Multigrid Multilevel Monte Carlo for Efficient Trace Estimation in Lattice QCD Simulations

Jose Jimenez-Merchan, A. Frommer, F. Knechtli, G. Ramirez-Hidalgo, J.A. Urrea-Nino

Bergische Universität Wuppertal

FOR 5269

BERGISCHE UNIVERSITÄT WUPPERTAL

Challenge: Variance reduction on trace estimation.

A slide on deflation methods.

Our main tool: Algebraic multigrid (AMG).

Our approach: Multigrid multi level monte carlo

Numerical Experiments

Conclusions and outlook

Outline

Challenge: Variance reduction on trace estimation.

A slide on deflation methods.

Our main tool: Algebraic multigrid (AMG).

Our approach: Multigrid multi level monte carlo

Numerical Experiments

Conclusions and outlook

Trace Estimation in LQCD.
The Hutchinson's method.

Trace estimation is an ubiquitous problem in Lattice QCD.
Given $A \in \mathbb{C}^{n\times n}$, large and sparse. $(10^4 < n < 10^8)$

Estimate the trace $\hat{\text{tr}}(A^{-1}) \approx tr(A^{-1})$.

With **Hutchinson's method** we get an

Estimate

$$\hat{\text{tr}}(A^{-1}) = \frac{1}{N}\sum_i \left(x^{(i)}\right)^\dagger A^{-1} x^{(i)}. \quad (1)$$

With random vectors following the distribution:

$$x \in \mathbb{R}^n : x_i \in \{-1, 1, -i, i\} \quad \text{with equal probability } \frac{1}{4}. \quad (2)$$

The problem of the variance

Estimate

$$\hat{\mathrm{tr}}(A^{-1}) = \frac{1}{N}\sum_i \left(x^{(i)}\right)^\dagger A^{-1} x^{(i)}.$$

Variance of the Estimate*

$$\mathbb{V}[\hat{\mathrm{tr}}(A^{-1})] = \frac{1}{2N}\overbrace{||\text{offdiag}(A^{-1})||_F^2}^{\mathbb{V}[\left(x^{(i)}\right)^\dagger A^{-1}x^{(i)}]}. \qquad (3)$$

In order to achieve an accuracy ϵ,

if $||...||_F^2$ is large $\rightarrow N$ is large.

*See [1] and [2].

Outline

Challenge: Variance reduction on trace estimation.

A slide on deflation methods.

Our main tool: Algebraic multigrid (AMG).

Our approach: Multigrid multi level monte carlo

Numerical Experiments

Conclusions and outlook

Tackling the variance in trace estimation.

Our work focuses on tackling the problem of large variances in the estimation of **the estimate** $\hat{\mathrm{tr}}(A^{-1})$.

Due to the relation between $|| \cdot ||_F^2$ and σ_i,

$$\mathbb{V}[\hat{\mathrm{tr}}(A^{-1})] = \frac{1}{2N}||\mathsf{offdiag}(A^{-1})||_F^2 = \frac{1}{2N}\left(\sum_{i=1}^{n}\frac{1}{\sigma_i^2} - \sum_{i=1}^{n}|A_{ii},|^{-2}\right)$$

we can **deflate** to get rid of the **smallest singular values** of the operator A to reduce the variance.

Deflation in a nutshell

In our context, the essence of **Deflation** [3] can be summarized as follows:

- **Idea:** Remove a part of the operator that contributes to most of the Variance.
- **How?** Split the trace in two terms:

$$\mathrm{tr}(A^{-1}) = \mathrm{tr}((I - \Pi)A^{-1}) + \mathrm{tr}(\Pi A^{-1}). \quad (4)$$

- **Stochastic term** with small variance (Hopefully).
- **Direct term**. Cheap to compute.

So one can be creative and smart when constructing Π.

Outline

Challenge: Variance reduction on trace estimation.

A slide on deflation methods.

Our main tool: Algebraic multigrid (AMG).

Our approach: Multigrid multi level monte carlo

Numerical Experiments

Conclusions and outlook

Our main tool: AMG solver

Estimate

$$\hat{\mathrm{tr}}(A^{-1}) = \frac{1}{N} \sum_i \left(x^{(i)}\right)^H A^{-1} x^{(i)}. \tag{5}$$

Our main deflation tool: AMG solver

Our Multigrid: DD-αAMG: [8]
Domain Decomposition
Aggregation-Based αdaptive
Algebraic Multigrid

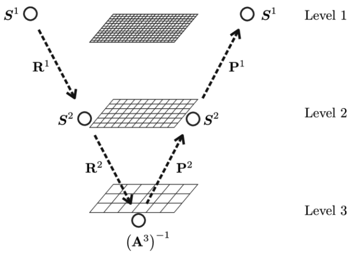


Figure: Picture 2

- **Coarse-grid operator**:

$$A_c = RAP, \quad R = P^\dagger.$$

- $\mathrm{range}(P)$ **contains the near kernel, i.e., many low modes**.
- Coarser grid $\rightarrow$ cheaper *solves*.

Outline

Our deflation methods: Deflation space for "free"

Mapping Multigrid to Multilevel Monte Carlo (a 2lvl method)

- Use $\mathrm{range}(P)$ since it contains the low modes of A.
- The inverse can be split as

$$A^{-1} = \underbrace{(A^{-1} - PA_c^{-1}P^\dagger)}_{M} + PA_c^{-1}P^\dagger.$$

- Regard $PA_c^{-1}P^\dagger$ as an approx. of A^{-1} on coarser grid.
- Hence, expect cancellation of the *problematic* modes and hence reduce the variance.
- Stochastically:

$$\hat{\mathrm{tr}}(A^{-1}) = \frac{1}{N}\sum_{i=1}^{N}\left[(x^{(i)})^\dagger M (x^{(i)})\right] + \mathrm{tr}(A_c^{-1}).$$

We can do the same at all Multigrid levels!

Expand so you cover all your MG levels: **MGMLMC**

$$\widehat{\mathrm{tr}}(A^{-1}) = \sum_{l=1}^{L-1} \widehat{\mathrm{tr}}(M_l) + \widehat{\mathrm{tr}}(A_L^{-1}).$$

with

$$M_l = \underbrace{A_l^{-1} - P_l A_{l+1}^{-1} P_l^{\dagger}}_{\text{Deflating L-1 levels}}$$

This is an oblique projection...

We can do the same at all Multigrid levels!

Expand so you cover all your MG levels: **MGMLMC**

$$\widehat{\mathrm{tr}}(A^{-1}) = \sum_{l=1}^{L-1} \widehat{\mathrm{tr}}(M_l) + \widehat{\mathrm{tr}}(A_L^{-1}).$$

with

$$M_l = \underbrace{A_l^{-1} - P_l A_{l+1}^{-1} P_l^{\dagger}}_{\text{Deflating L-1 levels}}$$

This is an oblique projection...

- Inversions are cheaper on coarser grid.

- Local coherence leads to larger deflation space.

- Other techniques require eigensolver to get deflation vectors.

- P is at hand from Multigrid solver.

Introducing an orthogonal Π

We can create an orthogonal projector Π from P. We get an

Orthogonal term $\quad O_l = (I - P_l P_l^\dagger) A_l^{-1} (I - P_l P_l^\dagger)$

When introducing it, a new term arises,

Full Rank term $\quad F_l = P_l^\dagger A_l^{-1} P_l - A_{l+1}$

so we call this method
Split MGMLMC

$$\widehat{\mathrm{tr}}(A^{-1}) = \sum_{l=1}^{L-1} \widehat{\mathrm{tr}}(O_l) \quad + \sum_{l=1}^{L-1} \widehat{\mathrm{tr}}(F_l) \quad + \widehat{\mathrm{tr}}(A_L^{-1}).$$

Outline

Challenge: Variance reduction on trace estimation.

A slide on deflation methods.

Our main tool: Algebraic multigrid (AMG).

Our approach: Multigrid multi level monte carlo

Numerical Experiments

Conclusions and outlook

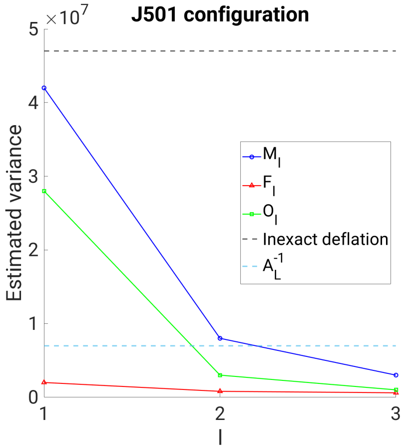
What we measure

$$V[\hat{\mathrm{tr}}(O_p)] = \frac{1}{2N} \underbrace{||\mathsf{offdiag}(O_p)||_F^2}_{\mathbb{V}[x^\dagger O_p x]}.$$

So in what follows, we present the sampled variance.

$$\mathbb{V}[x^\dagger O_p x] \approx \hat{\mathbb{V}}.$$

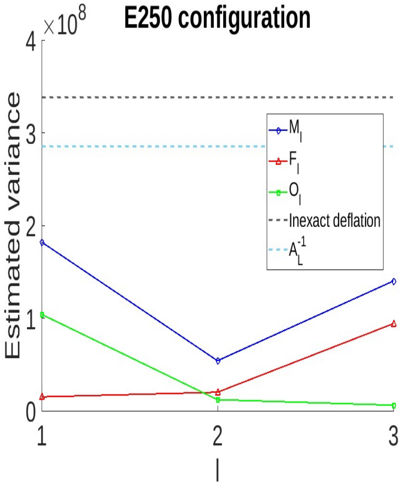
Runs on a 192×64^3 lattice. CLS collaboration



$$\widehat{\mathrm{tr}}(A^{-1}) = \sum_{l=1}^{L-1} \widehat{\mathrm{tr}}(M_l) + \widehat{\mathrm{tr}}(A_L^{-1}).$$

$$\widehat{\mathrm{tr}}(A^{-1}) = \sum_{l=1}^{L-1} \widehat{\mathrm{tr}}(O_l) + \sum_{l=1}^{L-1} \widehat{\mathrm{tr}}(F_l) + \widehat{\mathrm{tr}}(A_L^{-1}).$$

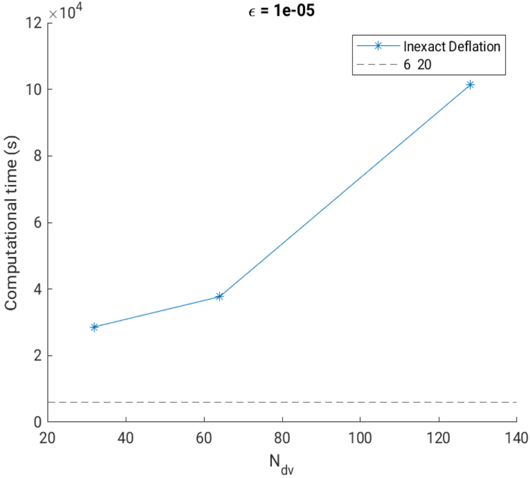
Runs on a 192×96^3 lattice. CLS collaboration



$$\widehat{\mathrm{tr}}(A^{-1}) = \sum_{l=1}^{L-1} \widehat{\mathrm{tr}}(M_l) + \widehat{\mathrm{tr}}(A_L^{-1}).$$

$$\widehat{\mathrm{tr}}(A^{-1}) = \sum_{l=1}^{L-1} \widehat{\mathrm{tr}}(O_l) + \sum_{l=1}^{L-1} \widehat{\mathrm{tr}}(F_l) + \widehat{\mathrm{tr}}(A_L^{-1}).$$

Cost 192×96^3 lattice



Work in progress.

We want to translate this results to traces of the form $tr(B(t)A^{-1}(t,t))$.

$B(t)$ is an operator which acts on spin, color and space indices,

For the operator $B = \sum_i \Gamma_i \nabla_i$:
We simply inserted our deflation on $A^{-1}(t,t)$.

Method	**Trace**	**Variance**
Hutchinson	12.1	8906
Inex-Deflation (k=32)	11.8	8660
MGMLMC	11.5	7896(1^{st}lvl) 497(2^{nd}lvl)

Work in progress: 16^4 Lattice in MATLAB.

Outline

Challenge: Variance reduction on trace estimation.

A slide on deflation methods.

Our main tool: Algebraic multigrid (AMG).

Our approach: Multigrid multi level monte carlo

Numerical Experiments

Conclusions and outlook

Conclusions and outlook

Takeaway

- Multigrid MLMC presents a better reduction of the variance when compared to inexact deflation.
- This, while saving the cost of constructing the deflation space.

Conclusions and outlook

Takeaway

- Multigrid MLMC presents a better reduction of the variance when compared to inexact deflation.
- This, while saving the cost of constructing the deflation space.

Notes and outlook

- **Note:** The MG hierarchy you have for a good solver is enough for our method.
- **Also tried:** Using deflation on top of our operators, does not bring further reduction of the variance.
- **Future posibilties:** How could we further improve Multigrid MLMC for $tr(B(t)A^{-1}(t,t))$? Deflate B? Probing using structure of B?.

Gracias!

Gracias!

[1] M. F. Hutchinson, "A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines," *Communications in Statistics - Simulation and Computation*, vol. 19, 1990.

[2] A. Frommer, M. N. Khalil, and G. Ramirez-Hidalgo, "A multilevel approach to variance reduction in the stochastic estimation of the trace of a matrix," *SIAM Journal on Scientific Computing*, vol. 44, no. 4, pp. A2536–A2556, 2022.

[3] A. S. Gambhir, A. Stathopoulos, and K. Orginos, "Deflation as a method of variance reduction for estimating the trace of a matrix inverse," *SIAM J. Sci. Comput.*, vol. 39, pp. A532–A558, 2017.

[4] A. Frommer, M. N. Khalil, and G. Ramirez-Hidalgo, "A multilevel approach to variance reduction in the stochastic estimation of the trace of a matrix," *SIAM Journal on Scientific Computing*, vol. 44, no. 4, pp. A2536–A2556, 2022.

[5] E. Romero, A. Stathopoulos, and K. Orginos, "Multigrid deflation for lattice qcd," *Journal of Computational Physics*, vol. 409, p. 109356, 2020.

[6] R. A. Meyer, C. Musco, C. Musco, and D. P. Woodruff, *Hutch++: Optimal Stochastic Trace Estimation*, pp. 142–155.

[7] A. Frommer, K. Kahl, S. Krieg, B. Leder, and M. Rottmann, "Adaptive aggregation-based domain decomposition multigrid for the lattice wilson-dirac operator," *SIAM Journal on Scientific Computing*, vol. 36, 2014.

[8] A. Frommer, K. Kahl, S. Krieg, B. Leder, and M. Rottmann, "Adaptive aggregation-based domain decomposition multigrid for the lattice wilson–dirac operator," *SIAM Journal on Scientific Computing*, vol. 36, no. 4, p. A1581–A1608, 2014.

Deflation techniques - comparison.

Deflation method	Main cost
Inexact Deflation. **With** k **singular vectors**.	▶ Precompute V_k (Inverse BPI) ▶ k solves.
Multigrid Deflation. With k singular vectors on coarser (smaller) grid.	▶ Precompute V_k^c on coarser grid. (cheaper BPI) ▶ Exact eigensolver for a matrix $C \in \mathbb{C}^{k \times k}$
Multigrid Multilevel Monte Carlo.	▶ Precompute Nothing. ▶ stochastics on coarser levels

MGMLMC a Projectors Perspective

What Projector do we use to deflate?

This is our multilevel construction:

$$\mathrm{tr}(A) = \sum_{l=1}^{L-1} \left[\mathrm{tr}(A_l - P_l A_{l+1} P_l^\dagger) \right] + \mathrm{tr}(A_L).$$

The matrix

$$\Pi_l = P_l A_{l+1}^{-1} P_l^\dagger A_l.$$

is an **oblique** projector,

$$\Pi_l^2 = \Pi \quad \text{and} \quad \Pi_l^\dagger \neq \Pi.$$

Furthermore:

$$A_l^{-1} = (I - \Pi_l) A_l^{-1} + \Pi_l A_l^{-1} = A_l^{-1} - P_l A_{l+1}^{-1} P_l^\dagger.$$

MGMLMC: an Orthogonal Projector

In our Multigrid construction, we have that $P_l^\dagger P_l = I_l$.
Then, we can proceed similarly as we did before:

$$A_l^{-1} = (I - P_l P_l^\dagger) A_l^{-1} + (P_l P_l^\dagger A_l^{-1}), \quad l = 1, ..., L.$$

Since

$$\mathrm{tr}((P_l P_l^\dagger) A_l^{-1}) = \mathrm{tr}(P_l^\dagger A_l^{-1} P_l) = \mathrm{tr}(P_l^\dagger A_l^{-1} P_l - A_{l+1}^{-1}) + \mathrm{tr}(A_{l+1}^{-1}),$$

then

$$\mathrm{tr}(A^{-1}) = \mathrm{tr}((I - P_l P_l^\dagger) A_l^{-1}) + \mathrm{tr}(P_l^\dagger A_l P_l^{-1} - A_{l+1}^{-1}) + \mathrm{tr}(A_{l+1}^{-1})$$

MGMLMC: an Orthogonal Projector

$$\mathrm{tr}(A^{-1}) = \mathrm{tr}((I - P_l P_l^\dagger) A_l^{-1}) + \mathrm{tr}(P_l^\dagger A_l P_l^{-1} - A_{l+1}^{-1}) + \mathrm{tr}(A_{l+1}^{-1})$$

And we can do one more trick:

$$\mathrm{tr}(A_l^{-1}) = \mathrm{tr}(\underline{(I - P_l P_l^\dagger)} A_l^{-1} (I - P_l P_l^\dagger)) + \mathrm{tr}(P_l^\dagger A_l^{-1} P_l - A_{l+1}^{-1})$$

An **L-level decomposition of the trace** of a matrix A

$$\mathrm{tr}(A^{-1}) = \sum_{l=1}^{L-1} \Big[\mathrm{tr}((I - P_l P_l^\dagger) A_l^{-1} (I - P_l P_l^\dagger)) + \mathrm{tr}(P_l^\dagger A_l^{-1} P_l - A_{l+1}^{-1}) \Big] + \mathrm{tr}(A_L)$$