Optimizing Shift Selection in Multilevel Monte Carlo for Disconnected Diagrams in Lattice QCD

Travis Whyte [1,2] Andreas Stathopoulos [2] Eloy Romero [3]
Kostas Orginos [2,3]

[1]University of Cambridge, DAMTP

[2]William & Mary

[3]Thomas Jefferson National Accelerator Facility

arXiv:2212.04430

Stochastic Estimation of the Trace of the Lattice Dirac Operator

- Evaluation of the disconnected contributions of the flavor-separated Generalized Parton functions involves estimating Alexandrou et. al 2021

$$tr\Gamma W(z)P_z D^{-1} \tag{1}$$

with $W(z)$ the Wilson line and P_z a permutation operator that displaces the inverse in the z-direction.

- Trace estimate computed via Hutchinson's and given by

$$t(\Gamma W(z)P_z D^{-1}) = \frac{1}{N_s}\sum_{i=0}^{N_s} z_i^{\dagger}\Gamma W(z)P_z D^{-1} z_i \tag{2}$$

with variance

$$Var(t(\Gamma W(z)P_z D^{-1}) = ||\Gamma W(z)P_z D^{-1}||_F^2 - \sum_{i=0}^{n} |(\Gamma W(z)P_z D^{-1})_{ii}|^2 \tag{3}$$

Frequency Splitting

Giusti et. al 2019

- Splits the high and low frequency modes of the inverse by creating a telescoping series of inverses separated by a set of real shifts, σ, with $0 < \sigma_1 < ... < \sigma_L$

$$D^{-1} = D^{-1} - (D+\sigma_1 I)^{-1} + (D+\sigma_1 I)^{-1} - ... - (D+\sigma_L I)^{-1} + (D+\sigma_L I)^{-1} \tag{4}$$

- Use the "One End Trick" to turn a difference of inverses into a product Boucaud et. al. 2008

$$(D + \sigma_l I) - (D + \sigma_{l+1} I) = (\sigma_l - \sigma_{l+1}) I \tag{5}$$

$$(D + \sigma_l I)^{-1} - (D + \sigma_{l+1} I)^{-1} = (\sigma_{l+1} - \sigma_l)(D + \sigma_{l+1} I)^{-1}(D + \sigma_l I)^{-1} \tag{6}$$

- Allows us to expand D^{-1} as a telescoping series in terms of products of inverses

Frequency Splitting

- Rewrite Equation (4) as

$$D^{-1} = \sum_{l=0}^{L-1} (\sigma_{l+1} - \sigma_l)(D + \sigma_l I)^{-1}(D + \sigma_{l+1} I)^{-1} + (D + \sigma_L I)^{-1} \quad (7)$$

- For brevity, let $\hat{\Gamma} = \Gamma W(z) P_z$. Taking the trace gives

$$tr(\hat{\Gamma} D^{-1}) = \sum_{l=0}^{L-1} (\sigma_{l+1} - \sigma_l) tr(\hat{\Gamma}(D + \sigma_l I)^{-1}(D + \sigma_{l+1} I)^{-1}) + tr(\hat{\Gamma}(D + \sigma_L I)^{-1}) \quad (8)$$

- But FS goes further! Multiplication of $\hat{\Gamma}$ on the left leaves the singular spectra of $(D + \sigma_l I)^{-1}(D + \sigma_{l+1} I)^{-1}$ unchanged.

Frequency Splitting

- Use the cyclic property of the trace and the fact that $[(D+\sigma_l I)^{-1}, (D+\sigma_{l+1} I)^{-1}] = 0$ to yield

$$tr\hat{\Gamma}D^{-1} = \sum_{i=0}^{L-1}(\sigma_{l+1}-\sigma_l)tr(D+\sigma_l I)^{-1}\hat{\Gamma}(D+\sigma_{l+1}I)^{-1} + tr\hat{\Gamma}(D+\sigma_L I)^{-1} \tag{9}$$

- Insertion of $\hat{\Gamma}$ changes singular spectra of product terms, reducing the variance! Terms within the summation known as the "split-even" estimator
- The trace estimator is given by

$$t(\hat{\Gamma}D^{-1}) = \sum_{l=0}^{L-1}\frac{1}{N_l}\sum_{s=0}^{N_l} z_{s,l}^{\dagger}(\sigma_{l+1}-\sigma_l)(D+\sigma_l I)^{-1}\hat{\Gamma}(D+\sigma_{l+1}I)^{-1}z_{s,l} + \frac{1}{N_L}\sum_{s=0}^{N_L} z_{s,L}^{\dagger}\hat{\Gamma}(D+\sigma_L I)^{-1}z_{s,L} \tag{10}$$

Multi Level Monte Carlo

Giles 2015

Given a sequence $X_0, ..., X_{L-1}$ that approximates the variable X_L that you want to estimate, we have

$$E[X_L] = E[X_0] + \sum_{l=1}^{L} E[X_l - X_{l-1}] \tag{11}$$

The total computational cost of the trace estimation is given by

$$C_{ML} = \epsilon^{-2}(\sum_{l=0}^{L} \sqrt{C_l V_l})^2 \tag{12}$$

ϵ^{-2} is a target variance, and C_l and V_l are the cost and variance of the *lth* level, respectively. In contrast to the total cost of the single level trace estimation of $\hat{\Gamma} D^{-1}$

$$C_{SL} = \epsilon^{-2} CV \tag{13}$$

Multilevel Monte Carlo

- In the context of FS, the V_l given by the estimator variance each term in the multilevel trace estimator and C_l the cost of solving the associated linear equations of the level l. Ignoring the multiplicative factor of $(\sigma_{l+1} - \sigma_l)$ for now, let

$$t_{l,l+1} = t((D + \sigma_l I)^{-1} \hat{\Gamma} (D + \sigma_{l+1} I)^{-1}) \quad (14)$$

$$t_L = t(\hat{\Gamma}(D + \sigma_L I)^{-1}) \quad (15)$$

Then

$$V_l = Var(t_{l,l+1}) = E[t^*_{l,l+1} t_{l,l+1}] - E[t_{l,l+1}]^* E[t_{l,l+1}] \quad (16)$$

$$V_L = Var(t_L) = E[t^*_L t_L] - E[t_L]^* E[t_L] \quad (17)$$

Multilevel Monte Carlo

- The variance of the multilevel trace estimator is then given by

$$V_{ML} = \sum_{l=0}^{L} \frac{V_l}{N_l} \tag{18}$$

with $N_l = \mu\sqrt{\frac{V_l}{C_l}}$ and the Lagrangian multiplier $\mu = \epsilon^{-2}(\sum_{l=0}^{L} \sqrt{V_l C_l})$

Challenges

- No a priori way to know the shifts that minimize the multilevel cost
- Testing many different shifts to find an approximate minimum of the multilevel cost too expensive
- The optimal shifts, in general, are different for each combination of Γ and P_z.

Sampling the Variances

Can we predict variances of the form of V_l and V_L with only a few samples through interpolation to find shifts that approximately minimize the multilevel cost function?

Need to define three different sets of shifts:

- Sampling set: A set of m real shifts $\hat{s}$ used to sample the estimator variances of the form V_l and V_L with $\hat{s}_0 = 0 < \hat{s}_1 < \ldots < \hat{s}_{m-1}$
- Evaluation set: A set of n real shifts s used to evaluate interpolating polynomials with $s_0 = 0 < s_1 < \ldots < s_{n-1} = \hat{s}_{m-1}$
- Optimal set: The set of L shifts chosen from s such that

$$\sigma = \underset{1 \le j_1 < j_2 < \ldots < j_L \le n}{\arg\min} C_{ML}(s_0, s_{j_1}, \ldots, s_{j_L}). \tag{19}$$

Sampling the Variances

In order to obtain estimates of the variances of the form V_l and V_L, we need to solve linear equations of the form

$$(D + \hat{s}_i I)^{-1} x = z \qquad \textit{for } i = 0, ..., m-1 \tag{20}$$

with z a random noise vector for N_s noise vectors.
Then compute

$$t_{ij} = t((D + \hat{s}_i I)^{-1} \hat{\Gamma} (D + \hat{s}_j I)^{-1}) \qquad \textit{for } i = j = 0, ..., m-1 \tag{21}$$

$$t_j = t(\hat{\Gamma}(D + \hat{s}_j I)^{-1}) \qquad \textit{for } j = 0, ..., m-1 \tag{22}$$

And compute the variance as in Equations (16) and (17) to introduce the shift-dependent functions

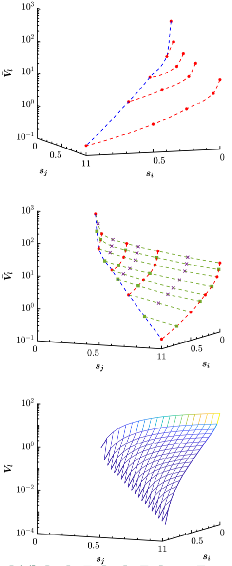
$$\bar{V}_l(\hat{s}_i, \hat{s}_j) = Var(t_{ij}) \tag{23}$$

$$\bar{V}_L(\hat{s}_j) = Var(t_j) \tag{24}$$

Interpolating the Variances

Input: The sampled variances $\bar{V}_l$, the m sampled shifts, $\hat{s}$ and n evaluation shifts, s.
Output: Predicted variances V_l.

for $j = 1 : m-1$
    $w = ln(\bar{V}_l(\hat{s}_i, \hat{s}_j))$ with $i = 0, \ldots, j$
    $q = Interpolate(\hat{s}_{0,\ldots,j}, w)$
    for $k = 0 : n-1$
        $V_l(s_k, \hat{s}_j) = e^{q(s_k)}$
    end
end
$w = ln(\bar{V}_l(\hat{s}_i, \hat{s}_j))$ with $i = j = 0, \ldots, m-1$
$q = Interpolate(\hat{s}_{0,\ldots,m-1}, w)$
for $k = 0 : n-1$
    $V_l(s_k, s_k) = e^{q(s_k)}$
end
for $i = 0 : n-1$
    $w = ln(\bar{V}_l(s_i, \hat{s}_j))$ with $j = 0, \ldots, m-1$
    $q = Interpolate(\hat{s}_{0,\ldots,m-1}, w)$
    for $k = 0 : n-1$
        $V_l(s_i, s_k) = (s_k - s_i)^2 e^{q(s_k)}$
    end
end



Parameters of the Calculation

- $32^3 \times 64$ lattice, Wilson-Clover action with $m_q = -0.2390$ and Stout-link smearing ($m_\pi = 358 MeV$)
- Sampling set $\hat{s} = [0,\ 0.05,\ 0.25,\ 0.5,\ 1.00]$
- Evaluation set s $= [logspace(-5, -2, 4)\ logspace(log10(1e-2+1e-3), 0, 76)]$
- Full spin-color dilution and $p8k7$ probing vectors with 5 Z_4 noise vectors to estimate V_l and V_L
  - Results in 960 inversions per shift in the sampling set, so 4800 inversions
- Solver is even-odd MG preconditioned block FGMRES, so our level cost C_l is given by the number of outer iterations of FGMRES
- Test optimization for γ_3, $\gamma_5\gamma_4$ and for displacements of size $p = 0, ..., 8$.
- Use HPE for terms $V_L(\hat{s}_j)$ when $m_q + \hat{s}_j > 0$

Accuracy of Interpolation

Introduce V_{total}, the sum of the variance of the estimators at each level

$$V_{total} = \sum_{l=0}^{L} V_l \qquad (25)$$

p	Pred. V_{total}	Est. V_{total}	Pred. C_{FS} ($\times 10^5$)	Est. C_{FS} ($\times 10^5$)
0	4.9504	5.2968	0.2921	0.3422
1	82.1364	99.4092	1.4293	1.7824
2	20.8019	23.7536	0.7521	0.8781
3	4.4729	4.6869	0.2371	0.2665
4	1.1335	1.1263	0.0680	0.0742
5	0.3491	0.3578	0.0215	0.0245
6	0.1469	0.1528	0.0084	0.0094
7	0.0826	0.0887	0.0041	0.0052
8	0.0410	0.0367	0.0030	0.0030

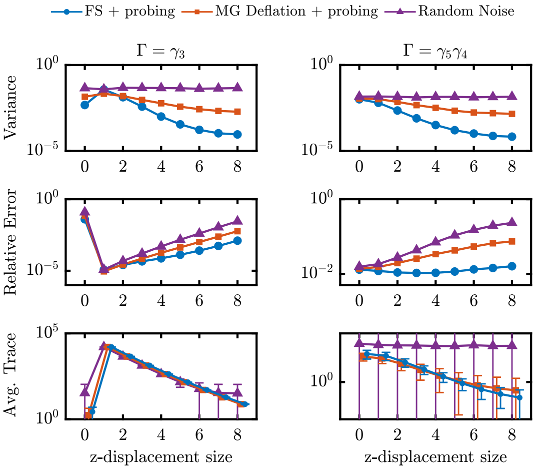
Table: The predicted and estimated V_{total} as well as the predicted and estimated C_{FS} while optimizing for $\Gamma = \gamma_3$ for all displacements of size p.

Accuracy of Interpolation

p	Pred. V_{total}	Est. V_{total}	Pred. C_{FS} $(\times 10^4)$	Est. C_{FS} $(\times 10^4)$
0	18.8176	21.6828	4.8145	5.7190
1	9.3321	9.7361	3.9570	4.4446
2	2.4040	2.5149	1.4491	1.6664
3	0.7998	0.8279	0.4895	0.5648
4	0.3110	0.3111	0.1812	0.2080
5	0.1509	0.1443	0.0796	0.0875
6	0.0581	0.0464	0.0408	0.0389
7	0.0356	0.0279	0.0253	0.0227
8	0.0320	0.0234	0.0217	0.0185

Table: The predicted and estimated V_{total} as well as the predicted and estimated C_{FS} while optimizing for $\Gamma = \gamma_5\gamma_4$ for all displacements of size p.

Comparison to MG Deflation



- Shifts used in FS come from an optimization of γ_3, P_4
- $\sigma_1 = 10^{-5}$
  $\sigma_2 = 0.053$
  $\sigma_3 = 0.146$
  $\sigma_4 = 0.360$
  $\sigma_5 = 0.618$
  $\sigma_6 = 1.00$
- Variance calculated at equal total computational cost of the methods

Conclusions

- FS can give some great speedups in conjunction with probing for large displacements of the lattice over MG deflation
- We can obtain a refined set of shifts through interpolation that reliably predicts variance and multilevel cost
- The shifts coming from an optimization of one configuration can be used for other configurations from the same ensemble with little penalty to performance
- Combine with other variance reduction methods? Most of the variance contained in the term $tr\hat{\Gamma}(D + \sigma_L I)^{-1}$, so possibly use other methods to reduce the variance of that term, such as polynomials, deflation etc.

Acknowledgements

This work was partially performed using compute resources at W&M. Special thanks to David Richards for the compute time on the KNL nodes at JLab and the organizers of the workshop for inviting me to give this talk!

Variance Reduction Methods

Single Level Methods

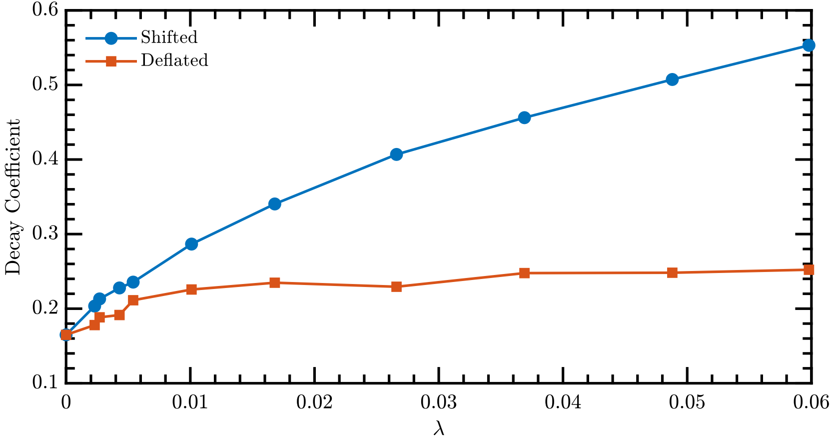
- "Spin-Color" dilution - removes correlation between individual matrix elements W.Wilcox 1999, J. Foley et al. 2005
- Probing - removes heaviest elements closest to the main (or displaced) diagonal Tang et. al 2012, Stathopoulos et. al 2013, Switzer et. al. 2021
- Deflation - removes contributions of largest singular values of the inverse from the variance Baral et. al 2016, Gambhir et. al. 2016, Romero et. al. 2020
- Polynomial Subtraction - approximates the matrix inverse via a polynomial of the matrix Liu et. al. 2014
- And combinations!

Multilevel Methods

- E.g. Chebyshev polynomials, multigrid, Frequency Splitting Hallmann and Troester 2021, Frommer et. al 2021, Giusti et al 2019

Frequency Splitting

Motivation: Shifting the Wilson-Dirac operator drastically decreases the decay of the offdiagonal elements of D^{-1} compared to deflation



Classical Probing

- Classical Probing eliminates elements that correspond to distances up to k by computing a distance-k coloring of the graph of A (which is the same as the distance-1 coloring of A^k), i.e. the heaviest elements near the main diagonal
- Orthogonal set of probing vectors, $z_j = 1, ..., c$ then formed as

$$z_j(i) = \begin{cases} 1 & \text{if color}(i) = j \\ 0 & \text{otherwise} \end{cases}. \tag{26}$$

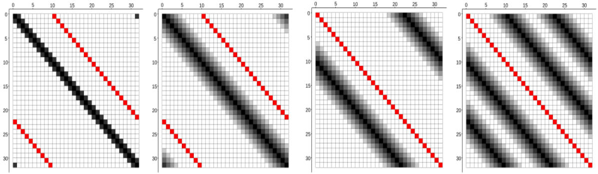
- Remove the deterministic bias by performing the Hadamard product with a noise vector z_0

$$Z = [z_0 \odot z_1, z_0 \odot z_2, ..., z_0 \odot z_c] \tag{27}$$

Probing for Lattice Displacements

Switzer et. al 2021

- As the lattice is displaced, the trace of D^{-1} becomes very small due to the decay of offdiagonal elements and the variance increases as the (previous) main diagonal becomes included in the offdiagonal elements
- Probing then has to target the neighbor hood of the displaced diagonal
- The coloring performed on the symmetric part of $P_z A^k$, given by $P_z A^k + (P_z A^k)^T$. For a node $x = [x_1, ..., x_4]$ in the lattice, this corresponds to a distance-k coloring of the neighborhoods centered at $x^+ = [x_1, x_2, x_3 + p, x_4]$ and $x^- = [x_1, x_2, x_3 + p, x_4]$



(a) Matrix A, 1D torus (b) Matrix of A^4 (c) Displace by 10 (d) Symmetrized

Selection of Probing Vectors

Relative error at large displacements is large due to the trace now being small in magnitude and the displaced trace now contributing to the variance. We then choose probing vectors that target large displacements

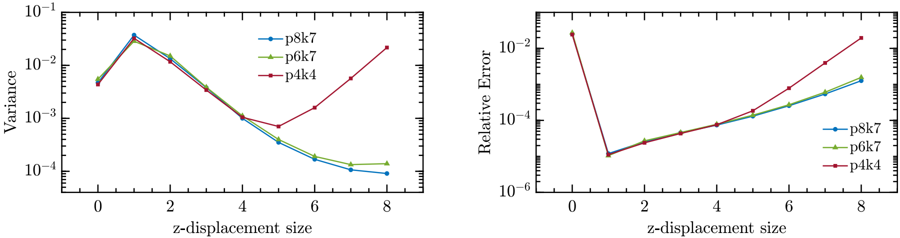


Figure: The FS variance given by Equation (18) (left) and relative error (right) using $\Gamma = \gamma_3$ for each set of chosen probing vectors.

Shift Selection: Evaluation Set Discretization

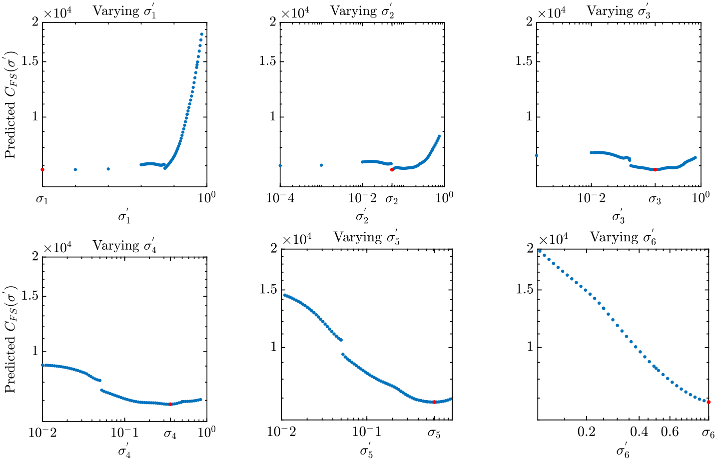


Figure: Slices of the $6D$ manifold of the predicted minimum cost, where we vary one shift and let the others take on the value that minimizes Equation (12). The shifts that approximately minimize Equation (12) are given in red.

Shift Selection: The number of shifts

Need to choose both the number of shifts to use and the discretization of the evaluation set, s

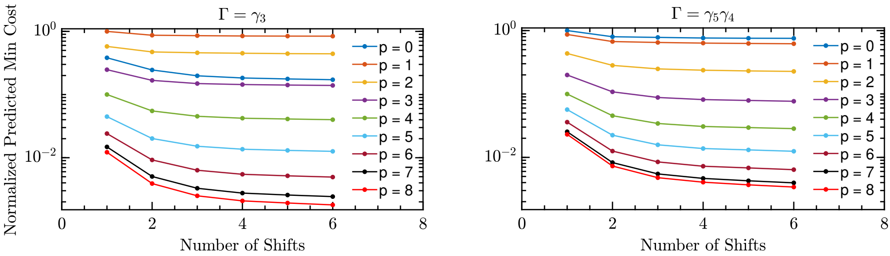


Figure: The normalized predicted minimum cost for all displacements of magnitude p in the z direction as a function of the number of chosen shifts.

Multiple Configurations

γ_3		
Displacement	Mean V_{total}	Rel. Std. Dev. V_{total}
0	5.4108	0.0052
1	41.4419	0.0029
2	16.0299	0.0038
3	4.7654	0.0047
4	1.1777	0.0056
5	0.3883	0.0058
6	0.1772	0.0061
7	0.1123	0.0083
8	0.0932	0.0100

$\gamma_5\gamma_4$		
Displacement	Mean V_{total}	Rel. Std. Dev. V_{total}
0	11.3365	0.0022
1	7.4778	0.0024
2	2.6171	0.0023
3	0.8827	0.0036
4	0.3455	0.0065
5	0.1727	0.0080
6	0.1079	0.0106
7	0.0826	0.0114
8	0.0733	0.0130

	Configuration Number				
	1	2	3	4	5
Est. Speedup	4.8436	5.4360	4.8494	4.5541	5.0838
	Configuration Number				
	6	7	8	9	10
Est. Speedup	3.4911	4.9955	4.5245	4.5861	5.7280

$$Est.\ Speedup = \frac{Est.\ Wallclock\ Time\ FS}{Est.\ Wallclock\ Time\ MG\ Deflation} \tag{28}$$

Recursive Frequency Splitting

- The MLMC analysis lets us know if we can push Frequency Splitting even further. Can we create operators that are a product of three inverses?
- The gist of it: Use the One End Trick on the split-even operators once again
- The trace that we want to compute takes the following form

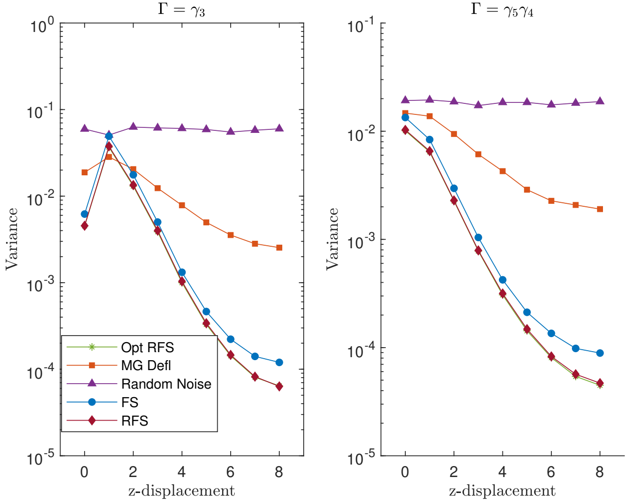
$$
\begin{aligned}
trace \hat{\Gamma} D^{-1} = & \sum_{l=0}^{L-1} (\sigma_{l+1} - \sigma_l)^2 trace (D + \sigma_{l+1} I)^{-2} \hat{\Gamma} (D + \sigma_l I)^{-1} \\
& + \sum_{l=1}^{L} (\sigma_l - \sigma_{l-1}) trace (D + \sigma_l I)^{-1} \hat{\Gamma} (D + \sigma_l I)^{-1} \\
& + trace \hat{\Gamma} (D + \sigma_L I)^{-1}.
\end{aligned}
\tag{29}
$$

Recursive Frequency Splitting

Pros and Cons

- Variance of the terms in the first sum of Equation (15) have variance proportional to $(\sigma_{i+1} - \sigma_i)^4$ rather than $(\sigma_{i+1} - \sigma_i)^2$
- Solver cost of terms in the second sum reduced by a factor of 2 since with full-spin color dilution we get the conjugate solution for free. Terms within the second summation also have less variance than the normal FS split-even operator due to the shifts being the same!
- More costly to optimize as calculating $(D + \sigma_{l+1} I)^{-2}$ is required
- Optimization now more difficult due to there being three types of terms, but as we will see preliminary results suggest optimization of RFS may not be necessary.

Preliminary Results for One Configuration



Hopping Parameter Expansion

HPE

- Taking the inverse gives

$$(D+\sigma_j I)^{-1} = (I - \frac{1}{2}A^{-1}H)^{-1}A^{-1} = \sum_{i=0}^{\infty}(\frac{1}{2}A^{-1}H)^i A^{-1} \tag{30}$$

- Separating this out to the $k-1$ power gives

$$(D+\sigma_j I)^{-1} = \sum_{i=0}^{k-1}(\frac{1}{2}A^{-1}H)^i A^{-1} + \sum_{i=k}^{\infty}(\frac{1}{2}A^{-1}H)^i A^{-1} \tag{31}$$

- The trace of the first term can be calculated exactly with an appropriate set of probing vectors. Due to laplacian structure of D, the trace is identically zero when $k-1$ is less than your displacement.

Hopping Parameter Expansion

HPE

- The trace of the second term of Equation (11) can be estimated stochastically by noting that

$$\sum_{i=k}^{\infty}(\frac{1}{2}A^{-1}H)^{i}A^{-1} = (\frac{1}{2}A^{-1}H)^{k}\sum_{i=0}^{\infty}(\frac{1}{2}A^{-1}H)^{i}A^{-1} = (\frac{1}{2}A^{-1}H)^{k}(D+\sigma_j I) \quad (32)$$

- The factor of $(\frac{1}{2}A^{-1}H)^{k}$ greatly reduces the variance, and Equation (12) is the only source of variance when estimating the trace of $(D + \sigma_j I)$.