025; published 17 June 2025)

Distillation is a quark-smearing method for the construction of a broad class of hadron operators useful in lattice QCD computations and defined via a projection operator into a vector space of smooth gauge-covariant fields. A new orthonormal basis for this space is constructed which builds in locality. This basis is useful for the construction of stochastic methods to estimate the correlation functions computed in Monte Carlo calculations relevant for hadronic physics.

DOI: [10.1103/mv6l-m8rq](https://doi.org/10.1103/mv6l-m8rq)

I. INTRODUCTION

Distillation [1] is a useful framework for computing correlation functions between a wide range of hadron-interpolating operators in lattice QCD. It is a quark smearing acting on the fields on a time slice of the lattice by applying the projection operator into Ω_D , the space spanned by the modes of the three-dimensional gauge-covariant Laplace operator on that slice. This technique has proved successful for meson spectroscopy (see for example Refs. [2–9]) where distilled operators have good overlap onto the low-lying spectrum of the Hamiltonian while enabling a broad range of creation operators resembling multihadron states to be employed. Good overlap [10] onto these states is important in calculations of scattering and hadron decays in Euclidean field theory calculations via the Lüscher method [11,12] (see Ref. [13] for a review).

The disadvantage of the method is seen when interpolating operators with more constituent quark fields such as for baryons [14–16] or tetraquarks [17,18] are used. Evaluating the diagrams following from Wick contractions of the quarks requires $\mathcal{O}(n_D^{d+1})$ arithmetic operations, where n_D is

the rank of distillation space and d the number of constituent fermion fields in the operator or equivalently $\mathcal{O}(V_3^{d+1})$, assuming the number of distillation vectors is proportional to the spatial volume V_3 . One established approach to reduce computational costs is to sample with random vectors in Ω_D and construct unbiased estimators for correlation functions [19]. The statistical noise added by this sampling is further reduced using dilution schemes [20] and effectiveness of the combination has been demonstrated in a range of calculations. These remain expensive so it is worth pursuing alternatives.

In this paper, a new basis for distillation space is constructed as a potential starting point in the design of stochastic schemes which reduce variance. The basis vectors are localized near defined positions on a coarse spatial grid and so the corresponding local hadron operators exhibit an approximate sparsity in their tensor distillation-space structure; the components where all constituents are near the same basis grid point are substantially larger than others. This basis and its numerical evaluation is described in Sec. II. The sparsity can be exploited in stochastic representations of the correlation function and first tests of these cost reductions are described in Secs. III and IV. Sections V and VI discuss our findings and summarize this work.

II. A LOCALIZED BASIS FOR DISTILLATION SPACE

Distillation is a linear operator acting on the fermion fields on a single time slice of the lattice which projects into

*Contact author: nicolas.lang@ific.uv.es

†Contact author: edwards@jlab.org

‡Contact author: mjp@maths.tcd.ie

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI. Funded by SCOAP³.

the small vector space of smooth modes spanned by the gauge-covariant three-dimensional Laplace operator. The distillation operator can be written

$$\square(t) = V(t)V^\dagger(t), \quad (1)$$

where $V(t)$ is the matrix whose n_D columns are the normalized eigenvectors with smallest eigenvalues of the Laplace operator on time slice t . The corresponding n_D -dimensional vector space of smooth modes is labeled Ω_D . The operator of Eq. (1) is invariant under any $U(n_D)$ global unitary transformation of the basis vectors, $V' = VU$ with $UU^\dagger = I$.

The two-point correlation function for isovector meson creation operator Φ formed from distilled fields on time slices t_0 and t_1 is written

$$C(t_1, t_0) = \text{Tr} \Phi(t_1) \tau(t_1, t_0) \Phi(t_0) \tau(t_1, t_0), \quad (2)$$

after defining the $n_D \times n_D$ matrices,

$$\begin{aligned} \Phi(t) &= V^\dagger(t) \Gamma(t) V(t), \\ \tau(t_1, t_0) &= V^\dagger(t_1) D^{-1}(t_1, t_0) V(t_0). \end{aligned} \quad (3)$$

$\Phi(t)$ encodes the spatial and Dirac structure of the operator and is discussed further below. τ is commonly referred to as the *perambulator* and describes quark propagation between the distillation spaces on the two time slices. It is clear from Eq. (2) the computation of the meson two-point function involves summing over a large number of contributions from smooth modes on the source and sink time slices.

The smearing operator acts locally, spreading the field out at the confining scale of hadrons and so can be regarded as a low-rank and approximately sparse matrix. The sparsity is not seen in the basis of eigenvectors of the Laplace operator, which cover the entire three-dimensional time slice. Methods to reduce the computational cost of evaluating correlation functions by exploiting locality can be constructed by working in a basis where this locality is more apparent.

The distillation operator is a low-rank projection into the smooth space that is invariant under spatial translations, so its application to n_D point sources evenly distributed on the time slice forms a spanning set for Ω_D . This spanning set is localized by construction. The locations of the source points, where three independent colored degrees of freedom are supported is constrained such that the number of sources matches n_D . For example, defining a mesh of 4^3 source points evenly spread over the time slice would span a $n_D = 3 \times 4^3 = 192$ distillation space. The vectors at source point x can be constructed to be covariant under gauge transformations by simply copying only the three entries at site x from the lowest three Laplace eigenvectors into the sources. Define Q as the matrix with this grid of point vectors as its columns. Now applying the distillation

operator yields column matrix $\bar{W}$, a new set of n_D vectors which span Ω_D ;

$$\bar{W} = \square Q = V A_0 \quad \text{with} \quad A_0 = V^\dagger \bar{W} = V^\dagger Q. \quad (4)$$

Here, the square matrix A_0 represents the mapping from the orthonormal eigenvector basis V to the new localized spanning set $\bar{W}$. Since the columns in $\bar{W}$ do not necessarily form an orthonormal basis, A_0 is not always unitary.

Developing software and algorithms is greatly simplified if an orthonormal basis for this spanning set can be used, but locality of the point source construction needs to be preserved. Traditionally, the Gram-Schmidt method is used, where basis vector e_k is explicitly orthogonalized against the preceding elements $e_j, j = 1 \dots k-1$ by subtraction. However this construction is not permutation invariant and destroys the locality added into the spanning set at the start of the process. An alternative method introduces a fictitious flow-time coordinate, s along which $A(s)$ evolves and by setting $A_0 = A(0)$ to be the initial condition of a flow equation whose fixed points are the desired unitary matrices in distillation space, so

$$\lim_{s \rightarrow \infty} A(s)^\dagger A(s) = I_D. \quad (5)$$

Unitary matrices are the set of zeros of a non-negative action, S acting on complex $n_D \times n_D$ matrices and given by

$$S(A) = \frac{1}{2} \text{tr} \Delta^2(A) \quad \text{with} \quad \Delta(A) = I - A A^\dagger. \quad (6)$$

A flow can be defined by the first-order differential equation in s ,

$$\frac{dA}{ds} = -\partial_* S \quad \text{with} \quad [\partial_* S]_{ij} = \frac{\partial S}{\partial [A^*]_{ij}} = \frac{\partial S}{\partial [A^\dagger]_{ji}}. \quad (7)$$

This flow follows the steepest-descent path for the action S . Taking the definition in Eq. (6), the resulting first-order differential equation becomes

$$\frac{dA}{ds} = (I - A(s) A^\dagger(s)) A(s), \quad (8)$$

and since for unitary matrices the right-hand side vanishes, they are fixed points of this flow. For the typical distillation-space dimensions used in spectroscopy calculations, evolving these flow equations to their fixed point to find an orthonormal basis is not computationally costly. Machine precision is easily achieved using the fourth-order Runge-Kutta scheme accelerated with an adaptive step size.

The advantage of this flow-based scheme over Gram-Schmidt is seen by considering two initial conditions for the flow A_0 and A'_0 , related by

$$A'_0 = P_L A_0 P_R, \quad (9)$$

with $P_L P_L^\dagger = I_{\mathcal{D}}$ and $P_R P_R^\dagger = I_{\mathcal{D}}$ defining a pair of arbitrary unitary transformations. Simple examples for $P_{L,R}$ would be given by a permutation of the source points used to form the local set of vectors spanning $\Omega_{\mathcal{D}}$. The flow equation and Eq. (9) imply $A'(s) = P_L A(s) P_R$, for all s so the resulting orthonormal basis vectors inherit the same transformation property linking A_0 and A'_0 , including permutations. This is seen empirically to maintain locality for the orthonormal basis vectors generated in contrast to using Gram-Schmidt.

The new orthonormal basis set is found by setting

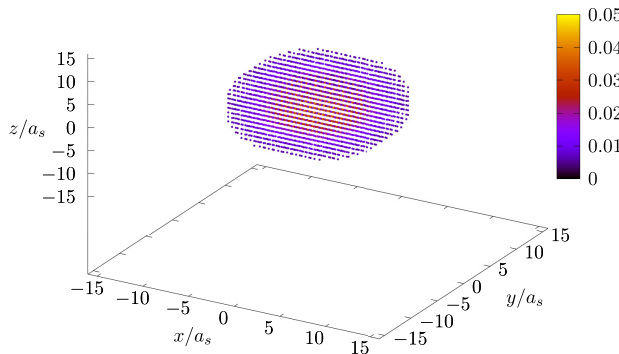
$$W = VU \quad \text{with} \quad U = \lim_{s \rightarrow \infty} A(s) \quad (10)$$

after evolving the flow equation to its fixed point. Now the distillation smearing operator can be expressed equivalently as either $\square = VV^\dagger$ or $\square = WW^\dagger$.

The time-slice color field $w_a^{(i)}(\underline{x}, t)$ is formed from the i th column of W and all $n_{\mathcal{D}}$ of these fields are seen to be localized close to their source point on time slice t after computing the local norm,

$$\rho^{(i)}(\underline{x}, t) = \left(\sum_{a=1}^3 w_a^{(i)*}(\underline{x}, t) w_a^{(i)}(\underline{x}, t) \right)^{\frac{1}{2}}. \quad (11)$$

Figure 1 shows the gauge-invariant sitewise norm $\rho^{(k)}(\underline{x}, t)$ averaged over an ensemble of representative gauge configurations. The locality of the resulting field is seen as the field takes large values close to the point where the source was anchored.



Smeared quark fields, χ are now given in terms of bare fields, ψ by

$$\chi_a^a(\underline{x}, t) = \square^{ab}(\underline{x}, \underline{y}; t) \psi_a^b(\underline{y}, t), \quad (12)$$

where $\underline{x}, \underline{y}$ label spatial sites, a, b color components and α the Dirac spinor index. Effective hadron creation operators are made from smeared quarks so for example an interpolating field for meson M with momentum $\underline{p}$ is defined by $\Gamma_M(\underline{p})$, a general gauge-covariant linear operator on time slice t which reads,

$$\mathcal{O}_M^\dagger(t) = \bar{\chi}_\alpha^a(\underline{x}, t) \Gamma_{M,\alpha\beta}^{ab}(\underline{x}, \underline{y}; \underline{p}, t) \chi_\beta^b(\underline{y}, t). \quad (13)$$

Expanding this expression in terms of the unsmeared fields using the new localized distillation space basis vectors, $W(t)$ while suppressing spatial and color indices and keeping Dirac indices explicit gives

$$\begin{aligned} \mathcal{O}_M^\dagger(t) &= \bar{\psi}_\alpha(t) \square(t) \Gamma_{M,\alpha\beta}(\underline{p}, t) \square(t) \psi_\beta(t) \\ &= \bar{\psi}_\alpha(t) W(t) W^\dagger(t) \Gamma_{M,\alpha\beta}(\underline{p}, t) W(t) W^\dagger(t) \psi_\beta(t) \\ &= \bar{\psi}_\alpha(t) W(t) \Phi_{M,\alpha\beta}(\underline{p}, t) W^\dagger(t) \psi_\beta(t), \end{aligned} \quad (14)$$

where $\Phi_M(t)$ encodes the spatial, color and Dirac structure of operator Γ_M . This can be decomposed efficiently into a sum of basis distillation-space matrices using

$$\Phi_{M,\alpha\beta}(\underline{p}, t) = \sum_k \phi^k(\underline{p}, t) S_{M,\alpha\beta}^k$$

$$\text{with } \phi^k(\underline{p}, t) = W^\dagger(t) \mathcal{D}^k(\underline{p}, t) W(t), \quad (15)$$

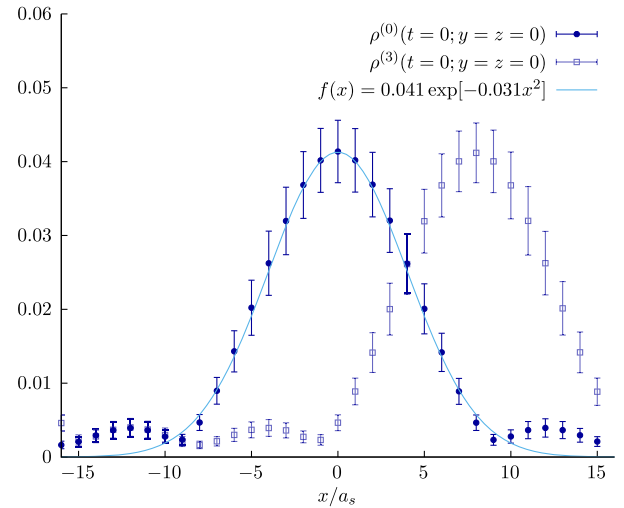


FIG. 1. Left: Spatial distribution of $\rho^{(0)}(\underline{x}, t=0)$. Elements which are smaller than one-eighth of the maximum are suppressed in the plot for improved readability. Right: Slice along the x -axis of the spatial distribution of $\rho^{(i)}(\underline{x}, t=0)$ for $i=0$ and $i=3$. For $i=0$ a Gaussian fit was performed and the blue line represents the fit result. The distributions correspond to the ensemble average over the magnitudes.

which factorizes the operator into independent spinor matrices, S_M^k , generated using the relevant Clebsch-Gordan coefficients depending on the momentum and quantum numbers of meson M and ϕ^k the distillation-space representations of $\mathcal{D}^k$, derivative-based operators with momentum $\underline{p}$. These are called *elementals*.

The magnitude of individual entries in these elementals depend strongly on the choice of basis and the aim of the new localized basis, W is to introduce more structure to these matrices. As an example, for a gauge-invariant operator to create a meson with momentum $\underline{p}$ made from smeared fields contracted without any derivatives applied, the elemental is evaluated using

$$\phi_{ij}^{(0)}(\underline{p}, t) = \sum_{\underline{x}, a} w_a^{(i)*}(\underline{x}, t) e^{i\underline{p} \cdot \underline{x}} w_a^{(j)}(\underline{x}, t). \quad (16)$$

In this expression, i and j label the new local basis vectors for distillation space. This corresponds to an inner product which has smaller magnitude for those entries where the localization anchor points are further apart and so ϕ has larger entries on its diagonal. Note if $\underline{p} = 0$, this matrix is

the identity (in any basis) due to the orthonormality of the spanning vectors. Using Laplace eigenmodes as basis vectors however, the matrix is dense for any nonzero momentum. Figure 2 illustrates the difference between these elementals in the two different bases and the block-diagonal structure is seen clearly.

For baryon creation operators, the benefits from localization are more significant. Now the creation operator using distilled fields can be written

$$\begin{aligned} \Phi_{\alpha\beta\gamma}^B(x, t) &= \sum_n \phi^{B,n}(x, t) S_{\alpha\beta\gamma}^n, \\ \phi_{ijk}^{B,n}(x, t) &= \epsilon_{abc} (\mathcal{D}_n^1 w^{(i)}(t))_x^a (\mathcal{D}_n^2 w^{(j)}(t))_x^b (\mathcal{D}_n^3 w^{(k)}(t))_x^c. \end{aligned} \quad (17)$$

Even for the simplest baryon elemental containing no derivatives, all entries take nonzero values and in general they have similar magnitudes. If these entries are large only when the three anchor points are close then the number of large entries is reduced from $\mathcal{O}(n_D^3)$ to just $\mathcal{O}(n_D)$. This effect is seen in Fig. 3.

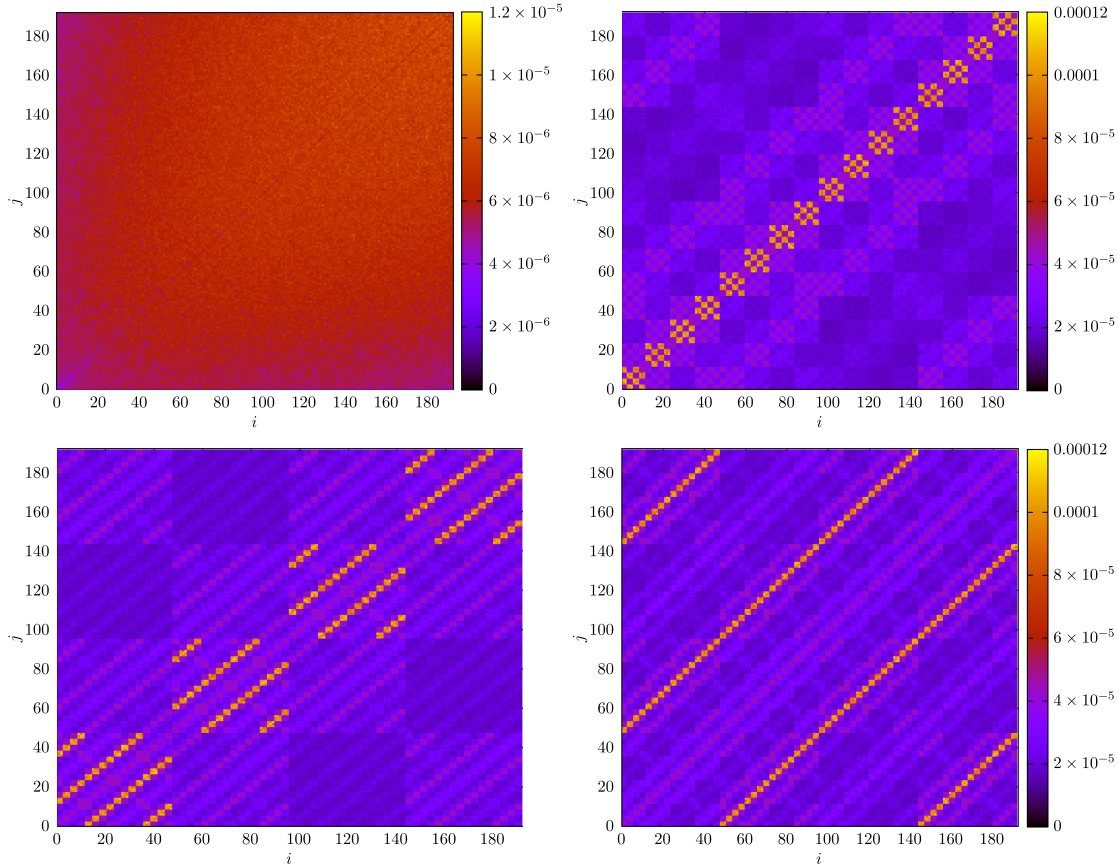


FIG. 2. Magnitudes of complex elements of meson elementals, $|\phi_{ij}^M(\underline{p} = 0, t)|$, constructed from nonrotated $v^{(i)}$ (top-left panel) and rotated $w^{(i)}$ (all other panels). Spatial derivatives in the x -direction (top left and right), y -direction (bottom-left) and z -direction (bottom-right) have been applied to the vectors as described in the text. The magnitudes of the matrix elements have been averaged over the ensemble of gauge configurations.

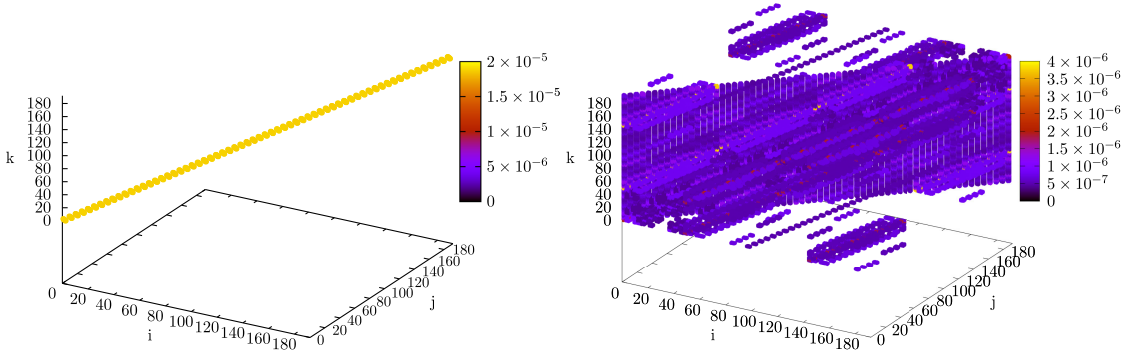


FIG. 3. Magnitudes of complex elements of baryon elementals, $|\phi_{ijk}^B(\underline{p} = 0, t)|$ constructed from rotated $w^{(i)}(t)$ with trivial spatial structure (left) and a spatial derivative in the z -direction applied to the last quark (right). Elements which are smaller than one-tenth of the maximum are suppressed in the plot for improved readability.

Formally, the evaluation of the correlation function for these hadronic creation operators proceeds in exactly the same way as the original distillation algorithm. The correlation function for meson M between a distilled source at time t_0 and sink at t_1 is written

$$C_M(t_0, t_1) = \langle \text{Tr} \Phi(t_0) \tau(t_0, t_1) \Phi(t_1) \tau(t_1, t_0) \rangle, \quad (18)$$

with $\tau(t_0, t_1)$ the perambulator, which describes quark propagation between the distillation spaces on time slices t_0 and t_1 ,

$$\tau_{\alpha\beta}(t_0, t_1) = W^\dagger(t_0) [M^{-1}]_{\alpha\beta}(t_0, t_1) W(t_1). \quad (19)$$

In this expression, the Dirac indices α, β are explicitly given to emphasize this structure is inherited solely from the inverse of the lattice Dirac operator, M . In general, since the inverse Dirac operator appears in the expression for the perambulator we will not assume using a localized basis for distillation space will lead to extra structure in the perambulator although this is worth further investigation. Following a similar procedure, two-point functions for the baryons can be expressed in terms of contractions of baryon elementals and perambulators.

As a trace over products of matrices in distillation space, Eq. (18) expresses the correlation function as a large sum over a number of distillation-space indices. In the following section, the observation that localization isolates large contributions in a small number of terms can be exploited to define stochastic representations with reduced variance compared to a similar scheme using the standard Laplace eigenvector basis for distillation space.

III. STOCHASTIC COMPUTATION OF CORRELATION FUNCTIONS

Consider σ , the sum of a set $\{a_1, a_2, \dots, a_n\}$ of complex numbers. This sum can be estimated stochastically. The *Hansen-Hurwitz* estimator [21] is a random number defined by

$$A = \begin{cases} a_1/p_1 & \text{with probability } p_1, \\ a_2/p_2 & \text{with probability } p_2, \\ \vdots & \\ a_n/p_n & \text{with probability } p_n. \end{cases} \quad (20)$$

$\mathbb{E}[A] = \sigma$ follows immediately from the definition for any set of normalized probabilities, $0 < p_k \leq 1$ with $\sum_k p_k = 1$. The variance of the estimator is

$$\text{Var}[A] = \sum_{k=1}^n p_k \left| \frac{a_k}{p_k} - \sigma \right|^2, \quad (21)$$

and the set of probabilities that minimizes this variance is found to be

$$\tilde{p}_k = \frac{|a_k|}{\sum_{j=1}^n |a_j|}. \quad (22)$$

This choice implies an importance sampling where elements with large magnitude are more likely to be selected. With $p_k = \tilde{p}_k$ and writing $a_k = |a_k| e^{i\phi_k}$, the variance of the Hansen-Hurwitz estimator becomes

$$\text{Var}[A] = \sum_{k=1}^n \tilde{p}_k \left| \sum_{j=1}^n |a_j| (e^{i\phi_k} - e^{i\phi_j}) \right|^2. \quad (23)$$

The variance is zero for the optimal sampling probabilities if all entries in the sum have the same complex phase. The computation of hadronic two-point functions using distillation gives examples of large sums over complex variables and with the structure provided by the sparse basis introduced above, a simple importance sampling method might provide an efficient approach to the evaluation of these correlation functions.

A. Estimating hadron correlation functions

Correlation functions between operators constructed in distillation spaces on two time slices are a tensor contraction involving perambulators and source operators projected into the distillation space on a given time slice.

To start with the simplest example, consider again a connected meson correlation function as seen in Eq. (18), and expand its evaluation on a single gauge-field background as an explicit sum over sets of indices;

$$c_{(AB)}(t_0, t_1) = \sum_{abcd} \Phi_{ab}^{(A)}(t_0) \tau_{bc}(t_0, t_1) \Phi_{cd}^{(B)}(t_1) \tau_{da}(t_1, t_0). \quad (24)$$

In this expression, each composite $\{a, b, c, d\}$ indicates a set of indices spanning a spin and distillation-space label to be summed over. As seen in Sec. III, the Hansen-Hurwitz estimator of this sum is defined by assigning a probability p_{abcd} to every term and the importance sampling optimum is found by setting

$$\tilde{p}_{abcd}(t_0, t_1) = \frac{|\Phi_{ab}^{(A)}(t_0) \tau_{bc}(t_0, t_1) \Phi_{cd}^{(B)}(t_1) \tau_{da}(t_1, t_0)|}{\sum_{abcd} |\Phi_{ab}^{(A)}(t_0) \tau_{bc}(t_0, t_1) \Phi_{cd}^{(B)}(t_1) \tau_{da}(t_1, t_0)|}. \quad (25)$$

Fully realizing this optimum is impractical. To obtain these probabilities we would need to compute the full contraction, defeating the purpose of the sampling. Another difficulty is that the perambulator in the definition would couple the sampling of elements on distinct time slices and prevent any data reuse. A simpler implementation is needed that captures the broad features of the importance sampling method, exploits the sparsity of the localized basis and can be reused in a flexible way.

One simplification is to define a sampling over the indices in the source and sink operators independently, which ignores structure in the perambulators and does not depend on the time index of the creation operators. By averaging over time slices and gauge configurations, we can significantly reduce the sensitivity of the sampling probabilities to fluctuations of the gauge field. Our final choice for the sampling probabilities reads,

$$\tilde{p}_{abcd}^{AB} = \underbrace{\left\langle \frac{1}{N_t} \sum_{t=0}^{N_t} \frac{|\Phi_{ab}^{(A)}(t)|}{\sum_{ij} |\Phi_{ij}^{(A)}(t)|} \right\rangle}_{\tilde{p}_{ab}^{(A)}} \times \underbrace{\left\langle \frac{1}{N_t} \sum_{t=0}^{N_t} \frac{|\Phi_{cd}^{(B)}(t)|}{\sum_{kl} |\Phi_{kl}^{(B)}(t)|} \right\rangle}_{\tilde{p}_{cd}^{(B)}}, \quad (26)$$

where $\langle \dots \rangle$ denotes the ensemble average. The factorization means the two pairs of indices $\{a, b\}$ and $\{c, d\}$ are

Algorithm 1 Meson connected two-point correlation function (Hansen-Hurwitz with replacement)

Draw a random index pair $\{a, b\}$ with probability $\tilde{p}_{ab}^{(A)}$
 Draw a random index pair $\{c, d\}$ with probability $\tilde{p}_{cd}^{(B)}$
return $\Phi_{ab}^{(A)}(t_0) \tau_{bc}(t_0, t_1) \Phi_{cd}^{(B)}(t_1) \tau_{da}(t_1, t_0) / (\tilde{p}_{ab}^{(A)} \tilde{p}_{cd}^{(B)})$

drawn independently and the joint probability of all indices factorizes so these samples can be reused in computations involving different time slices for the source and sink. This is done in Algorithm 1. To be close to the optimum, the modulus of matrix entries in the perambulator must depend weakly on their location within the matrix and this dense structure is observed approximately for large time separations.

The discussion to this point describes evaluation via sampling of a meson connected two-point correlation function. These ideas generalize to a broad class of correlation functions relevant for hadron physics. In the distillation framework, correlation functions are tensor contractions over sets of indices on the source and sink time slices with the indices coupled via perambulators. Their evaluation reduces to a sum over the distillation space and spin indices and so the Hansen-Hurwitz method, combined with the sparse basis can be broadly applied. In particular, the advantage for contractions involving baryons, tetraquarks, and beyond is potentially greater since these evaluations involve deeper nested sums. The evaluation of correlation functions involving baryons is described in the next section, where a hierarchy of efficient partial sums is needed.

B. Implementation of sparse tensor contractions

The term-by-term computation of the contraction performed in algorithm III A is inefficient since the necessary complex-number multiplications scale like $\mathcal{O}(n_D^{2d})$ in the large-sample limit. However this can be reduced to $\mathcal{O}(n_D^{d+1})$ if the contractions are performed sequentially. In the case of the meson connected two-point correlation function we compute the *temporary* tensors,

$$\begin{aligned} \Phi_{ac}^{(A,1)}(t_0, t_1) &= \sum_b \Phi_{ab}^{(A)}(t_0) \tau_{bc}(t_0, t_1), \\ \Phi_{dc}^{(A,2)}(t_0, t_1) &= \sum_a \tau_{da}(t_1, t_0) \Phi_{ac}^{(A,1)}, \end{aligned} \quad (27)$$

and subsequently the full contraction

$$c_{(AB)}(t_0, t_1) = \sum_{cd} \Phi_{dc}^{(A,2)}(t_0, t_1) \Phi_{cd}^{(B)}(t_1). \quad (28)$$

This computation can be combined with sampling in the following way. First construct two vectors $\underline{s}^{(A)}$ and $\underline{s}^{(B)}$ of dimension n_s with entries

$$\begin{aligned} s_i^{(A)} &= (a, b) \quad \text{with probability } \tilde{p}_{ab}^A, \\ s_i^{(B)} &= (c, d) \quad \text{with probability } \tilde{p}_{cd}^B, \end{aligned} \quad (29)$$

called samples. Now define the index-pair multiplicities in sample $\underline{s}$ by

$$m_{\underline{s}}(a, b) = \sum_{i=1}^{n_s} \delta_{s_{i,1}a} \delta_{s_{i,2}b}, \quad (30)$$

where $s_{i,1}$ ($s_{i,2}$) references the first (second) component of the i th entry in sample $\underline{s}$. It will also be useful to define the indicator function,

$$I_{\underline{s}}(a, b) = \Theta(m_{\underline{s}}(a, b) - 1), \quad (31)$$

and *marginal* indicator function

$$I_{\underline{s}}(\cdot, b) = \Theta\left(\sum_{a=1}^{n_D} m_{\underline{s}}(a, b) - 1\right) \quad (32)$$

[analogously $I_{\underline{s}}(a, \cdot)$], where $\Theta(\dots)$ is the Heaviside step function. Now construct the weighted source and sink tensors

$$\begin{aligned} \tilde{\Phi}_{ab}^{(A)} &= \frac{1}{\tilde{p}_{ab}^A} m_{\underline{s}^{(A)}}(a, b) \cdot \Phi_{ab}^{(A)} \quad \text{and} \\ \tilde{\Phi}_{cd}^{(B)} &= \frac{1}{\tilde{p}_{cd}^B} m_{\underline{s}^{(B)}}(c, d) \cdot \Phi_{cd}^{(B)}. \end{aligned} \quad (33)$$

To construct the first sparse temporary define the sets

$$\begin{aligned} X^{(A)} &= \{(a, b) | a, b = 1, \dots, n_D; I_{\underline{s}^{(A)}}(a, b) = 1\} \quad \text{and} \\ X^{(B,1)} &= \{c | c = 1, \dots, n_D; I_{\underline{s}^{(B)}}(c, \cdot) = 1\}. \end{aligned} \quad (34)$$

Algorithm 2 then computes the first sparse temporary from the samples.

This algorithm manifestly scales like $\mathcal{O}(|X^{(A)}| \cdot |X^{(B,1)}|)$. Analogously, for the second sparse temporary we define the sets

Algorithm 2 Construction of first sparse temporary tensor from samples

```

Initialize  $\hat{\Phi}_{ab}^{(A,1)}(t_0, t_1) = 0$  for all  $a, b \in 1, \dots, n_D$ 
for  $(a, b) \in X^{(A)}$  do
  for  $c \in X^{(B,1)}$  do
     $\hat{\Phi}_{ac}^{(A,1)}(t_0, t_1) = \tilde{\Phi}_{ab}^{(A)}(t_0) \tau_{bc}(t_0, t_1)$ 
  end for
end for
return  $\hat{\Phi}^{(A,1)}(t_0, t_1)$ 

```

Algorithm 3 Construction of second sparse temporary tensor from samples

```

Initialize  $\hat{\Phi}_{dc}^{(A,2)}(t_0, t_1) = 0$  for all  $c, d \in 1, \dots, n_D$ 
for  $(c, d) \in X^{(B)}$  do
  for  $a \in X^{(A,1)}$  do
     $\hat{\Phi}_{dc}^{(A,2)}(t_0, t_1) = \tau_{da}(t_1, t_0) \tilde{\Phi}_{ac}^{(A,1)}(t_0, t_1)$ 
  end for
end for
return  $\hat{\Phi}^{(A,2)}(t_0, t_1)$ 

```

$$\begin{aligned} X^{(B)} &= \{(c, d) | c, d = 1, \dots, n_D; I_{\underline{s}^{(B)}}(c, d) = 1\} \quad \text{and} \\ X^{(A,1)} &= \{a | a = 1, \dots, n_D; I_{\underline{s}^{(A)}}(\cdot, a) = 1\}, \end{aligned} \quad (35)$$

and execute the sparse contraction as per Algorithm 3.

Lastly, the correlator can be computed from the second sparse temporary and the weighted sink, as done in Algorithm 4.

In the case of baryon operators such as the ones we will discuss in the next section, the samples are vectors of index triplets. For example, we may compute the contraction

$$\begin{aligned} c_{(AB)}(t_0, t_1) &= \sum_{\substack{abc \\ def}} \Phi_{abc}^{(A)}(t_0) \tau_{ad}^{(1)}(t_0, t_1) \tau_{be}^{(2)}(t_0, t_1) \tau_{cf}^{(3)}(t_0, t_1) \Phi_{def}^{(B)}(t_1), \end{aligned}$$

where the perambulator superscripts indicate that these may be of different quark flavors. Given weighted source and sink tensors $\tilde{\Phi}_{abc}^{(A)}$ and $\tilde{\Phi}_{def}^{(B)}$, to construct the first temporary we define the sets,

$$\begin{aligned} X^{(A)} &= \{(a, b, c) | a, b, c = 1, \dots, n_D; I_{\underline{s}^{(A)}}(a, b, c) = 1\} \\ \text{and } X^{(B,2)} &= \{d | d = 1, \dots, n_D; I_{\underline{s}^{(B)}}(d, \cdot, \cdot) = 1\}, \end{aligned} \quad (36)$$

where $I_{\underline{s}}(a, b, c)$ is the indicator function for the triplet (a, b, c) in the sample $\underline{s}$ and

$$I_{\underline{s}}(d, \cdot, \cdot) = \Theta\left(\sum_{e,f=1}^{n_D} I_{\underline{s}}(d, e, f) - 1\right) \quad (37)$$

is the marginal indicator function for the first index. Then algorithm 5 computes the first sparse temporary.

Algorithm 4 Computation of meson connected two-point correlation function from sparse temporaries

```

Initialize  $\hat{c}_{(AB)}(t_0, t_1) = 0$ 
for  $(c, d) \in X^{(B)}$  do
   $\hat{c}_{(AB)}(t_0, t_1) += \hat{\Phi}_{dc}^{(A,2)}(t_0, t_1) \tilde{\Phi}_{cd}^{(B)}(t_1)$ 
end for
return  $\hat{c}_{(AB)}(t_0, t_1) / n_s^2$ 

```

Algorithm 5 Construction of first sparse temporary tensor for a baryon contraction

```

Initialize  $\hat{\Phi}_{abc}^{(A,1)}(t_0, t_1) = 0$  for all  $a, b, c \in 1, \dots, n_D$ 
for  $(a, b, c) \in X^{(A)}$  do
    for  $d \in X^{(B,2)}$  do
         $\hat{\Phi}_{dbc}^{(A,1)}(t_0, t_1) = \hat{\Phi}_{abc}^{(A)}(t_0) \tau_{ad}^{(1)}(t_0, t_1)$ 
    end for
end for
return  $\hat{\Phi}^{(A,1)}(t_0, t_1)$ 
    
```

The other temporaries and full contraction follow the same pattern. This procedure generalizes to arbitrary tensor contractions. For connected two-point correlation functions the number of temporaries is equal to the number of open distillation-space indices on the source and sink tensors. The sets $X^{(A,r-1)}$ or $X^{(B,d-r)}$, which are inputs for the algorithm that builds the r th temporary, are constructed analogously using generalizations of the indicator functions defined in Eq. (37). The most general form must allow for permutations P of the quark lines and can be written

$$I_{\underline{s}}^{k,P}(i_{P(k+1)}, \dots, i_{P(d)}) = \Theta \left(\sum_{i_{P(1)}, \dots, i_{P(k)}=1}^{n_D} I_{\underline{s}}(i_1, \dots, i_d) - 1 \right), \quad (38)$$

where d is the dimension of the source tensor and $k < d$.

IV. RESULTS FROM FIRST TESTS

As a test case for this new contraction method we compute the single-baryon $c_{(N \rightarrow N)}(t) = \langle 0 | \mathcal{O}_N(t) \mathcal{O}_N^\dagger(0) | 0 \rangle$ correlator in the $I(J)^P = 1/2(1/2)^+$ channel. We also compute the correlators $c_{(\Delta \rightarrow \Delta)}(t) = \langle 0 | \mathcal{O}_\Delta(t) \mathcal{O}_\Delta^\dagger(0) | 0 \rangle$ and $c_{(N\pi \rightarrow \Delta)}(t) = \langle 0 | \mathcal{O}_\Delta(t) \mathcal{O}_{N\pi}^\dagger(0) | 0 \rangle$, where the operators interpolate at-rest single-hadron and two-hadron $I(J)^P = 3/2(1/2)^-$ states.

In the construction of the baryon operators we follow the procedure described in detail in Ref. [15]. As explained there, in any baryon operator the contraction of the color indices with the totally antisymmetric tensor requires the remaining quantum numbers to be symmetric under the interchange of two quarks for the operator to obey the Pauli principle. For simplicity and since all our calculations will be at zero overall momentum, we use two-component Pauli spinors in our constructions. Following the notation in Ref. [15], our nucleon operator is denoted as

$$\mathcal{O}_N^{J^P=\frac{1}{2}^+} = \left(N_M \otimes \left(\frac{1}{2} \right)_M \otimes \mathbf{1}_{L=0,S} \right)^{J^P=\frac{1}{2}^+}, \quad (39)$$

meaning that this operator is mixed-symmetric in spin and flavor and symmetric (and trivial) in space, giving an overall symmetric combination of quantum numbers. Since

the simplest nucleon operator has no derivatives, its distillation space projection is sparse. Only tensor components corresponding to the same coarse grid site have large magnitudes. These are confined to three-by-three-by-three cubes on the body diagonal of the tensor in index space (see left-hand panel of Fig. 3). Our Δ operator has the construction,

$$\mathcal{O}_\Delta^{J^P=\frac{1}{2}^-} = \left(\Delta_S \otimes \left(\frac{1}{2} \right)_M \otimes D_{L=1,M}^{[1]} \right)^{J^P=\frac{1}{2}^-}, \quad (40)$$

meaning it is flavor-symmetric under quark exchange and mixed-symmetric in spin and derivatives, again giving an overall symmetric combination. The angular momenta of the spinors and derivatives are combined using the standard Clebsch-Gordan formula of $SU(2)$ in order to obtain operators with good J in the continuum,

$$|[J, M]\rangle = \sum_{s,l} |[S, s]\rangle \otimes |[L, l]\rangle \langle Ss; Ll | JM \rangle. \quad (41)$$

For the Δ with total spin $J = M = \frac{1}{2}$, spin $S = \frac{1}{2}$ and angular momenta of the derivatives $L = 1$, there are two terms contributing in the sum with the pairs $(s = +\frac{1}{2}, l = 0)$ and $(s = -\frac{1}{2}, l = +1)$. Thus, for the baryon elemental construction from Eq. (17), there are two spatial structures, shown in Fig. 5. The $l = 0$ term involves a single gauge-covariant derivative D_z in the spatial z direction, while the $l = +1$ term involves the circular basis construction of $D_x + iD_y$, thus derivatives in the spatial x and y directions. This latter term is therefore denser in distillation space. Finally, both the nucleon and Δ operator subduce trivially into the G_1 lattice irreducible representation (irrep), and there are no additional reweightings of baryon elemental structures.

For the $I = 1$ pion operator we use the spatially trivial construction,

$$\mathcal{O}_\pi^{J^P=0^-}(\underline{p} = 0, t) = \sum_{\underline{x} \in \Lambda_3} \bar{u}(\underline{x}, t) \gamma_5 d(\underline{x}, t), \quad (42)$$

which leads to

$$\Phi_{\alpha\beta}^\pi(\underline{p} = 0, t) = W^\dagger(t) W(t) \cdot (\gamma_5)_{\alpha\beta} = \mathbf{1} \cdot (\gamma_5)_{\alpha\beta}. \quad (43)$$

The two-hadron $N\pi$ operator is built from the operator product of $\mathcal{O}_N^{J^P=\frac{1}{2}^+}$ and $\mathcal{O}_\pi^{J^P=0^-}$, each projected onto zero spatial momentum, and projected to definite spin and isospin. Since we only compute the rows of maximum J_3 and I_3 we have

$$\left\langle \frac{1}{2}, \frac{1}{2}; 0, 0 \left| \frac{1}{2}, \frac{1}{2} \right\rangle_J = 1 \quad (44)$$

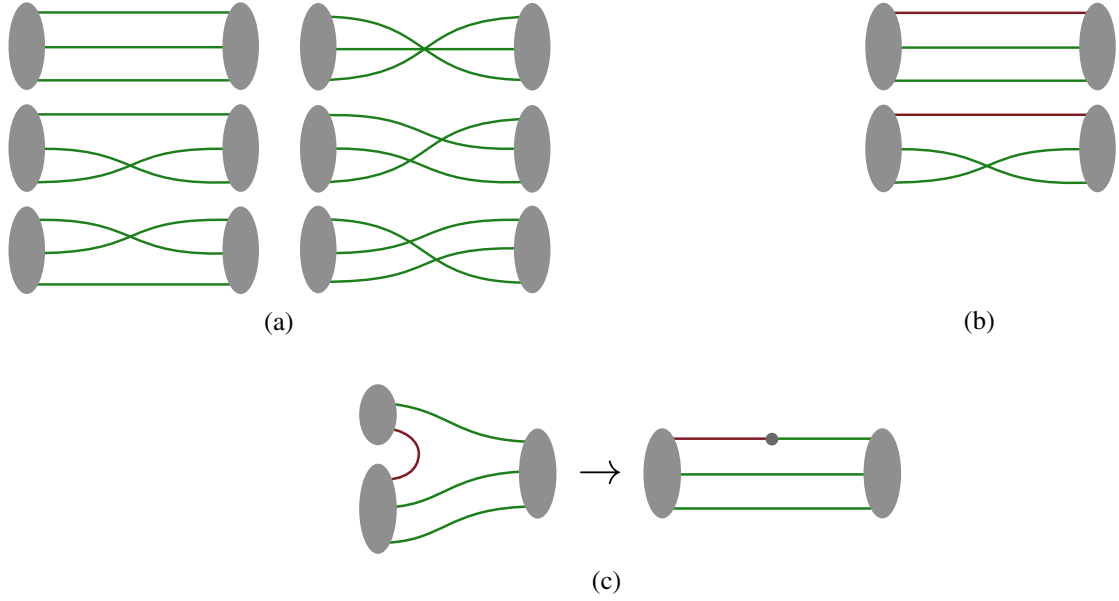


FIG. 4. Topologies of diagrams contributing to the correlators discussed in Sec. IV. Green and red lines correspond to two distinct light flavors of quark. The line with a dark circle on the bottom panel represents a pion contracted with quark propagators on both sides. (a) $c_{(\Delta \rightarrow \Delta)}$. (b) $c_{(N \rightarrow N)}$. (c) $c_{(N\pi \rightarrow \Delta)}$ (+ permutations as in panel 4a)

and

$$\left\langle \frac{1}{2}, \frac{1}{2}; 1, 1 \left| \frac{3}{2}, \frac{3}{2} \right\rangle_I = 1, \quad (45)$$

which simplifies the construction. The resulting operator $O_{N\pi}^{J^P=\frac{1}{2}^-}$ subduces one-to-one in the G_1 lattice irrep.

The topologies of the possible diagrams appearing in $c_{(N \rightarrow N)}$, $c_{(\Delta \rightarrow \Delta)}$ and $c_{(N\pi \rightarrow \Delta)}$ are shown in Fig. 4. Note that $c_{(N\pi \rightarrow \Delta)}$ has a quark-annihilation line at the source. Since the pion at zero momentum is trivial in distillation space, we can simplify the calculation by defining a pion perambulator, given by

$$\begin{aligned} \tilde{\tau}_{\alpha\delta}(t, 0) &= \tau_{\alpha\beta}(t, 0) \Phi_{\beta\gamma}^\pi(\underline{p} = 0, 0) \tau_{\gamma\delta}(0, 0) \\ &= \tau_{\alpha\beta}(t, 0) (\gamma_5)_{\beta\gamma} \tau_{\gamma\delta}(0, 0). \end{aligned} \quad (46)$$

With this definition the topologies of the diagrams appearing in $c_{(N\pi \rightarrow \Delta)}$ are identical to those in $c_{(\Delta \rightarrow \Delta)}$, but in this case source and sink operators differ. This is illustrated for the first of the six diagrams in Fig. 5(c). The pion perambulator is represented by a line adorned with a filled circle.

For each diagram topology all quark-spin combinations corresponding to $S = \frac{1}{2}$ need to be computed. This amounts to 36 individual spin contributions to each of the six diagram topologies of $c_{(\Delta \rightarrow \Delta)}$. However, after performing

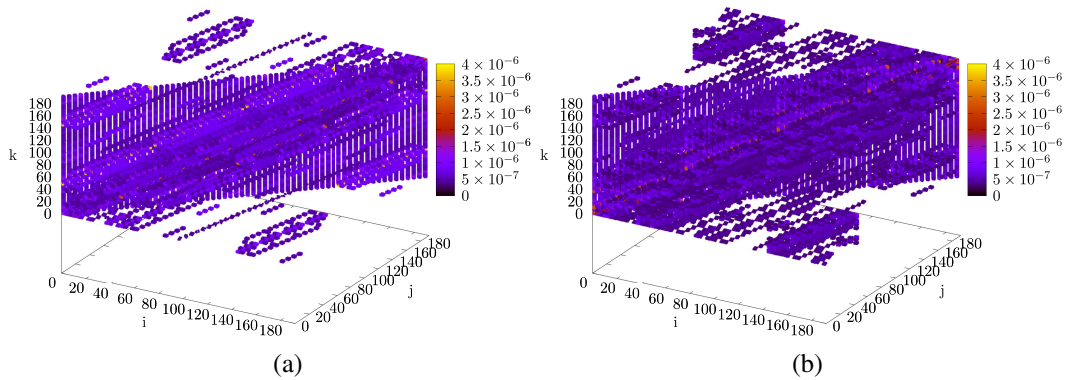


FIG. 5. Sampling probabilities of the two spin components of $O_{\Delta}^{J^P=\frac{1}{2}^-}$ that are distinct in terms of their weight densities. Panel (a) shows the weights with a D_z angular momentum structure ($l = 0$) and panel (b) shows weights with the angular momentum structure $D_x + iD_y$ ($l = +1$). The other four weight tensors can be obtained by index permutations. Only tensor entries larger than one tenth of the maximum are shown. (a) $S = \frac{1}{2}$, $l = 0$. (b) $S = \frac{1}{2}$, $l = +1$

the spin sum, the diagrams in Fig. 5(a) are all identical such that we have to compute 36 distillation-space sums overall for the Δ -to- Δ correlator. In the case of the nucleon, due to the flavor structure the operator can be constructed in a way that only two quark-spin configurations are needed on each end of the correlator. Together with the two distinct diagram topologies shown in Fig. 5(b), this amounts to eight full contractions in distillation space. For $c_{(N\pi\rightarrow\Delta)}$, the six diagram topologies are again identical after summing over spins, such that a total of twelve distillation-space contractions are required considering the two spin-configurations of the $N\pi$ operator and six of the Δ operator. Note that there are internal spin-sums appearing in the contractions of the operators with the perambulators that we need to keep track of.

The computation of the correlator is carried out on an ensemble of 482 gauge configurations over 32 time slices using a single time source. The ensemble has been generated using a tree-level Symanzik-improved

anisotropic action for the gauge fields and a tree-level, tadpole-improved Sheikholeslami-Wohlert action with $N_f = 2 + 1$ dynamical quark flavors for the fermion sector, where stout-smearing has been applied to the spatial gauge links [22]. On each gauge configuration and time slice, and for each configuration of quark spins, we calculate the Hansen-Hurwitz estimate of the correlator, given by

$$\hat{c}_{(AB)}(t_0, t_1) = \frac{1}{n_s^2} \sum_{\substack{(abc) \in s^{(A)} \\ (def) \in s^{(B)}}} \frac{[c_{(AB)}(t_0, t_1)]_{abcdef}}{\tilde{p}_{abc}^{(A)} \times \tilde{p}_{def}^{(B)}}, \quad (47)$$

from two samples $s^{(A)}$ and $s^{(B)}$ of lengths n_s drawn from the distributions $\tilde{p}^{(A)}$ and $\tilde{p}^{(B)}$, respectively, where A and B are placeholders for the aforementioned operators. For example, for the first diagram in Fig. 5(b) we compute

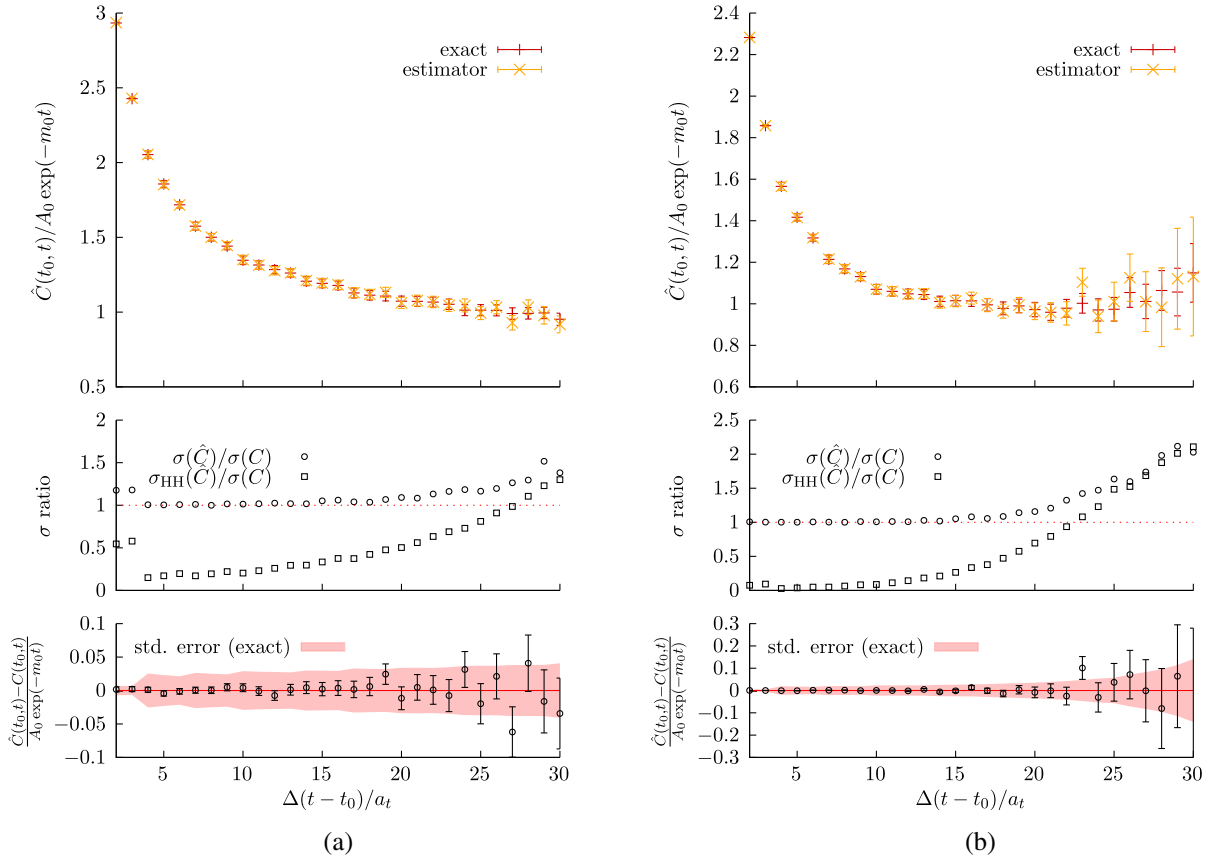


FIG. 6. $C_{(N \rightarrow N)}$ and $C_{(\Delta \rightarrow \Delta)}$ computed with exact and sampled distillation using sampling with replacement and the Hansen-Hurwitz estimator. The correlators are rescaled by the leading exponential obtained from a fit to the exact data. In each plot the middle panel shows the ratio of the standard error between the exact and the sampled correlator as well as an estimate of the contribution to the error that is solely due to the sampling procedure. The bottom panel shows the absolute difference between correlator values computed by the two methods divided by the leading exponential. The error bars in the bottom panels represent the sampling error only. (a) $C_{(N \rightarrow N)}$, $n_s = 5 \times 10^4$. (b) $C_{(N \rightarrow N)}$, $n_s = 8 \times 10^5$.

$$\begin{aligned}
& [c_{(N \rightarrow N)}(t_0, t_1)]_{abc def} \\
&= \frac{(\Phi^{(N)})_{abc}^\dagger(t_0) \tau_{ad}(t_0, t_1) \tau_{be}(t_0, t_1) \tau_{cf}(t_0, t_1) \Phi_{def}^{(N)}(t_1)}{\tilde{P}_{abc}^{(N)} \times \tilde{P}_{def}^{(N)}}
\end{aligned} \tag{48}$$

for all index tuples (abc) and (def) contained in two distinct samples drawn from $\tilde{P}^{(N)}$, where $\Phi^{(N)}(t)$ is the distillation-space projection of $\mathcal{O}_N^{J^P=\frac{1}{2}^+}$ on time slice t for the given set of quark spins. An unbiased estimator for the variance of $\hat{c}_{(AB)}(t_0, t_1)$ is given by

$$\begin{aligned}
& \hat{\sigma}_{\text{HH}}^2[\hat{c}_{(AB)}(t_0, t_1)] \\
&= \frac{1}{n_s^2(n_s^2 - 1)} \sum_{\substack{(abc) \in s^{(A)} \\ (def) \in s^{(B)}}} \left\{ \frac{|[c_{(AB)}(t_0, t_1)]_{abc def}|^2}{(\tilde{P}_{abc}^{(A)} \times \tilde{P}_{def}^{(B)})^2} \right. \\
&\quad \left. - |\hat{c}_{(AB)}(t_0, t_1)|^2 \right\}
\end{aligned} \tag{49}$$

and computed in the same way as the correlator using the method presented in Sec. III B. Since the nucleon is less dense in distillation space than the Δ , $\hat{c}_{(N \rightarrow N)}$ can be computed with a smaller sample size than the correlators involving the Δ operator. We choose a sample size $n_s = 5 \times 10^4$ for the former whereas $\hat{c}_{(\Delta \rightarrow \Delta)}$ and $\hat{c}_{(N\pi \rightarrow \Delta)}$ are sampled with $n_s = 8 \times 10^5$. In all cases we choose symmetric sample sizes for the two samples corresponding to the source and sink operators. Furthermore sampling is only applied in distillation space such that sums over Dirac space are always computed exactly. Distillation-space samples are drawn individually for each set of Dirac components. We also compute the exact contractions using full distillation in order to have a reference value. The results of these computations are shown in Fig. 6 and 7. We denote the ensemble averages with capital letters, i.e., $C_{(AB)}(t_0, t) = \langle c_{(AB)}(t_0, t) \rangle$, and capital letters with hats denote ensemble averages of estimators. $\sigma_{\text{HH}}(\hat{C})$ denotes the ensemble average of the standard deviation of the Hansen-Hurwitz estimator whereas $\sigma(\hat{C})$ denotes the full standard deviation of the correlator including sampling and gauge noise. Ensemble quantities are estimated using jackknife resampling.

For all three correlators we are able to obtain an unbiased estimate from the sampling procedure. We find that $\hat{C}_{(N \rightarrow N)}$ estimates the exact result with good precision across all time slices. The additional noise introduced by the Hansen-Hurwitz sampling is subleading with respect to the gauge noise [see middle panel of Fig. 6(a)]. $\hat{C}_{(\Delta \rightarrow \Delta)}$ has good precision up to $(t - t_0)/a_t \approx 22$, at which point a plateau has formed. $\hat{C}_{(N\pi \rightarrow \Delta)}$ is substantially more noisy and there is a significant contribution to the noise due to the

Hansen-Hurwitz sampling even at early time slices [see Fig. 7]. We suspect that cancellations due to complex phases make the sampling less efficient and that this effect may be enhanced in the case of unequal source and sink operators.

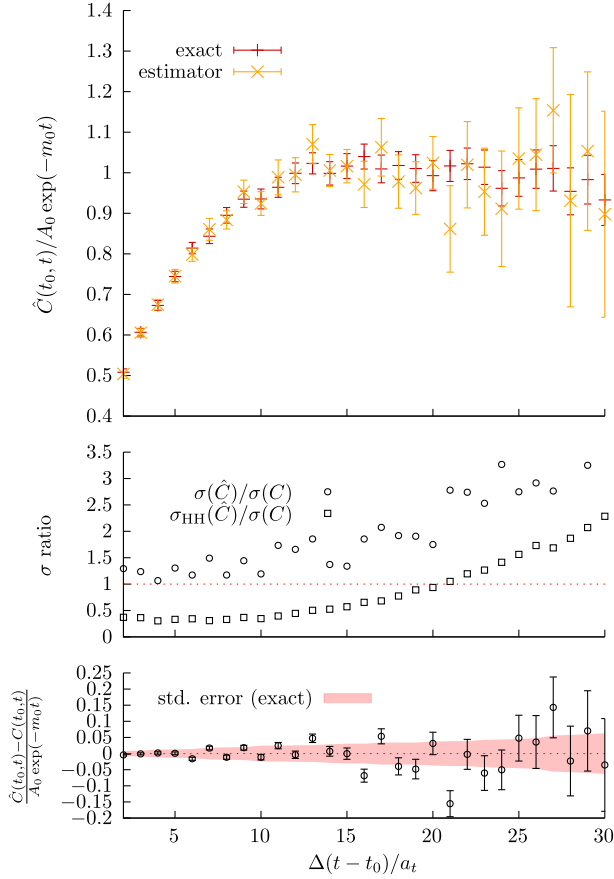
V. DISCUSSION

Having demonstrated the viability of the localized distillation method, we dedicate this section to some general observations regarding the efficiency and scaling of the algorithm. We do not attempt to make precise statements about the actual computer time but rather identify the conditions under which localized distillation can have substantial advantages over standard distillation.

A. Efficiency of the method for the evaluation of baryon correlation functions

For baryons, the most expensive part of the sampling method is the construction of the second temporary. To see this, recall the algorithm to construct the r th temporary scales like $\mathcal{O}(|X^{(A, r-1)}| \cdot |X^{(B, d-r)}|)$, where $X^{(A, k)}$ ($X^{(B, k)}$) is the set of distinct index tuples in the sample of $\mathcal{O}^A$ ($\mathcal{O}^B$) marginalized over the k outermost (innermost) index-dimensions. As a simple example, consider $n_D = 2$, $d = 3$. Assume for simplicity that the same sample of indices $\underline{s} = [(1, 1, 1), (1, 2, 1), (2, 2, 1), (1, 1, 1), (1, 2, 2)]$ has been drawn for operators $\mathcal{O}^A$ and $\mathcal{O}^B$. Since we can factor multiply-occurring indices into the weights, we construct the set of unique indices $X^B = X^A \equiv X^{A,0} = \{(1, 1, 1), (1, 2, 1), (1, 2, 2), (2, 2, 1)\}$. The first marginal set for operator $\mathcal{O}^A$ is then obtained by marginalizing over the outermost index and given by $X^{(A,1)} = \{(1, 1), (1, 2), (2, 2)\}$. The second marginal set for A is $X^{(A,2)} = \{(1), (2)\}$. For $\mathcal{O}^B$ one has to marginalize over the innermost indices, so $X^{(B,1)} = \{(1, 1), (2, 1), (2, 2)\}$ and $X^{(B,2)} = \{(1), (2)\}$. Note that the maximum size of the sets $X^{(A, k)}$ and $X^{(B, k)}$ is n_D^{d-k} . We can then define the *sampling occupancy* of each set as $\rho_k^A = |X^{(A, k)}|/n_D^{d-k}$ and analogously ρ_k^B . In this example we would obtain $\rho_0^A = \rho_0^B = \frac{1}{2}$, $\rho_1^A = \rho_1^B = \frac{3}{4}$ and $\rho_2^A = \rho_2^B = 1$. Defining the cost coefficient $\zeta_r = \rho_{r-1}^A \cdot \rho_{d-r}^B$, the cost of the computation of the r th temporary is given by $\mathcal{O}(\zeta_r n_D^{d+1})$. In our example we would get $\zeta_1 = \frac{1}{2}$, $\zeta_2 = \frac{9}{16}$ and $\zeta_3 = \frac{1}{2}$. This illustrates that the cost coefficient pertaining to the second temporary is largest. The fact that we assumed identical samples on both sides does not affect this point. For a given sample size n_s with replacement, the occupancies can be obtained directly from the weights. The expected value for the size of the set X^A is

$$\mathbb{E}[|X^{(A)}|] = n_D^d - \sum_{i_1, \dots, i_d=1}^{n_D} (1 - \tilde{p}_{i_1, \dots, i_d}^{(A)})^{n_s}. \tag{50}$$


 FIG. 7. Like Fig. 6 but showing $C_{(N\pi\rightarrow\Delta)}$ with $n_s = 8 \times 10^5$.

For the marginal sets we define the marginal probabilities

$$\tilde{p}_{i_1, \dots, i_{d-k}}^{(A,k)} = \sum_{i_{d-k+1}, \dots, i_d=1}^{n_D} \tilde{p}_{i_1, \dots, i_d}^{(A)}, \quad (51)$$

such that

$$\mathbb{E}[X^{(A,k)}] = n_D^{d-k} - \sum_{i_1, \dots, i_{d-k}=1}^{n_D} (1 - \tilde{p}_{i_1, \dots, i_{d-k}}^{(A,k)})^{n_s}. \quad (52)$$

The marginal probabilities for operator $\mathcal{O}^B$ are constructed analogously (summing over the innermost indices). If $\mathcal{O}^A = \mathcal{O}^B$ and the operator has no derivatives (and is therefore antisymmetric under index permutations) they are equal.

Figure 8 shows $\tilde{p}^{(N,1)}$ for the nucleon operator (cf. the left-hand panel of Fig. 3 for $\tilde{p}^N$). Clearly the probability mass function becomes flatter when marginalized over an increasing number of indices. In $\tilde{p}^{N,2}$ all the original structure is lost and the measured optimal probabilities are independent of the index in the localized basis. Figure 9 shows the expected occupancies ρ_k^N of the index sets and the cost coefficients ζ_k of the temporaries as a function of

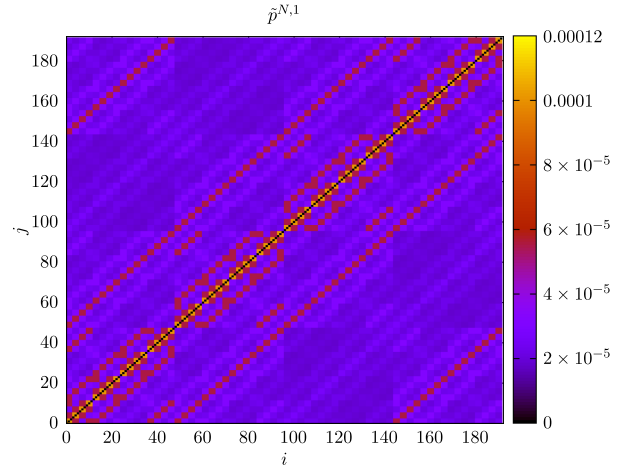


FIG. 8. The distribution of probability weights marginalized over one index for the evaluation of the nucleon correlation function.

the sample size in the computation of the nucleon-nucleon correlator. Note that ρ_1^N reaches its plateau just below 1 as the contraction of the color components with the epsilon tensor in the construction of the nucleon operator causes the probabilities for any pair of identical indices to vanish. From the right-hand side of Fig. 9 we see the cost coefficient of the second temporary plateaus much earlier compared to the other two. Beyond a sample size of $n_s \approx 10^5$ local distillation does not grant any substantial efficiency gain over standard distillation in this scenario.

B. Dependence on the number of constituent quarks and the physical volume

The previous section demonstrated that within the sequence of temporaries, the central ones will be the most expensive to compute. Explicitly, if the number of constituent quarks d is odd, the $[(d+1)/2]$ th temporary will be the most expensive. If d is even, the $d/2$ th and $(d/2+1)$ st temporaries will be the most expensive. For example, in the case of a two-point tetraquark correlator, this corresponds to the second and third temporaries. Based on our baryon results, we attempt to make a crude prediction for the cost reduction of a tetraquark calculation using the localized distillation basis and our sampling method.

Tetraquark operators are typically constructed by contracting a diquark and an antidiquark to a color singlet. The two possible colorless contractions are $\bar{\mathbf{3}} \otimes \mathbf{3}$ and $\mathbf{6} \otimes \bar{\mathbf{6}}$. Here we use operators made of (anti)diquarks in the $\bar{\mathbf{3}} (\bar{\mathbf{3}})$ representation constructed from locally distilled quark fields with no derivatives between them. The quark and antiquark color indices of this construction will be pairwise totally antisymmetric while their spatial components are

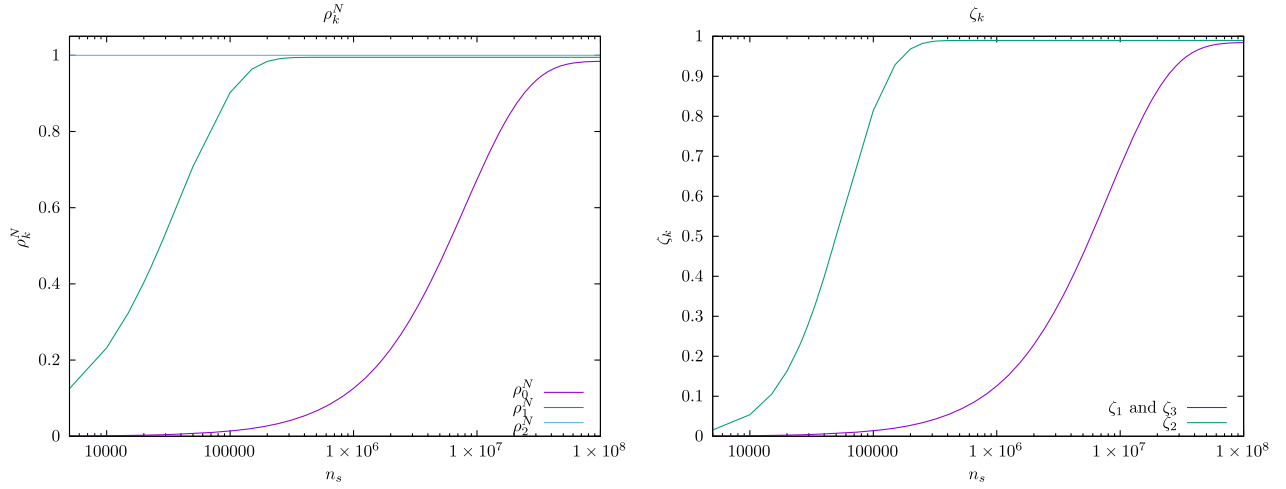


FIG. 9. Expected index-set occupancies (left) and the cost factors of the three temporaries (right) of the nucleon operator.

symmetric. In local distillation, this translates to a tensor of probabilities p_{abcd}^T that is symmetric under distillation-index permutations and zero for any pair of identical quark or antiquark indices. Marginalizing over one of the indices we expect to obtain a three-index probability tensor similar to the nucleon so we use this to estimate the scaling of our algorithm for tetraquarks. To estimate the cost factors of the 2nd and 3rd temporary in the computation of the Wick contractions for the connected tetraquark 2-point correlator, we need the marginalized probabilities $\tilde{p}^{(T,1)} \approx \tilde{p}^{(N,0)}$ and $\tilde{p}^{(T,2)} \approx \tilde{p}^{(N,1)}$. The cost factors $\zeta_2 = \zeta_3$ can then be calculated as described in the previous section. Figure 10 shows the expected scaling of this hypothetical computation with the sample size. The final question is which sample size would be required for a compact tetraquark operator to achieve a sampling noise smaller than the gauge noise. The answer is not trivial and needs a dedicated investigation, but it is reasonable to assume the necessary sample size scales like M^{d-1} , where M is the number of sources that significantly overlap with a given coarse grid site. This number will be larger when derivatives are applied to the sources. In the case of the nucleon we found a calculation with $n_s = 5 \times 10^4$ still results in a good signal-to-noise ratio. We can approximate $M \approx \sqrt{n_s/n_D} \approx 16$ (we divide by n_D because the location of the first quark source is arbitrary). Since both the nucleon and the compact tetraquark operators are ultra-local we expect their value of M to be similar. Our rough estimate for the minimum sample size needed to compute a tetraquark correlation function is then $n_s^T = n_D \times M^3 \approx 800,000$. We see from Fig. 10 we may expect a speed-up by a factor of ≈ 10 with respect to full distillation for this calculation.

Apart from the number of constituent quarks the physical volume, which correlates with the number of grid points, is expected to affect the sparsity of the tensors and thereby the efficiency of the sampling method. The sparse basis is

constructed by first defining a coarse grid of source points. If the three-dimensional volume of the lattice is increased and the physical separation between sites on the basis grid is fixed, this leads directly to a linear increase in the number of sites on the grid. To support this increase in grid points, the corresponding dimension of the distillation space must grow in the same proportion. This matches the observed dependence of the space needed to maintain the same level of smearing in a larger volume. It is also commonly noted that a fixed size of distillation space leads to a constant level of smearing as the lattice spacing is reduced and this corresponds to maintaining a fixed distance in physical units between basis grid points. Increasing the physical volume and hence the number of coarse grid sites would decrease the fraction of overlapping sources corresponding to tensor elements with large magnitudes and is therefore expected to give a greater advantage to local over standard distillation.

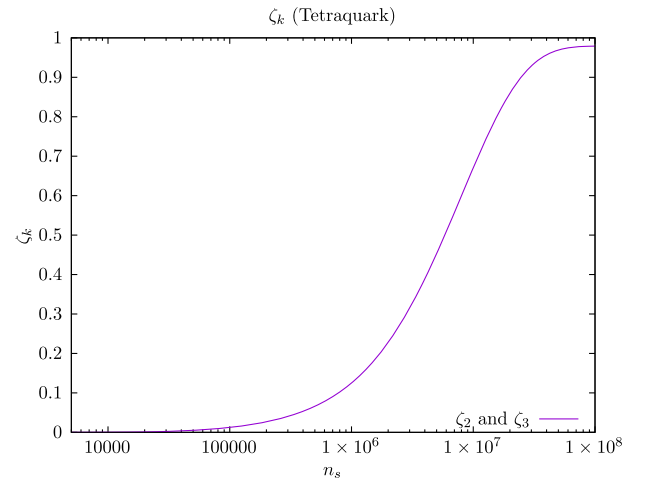


FIG. 10. Expected cost coefficients of the second and third temporary in the computation of the connected tetraquark 2-point correlator.

VI. CONCLUSION

In this paper, a method to construct a new basis for distillation space, the vector space spanned by the lowest eigenvectors of the gauge-covariant three-dimensional Laplacian on each time slice, is introduced. The new basis vectors are more localized. Methods to exploit this locality to accelerate evaluation of the correlation functions needed in hadron spectroscopy calculations are then described and tested.

The eigenvectors of the gauged Laplacian themselves form an orthonormal basis but there is no natural locality in these vectors by construction. The eigenvectors of the free Laplacian, for example, are the Fourier modes and so not localized at all. Something analogous to a gauge-covariant wavelet basis combining smoothness and locality would be useful and such a scheme was presented and tested in Sec. II.

The method works by applying the distillation space projection operator to point sources well-separated on a large grid over the time slice. Empirically, the resulting vectors stay localized to the sources due to both confinement and the truncation to a few low-eigenvalue modes in the distillation operator itself. These vectors are no longer orthonormal, however, and a simple reorthogonalization scheme such as Gram-Schmidt destroys locality. Orthonormality is introduced by defining a flow equation in the space of square matrices whose stable fixed point is the subset of unitary matrices. In tests, it is seen that flowing to the fixed subset makes changes to the matrix, and subsequently the new basis vectors, which are small enough to preserve locality. For hadron operators in the new basis, which are colorless tensors made from the vector components, locality manifests as approximate sparseness where a small subset of entries have much larger magnitude compared to the bulk. Correlation functions are unchanged since all that is being exploited is a basis transformation. The first tests of these ideas in Sec. II demonstrated the flow method works well. The sparseness of the tensors suggests a stochastic importance sampling method might converge quickly and reduce the numerical cost of tensor contractions.

In Sec. III, stochastic representations of sums were introduced, based on the Hansen-Hurwitz estimator. This uses importance sampling to translate the numerical problem of evaluating a large sum, which in this case is over distillation space basis vectors, into a more efficient Monte Carlo method with minimal variance. The variance reduction is enhanced by ensuring a small subset of terms contributes the bulk of the signal and this is enabled in the new basis which generates sparse tensors by exploiting locality.

The expense of using a distillation space of size n_D in a direct computation of correlation functions grows rapidly as the number of constituent quark fields increases. For meson correlation functions, the cost grows in proportion to n_D^3 while for baryons, this increases to n_D^4 . The price tag of distillation has restricted calculations of baryons and

tetraquarks to small vector spaces, which are often far from the best choice set by the physics. The natural way to perform the direct evaluation is to store temporary tensors after summing a subset of indices, which for a simple source-sink construction involving d constituent quark fields reduces the scaling from the naive n_D^{2d} to the much more tractable n_D^{d+1} quoted above. We constructed a Hansen-Hurwitz estimator which preserves this scaling behavior in the large sample-size limit. Some first tests of this estimator applied to baryon correlation functions were presented in Sec. IV and showed that good results can be achieved by summing a small fraction of terms contributing to the correlation function.

The bottleneck of our algorithm, in the case of baryons, is the construction of the second temporary tensor, which quickly approaches the $\mathcal{O}(n_D^4)$ limit as the sample size is increased. A detailed discussion of this effect was presented in Sec. V. Some crude extrapolations of our baryon results suggest that this saturation of the limit may become less of a concern as larger distillation spaces and operators with more constituent quarks are used. The example of a hypothetical tetraquark calculation was discussed in Sec. VB.

The index sampling presented in this article is the most straightforward implementation of importance sampling. We leave it to future work to investigate more efficient index-sampling schemes that might better exploit the approximate sparsity of the distillation-space operators, for example by incorporating known symmetries of the tensors or by making use of stratified sampling techniques. The effect of phase cancellations on the variance and whether it can be mitigated needs to be studied in more detail. Different kinds of coarse grid embeddings may be tested. We did not consider sparsity in disconnected insertions, but it is a topic that could be investigated in the future. We also did not report on tests with nonzero momentum operators as the basis for our application in Sec. IV did not require such constructions. In the future, further studies will be required.

While Sec. IV shows promising first test results, there is more work needed to build a robust algorithm and optimize the necessary work flows for physics calculations. The stochastic approach has the promise of better scaling at larger distillation spaces which would arise as larger physical volumes are used, as well as better performance as more constituent quarks are added, which has severely limited tetraquark calculations using exact evaluations in distillation to date.

The software codes Chroma [23], QUDA [24,25], QPhiX [26], QOPQDP [27,28], and Redstar [29] were used.

ACKNOWLEDGMENTS

We thank our colleagues within the Hadron Spectrum Collaboration [30], in particular Christopher Thomas for comments on the manuscript and Frank Winter and Antoni Woss for discussions at an early stage of this work. R. G. E.

acknowledges support from the U.S. Department of Energy Contract No. DE-AC05-06OR23177, under which Jefferson Science Associates, LLC, manages and operates Jefferson Lab. N.L. acknowledges support by the EU H2020 research and innovation programme under the Staff Exchange Grant Agreement No. 101086085-ASYMMETRY, as well as by the Spanish Ministerio de Ciencia e Innovación Project No. PID2020-113644GB-I00 and by Generalitat Valenciana through the grant PROMETEO/2019/083. This work was supported by the STRONG-2020 project, funded by the European Community Horizon 2020 research and innovation programme under Grant Agreement No. 824093. This work also used clusters at Jefferson Laboratory under the USQCD Initiative and the LQCD ARRA project, and the authors acknowledge support from the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research and Office of Nuclear Physics, Scientific Discovery through Advanced Computing (SciDAC) program, and the U.S. Department of Energy Exascale Computing Project. This research was supported in part under an ALCC award, and used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. This research is also part of the Blue Waters sustained-petascale computing project, which is supported by the National

Science Foundation (Awards No. OCI-0725070 and No. ACI-1238993) and the state of Illinois. Blue Waters is a joint effort of the University of Illinois at Urbana-Champaign and its National Center for Supercomputing Applications. This work is also part of the PRAC “Lattice QCD on Blue Waters.” This research used resources of the National Energy Research Scientific Computing Center (NERSC), a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DEAC02-05CH11231. The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources that have contributed to the research results reported within this paper. Gauge configurations were generated using resources awarded from the U.S. Department of Energy INCITE program at Oak Ridge National Lab, NERSC, the NSF Teragrid at the Texas Advanced Computer Center and the Pittsburgh Supercomputer Center, as well as at Jefferson Lab.

DATA AVAILABILITY

The data that support the findings of this article are not publicly available upon publication because it is not technically feasible and/or the cost of preparing, depositing, and hosting the data would be prohibitive within the terms of this research project. The data are available from the authors upon reasonable request.

-
- [1] M. Peardon, J. Bulava, J. Foley, C. Morningstar, J. Dudek, R. G. Edwards, B. Joo, H.-W. Lin, D. G. Richards, and K. J. Juge (Hadron Spectrum Collaboration), *Phys. Rev. D* **80**, 054506 (2009).
 - [2] J. J. Dudek, R. G. Edwards, M. J. Peardon, D. G. Richards, and C. E. Thomas, *Phys. Rev. Lett.* **103**, 262001 (2009).
 - [3] J. J. Dudek, R. G. Edwards, B. Joo, M. J. Peardon, D. G. Richards, and C. E. Thomas, *Phys. Rev. D* **83**, 111502 (2011).
 - [4] L. Liu, G. Moir, M. Peardon, S. M. Ryan, C. E. Thomas, P. Vilaseca, J. J. Dudek, R. G. Edwards, B. Joo, and D. G. Richards (Hadron Spectrum Collaboration), *J. High Energy Phys.* **07** (2012) 126.
 - [5] N. Lang and D. J. Wilson (Hadron Spectrum Collaboration), *Phys. Rev. Lett.* **129**, 252001 (2022).
 - [6] D. J. Wilson, C. E. Thomas, J. J. Dudek, and R. G. Edwards (Hadron Spectrum Collaboration), *Phys. Rev. Lett.* **132**, 241901 (2024).
 - [7] H. Yan, M. Garofalo, M. Mai, U.-G. Meißner, and C. Urbach, *Phys. Rev. Lett.* **133**, 211906 (2024).
 - [8] H. Li, C. Shi, Y. Chen, M. Gong, J. Liang, Z. Liu, and W. Sun, *arXiv:2402.14541*.
 - [9] P. Boyle, F. Erben, V. Gülpers, M. T. Hansen, F. Joswig, N. P. Lachini, M. Marshall, and A. Portelli, *Phys. Rev. Lett.* **134**, 111901 (2025).
 - [10] F. Knechtli, T. Korzec, M. Peardon, and J. A. Urrea-Niño, *Phys. Rev. D* **106**, 034501 (2022).
 - [11] M. Luscher, *Nucl. Phys.* **B354**, 531 (1991).
 - [12] M. Luscher and U. Wolff, *Nucl. Phys.* **B339**, 222 (1990).
 - [13] R. A. Briceno, J. J. Dudek, and R. D. Young, *Rev. Mod. Phys.* **90**, 025001 (2018).
 - [14] S. Basak, R. G. Edwards, G. T. Fleming, U. M. Heller, C. Morningstar, D. Richards, I. Sato, and S. Wallace, *Phys. Rev. D* **72**, 094506 (2005).
 - [15] R. G. Edwards, J. J. Dudek, D. G. Richards, and S. J. Wallace, *Phys. Rev. D* **84**, 074508 (2011).
 - [16] J. R. Green, A. D. Hanlon, P. M. Junnarkar, and H. Wittig, *Phys. Rev. Lett.* **127**, 242003 (2021).
 - [17] G. K. C. Cheung, C. E. Thomas, J. J. Dudek, and R. G. Edwards (Hadron Spectrum Collaboration), *J. High Energy Phys.* **11** (2017) 033.
 - [18] S. Collins, A. Nefediev, M. Padmanath, and S. Prelovsek, *Phys. Rev. D* **109**, 094509 (2024).

- [19] C. Morningstar, J. Bulava, J. Foley, K. J. Juge, D. Lenkner, M. Peardon, and C. H. Wong, *Phys. Rev. D* **83**, 114505 (2011).
- [20] J. Foley, K. Jimmy Juge, A. O’Cais, M. Peardon, S. M. Ryan, and J.-I. Skullerud, *Comput. Phys. Commun.* **172**, 145 (2005).
- [21] M. H. Hansen and W. N. Hurwitz, *Ann. Math. Stat.* **14**, 333 (1943).
- [22] C. Morningstar and M. Peardon, *Phys. Rev. D* **69**, 054501 (2004).
- [23] R. G. Edwards and B. Joo (SciDAC, LHPC, and UKQCD Collaborations), *Nucl. Phys. B, Proc. Suppl.* **140**, 832 (2005).
- [24] M. A. Clark, R. Babich, K. Barros, R. C. Brower, and C. Rebbi (QUDA Collaborations), *Comput. Phys. Commun.* **181**, 1517 (2010).
- [25] R. Babich, M. A. Clark, and B. Joo, in *Proceedings of Supercomputing 2010 (SC 10)* (2010), [arXiv:1011.0024](#).
- [26] B. Joó, D. D. Kalamkar, K. Vaidyanathan, M. Smelyanskiy, K. Pamnany, V. W. Lee, P. Dubey, and W. Watson, *Lect. Notes Comput. Sci.* **7905**, 40 (2013).
- [27] J. C. Osborn, R. Babich, J. Brannick, R. C. Brower, M. A. Clark, S. D. Cohen, and C. Rebbi, *Proc. Sci. LATTICE2010* **(2010)** 037 [[arXiv:1011.2775](#)].
- [28] R. Babich, J. Brannick, R. C. Brower, M. A. Clark, T. A. Manteuffel, S. F. McCormick, J. C. Osborn, and C. Rebbi, *Phys. Rev. Lett.* **105**, 201602 (2010).
- [29] J. Chen, R. G. Edwards, and W. Mao, in *Platform for Advanced Scientific Computing* (Association for Computing Machinery, New York, NY, 2023).
- [30] www.hadspec.org.