Multigrid deflation for Lattice QCD

Eloy Romero[a,*], Andreas Stathopoulos[a], Kostas Orginos[b,c]

[a]*Department of Computer Science, The College of William & Mary, USA*
[b]*Department of Physics, The College of William & Mary, USA*
[c]*Jefferson Laboratory, USA*

ARTICLE INFO

ABSTRACT

Computing the trace of the inverse of large matrices is typically addressed through statistical methods. Deflating out the lowest eigenvectors or singular vectors of the matrix reduces the variance of the trace estimator. This work summarizes our efforts to reduce the computational cost of computing the deflation space while achieving the desired variance reduction for Lattice QCD applications. Previous efforts computed the lower part of the singular spectrum of the Dirac operator by using an eigensolver preconditioned with a multigrid linear system solver. Despite the improvement in performance in those applications, as the problem size grows the runtime and storage demands of this approach will eventually dominate the stochastic estimation part of the computation.

In this work, we propose to compute the deflation space in one of the following two ways. First, by using an inexact eigensolver on the Hermitian, but maximally indefinite, operator $A\gamma_5$. Second, by exploiting the fact that the multigrid prolongator for this operator is rich in components toward the lower part of the singular spectrum. We show experimentally that the inexact eigensolver can approximate the lower part of the spectrum even for ill-conditioned operators. Also, the deflation based on the multigrid prolongator is more efficient to compute and apply, and, despite its limited ability to approximate the fine level spectrum, it obtains similar variance reduction on the trace estimator as deflating with approximate eigenvectors from the fine level operator.

1. Introduction

To estimate *ab initio* elementary particle properties, Lattice Quantum Chromodynamics (LQCD) involves the evaluation of correlation functions expressed as traces over products of propagators with constant matrices. The

*Corresponding author
e-mail:* eromeroalcalde@wm.edu (Eloy Romero), andreas@cs.wm.edu (Andreas Stathopoulos), knorgi@wm.edu (Kostas Orginos)

propagators include the inverse of the Dirac operator defined on a 4- or 5-dimensional lattice. In practice, the resulting matrix size makes it impossible to compute these traces directly by evaluating the propagators at every point in the lattice. The general strategy is to replace the exact computations by stochastic estimations and control its variance. For the case study in this paper, the propagators are closed, which leads to the evaluation of the trace on the inverse of the Dirac operator, A^{-1}.

There are several methods to estimate the trace of an inverse matrix. The simplest method is to estimate the trace as a multiple of the average diagonal element of A^{-1}, $\text{E}[N\mathbf{e}_i^\dagger A^{-1}\mathbf{e}_i]$, where $\mathbf{e}_i$ is the i-th column of the identity matrix, and N is the dimension of A. The variance of this estimator grows with N^2 and can be very large even when the diagonal elements of A^{-1} have a small variance. Sophisticated versions of this approach can be found in [1, 2].

The standard tool for this problem is the Hutchinson method [3] which estimates the trace of A^{-1} as

$$t(A^{-1}) = \text{E}[\mathbf{x}^\dagger A\mathbf{x}] = \frac{1}{s}\sum_{j=1}^{s}\mathbf{x}_j^\dagger A^{-1}\mathbf{x}_j, \tag{1}$$

where $\mathbf{x}$ is a random vector whose elements x_i satisfy $\text{E}[x_i^\dagger x_j] = \delta_{i,j}$.

If the $\mathbf{x}_i$ are Rademacher vectors, i.e., following the discrete uniform distribution with values ± 1, the trace estimator can also be written as $t(A^{-1}) = \frac{1}{s}A^{-1} \odot XX^\dagger$, where X is a matrix with all the $\mathbf{x}_i$ as columns and $\odot$ is the element-wise product of matrices. Then, the variance associated with the trace estimator $t(A^{-1})$ is

$$\text{Var}[t(A^{-1})] = \frac{2}{s}\|A^{-1}\|_F^2 - \frac{2}{s}\sum_{i=1}^{N}|(A^{-1})_{i,i}|^2. \tag{2}$$

In practice $A^{-1}\mathbf{x}_j$ is computed approximately as $A_\xi^{-1}(\mathbf{x}_j)$ by an iterative solver that stops when $\|A A_\xi^{-1}(\mathbf{x}_j) - \mathbf{x}_j\| \le \xi\|\mathbf{x}_j\|$. The bias of the resulting operator can be corrected by an independent estimation of the trace of $A^{-1} - A_\xi^{-1}$. This trick is known as the Truncated Solver Method [4].

As in any Monte-Carlo process, the variance reduces linearly with the number of samples, s. Every sample involves solving a linear system with the Dirac operator, and for cases where the variance is too large, collecting enough samples can be expensive. Partitioning and deflation can reduce the variance further.

A deterministic way to reduce the variance is by partitioning the matrix rows and columns into k groups and approximating the trace of the diagonal blocks [5, 6, 7]. As k approaches N, the variance reduces to zero. In practice, it is used to remove known strongly coupled components on A^{-1} by setting them on different groups. Examples of this technique are spin-color dilution (separating strongly coupled spin-color components) and partitioning (separating close spacetime components). The most basic way of lattice spacetime partitioning is the red-black (immediate neighbors have the opposite color to the node). Hierarchical probing [8] is a sophisticated extension of this approach that identifies a series of nested colorings at increasing lattice distance. The nesting ensures that, in case of increasing the number of partitions, only linear systems associated with the new colors have to be solved. The drawback of this technique is that the variance is significantly reduced only at the particular number of partitions $k = 2^{Di}$, for $i = 0, 1, \ldots$ and D being the number of dimensions of the lattice. Those particular values of k are called the closing colors.

For similar configurations to the $32^3 \times 64$ lattice used in this paper, and with spin-color dilution in all cases, the red-black partition reduces the variance of the original trace estimator by 3-fold, while using higher-level hierarchical probing with 512 vectors reduces the variance by 10-fold [9, 10]. The latter reduction in variance is modest compared with merely using Hutchinson with the same number of random vectors. In practice, partitioning is combined with other techniques and especially deflation.

The initial motivation for deflation is that singular values near zero of the Dirac operator are responsible for the large variance in the trace estimator. Equation (2) relates the variance with the spectrum of the traceless matrix inverse through the Frobenius norm, which is the addition of all singular values squared. The singular value spectrum of the traceless matrix is a small perturbation of the original spectrum if the largest singular values of A^{-1} are far greater than the diagonal elements of A^{-1} in magnitude, which is the case for Dirac operators. As shown in [10], if the left and the right singular vectors of a non-Hermitian matrix A are statistically independent standard random unitary matrices, then the variance of $t(A^{-1})$ is approximately $\sum_{i=1}^{N}\sigma_i^{-2}$ (Theorem 2.6 in [10]), where σ_i are the singular values of A. The authors also show various examples where the variance is close to that estimation, even when the assumptions of the theorem on the singular vectors are slightly relaxed.
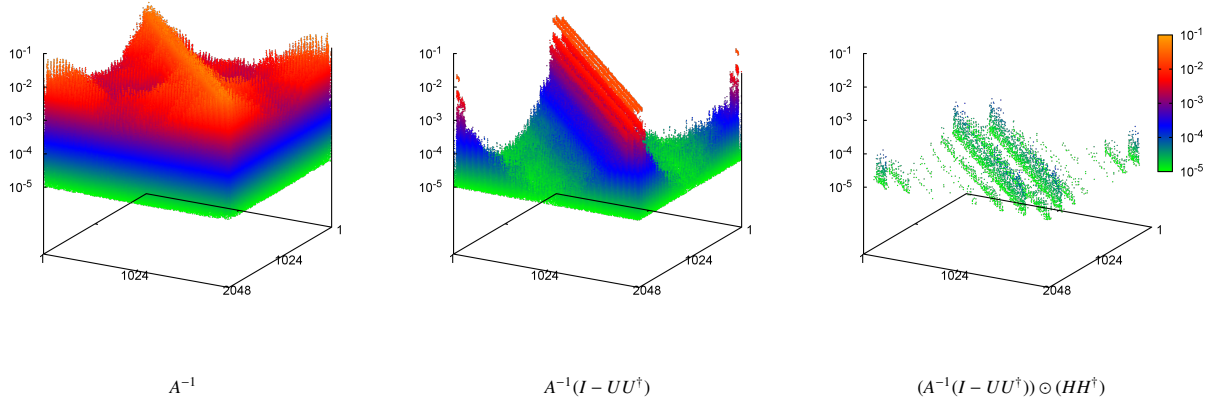
$$A^{-1} \qquad\qquad A^{-1}(I - UU^{\dagger}) \qquad\qquad (A^{-1}(I - UU^{\dagger})) \odot (HH^{\dagger})$$

Fig. 1. Absolute value of the elements of the inverse of a Wilson 2D 32^2 (left), after deflating the space associated to the 100 smallest singular values (middle), and, besides that, applying spin-color dilution and 32 basis of hierarchical probing (right).

The above suggests that the trace estimator has less variance if applied to an operator removing the near-null part of the spectrum (deflation). The strategy is, therefore, to split the trace into two parts, $A^{-1}P$ and $A^{-1}(I - P)$, so that the trace of $A^{-1}P$ can be efficiently computed, and the variance of estimating the trace of $A^{-1}(I - P)$ is far smaller than the variance with A^{-1}. If the singular values of A^{-1} decay *quickly*, only a small percentage of the spectrum needs to be deflated to reduce sufficiently the variance on the trace estimator of $A^{-1}(I - P)$. For the trace of A^{-1}, the left singular vectors of A should be used in P [10].

An interesting synergy occurs when combining partitioning and deflation. The lower part of the operator's spectrum has the dominant contribution to the magnitude of the elements of A^{-1}, mainly affecting the matrix elements that correspond to long distances in the lattice (see Fig. 1, left). After removing part of the lower spectrum, the remaining eigenvectors contribute to more localized regions of the matrix (see middle graph), for instance, among certain sub-diagonals of the matrix, that can be better addressed by partitioning (see right graph). For the same lattice $32^3 \times 64$ that hierarchical probing reduces the variance by 10-fold over the original Hutchinson method; also deflating 1000 approximate singular values gives a 100-fold total variance reduction [10]. In [4], they reported a speedup of two in cost by combining both techniques. They also considered the acceleration stemming from solving linear systems with the deflated operator, $A^{-1}(I - P)$, which is a well-known effect [11, 12, 13].

To compute the deflation space, previous efforts computed the eigenpairs of the Hermitian Dirac operator, $A\gamma_5$, usually with Chebyshev-accelerated Arnoldi [14, 4]. [10] briefly discussed alternatives to compute the deflation space. In particular, they discourage the use of methods such as eigCG [15] or Lanczos because they have to be used unpreconditioned, and thus the convergence is slow. Instead they show results using the state-of-the-art preconditioned eigensolver PRIMME [16][1] on $A^{\dagger}A$. The preconditioner employed is based on the adaptive algebraic multigrid (AMG) [13] as implemented in QOPQDP[2]. In this paper, we consider a different strategy for preconditioning that seems more effective.

Because the number of singular vectors required scales as a small percentage of the lattice volume, for large-size lattices the cost of computing and storing the singular vectors becomes unaffordable. Iterative eigensolvers, e.g., Krylov and Davidson-type, access the operator at least once for every singular vector computed, yielding a quadratic cost in terms of lattice volume[3]. Similarly, storage and the application of the deflation basis involve quadratic terms. In addition, orthogonalization of eigenvectors introduces a cubic cost on the number of singular vectors, and thus on the lattice volume, albeit with a much smaller constant because of more efficient memory access patterns.

---

[1]For an introduction to preconditioned eigensolvers see [17] and section 11 in [18].

[2]http://usqcd-software.github.io/qopqdp/

[3]The block version of the eigensolvers can reduce the number of accesses and the computational time by a constant factor, but it does not remove the quadratic term of the cost.

In LQCD, the Dirac operator becomes more ill-conditioned as the fermion mass is lighter and, as a consequence, linear system solvers take more iterations (critical slowing down). The adoption of multigrid techniques has accelerated the solution of the linear systems dramatically, not only keeping the number of iterations independent of the operator's conditioning but also making the cost of the solvers scale linearly to the lattice volume [13, 19].

In this work we propose the use of multigrid directly into reducing the variance of the trace estimator to overcome the scalability limitations of deflation. The approach is described in section 2. Although we compute the lower part of the singular spectrum of a much smaller dimension operator, its performance continues to be critical, as we show in section 3. In section 4, we evaluate the variance reduction of the new approach experimentally. We offer some conclusions in Section 5.

2. Multigrid Variance reduction

To establish notation, we show the two-grid method for solving $A\mathbf{x} = \mathbf{b}$ for a Hermitian matrix A. A two-grid iterative method alternates between the following two steps:

1. Coarse grid correction: $\mathbf{x}_i^c = \mathbf{x}_i + V(V^\dagger A V)^{-1} V^\dagger (\mathbf{b} - A\mathbf{x}_i)$
2. Smoothing: $\mathbf{x}_{i+1} = \mathbf{x}_i^c + M^{-1}(\mathbf{b} - A\mathbf{x}_i^c)$.

In step 1, the method projects the residual vector to a lower dimension (coarsening), corrects it, and prolongates it back. The matrices V and $V^\dagger$ are the multigrid's prolongator (or interpolator) and restrictor respectively. In step 2, the new residual vector $\mathbf{b} - A\mathbf{x}_i^c$ is *smoothed* by applying few iterations of a linear system solver, e.g., GCR, on A which we represent by the operator M^{-1}. The goal of the coarse grid is to correct directions that the fine grid smoothing struggles to fix. One of the most relevant distinctions between multigrid variants is the coarsening.

In adaptive multigrid for LQCD, the setup starts with the approximate solutions of $A\mathbf{y}^{(i)} = 0$ for n different random initial guesses. The returned n solution vectors, Y, are rich in directions of singular vectors with singular values close to the origin, because linear solvers have difficulty resolving these lower spaces. Interestingly, these singular vectors on the lower part of the spectrum look similar on a local lattice scale, and nearby singular vectors can be generated by combining local pieces from different singular vectors [20, 21]. Taking advantage of that, the adaptive multigrid method partitions the lattice into a set of m compact subdomains that maintain these "local" features. Then, each of the solution vectors is broken into $j = 1, \ldots, m$ vectors, each with support only on the subdomain $\mathbb{D}_j$. This process is referred to as *blocking*,

$$[Y]_D = [D_1 Y, \quad D_1 Y, \quad \ldots, \quad D_m Y], \text{ with } \mathbb{D}_i \cap \mathbb{D}_j = \emptyset \ \forall i \neq j,$$

where D_i is the projector onto the domain $\mathbb{D}_i$. For numerical stability, implementations work with an orthonormal basis V of the blocked solutions $[Y]_D$.

To see the role of deflation in multigrid, we rewrite the previous two steps in terms of the oblique projector $P = AV(V^\dagger A V)^{-1} V^\dagger$ on the space of V,

$$\mathbf{x}_{i+1} = \mathbf{x}_i + A^{-1} P(\mathbf{b} - A\mathbf{x}_i) + M^{-1}(I - P)(\mathbf{b} - A\mathbf{x}_i).$$

The coarse grid correction is interpreted as the exact solution of the components of the residual vector on P, and smoothing takes care of the remaining components of the residual vector. The deflation view suggests that, in general, if V approximates the lower part of A's spectrum, then the effective condition number of $A^{-1}(I - P)$ is smaller than that of A^{-1}, accelerating the convergence of the iterative method M^{-1}.

In the case of the trace estimation, the deflation splits the trace of A^{-1} into the trace of $A^{-1}(I - P)$ and the trace of $A^{-1}P$, so that the trace of $A^{-1}P$ can be computed efficiently, and the variance of estimating the trace of $A^{-1}(I - P)$ is far smaller than with A^{-1}. As with solving linear systems, we are interested in P approximating the lower part of A's spectrum.

2.1. Choice of projector

If P is the oblique projector defined above and V is of much smaller rank than N, the trace of $A^{-1}P$ can be computed efficiently since trace $A^{-1}AV(V^\dagger AV)^{-1}V^\dagger$ = trace $V(V^\dagger AV)^{-1}V^\dagger$ = trace$(V^\dagger AV)^{-1}$, which is the trace of a smaller matrix. If V is an invariant subspace, P becomes the orthogonal spectral projector. If V is far from an invariant space, the variance of the trace estimator $t(A^{-1}(I - P))$ is usually smaller with the orthogonal projector than with the oblique projector. Unfortunately, there is no efficient way to approximate the trace of $A^{-1}P$ if $P = VV^\dagger$, unless V is

close enough to left singular space of A so that the errors on the singular values are smaller than the standard deviation of the trace estimator. In this work, we propose the use of the oblique projector but also spend the time to compute and report the variance of $t(A^{-1}(I - P))$ for the orthogonal projector as a rough estimate of the lower bound to the variance achievable by any projector of that rank.

If A is not a Hermitian positive definite matrix, and V is not close to a left singular space, the coarse operator $V^\dagger A V$ may have singular values that are close to zero but do not correspond to singular values of A, which are known as spurious values. In addition to making $V^\dagger A V$ ill-conditioned, these small spurious values increase the variance of the $t(A^{-1}(I - P))$. One can consider variants of the oblique projector that may prevent the directions in V that are far from an eigenspace from harming the final variance,

$$\text{trace } A^{-1} = \text{trace } A^{-1}(I - AV(V^\dagger K A V)^{-1} V^\dagger K) + \text{trace } V^\dagger K V(V^\dagger K A V)^{-1}.$$

Setting $K = A^{-1}$ avoids the problems of spurious values and makes P the orthogonal projector on V. If A is not Hermitian, another obvious choice for K is $A^\dagger$, but it makes the coarse operator more expensive by introducing more nonzero elements. In the results presented in Section 4, we use $K = \gamma_5$, that does not protect against the directions in V that are far from eigenspaces and may result in small spurious eigenvalues, but it is somehow optimal as we explain in the following. For the spurious values, we supplement this technique with filtering of the prolongator space, as explained in the next section.

If A^{-1} is not Hermitian, one can reduce the variance by evaluating the trace of the symmetric operator $(A^{-1}+A^{-T})/2$ instead, which has the same trace and less or equal variance. So symmetric and Hermitian matrices are optimal for variance reduction under *symmetrization*. The spin-color dilution probing basis H transforms any γ_5-Hermitian matrix into a Hermitian, and we wish that the deflated operator be also γ_5-Hermitian. It is easy to see that the operator with the projector for $K = \gamma_5$, $A^{-1}(I - AV(V^\dagger \gamma_5 A V)^{-1} V^\dagger \gamma_5)$, is γ_5-Hermitian.

2.2. Filtering

The blocking of the approximate solutions to the n equations $A\mathbf{y}_i = 0$ (the $\mathbf{y}_i$ are often called the *null vectors*) produces approximations to a large part of the lowest singular spectrum of A. However, the quality of these approximations deteriorates rapidly beyond a few lowest singular vectors. Still, using these approximations as an orthogonal projector would effectively reduce the variance of the trace estimator. With the oblique projector, there is an optimal size for the deflation space beyond which the interior singular vectors are not accurate at all, and variance increases. We have confirmed this experimentally with Laplacian matrices and the ensembles presented in Section 4.

It is not clear how to identify this optimal size of the deflation basis *a priori*. However, this is not necessary since the application of the oblique deflation operator can be decoupled from the inversions of $A^{-1}\mathbf{z}_i$ in the Hutchinson method. Specifically, we propose to compute and store first a sufficiently large rank of the lower part of the spectrum of the coarse operator $V^\dagger A V$, say $\bar{U}\Sigma\bar{V}^\dagger$. The deflated operator in the Hutchinson method would then involve $A^{-1}(I - P)\mathbf{z}_i = A^{-1}\mathbf{z}_i - V\bar{V}(\bar{U}^\dagger V^\dagger K A V \bar{V})^{-1}\bar{U}^\dagger V^\dagger K\mathbf{z}_i$, for the entire $\bar{U}, \bar{V}$ or some column subset of them. After completing the inversions with the noise vectors $Z = \{\mathbf{z}_i, i = 1, \ldots, n\}$, the deflation part can be recomputed with different number of triplets, and the optimal rank of P can be identified. This *a posteriori* analysis can be done efficiently without storing any vector of lattice dimension, just the matrices, $Z^\dagger V\bar{V}$, and $\bar{U}^\dagger V^\dagger K Z$.

Finally, we note the difference between our approach, which computes the eigenspaces from the coarse space, and the approach in [21] where the computed fine grid eigenspace is stored compactly using the prolongator.

3. Computation of the deflation space

The γ_5-Hermiticity of the Wilson operator as well as of the coarse operators built by multigrid allows for several algorithmic alternatives for computing the smallest singular values and their vectors. First, we examine the relative performance of these alternatives. However, as the density of singular values near zero increases with lattice volume, the eigenvalue problems become particularly challenging for iterative methods, and complementary techniques must be employed.

3.1. Choice of eigensolver

A common approach to approximate the smallest singular values of A is to use some variant of the Lanczos method on either $A^\dagger A$ or $A\gamma_5$. However, without preconditioning, Lanczos converges slowly for the operators that we

Table 1. Number of inversions in computing the 10 smallest singular vectors on a Schwinger model Wilson 32^2 with several approaches. The eigensolvers mark an eigenpair $(\lambda, \mathbf{x})$ as converged when $\|\gamma_5 A^{-1}\mathbf{x} - \lambda^{-1}\mathbf{x}\| \le 10^{-2}\|A^{-1}\|$. The inverter used in all cases, A_ξ^{-1}, is QMR stopping at relative tolerance $\xi = 10^{-2}$. GD+k is from PRIMME, and Lanczos is Matlab's `eigs`, both with default settings.

Approach	# Inversions
GD+k on $A^\dagger A$ with preconditioner $A_\xi^{-1}A_\xi^{-1\dagger}$	60
GD+k on $A_\xi^{-1}A_\xi^{-1\dagger}$	34
GD+k on $A\gamma_5$ with preconditioner $\gamma_5 A_\xi^{-1}$	27
GD+k on $\gamma_5 A_\xi^{-1}$	23
Lanczos on $A_\xi^{-1}A_\xi^{-1\dagger}$	100
Lanczos on $\gamma_5 A_\xi^{-1}$	25

are working with. Alternatively, the shift-and-invert Lanczos has much faster convergence at the expense of solving a linear system of equations per (outer) iteration. We consider the inexact Lanczos variant, in which the linear systems are solved approximately with a relative tolerance ξ. The approximation also limits the accuracy of the computed eigenspace by ξ.

Besides inexact Lanczos, we also consider Generalized Davidson, a preconditioned method that offers a few more techniques for improving convergence and robustness (the GD+k method) [22]. We test both the inexact GD+k variant on the operator A_ξ^{-1} and the exact GD+k on the operator $A\gamma_5$ or $A^\dagger A$, while providing the solution of linear systems as a preconditioner instead.

Table 1 shows a comparison of the various eigensolver alternatives on a Schwinger model Wilson operator[4], a 32^2 lattice, and $\beta = 6$. The experiments are performed in MATLAB and, although the problem size is small, they are indicative of the performance of these methods on much larger problem sizes. We observe that working with $A^\dagger A$ involves more solutions of linear systems than working with $A\gamma_5$. For the latter operator, inexact Generalized Davidson performs slightly better than Generalized Davidson with preconditioner and inexact Lanczos.

3.2. Eigensolver tuning

The Dirac operators that we work with exhibit a fast decay of the singular value spectrum, so most of the eigenmethods on $\gamma_5 A_\xi^{-1}$ obtain good approximations within an average of two iterations per singular triplet. The singular triplets closest to the origin need more inversions (outer steps), and each of those inversions takes many more iterations than for more interior triplets. Therefore tuning the eigensolver and the multigrid inverter can be performed early before computing thousands of singular triplets.

Because $\gamma_5 A_\xi^{-1}$ and the corresponding coarse grid operator are indefinite Hermitian matrices, they still may suffer from many spurious eigenvalues near zero. In some cases, this may cause the inverter to stagnate at an error level well above the requested ξ. A common way to avoid stagnation is to shift the operator of the linear systems to reduce its condition number. Shifting the operator by the diagonal $i\gamma_5$ moves the eigenvalues away from the origin (see Fig 2, left), while the eigenvectors of $\gamma_5 A^{-1}$ and $(A\gamma_5 + i\tau I)^{-1}$ have the same order with respect to the distance of their eigenvalues from the origin (see Fig 2, right).

The new operator after the shifting is no longer Hermitian but is normal. One option is to modify a Hermitian eigensolver to support normal operators. Specifically, the projected Rayleigh-Ritz problem at every iteration needs to be solved with a Schur decomposition instead of a Hermitian eigensolver. Another option is to use non-Hermitian eigensolvers. We followed the former approach, making the proper changes in PRIMME. When the inverter fails, it happens in the first few iterations of the eigensolver when spurious eigenvalues near zero are present. In that case, we rebuild the multigrid inverter with a larger shift, and the eigensolver is restarted.

A challenge with any eigensolver that computes a large number of eigenpairs is the cost of orthogonalization. The number of floating-point operations associated with orthogonalization grows as $O(N\times\#SV\times\text{iterations})$. In addition, on massively parallel environments, orthogonalization involves between $O(\#SV)$ and $O(\#SV^2)$ global synchronizations depending on the algorithm. It is clear, therefore, that for sufficiently many singular values (#SV) orthogonalization

---

[4]Generated with `quantum-mg`, source code at `https://github.com/weinbe2/quantum-mg`.

Table 2. Configuration of the Dirac-Wilson operator and the multigrid solver used in the experiments. The Preconditioned settings build the multigrid for the odd-even operator for solving linear systems. The Unpreconditioned multigrid is used to obtain the singular vectors of the original operator.

Configuration	Dimension fine grid (N)	Preconditioned		Unpreconditioned		
		Blocking per level (m)	Null vectors per level (n)	Blocking per level	Null vectors per level	Dimension of coarsest
$32^3 \times 64$ $\beta = 6.3$ $\mu = -0.2390$	25,165,824	4^4, 2^4	24, 32	4^4 4^4, 2^4 4^4 4^4, 2^4	24 24, 32 48 48, 64	393,216 32,768 786,432 65,536
$64^3 \times 128$ $\beta = 6.3$ $\mu = -0.2416$	402,653,184	4^4, $2^3 \times 4$	24, 32	4^4 4^4, 2^4	48 48, 64	12,582,912 1,048,576

Table 3. Running times for computing the partial singular decomposition on Wilson operators and coarse operators listed on Table 2.

Configuration	# Inversions	Inverter time	Ortho. time	Total time
Wilson $32^3 \times 64$ on 22 nodes:				
1024 SV on A $\tau = 0.00$	1,429	655	254	957
nv 24 & 1024 SV on coarse op. $\tau = 0.04$	7,544	1,770	16	1,797
nv 24, 32 & 1024 SV on coarse op. $\tau = 0.04$	6,664	90	2	97
nv 48 & 1024 SV on coarse op. $\tau = 0.04$	7,794	8,570	92	8,826
nv 48, 64 & 1024 SV on coarse op. $\tau = 0.04$	7,725	193	22	364
Wilson $64^3 \times 128$ on 256 nodes:				
1024 SV on A $\tau = 0.02$	3,821	5,874	860	6,952
nv 48, 64 & 1024 SV on coarse op. $\tau = 0.02$	6,675	586	119	807
nv 48, 64 & 4096 SV on coarse op. $\tau = 0.02$	21,865	1,919	1,036	3,122

time eventually dominates the eigensolver performance for LQCD. We demonstrate the performance of the GD+k eigensolver on the configurations of two large lattices shown in Table 2.

In Table 3 we can see that orthogonalization time starts to dominate for the coarse operator of the Wilson $64^3 \times 128$ case. Going from 1024 to 4096 singular triplets, the number and the cost of inversions increases by a factor of 3.3. However, the orthogonalization time increases by a factor of 8.8, making it a third of the total computation time. Block variants of iterative methods, including GD+k, can utilize better optimized linear algebra library functions and thus may reduce the orthogonalization time by a small factor. However, orthogonalization remains the scalability bottleneck.

The motivation to find the deflation space from a coarse operator is twofold: compute a much larger space and do all the operations on matrices of much smaller dimension. However, the benefits are not commensurate. First, the coarse level operators $V^\dagger A V$ are typically much denser in nonzeros per row than A, so despite a significant reduction in the dimension of the matrix, the time savings are moderate. For example, consider the $32^3 \times 64$ lattice in Table 2. The first coarse level (24 null vectors) is 12 times smaller than the fine level. Yet, the cost per inversion is only half. Second, although the floating-point operations in orthogonalization reduce linearly with the dimension, the number of processors and synchronizations remains the same. Therefore the parallel performance of the orthogonalization kernel deteriorates and its time does not reduce. In the examples of Table 2, the matrix of the coarsest levels (with 48, 64 null vectors) is 384 times smaller than the fine level. However, when solving for 1024 singular vectors, the orthogonalization times per iteration reduce only by factors of 62 and 12 for the lattices $32^3 \times 64$ and 64×128 respectively.

To conclude, even though the inverter on A spends most of the time on the coarsest level, the eigensolver on the coarse operator spends significantly less time on the inverter.

4. Experimental Performance of the Estimator

Our benchmark deflation test is the use of the orthogonal projector of the approximate left singular vectors of A, computed with an inexact eigensolver as described in Section 3 and Algorithm 1 below. Our proposed method uses an
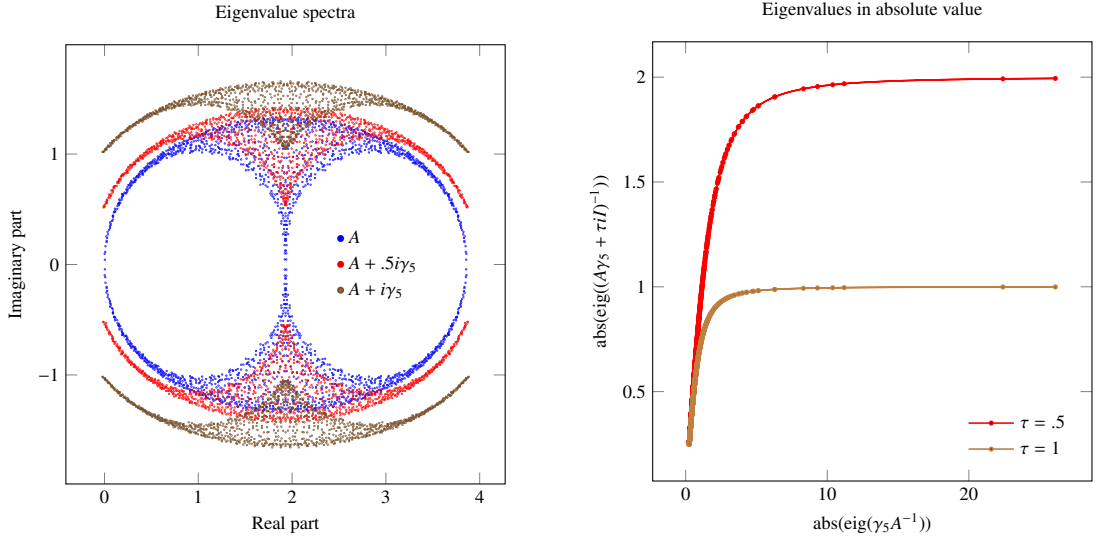
Fig. 2. Eigenvalue spectrum after shifting a Schwinger model Wilson 32^2, A, with $\tau i\gamma_5$ for values of τ of 0.5 and 1 (left). Transformation of the eigenvalue spectrum of the inverse of the Hermitian operator, $\gamma_5 A^{-1}$, after shifting, $(A\gamma_5 + \tau iI)^{-1}$ (right).

oblique projector to deflate the subspace of a multigrid prolongator that corresponds to the lower part of the singular spectrum of the coarse operator as described in Algorithm 2.

Input: Operator, A; deflation space rank, k; tolerance, ξ; shift, τ; partition basis, H

1 Compute deflation space: $U = \text{eigs}(\gamma_5(A + i\tau\gamma_5)_\xi^{-1}, \text{largest magnitude}, k, \xi)$

2 Compute the deterministic part: $t_0 = \sum_i \mathbf{u}_i^\dagger A_\xi^{-1} \mathbf{u}_i$

3 Compute the stochastic part: $t_1 = t(A_\xi^{-1}(I - UU^\dagger) \odot HH^\dagger)$

4 Return $t_0 + t_1$

Algorithm 1: Variance reduction for trace estimation using an orthogonal projector of approximate singular vectors of A

Input: Operator, A; deflation space rank, k; tolerance, ξ; shift, τ; partition basis, H

1 Compute multigrid prolongator W; $\gamma_5 W = W\gamma_5^W$

2 Compute deflation: $\bar{U} = \text{eigs}(\gamma_5^W(W^\dagger AW + i\tau\gamma_5^W)_\xi^{-1}, \text{largest magnitude}, k, \xi)$

3 Factorize $\hat{U}\Lambda\hat{U}^\dagger = \bar{U}^\dagger\gamma_5^W W^\dagger AW\bar{U}$, and set $U = \bar{U}\hat{U}$

4 Compute the direct part: $t_0 = \sum_i \mathbf{u}_i^\dagger \gamma_5 \mathbf{u}_i \lambda_i^{-1}$

5 Compute the stochastic part: $t_1 = t(A_\xi^{-1}(I - AW\gamma_5 U\Lambda^{-1}U^\dagger W^\dagger) \odot HH^\dagger)$

6 Return $t_0 + t_1$

Algorithm 2: Variance reduction for trace estimation using an oblique projector from approximate singular subspace of the coarse operator

---

The implementation is developed within QOPQDP, a library in the USQCD suite which provides several operators and linear system solvers. We only use the Wilson clover operator and the multigrid even-odd solver. We instantiate two multigrid variants as explained in the caption of Table 2. The preconditioned multigrid variant is based on the odd-even operator and is used for solving the linear systems in the stochastic approximation of the trace, and also for the inversions of the inexact eigensolver of Algorithm 1. The unpreconditioned multigrid variant is used for generating the prolongators and the coarse operators from which the deflation space will be computed in Algorithm 2. At coarse levels, the solution of the linear systems is not accelerated with a multigrid recursion. Instead, even-odd preconditioned GCR is employed. We extended the implementation of QOPQDP solvers to support the implicit $i\tau\gamma_5$ shifting of the operator and the access to the coarse operators from PRIMME.
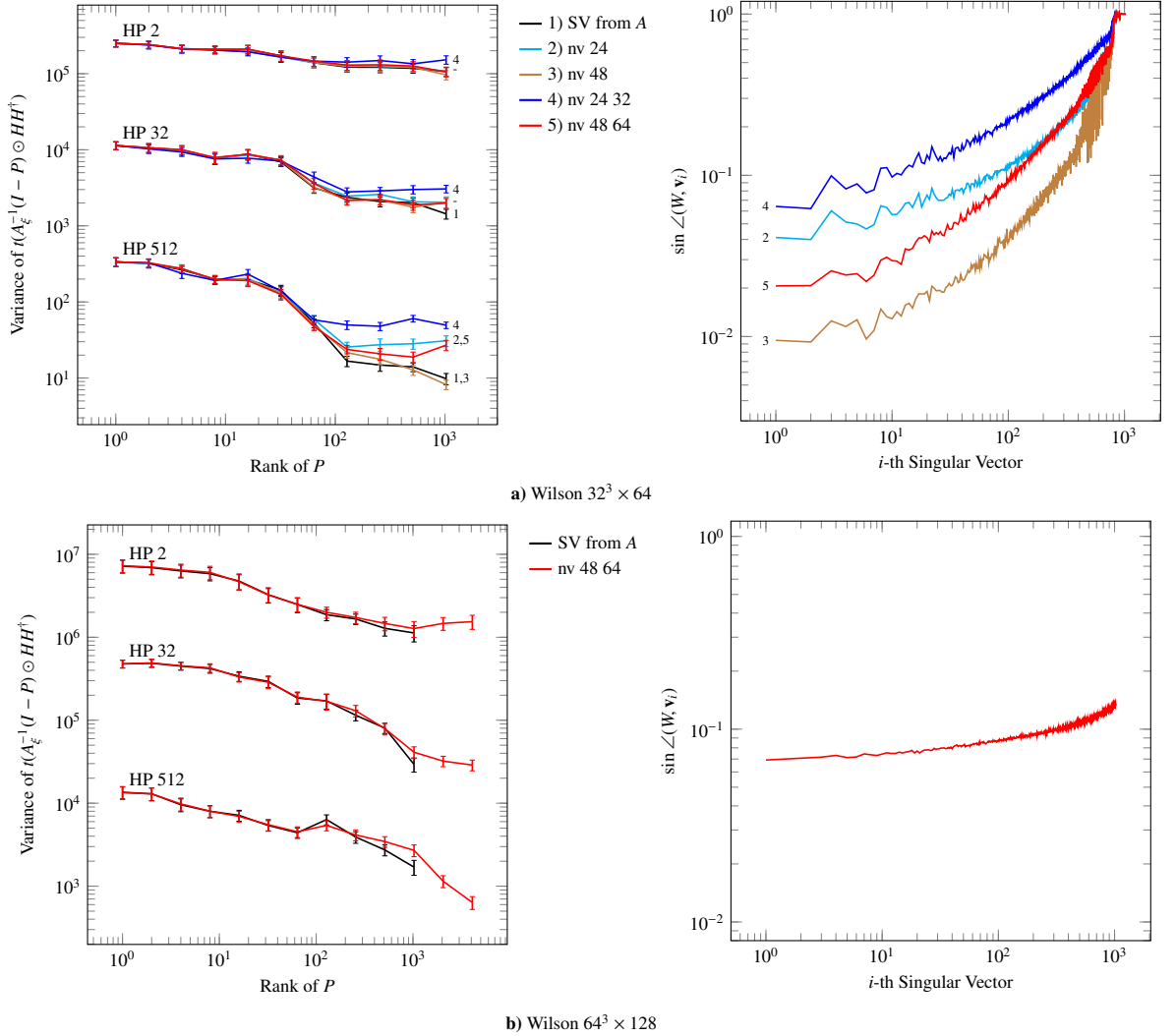
Fig. 3. Experimental estimation of the variance of $t(A_\xi^{-1}(I - P) \odot HH^\dagger)$ as the rank of the deflation projector P increases (left), and the sine of the angles between the approximate singular vectors from the operator and the subset of the prolongators used for deflation (right), for Wilson $32^3 \times 64$ (top) and Wilson $64^3 \times 128$ (bottom). Error bars indicate the jackknife error. The tolerance is $\xi = 10^{-2}$.

For the stochastic estimation step, $t(A_\xi^{-1}(I - P) \odot HH^\dagger)$, our implementation follows a blocked approach in which the 12 vectors of spin-color dilution of each hierarchical probing vector $\mathbf{h}_i$ are processed together. This can potentially accelerate the solution of the linear systems, and the application of the projectors for Algorithm 1, where the orthogonal projectors can employ optimized BLAS matrix-matrix multiplication routines.

Tuning the prolongators configuration is crucial to get good variance reduction. In the examples tested, we found that the number of null vectors used for setting up the multigrid levels should be twice as many for Algorithm 2 than the ones used only to solve linear systems. For variance reduction, we configure the solver used on the null vectors to stop at a relative residual norm tolerance of 10^{-2} or after 200 iterations.

4.1. Variance Reduction

We show that deflating with a prolongator subspace reduces the variance of the trace estimator, and the effectiveness is related to how close the prolongator subspace is to the lower part of the operator's spectrum. As a base case for comparison, we consider the undeflated matrix after 2 hierarchical probing (HP) vectors have been applied with full spin-color dilution, i.e., 24 inversions. The best results are obtained after 512 HP vectors with the maximum number of vectors deflated.

Table 4. Times on NERSC's Edison estimating the trace from 2 noise vectors ($s = 2$) solving the linear systems with a tolerance $\xi = 10^{-2}$ using a partitioning of full spin-color dilution and 512 hierarchical proving vectors. Approaches detailed in algorithms 1 and 2 are compared.

Computing deflating space P		Computing $t(A_\xi^{-1}(I - P) \odot HH^\dagger)$		Variance
		Applying P	Solving LS	
Wilson $32^3 \times 64$ on 22 nodes:				
1024 SV on A	957s	353s	2,400s	10
nv 24 & 1024 SV on coarse op.	269s	297s	2,400s	35
nv 24 32 & 1024 SV on coarse op.	97s	303s	2,400s	55
nv 48 & 1024 SV on coarse op.	5489s	318s	2,400s	8
nv 48 64 & 1024 SV on coarse op.	430s	326s	2,400s	18
Wilson $64^3 \times 128$ on 256 nodes:				
1024 SV on A	6,952s	725s	5,903s	1,700
nv 48 64 & 4096 SV on coarse op.	3,122s	1,126s	5,903s	640

In Fig. 3(a) we show results on the Wilson $32^3 \times 64$ lattice. Deflating the 1024 smallest singular triplets computed with Algorithm 1 with tolerance $\xi = 10^{-2}$ and 512 HP, spin-color diluted vectors reduces the variance more than four orders of magnitude. The top-left graph in Fig. 3(a) shows the effect of prolongators with the four different configurations listed in Table 2. For the HP2 and HP32 cases, deflation does not have as much of an effect because the nearest neighbor connections that HP has not removed yet dominate. For 512 HP vectors, we see that deflation from the first coarse level is very close to the benchmark fine grid deflation (curves 1 and 3). Deflating with singular spaces from the second level is also close to the benchmark up to 100 singular vectors but deflating with more does not have an additional benefit.

An explanation for the limited benefit at the coarsest level is seen in the top-right graph of Fig. 3 that shows the sines of the angles between the computed singular vectors of A and the prolongator subspaces. The prolongators with smaller angles tend to perform better. The graph also shows how fast these angles increase in the interior of the spectrum. All prolongator subspaces have almost no component to singular spaces after the 1000th lowest one, and we should not expect a further reduction on the variance estimation by increasing the rank. Finally, notice that prolongators generated with more null vectors (curves 3 and 5) yield a smaller variance.

In Fig. 3(b) we show a similar experiment on the Wilson $64^3 \times 128$ lattice. We limit the number of computed triplets from the original operator to 1024 due to memory limitations on Edison's nodes, but we let the multigrid deflation use up to a rank of 4096. As in the previous case, both deflation spaces have a similar effect on the variance for the same rank. Deflation achieves a speedup of 20 over the undeflated HP512 experiment.

The second operator is 16 times larger than the first, so it expected that many more vectors are needed to get a similar deflation effect as on the first operator. At the same time, the prolongator subspace from the second coarse level of the second operator has more components on the inner part of the spectrum. This fact may indicate that the effectiveness of the multigrid deflation also scales with the operator dimensions. Further experiments are required to confirm these claims.

Finally, we give a rough estimate on the speedup of these techniques over performing simply a Hutchinson method with random vectors and no HP or dilution. Both experiments start at 24 inversions and obtain a little more than 4 orders of magnitude improvements with 512*12 = 6144 inversions. This corresponds to speedups of about 50-100 over the Hutchinson method.

4.2. Performance

Table 4 reports the timings for the computational kernels involved in estimating the trace with just two noise vectors. The most expensive operations are the computation of the deflation space and the inversions in the stochastic estimation of the trace. Notice that applying the deflation for approximate singular vectors from the original matrix is usually cheaper than applying the oblique projector in multigrid deflation, even though the latter prolongator requires less storage. This is because the oblique projector involves a multiplication with the original operator, and because the orthogonal projector is implemented in a block fashion, which is ten times faster than applying the deflation vector by vector. Using a block version of the oblique projectors, i.e., applying the operator and the prolongators on multiple vectors at once, may also accelerate the computation significantly.

5. Summary

In this work, we show that, in the context of reducing variance in the stochastic evaluation of disconnected loops in LQCD, inexact eigensolvers are efficient in approximating the lower part of the singular value spectrum when a fast linear system solver is available. The Dirac operator can be shifted if the problem is too ill-conditioned. However, computing and storing a fraction of the lattice volume of the singular vectors is computationally prohibitive.

We show that the prolongators generated by adaptive multigrid solvers without even-odd preconditioning have limited accuracy on the lower part of the spectrum, and the accuracy degrades quickly into the inner part of the spectrum. However, the decay in the accuracy may be slower on larger Dirac operators. In addition, proper tuning of the multigrid configuration can give enough accuracy to make the prolongators suitable for reducing the variance of trace estimators on the ensembles we have tested and with significant computational gains over the fine-grid deflation. As noted before, for large matrices, the speedup over plain Hutchinson method without hierarchical probing can reach two orders of magnitude.

Acknowledgements

References

[1] J. M. Tang, Y. Saad, A probing method for computing the diagonal of a matrix inverse, Numerical Linear Algebra with Applications 19 (2012) 485–501.

[2] L. Wu, J. Laeuchli, V. Kalantzis, A. Stathopoulos, E. Gallopoulos, Estimating the trace of the matrix inverse by interpolating from the diagonal of an approximate inverse, Journal of Computational Physics 326 (2016) 828 – 844.

[3] M. Hutchinson, A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines, Communications in Statistics - Simulation and Computation 19 (1990) 433–450.

[4] G. S. Bali, S. Collins, A. Schfer, Effective noise reduction techniques for disconnected loops in lattice qcd, Computer Physics Communications 181 (2010) 1570 – 1583.

[5] S. Bernardson, P. McCarty, C. Thron, Monte Carlo methods for estimating linear combinations of inverse matrix entries in lattice QCD, Computer Physics Communications 78 (1994) 256 – 264.

[6] J. Foley, K. J. Juge, A. O. Cais, M. Peardon, S. M. Ryan, J.-I. Skullerud, Practical all-to-all propagators for lattice qcd, Computer Physics Communications 172 (2005) 145 – 162.

[7] R. Babich, R. C. Brower, M. A. Clark, G. T. Fleming, J. C. Osborn, C. Rebbi, D. Schaich, Exploring strange nucleon form factors on the lattice, Phys. Rev. D 85 (2012) 054510.

[8] A. Stathopoulos, J. Laeuchli, K. Orginos, Hierarchical probing for estimating the trace of the matrix inverse on toroidal lattices, SIAM Journal on Scientific Computing 35 (2013) S299–S322.

[9] J. Green, S. Meinel, M. Engelhardt, S. Krieg, J. Laeuchli, J. Negele, K. Orginos, A. Pochinsky, S. Syritsyn, High-precision calculation of the strange nucleon electromagnetic form factors, Phys. Rev. D 92 (2015) 031501.

[10] A. S. Gambhir, A. Stathopoulos, K. Orginos, Deflation as a method of variance reduction for estimating the trace of a matrix inverse, SIAM Journal on Scientific Computing 39 (2017) A532–A558.

[11] A. Stathopoulos, K. Orginos, Computing and deflating eigenvalues while solving multiple right-hand side linear systems with an application to quantum chromodynamics, SIAM Journal on Scientific Computing 32 (2010) 439–462.

[12] A. Abdel-Rehim, R. Morgan, D. Nicely, W. Wilcox, Deflated and restarted symmetric lanczos methods for eigenvalues and linear equations with multiple right-hand sides, SIAM Journal on Scientific Computing 32 (2010) 129–149.

[13] R. Babich, J. Brannick, R. C. Brower, M. A. Clark, T. A. Manteuffel, S. F. McCormick, J. C. Osborn, C. Rebbi, Adaptive multigrid algorithm for the lattice Wilson-Dirac operator, Phys. Rev. Lett. 105 (2010) 201602.

[14] H. Neff, N. Eicker, T. Lippert, J. W. Negele, K. Schilling, Low fermionic eigenmode dominance in qcd on the lattice, Physical Review D 64 (2001) 114509.

[15] A. M. Abdel-Rehim, A. Stathopoulos, K. Orginos, Extending the eigCG algorithm to nonsymmetric Lanczos for linear systems with multiple right-hand sides, Numerical Linear Algebra with Applications 21 (2014) 473–493.

[16] A. Stathopoulos, J. R. McCombs, PRIMME: PReconditioned Iterative MultiMethod Eigensolver: Methods and software description, ACM Transactions on Mathematical Software 37 (2010) 21:1–21:30.

[17] A. V. Knyazev, Preconditioned eigensolversan oxymoron, Electron. Trans. Numer. Anal 7 (1998) 104–123.

[18] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, H. van der Vorst, Templates for the Solution of Algebraic Eigenvalue Problems, Society for Industrial and Applied Mathematics, 2000. doi:10.1137/1.9780898719581.

[19] J. Brannick, R. C. Brower, M. Clark, J. C. Osborn, C. Rebbi, Adaptive multigrid algorithm for lattice qcd, Physical review letters 100 (2008) 041601.

[20] M. Lüscher, Local coherence and deflation of the low quark modes in lattice qcd, Journal of High Energy Physics 2007 (2007) 081.

[21] K. Clark, C. Jung, C. Lehner, Multi-grid lanczos, PoS, The 35th International Symposium on Lattice Field Theory, LAT2017 175, 14023 (2017).

[22] A. Stathopoulos, Nearly optimal preconditioned methods for Hermitian eigenproblems under limited memory. Part I: Seeking one eigenvalue, SIAM Journal on Scientific Computing 29 (2007) 481–514.