Deflation for Monte-Carlo estimation of the trace of a matrix inverse

Andreas Stathopoulos   Eloy Romero Alcalde
also with A. S. Gambhir and K. Orginos
Special mention to E. Weinberg and R. C. Brower

Computer Science Department, College of William & Mary, USA

QCDNA'17

Acknowledgment: DOE SciDAC, DOE ECP

Estimation of $\operatorname{tr} A^{-1}$

- Goal: estimate $\operatorname{tr} A^{-1}$ or $\operatorname{tr} \Gamma A^{-1}$ for large-dimension squared matrix A
- Application: disconnected diagrams
- Approach: Monte-Carlo (Hutchinson, 1989)
- Cost: depends on $\operatorname{Var}(x^\dagger A^{-1} x)$

$$\operatorname{tr} A^{-1} = E[x^\dagger A^{-1} x], \text{with } E[\bar{x}_i x_j] = \delta_{i,j}$$

Monte-Carlo Trace

for $n = 1, 2 \ldots$

1. $x \leftarrow \text{rand(N,1)}$
2. $q_i \leftarrow x^\dagger A^{-1} x$
3. Stop if $\operatorname{Var}(q)/n$ is small

Return mean of q

Variance reduction techniques

- Noise vectors
  If using Gaussian noise:
  $$\mathrm{Var}(x^\dagger A^{-1} x) = 2\|A^{-1}\|_F^2$$
  If using $Z_4 = \{-1, 1, -i, i\}$:
  $$\mathrm{Var}(x^\dagger A^{-1} x) = \|A^{-1} - \mathrm{diag}(A^{-1})\|_F^2 = \|A^{-1}\|_F^2 - \|\mathrm{diag}(A^{-1})\|_F^2$$
- Deflation
  $$\mathrm{tr}\, A^{-1} = \underbrace{\mathrm{tr}\, A^{-1} P}_{\text{direct}} + \underbrace{\mathrm{tr}\, A^{-1}(I - P)}_{\text{stocastic}},$$
  with P being low-rank. We hope
  $\mathrm{Var}(x^\dagger A^{-1}(I - P)x) < \mathrm{Var}(x^\dagger A^{-1} x)$

Deflation

$$\operatorname{tr} A^{-1} = \underbrace{\operatorname{tr} A^{-1} P}_{\text{direct}} + \underbrace{\operatorname{tr} A^{-1}(I - P)}_{\text{stocastic}},$$

with P being low-rank. We hope
$\mathrm{Var}(x^\dagger A^{-1}(I-P)x) < \mathrm{Var}(x^\dagger A^{-1} x)$

Source of P:

- Few accurate singular vectors/eigenvectors (future work) from A corresponding to the lowest modes; good variance reductions, expensive to compute and store
- Spatially-blocked basis that represents a significant chunk of the lowest modes, but most of them inaccurately (Multigrid prolongators); computing $\operatorname{tr} A^{-1} P$ can be expensive, cheap to compute and store

SVD deflation

Singular Value Decomposition

$$A = \sum_i u_i \sigma_i v_i^\dagger, \quad A v_i = u_i \sigma_i, \qquad \sigma_i \in R^+, \quad u_i^\dagger u_j = v_i^\dagger v_j = \delta_{ij}$$

- Let $P = \sum_{i=1}^k u_i u_i^\dagger$, with the k smallest σ_i, then

$$\operatorname{tr} A^{-1} = \overbrace{\operatorname{tr} A^{-1} P}^{\text{direct}} + \overbrace{\operatorname{tr} A^{-1}(I-P)}^{\text{stocastic}} = \sum_{i=1}^{k} u_i^\dagger v_i \sigma_i^{-1} + \operatorname{tr} \sum_{i=k+1}^{n} u_i \sigma_i^{-1} v_i^\dagger$$

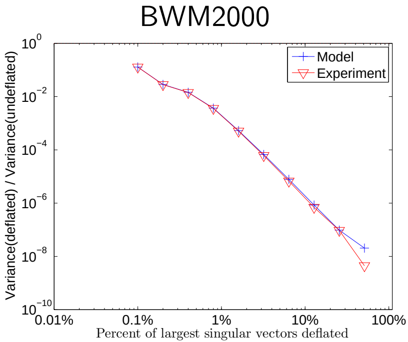
- P reduces the variance when using Gaussian noise:

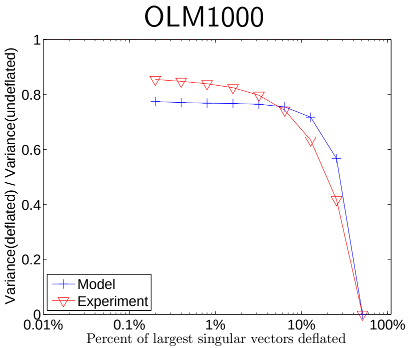
$$\operatorname{Var} x^\dagger A^{-1}(I-P) = 2\|A^{-1}(I-P)\|_F^2 = 2\sum_{i=k+1}^{n} \sigma_i^{-2} \le 2\sum_{i=1}^{n} \sigma_i^{-2} = 2\|A^{-1}\|_F^2$$

- P reduces the variance if using Z_4 noise and u_i, v_j are independent random unitary vectors (Corollary 2.7, A.S. Gambhir, A. Stathopoulos, K. Orginos, 2017)

$$\operatorname{Var} x^\dagger A^{-1}(I-P)x \le \operatorname{Var} x^\dagger A^{-1} x$$

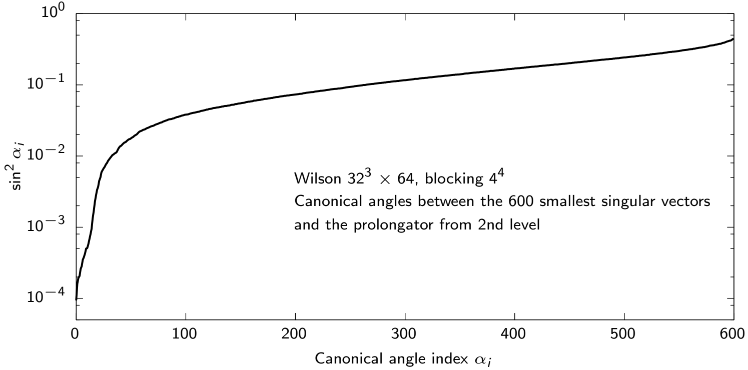
SVD deflation: example





Multigrid prolongators

- Multigrid prolongators exploit that the lowest n modes of A can be well represented on a spatially-blocked basis out of the lowest $k \ll n$ modes
- Similarly the singular vectors corresponding to the smallest singular values can be well represented on a spatially-blocked basis out of the lowest $k \ll n$ modes



Oblique projectors on prolongator

$$\operatorname{tr} A^{-1} = \underbrace{\operatorname{tr} A^{-1} P}_{\text{direct}} + \underbrace{\operatorname{tr} A^{-1}(I-P)}_{\text{stocastic}},$$

- If $P = P_i P_i^\dagger$ for the prolongator P_i of rank k, $\operatorname{tr} A^{-1} P$ can be computed either with k inversions or MC (future work)
- Alternative 1: $P = AP_i(P_i^\dagger A P_i)^{-1} P_i^\dagger$

$$\operatorname{tr} A^{-1} P = \operatorname{tr} P_i (P_i^\dagger A P_i)^{-1} P_i^\dagger = \operatorname{tr}(P_i^\dagger A P_i)^{-1}$$

$\operatorname{tr} A^{-1} P$ can be computed efficiently, but poor variance reduction

- Alternative 2: if $AP_i V_c \approx P_i U_c \Sigma_c$, then

$$P = AP_i V_c (U_c^\dagger P_i^\dagger A P_i V_c)^{-1} U_c^\dagger V_c^\dagger P_i^\dagger \approx P_i U_c U_c^\dagger P_i^\dagger$$

This can work well if $P_i U_c, P_i V_c$ are approximated singular vectors on A. Hint: the small singular values are well represented in P_i, but both left and right approximate singular vectors are needed

Left and right singular vectors in prolongators

- The presence of v_i, u_i on P_i is necessary (but not sufficient) to correlate the smallest part of the singular value spectrum of $P_i^\dagger A P_i$ with A

- If A is γ-Hermitian, then γA is Hermitian, and $\gamma A v_i = v_i \lambda_i$
- Singular Value Decomposition of a γ-Hermitian matrix

$$A = \sum_i \overbrace{\gamma v_i \mu_i}^{u_i} \sigma_i v_i^\dagger, \quad A v_i = \gamma v_i \mu_i \sigma_i, \quad \sigma_i \in R^+, \mu_i = \pm 1, v_i^\dagger v_j = \delta_{ij}$$

- $\text{span}\{v_i, u_i\} = \text{span}\{v_i, \gamma v_i\}$ forms a subspace that has chirality-split basis; also after chirality-splitting v_i, the basis expands v_i and u_i
- Assuming that the near null space found at the first step of creating the prolongators has good approximations of v_i corresponding to the smallest σ_i, chirality-splitting will put also the u_i on the prolongators

PRIMME

- Solver for singular value problems and Hermitian eigenproblems
- Efficient for computing a few values and vectors
- No dependencies but BLAS and LAPACK
- Support for MPI and, soon, GPUs
- BSD
- Based on Davidson-type methods, which allows acceleration of the convergence by using:
  - Preconditioning
  - Many initial guesses

`https://github.com/primme/primme`

Comparative of projectors

Wilson $32^3 \times 64$, blocking 4^4 (A.S. Gambhir, A. Stathopoulos, K. Orginos 2017):

Operator	rank(P)	Var(Op.)	Compute P
A^{-1}		20e4	
$A^{-1}(I - UU^\dagger)$	600	1e4	6126s
$A^{-1}(I - AP_1(P_1^\dagger AP_1)^{-1}P_1^\dagger)$		18e4	
$A^{-1}(I - AP_2(P_2^\dagger AP_2)^{-1}P_2^\dagger)$		19e4	
$A^{-1}(I - AP_1 V_c \Sigma_c^{-1} U_c^\dagger P_1^\dagger)$	1000 (1st level)	4e4	683s
$A^{-1}(I - AP_2 V_c \Sigma_c^{-1} U_c^\dagger P_2^\dagger)$	1000 (2nd level)	4e4	67s

Probing

- Ignore off-diagonal of A^{-1}: spin-color dilution, probing

Partition of $B = A^{-1}$ into k domains:

$$B = \begin{bmatrix} B_{1,1} & B_{1,2} & \cdots & B_{1,k} \\ B_{2,1} & B_{2,2} & \cdots & B_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ B_{k,1} & B_{k,2} & \cdots & B_{k,k} \end{bmatrix}$$

$$\operatorname{tr} B = \sum_i \operatorname{tr} B_{i,i}$$

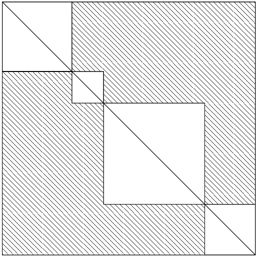
We hope

$$\operatorname{Var} x^\dagger B x \gg \sum_i \operatorname{Var} x^\dagger B_{i,i} x$$

Problem: determine how many partitions are required to reduce the variance enough

Hierarchical Probing

B



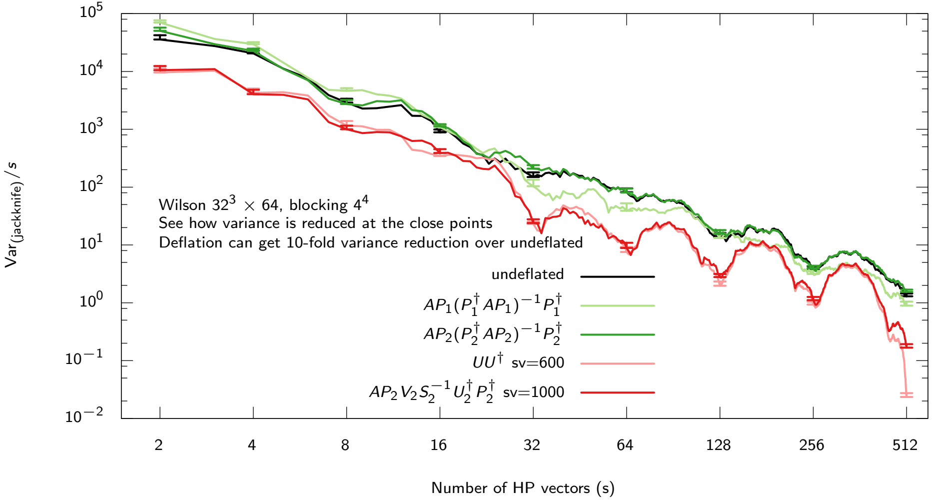
p_1	p_2	p_3	p_4
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
0	1	0	0
0	1	0	0
0	0	1	0
0	0	1	0
0	0	1	0
0	0	1	0
0	0	1	0
0	0	1	0
0	0	0	1
0	0	0	1
0	0	0	1

$$\begin{aligned} \operatorname{tr} B &= E[x^\dagger B x] \\ &= \sum_{i=1}^{k} E[(x \odot p_i)^\dagger B (x \odot p_i)] \end{aligned}$$

Hierarchical probing:

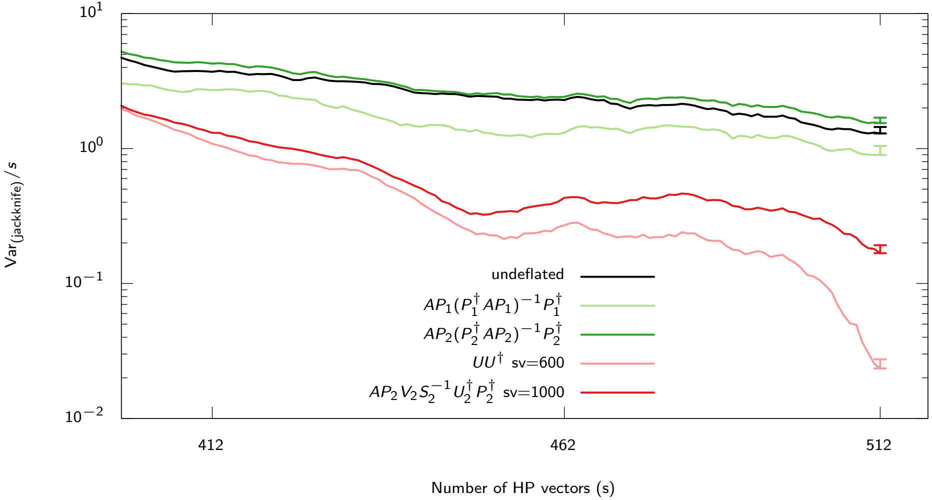
- Use Hadamard basis instead of structural basis
- The span of the first 2^k HP vectors coincides with a 2^k partition of the matrix

Results



10^5
10^4
10^3
10^2
10^1
10^0
10^{-1}
10^{-2}
$\mathrm{Var}_{(\mathrm{jackknife})}/s$
Wilson $32^3 \times 64$, blocking 4^4
See how variance is reduced at the close points
Deflation can get 10-fold variance reduction over undeflated
undeflated
$AP_1(P_1^\dagger AP_1)^{-1}P_1^\dagger$
$AP_2(P_2^\dagger AP_2)^{-1}P_2^\dagger$
$UU^\dagger$ sv=600
$AP_2V_2S_2^{-1}U_2^\dagger P_2^\dagger$ sv=1000
2
4
8
16
32
64
128
256
512
Number of HP vectors (s)

Results



10^1
10^0
10^{-1}
10^{-2}
Var$_{\text{(jackknife)}}$ / s
412
462
512
Number of HP vectors (s)
undeflated
$AP_1(P_1^\dagger AP_1)^{-1}P_1^\dagger$
$AP_2(P_2^\dagger AP_2)^{-1}P_2^\dagger$
$UU^\dagger$ sv=600
$AP_2 V_2 S_2^{-1} U_2^\dagger P_2^\dagger$ sv=1000

Summary

- We explore several projectors for deflation
  - $P = UU^\dagger$, 20 times better than undeflated, expensive to compute, high storage demand
  - $P = AP_i V_c \Sigma_c^{-1} U_c^\dagger P_i^\dagger$, 5 times better than undeflated, cheap to compute, low storage demand
- Hierarchical probing reduces further the variance