PROCEEDINGS OF SCIENCE

Computation of disconnected contributions to nucleon observables

Constantia Alexandrou
Department of Physics, University of Cyprus, P.O. Box 20537, 1678 Nicosia, Cyprus, and Computation-based Science and Technology Research Center, Cyprus Institute, 20 Kavafi Str., Nicosia 2121, Cyprus
E-mail: alexand@ucy.ac.cy

Vincent Drach
NIC, DESY, Platanenallee 6, D-15738 Zeuthen, Germany
E-mail: vincent.drach@desy.de

Karl Jansen
NIC, DESY, Platanenallee 6, D-15738 Zeuthen, Germany
E-mail: karl.jansen@desy.de

Giannis Koutsou
Computation-based Science and Technology Research Center, Cyprus Institute, 20 Kavafi Str., Nicosia 2121, Cyprus
E-mail: g.koutsou@cyi.ac.cy

Alejandro Vaquero Avilés-Casco[a*]
Computation-based Science and Technology Research Center, Cyprus Institute, 20 Kavafi Str., Nicosia 2121, Cyprus
E-mail: a.vaquero@cyi.ac.cy

We compare several methods for computing disconnected fermion loops contributing to nucleon three-point functions. The comparison is carried out using one ensemble of $N_f = 2+1+1$ twisted mass fermions with pion mass of 373 MeV. The complete set of operators up to one-derivative are examined by developing optimized code for mutli-GPUs. Simple guidelines are given as to the preferable method for each class of operators.

31st *International Symposium on Lattice Field Theory LATTICE 2013*
July 29 – August 3, 2013
Mainz, Germany

*Speaker.

1. Introduction

The evaluation of disconnected quark loops is of paramount importance for the computation of flavor singlet quantities, but, on the lattice, this requires the calculation of all- time-slice-to-all propagators, which cannot be carried out by inverting the Dirac matrix at all lattice points, so stochastic methods are tranditionally used to estimate the inverse matrix. Provided the number of stochastic noise vectors N_r needed are much less than the number of lattice points, then this method can be applied efficiently. To reduce the stochastic noise for operators requiring a large number of N_r, we applied the truncated solver method [1]; however, disconnected fermion loops are proned to gauge noise. Therefore, one needs a large number of statistics as well as other noise reduction techniques. In this work we analyze the efficiency of several variance reduction methods for twisted mass fermions, implemented on GPUs.

2. Stochastic methods

A direct computation of the inverse of the fermionic matrix, whose size ranges from $\sim 10^7$ to $\sim 10^9$ for the largest volumes considered nowadays, is not feasible with our current computer power. Nonetheless, we can calculate an unbiased stochastic estimate of the inverse by generating a set of N_r random sources $|\eta_j\rangle$, filling each component with $\mathbb{Z}_N$ noise with the following properties:

$$\frac{1}{N}\sum_{j=1}^{N_r}|\eta_j\rangle = O\left(\frac{1}{\sqrt{N_r}}\right), \qquad (2.1) \qquad \frac{1}{N_r}\sum_{j=1}^{N_r}|\eta_j\rangle\langle\eta_j| = \mathbb{I} + \mathscr{O}\left(\frac{1}{\sqrt{N_r}}\right). \quad (2.2)$$

The first property ensures that our estimate of the propagator is unbiased. The second one allows us to reconstruct the inverse matrix by solving for $|s_r\rangle$ in

$$M|s_r\rangle = |\eta_r\rangle \quad \longrightarrow \quad M_E^{-1} := \tfrac{1}{N_r}\textstyle\sum_{r=1}^{N_r}|s_r\rangle\langle\eta_r| \approx M^{-1}. \qquad (2.3)$$

The error in our estimate decreases as $\mathscr{O}\left(1/\sqrt{N_r}\right)$. $\mathbb{Z}_4$ noise sources were used for this work.

2.1 The Truncated Solver Method

The Truncated Solver Method (TSM) [1] is a way to increase N_r at a reduced computational cost. Instead of solving to high precision Eq. (2.3), we can obtain a low precision (LP) estimate where the inverter, a CG solver in this work, is truncated. The truncation criterion can be a large value of the residual $\hat{r}$, or a fixed number of iterations. This way we can increase the number of stochastic sources N_{LP} at a very small cost. However, the LP sources will produce a biased estimate of M_E^{-1}. This can be corrected by including a few high precision inversions together with the low precision ones, and calculating the difference as follows

$$M_{E_{TSM}} := \underbrace{\frac{1}{N_{\mathrm{HP}}}\sum_{j=1}^{N_{\mathrm{HP}}}\left[|s_j\rangle_{HP} - |s_j\rangle_{LP}\right]\langle\eta_j|}_{Correction} + \underbrace{\frac{1}{N_{\mathrm{LP}}}\sum_{j=N_{\mathrm{HP}}+1}^{N_{\mathrm{HP}}+N_{\mathrm{LP}}}|s_j\rangle_{LP}\langle\eta_j|}_{Biased\ \ estimate}, \qquad (2.4)$$

which requires N_{HP} high precision inversions and $N_{\mathrm{HP}} + N_{\mathrm{LP}}$ low precision inversions. If enough sources are used for the correction, the error of this improved estimator scales as $\propto 1/\sqrt{N_{\mathrm{LP}}}$.

In order to achieve optimal performance of the TSM we must tune several parameters. The first issue is to determine the truncation criterion for the low precision inversions. In our case, we choose as stopping condition a fixed value for the residual $|\hat{r}|_{\rm LP} \sim 10^{-2}$. The second parameter is the number of $N_{\rm HP}$ required to correct the bias introduced when using $N_{\rm LP}$ low precision vectors. To fix these parameters, we performed empirical test upon a reduced set of configurations. As shown in Fig. 1, different insertions behave in different ways and might require different tuning.

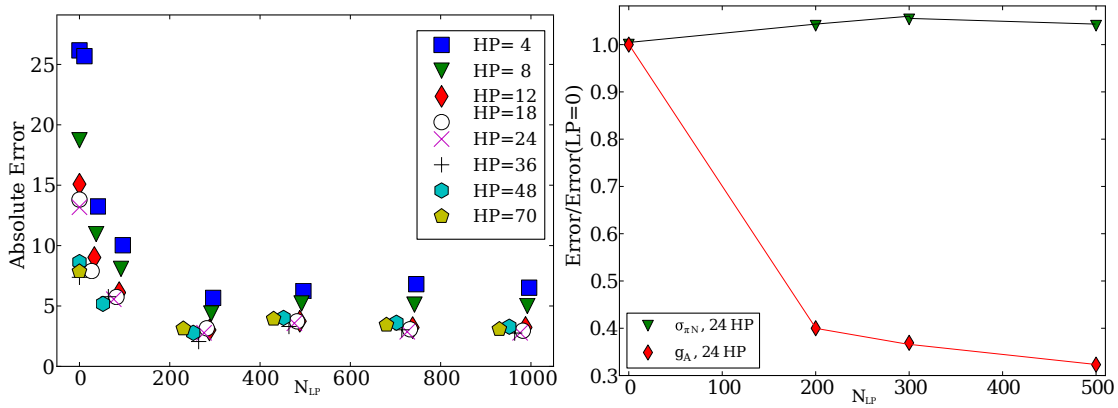

Figure 1: Left: Results on the error of the operator $i\bar{\psi}\gamma_3 D_3\psi$ versus $N_{\rm LP}$ for 50 measurements Right: Data for $\sigma_{\pi N}$ (black line) and g_A (red line) for 56400 measurements. The time of the operator insertion $t_{\rm ins} = 8$ and the sink time $t_{\rm s} = 16$ with th source taken at time zero.

2.2 The one-end trick

The twisted mass fermion formulation allows the use of the *one-end trick* [2, 3] to reduce the variance of the stochastic estimate of disconnected diagrams. If the operator X has an isovector-flavor structure in the twisted basis, then one can use the identity $M_u^{-1} - M_d^{-1} = -2i\mu a M_d^{-1}\gamma_5 M_u^{-1}$ to write the loop as

$$\frac{2i\mu a}{N_r}\sum_{r=1}^{N_r}\left\langle s_r^\dagger \gamma_5 X s_r\right\rangle = \mathrm{Tr}\left(M_u^{-1}X\right) - \mathrm{Tr}\left(M_d^{-1}X\right) + O\left(\frac{1}{\sqrt{N_r}}\right). \tag{2.5}$$

With this substitution the fluctuations are reduced by the μ factor, which should be small in a reasonable simulation. Also, there is an implicit sum of V terms in Eq. (2.5), which improves the signal to noise ratio from $1/\sqrt{V}$ to $V/\sqrt{V^2}$. Unfortunately this technique can only be applied to operators having a τ_3 flavor matrix in the twisted basis. For operators which do not have a τ_3 flavor matrix in the twisted basis, we can use instead

$$\frac{2}{N_r}\sum_{r=1}^{N_r}\left\langle s_r^\dagger \gamma_5 X \gamma_5 D_W s_r\right\rangle = \mathrm{Tr}\left(M_u^{-1}X\right) + \mathrm{Tr}\left(M_d^{-1}X\right) + O\left(\frac{1}{\sqrt{N_r}}\right). \tag{2.6}$$

However, this generalization lacks the μ-suppression factor, we thus expect that for this class of operators the fluctuations to be larger. Because of the volume sum introduced by our identities, the sources must have entries on all sites, which in turn means that we compute the fermion loop at all insertion times simultaneously.

2.3 Time-dilution

A well-known variance reduction technique is time-dilution [4], i.e. instead of filling up all the entries of the source vector, we decompose the whole space $\mathscr{R} = V \oplus \text{color} \oplus \text{spin}$ in S smaller subspaces $\mathscr{R} = \sum_{t=1}^{S} \mathscr{R}_i$, one per time-slice, and we define our noise sources at each time-slice. As the noise on one time-slice contributes to the signal only on this time-slice, but to the noise on all the other time-slices, time-dilution should reduce the stochastic error. In addition, one can apply the coherent source method [5] using noise vectors with entries on several time slices, as long as these time-slices are far enough from each other, so that they don't interfere with each other.

Time-dilution has a disadvantage for operators involving a time derivative, since additional inversions at time-slices $t-a$ and $t+a$ are needed, tripling the computer cost. Therefor time dilution is benchmarked only for ultra-local current insertions.

2.4 Hopping Parameter Expansion

Another technique to reduce the variance is the *Hopping Parameter Expansion* (HPE) [6]. The idea is to expand the inverse of the fermionic matrix in terms of the hopping parameter κ as:

$$M_u^{-1} = B - BHB + (BH)^2 B - (BH)^3 B + (BH)^4 M_u^{-1}, \tag{2.7}$$

where $B = (1 + i2\kappa\mu a\gamma_5)^{-1}$ and $H = 2\kappa\not{D}$, with H the hopping term. The first four terms in this expansion can be computed exactly, while the fifth term is calculated stochastically via

$$\frac{1}{N_r}\sum_{r=1}^{N_r}\left[X\,(BH)^4 s_r \eta_r^\dagger\right] = \mathrm{Tr}\left[X\,(BH)^4 M_u^{-1}\right] + O\left(\frac{1}{\sqrt{N_r}}\right). \tag{2.8}$$

All terms involved in Eq. (2.7) are computed in advance and don't depend on the gauge configuration for local operators, so they do not incur a serious computational overhead. If one expands the inverse M_u^{-1} to a higher order, then one would have to deal with terms like $(BH)^4 B$, involving the plaquette or, for high enough orders, with $(BH)^{2n} B$, involving $2n$-link structures.

3. Simulation details

In order to compare these methods with each other, we consider an ensemble of $N_f = 2+1+1$ twisted mass fermions with 4697 gauge configurations. The pion mass is $m_\pi = 373$ MeV, with the the strange and charm quark masses fixed to approximately their physical values. The lattice spacing of the ensemble is $a = 0.082(1)$ determined form the nucleon mass, and the volume $32^3 \times 64$, giving $m_\pi L \sim 5$. For the disconnected diagrams we use of the branch *discLoop* of the QUDA library [7, 8]. Details on the implementation can be found in Ref. [9].

4. Comparison of different methods

Efficiency of TSM: In Figs. 2 and 3 we show the nucleon σ-term, for which the application of the one-end trick brings the μ-noise suppression factor, and g_A, for which it does not, and that is therefore expected to be a more demanding quantity to compute. We show the disconnected contributions of the light sector, the strange and charm quark to both of these quantities.

	Light sector	Strange sector	Charm sector
R_E for σ-term	1.05	0.91	0.67
R_E for g_A	0.48	0.30	0.28
R_C	0.66	1.09	4.80
$R_C R_E^2$ for σ-term	0.73	0.90	2.15
$R_C R_E^2$ for g_A	0.15	0.098	0.38

Table 1: In the first column we give R_E and R_C as well as the quantity $R_C R_E^2$, which if less than one indicates an advantage of the TSM.

In Table 1 we compare the efficiency of TSM by giving the ratio R_E of the error when using TSM to that without TSM and the ratio R_C of the computer cost with TSM to the cost without TSM. We also give $R_C R_E^2$, which measures the ratio of efficiencies independently of the statistics and the error, therefore a value less than one indicates that the TSM is favorable. The one-end trick is implemented in all cases. For the light quark mass, the TSM is more efficient for both observables as the product $R_C R_E^2$ indicates. As the quark mass increases, the advantage of using the TSM is generally reduced. For the charm quark loops contributing to the σ-term the TSM ceases to be advantageous whereas for g_A the TSM is still useful in all range of masses.

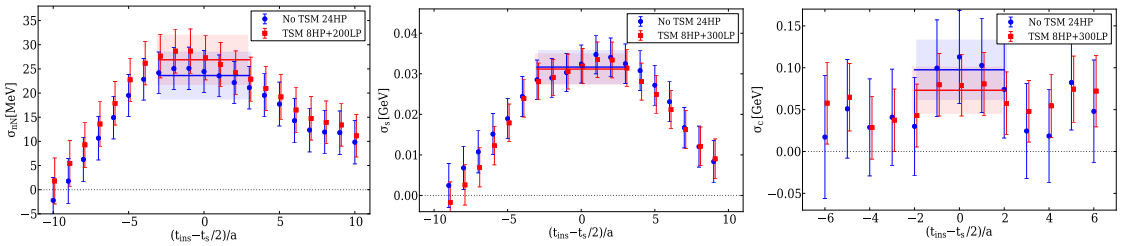

Figure 2: Comparison of the one-end trick with and without TSM for the disconnected contribution to $\sigma_{\pi N}$ (left, 56400 measurements), σ_s (center, 58560 measurements) and σ_c (right, 58560 measurements).

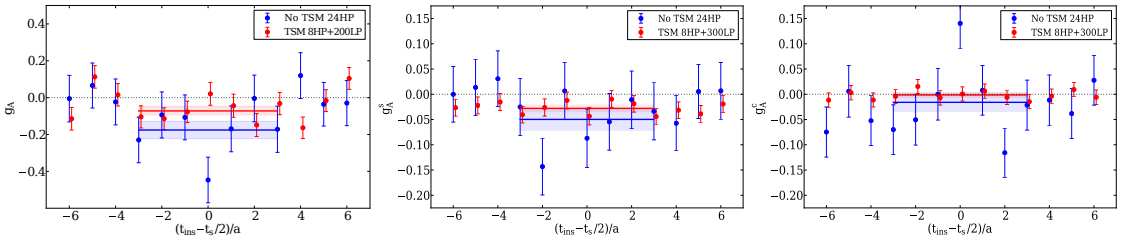

Figure 3: Comparison of the one-end trick with and without TSM for the disconnected contribution to isoscalar g_A (left), g_A^s (center) and g_A^c (right). Same statistics as in the previous figure.

For the case of strange quark loops, we also examine the efficiency of the TSM with respect to time-dilution, as well as whether including the HPE gives any additional benefit. The performance in this case can be assessed easily since the computational cost is roughly the same. As shown in Fig. 4, the TSM always reduces the error, and including HPE is a must, for it comes at virtually no cost, and nearly halves the error. However, we expect the HPE to perform worse (better) as we decrease (increase) the quark mass.

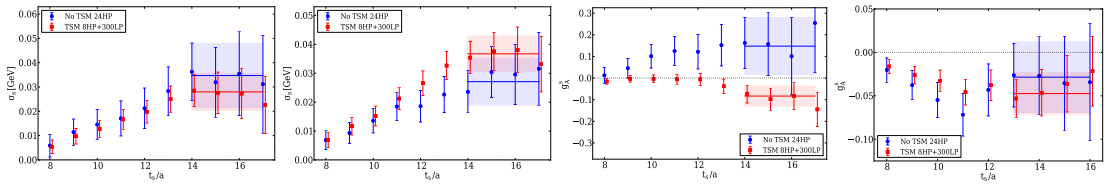

Figure 4: Comparison of time-dilution plus HPE, with and without the TSM, for the case of σ_s (first and second to the left) and g_A^s (first and second to the right). The operator insertion is $t_{\text{ins}} = 8a$ and the number of measurements 18628.

A way to measure the efficiency of the TSM is the ratio $R_{\text{HP/LP}}$, which is the number of LP inversions and source contractions one can compute using the time required for a HP inversion with the associated contractions. Thus, the value of this ratio depends not only on the time required for the inversions, but also includes time needed to perform all the contractions to obtain the loops. In Table 2 we give the ratio $R_{\text{HP/LP}}$ for the different methods and quark masses. A large value for this ratio means that the TSM is advantageous. For the light sector we find a big benefit since the inversions are much more time consuming than the contractions. For the charm sector the time needed for contractions and for a HP inversion are similar, and the TSM brings no benefit.

Method	Quark sector	$R^{Local}_{\text{HP/LP}}$	$R^{One-Deriv.}_{\text{HP/LP}}$
One-end trick	Light	~ 26.7	~ 10
One-end trick	Strange	~ 16.9	~ 5.8
One-end trick	Charm	~ 2.9	~ 1.4
Time-dilution	Strange	~ 20.7	—
Time-dilution + HPE	Strange	~ 19.1	—

Table 2: The $R_{\text{HP/LP}}$ ratio for the different methods for light, strange and charm quark loops. In the third column the ratio for all ultra-local operators is given and in the fourth column all one-derivative operators are also included in $R_{\text{HP/LP}}$.

Time-dilution plus HPE vs the one-end trick: Besides comparing the advantages of the TSM, it would be interesting to compare the one-end trick to time-dilution with HPE.

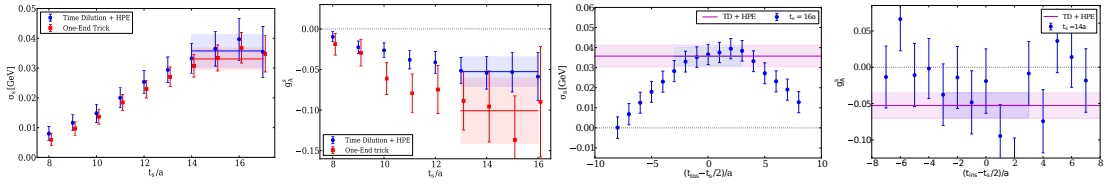

Figure 5: Comparison of results when using the one-end trick plus TSM ($N_{\text{HP}} = 24$ and $N_{\text{LP}} = 300$) to using time-dilution plus HPE plus TSM ($N_{\text{HP}} = 24$ and $N_{\text{LP}} = 300$), same statistics, for σ_s (leftmost) and g_A^s (second from the left). The number of measurement is 18628 and the current method was used with $t_{\text{ins}} = 8a$. In the first right and rightmost panels we show the same quantities, but computed using the fixed sink method for the one-end trick. The purple band is the value of the plateau when time-dilution is used with the fixed current method.

Results are shown in Fig. 5 for σ_s and g_A^s. For σ_s time-dilution gives larger errors as compared to the one-end trick for the same statistics, while for g_A^s considerably smaller errors are obtained with time-dilution. Nonetheless, with the one-end trick one obtains the quark loops at all time-

slices, yielding effectively more measurements. This also allows to vary the insertion time-slide and to fit to a plateau as shown by the blue band in the rightmost plot of Fig. 5. As can be seen, this plateau value has the same error as the one extracted from fitting the asymptotic behavior of the ratio computed using time-dilution with HPE (purple band). Therefore the one-end trick, having the advantage of yielding all time-slides, can perform as well as time-dilution with HPE, also in the case of g_A^s.

5. Conclusions

The computation of disconnected contributions has become feasible due to improvements in algorithms and computational power. In this work, we compare several different strategies to calculate disconnected diagrams by using the GPU-optimized library QUDA on its discLoop branch.

Our comparison shows that the one-end trick with the TSM is the optimal method for the computation of the light and strange quark loops with an ultra-local and one-derivative operator insertions, whereas for the charm quark loops, we prefer time-dilution with the HPE and TSM for ultra-local operators, and the one-end trick for one-derivative insertions. The last choice is justified by the increase in the number of inversions required to apply time-dilution.

Acknowledgments

A. Vaquero and K. Jansen are supported by funding from the Cyprus RPF under contract EPYAN/0506/08 and ΠΡΟΣΕΛΚΥΣΗ/ΕΜΠΕΙΡΟΣ/0311/16 respectively. This research was in part supported by the Research Executive Agency of the EU under Grant Agreement number PITN-GA-2009-238353 (ITN STRONGnet) and the infrastructure project INFRA-2011-1.1.20 number 283286, and the Cyprus RPF under contracts ΚΥ-ΓΑ/0310/02 and ΝΕΑ ΥΠΟΔΟΜΗ/ΣΤΡΑΤΗ/0308/31. Computer resources were provided by Cy-Tera at CaSToRC, Forge at NCSA Illinois (USA), Minotauro at BSC (Spain), and Jugene Blue Gene/P at the JSC, awarded under the 3rd PRACE call.

References

[1] G. Bali, S. Collins and A. Schäffer, PoSLaT**2007**, 141, arXiv:0709.3217.

[2] M. S. Foster and C. Michael, Phys. Rev. D**59** (1999), 074503, arXiv:hep-lat/9810021.

[3] C. McNeile and C. Michael, Phys. Rev. D**73** (2006), 074506, arXiv:hep-lat/0603007.

[4] S. Bernardson, P. McCarty and C. Thron, Comput. Phys. Commun. **78** (1993), 256.

[5] J. D. Bratt *et al.*, PoSLaT**2008**, 141, arXiv:0810.1933.

[6] C. Michael, M. S. Foster and C. McNeile, Nucl. Phys. Proc. Suppl. **83** (2000), 185, arXiv:hep-lat/9909036.

[7] M. A. Clark *et al.*, Comput. Phys. Commun. **181** (2010), 1517, arXiv:0911.3191.

[8] R. Babich *et al.*,SC 2011, arXiv:1109.2935.

[9] C. Alexandrou, G. Koutsou, K. Hadjiyiannakou, A. Strelchenko and A. Vaquero, PoSLaT**2013**, 411.