Multi-Grid Lanczos

M. A. Clark[1], *Chulwoo* Jung[2], and *Christoph* Lehner[2],[*]

[1] *NVIDIA Corporation, 2701 San Tomas Expressway, Santa Clara, CA 95050, USA*
[2] *Physics Department, Brookhaven National Laboratory, Upton, NY 11973, USA*

Abstract. We present a Lanczos algorithm utilizing multiple grids that reduces the memory requirements both on disk and in working memory by one order of magnitude for RBC/UKQCD's 48I and 64I ensembles at the physical pion mass. The precision of the resulting eigenvectors is on par with exact deflation.

1 Introduction

In recent years RBC/UKQCD has benefited significantly from the generation of the 2000 lowest eigenvectors of the preconditioned normal (z)Mobius Domain Wall Fermion Dirac operator for the light quarks on the 48I (a^{-1} = 1.7 GeV) and 64I (a^{-1} = 2.3 GeV) ensembles at near physical pion mass. These eigenvectors were used for deflation and volume averages over the low-mode space and were a key ingredient in the on-going $g - 2$ projects [1–3]; they have also found additional use in the calculation of ΔM_K [4].

The storage cost for these vectors is substantial with 9.3 TB and 36 TB per configuration for the 48I and 64I ensembles respectively. These high storage requirements both on disk and in RAM are addressed in this contribution allowing for usage of these methods at even larger volumes. Our approach makes deflation much more applicable to architectures with limited amounts of high-bandwidth memory such as GPUs and allows for running on small-scale clusters.

We note that related ideas were recently successfully used in the context of Monte-Carlo estimation of the trace of a matrix inverse [5].

2 Eigenvector compression

We first explore the compression of existing eigenvectors computed with a Chebyshev-accelerated implicitly restarted Lanczos (IRL) on the original lattice. To this end, we create a spatially-blocked basis out of the lowest N modes and write all eigenmodes in this basis [6]. For the figures shown below, we have used $N = 400$ for the 48I ensemble and $N = 250$ for the 64I ensemble. The blocking allows us to create a coarse-grid representation of the eigenmodes. Figs. 1 and 2 illustrate the efficacy of this blocking for the eigenvector compression. The squared relative error is the squared norm of the difference of original and reconstructed vector divided by the squared norm of the original vector. In all cases shown here, we only have a single block in the fifth dimension.
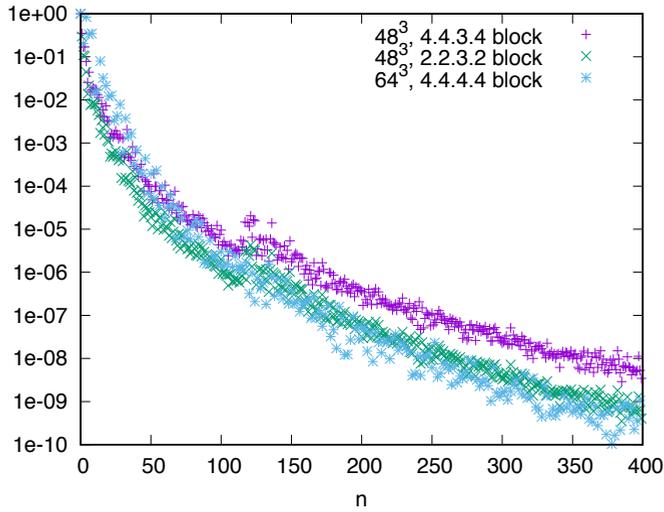
---

[*]Speaker, e-mail: clehner@bnl.gov

Figure 1. Squared relative error of eigenvector $n + 1$ when reconstructed from blocked eigenvectors $1, \ldots, n$.
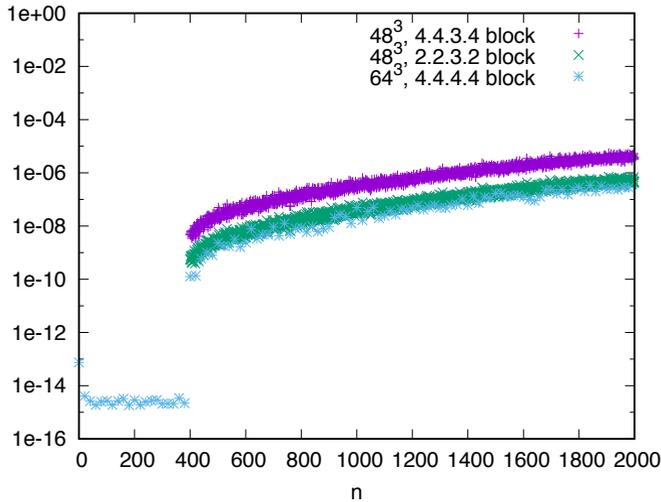

Figure 2. Squared relative error of eigenvector n when reconstructed from blocked eigenvectors $1, \ldots, 400$.

We furthermore reduce storage cost by expressing the eigenvectors in terms of a two-byte fixed-precision representation, where all spin-color elements for a given five-dimensional position share a common two-byte exponent. We use a single-precision representation for the first 100 basis vectors and this two-byte representation for $101, \ldots, N$ to reduce precision loss, see Fig. 3.

In the following, we show that the precision loss from this compression technique is minimal. The effects on a sloppy CG solve and on a full low-mode volume average are negligible, see Figs. 4 and 5.
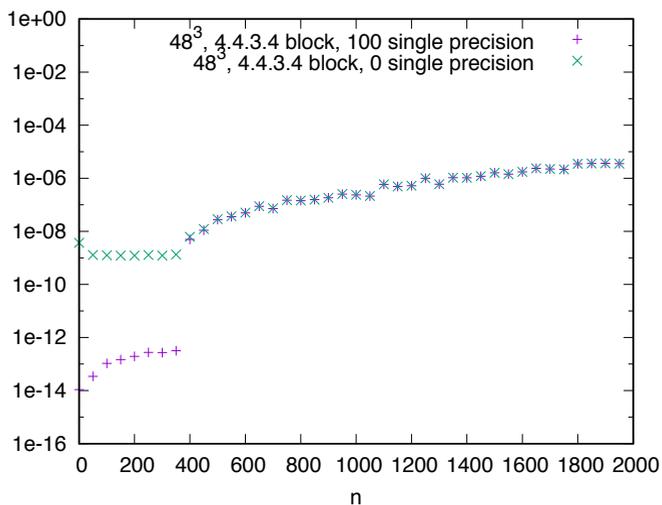
Figure 3. Effect of keeping the first 100 basis vectors in single precision instead of keeping all vectors in two-byte fixed-point precision.
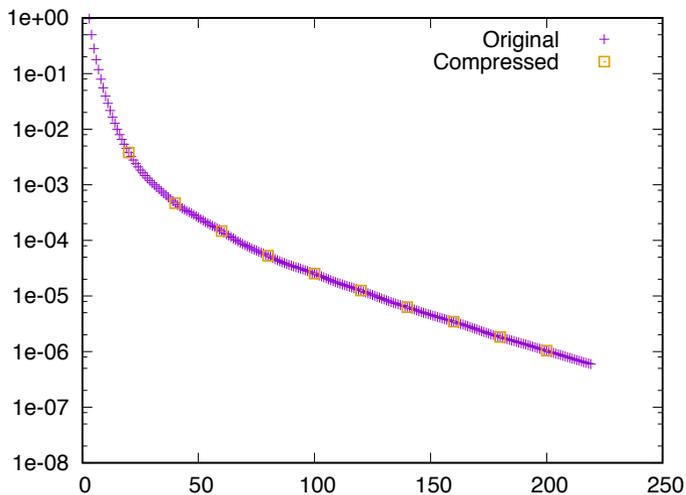

Figure 4. Squared CG residual as function of iteration number for point source on 48I ensemble.

In case of a single point source the effects become visible at long distances, see Fig. 6, however, are sufficiently small that the statistical advantage of a low-mode subtraction is not reduced.

The numerical experiments presented here were performed with an open-source stand-alone compression tool that is available at Ref. [7].
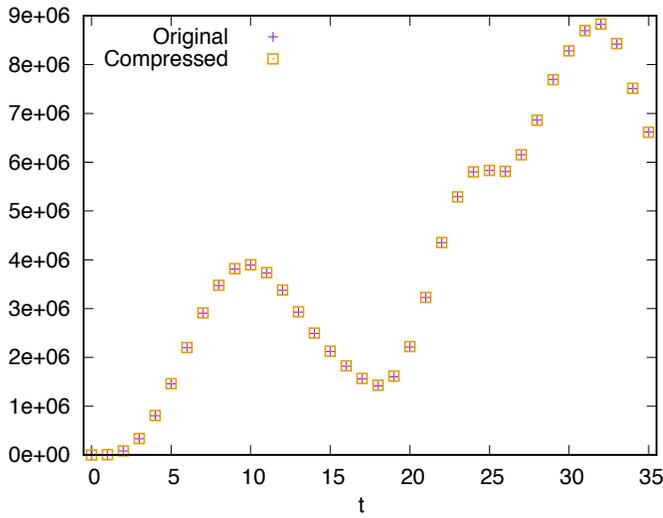
Figure 5. γ_0–γ_0 correlator $C(t)$ times t^4 on 48I ensemble for the full low-mode average on a single configuration.
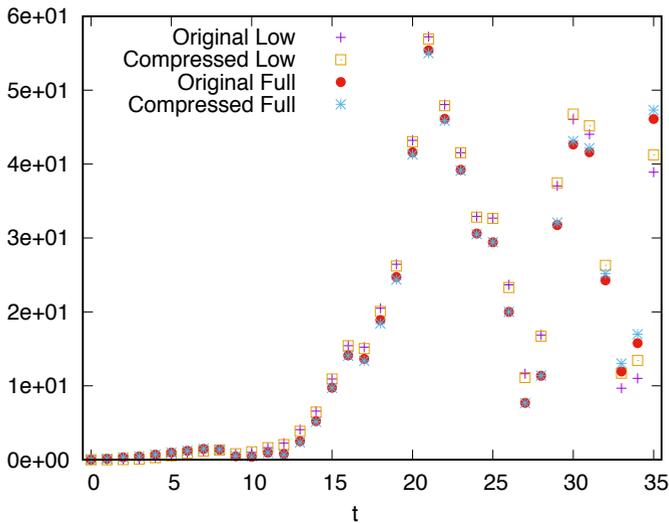

Figure 6. γ_0–γ_0 correlator $C(t)$ times t^4 on 48I ensemble and its low-mode approximation for a single point source on a single configuration.

3 Multi-Grid Lanczos

In this section we demonstrate that we can also generate the eigenvector data directly in its compressed representation. To this end, we have developed a Multi-Grid Lanczos method that is now publicly available at Ref. [8].
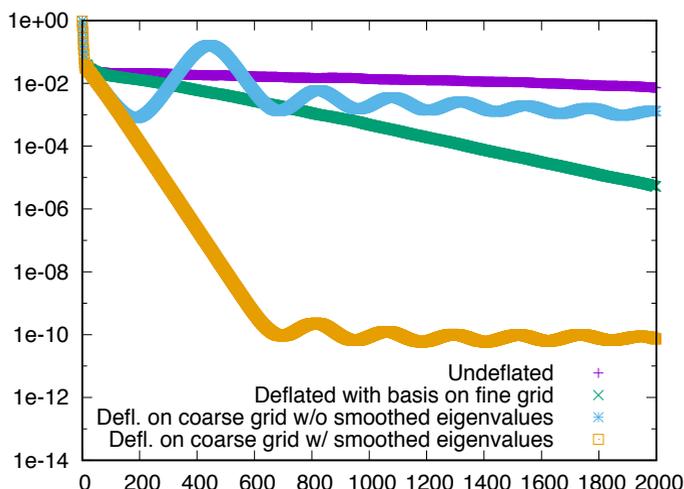
The basic steps are as follows:

Figure 7. Squared CG residual as function of iteration number for volume source on 64I ensemble using Multi-Grid Lanczos.

1. Compute the N basis vectors with a first round of Chebyshev-accelerated IRL. We have found significant precision benefits by creating a precise basis through the Lanczos algorithm compared to the use of an imprecise basis. The use of other methods such as the Jacobi-Davidson iteration to create the basis is currently being investigated.

2. For a given blocking, create a locally orthogonal basis using the results of step 1. This defines the mapping between coarse and fine grid.

3. Solve a second round of Chebyshev-accelerated IRL on the coarse Grid to obtain the full set of eigenvectors.

4. Reconstruct an approximation of the eigenvalues by locally inverting the Chebyshev polynomial of the Lanczos eigenvalues.

5. The first eigenvalues outside of the basis $N + 1, N + 2, \ldots$ may lack sufficient precision which we correct by smoothening the corresponding eigenvectors (currently with low-iteration CG) and then determining the precise fine-grid eigenvalues.

The importance of precise fine-grid eigenvalues is illustrated in Fig. 7.

4 Summary

By using both local coherence of eigenvectors [6] and a two-byte fixed-precision representation of eigenvectors we are able to reduce the memory footprint of the 48I eigenvectors by 85%, from 9.3 TB to 1.4 TB, and of the 64I eigenvectors by 90%, from 36 TB to 3.5 TB. Both a stand-alone compression tool [7] and a Multi-Grid Lanczos implementation [8] are available.

References

[1] T. Blum, N. Christ, M. Hayakawa, T. Izubuchi, L. Jin, C. Lehner, Phys. Rev. **D93**, 014503 (2016), `1510.07100`

[2] T. Blum, P.A. Boyle, T. Izubuchi, L. Jin, A. Jüttner, C. Lehner, K. Maltman, M. Marinkovic, A. Portelli, M. Spraggs, Phys. Rev. Lett. **116**, 232002 (2016), `1512.09054`

[3] T. Blum, N. Christ, M. Hayakawa, T. Izubuchi, L. Jin, C. Jung, C. Lehner, Phys. Rev. Lett. **118**, 022005 (2017), `1610.04603`

[4] Z. Bai, N.H. Christ, T. Izubuchi, C.T. Sachrajda, A. Soni, J. Yu, Phys. Rev. Lett. **113**, 112003 (2014), `1406.0916`

[5] E.R. Alcalde, A. Stathopoulos, *Deflation for Monte-Carlo estimation of the trace of a matrix inverse*, QCDNA X, Coimbra, Portugal, June 27-29, 2017.

[6] M. Luscher, JHEP **07**, 081 (2007), `0706.2298`

[7] C. Lehner, `https://github.com/lehner/eigen-comp/`

[8] The Multi-Grid code is available at `https://github.com/lehner/Grid/` which adds this feature to Peter Boyle's Grid library `https://github.com/paboyle/Grid/`.