PROCEEDINGS OF SCIENCE

Evaluation of disconnected contributions using GPUs

C. Alexandrou[ab], V. Drach[c], K. Hadjiyiannakou[b], K. Jansen[c], G. Koutsou[a], A. Strelchenko[a] and A. Vaquero*[a]

[a] *Computation-based Science and Technology Research Center (CaSToRC), The Cyprus Institute, 20 Constantinou Kavafi Street Nicosia 2121, Cyprus*

[b] *Departament of Physics, University of Cyprus, P.O. Box 20537, 1678 Nicosia, Cyprus*

[c] *NIC, DESY, Platanenallee 6, D-15738 Zeuthen, Germany*

We calculate on GPUs the disconnected diagrams associated with the nucleon form factors and moments of generalized parton distributions using Nf=2+1+1 twisted mass fermions. We employ the truncated solver method (TSM) for estimating the all-to-all propagators. Due to the fact that the TSM involves many low precision stochastic estimators, the usage of GPUs is essential to perform efficiently the contractions and the inversions.

The 30th International Symposium on Lattice Field Theory
Cairns Convention Centre, Cairns, Australia
Sunday, June 24 — Friday, June 29 2012

*Speaker.

PoS(Lattice 2012)184

1. Introduction

The evaluation of disconnected diagrams is of paramount importance for eliminating systematic errors in the determination of proton and neutron observables. These contribute significantly in the evaluation of the η' mass and strange content of the nucleon, and require a non-perturbative evaluation involving all-to-all propagators at a given time slice. This, and the inherent gauge noise associated with fermionic loops, explains why most hadron studies neglect these contributions.

Fortunately, in the recent years there has been progress in algorithms and an increase in computational power, making these computations feasible. On the algorithmic side, the introduction of improvements as the one-end trick, and the truncated solver method (TSM) had led to a significant reduction in the variance of disconnected computations. Using the properties of twisted mass fermions, one can further reduce the variance in isoscalar quantities by taking appropriate combinations of two flavors of twisted mass fermions. On the hardware side, GPU units provide a large speed-up in the evaluation of quark propagators and contractions. For the TSM, they provide an optimal platform for swiftly increasing the amount of measurement we can perform.

2. Methods for disconnected calculations

2.1 Stochastic estimation

The exact computation of all-to-all propagators for the lattice volumes of physical interest is outside the current computer power. The fermionic matrix size ranges from $\sim 10^6$ to $\sim 10^9$ in the largest volumes, thus an exact computation of the inverse would require an equal number of inversions, and the situation for timeslice-to-all propagators is equally unfeasible. A way to make progress is to compute an unbiased stochastic estimation of the propagator [1]: we generate a set of N sources $|\eta_j\rangle$ randomly, by filling each component of the source with a number, in our case a particular representation of the $\mathbb{Z}_2$ or $\mathbb{Z}_4$ group. Then the sources have the following properties:

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

The first property ensures that our estimation is unbiased. The second one allows us to reconstruct the inverse matrix by solving for $|s_j\rangle$ in $M|s_j\rangle = |\eta_j\rangle$ and calculating

$$M_E^{-1} := \frac{1}{N}\sum_{j=1}^{N}|s_j\rangle\langle\eta_j| \approx M^{-1}. \tag{2.2}$$

This way the computation becomes feasible, although it is still expensive due to the high number of inversions required to achieve a good estimate of M^{-1} in Eq. (2.2).

The deviation of the estimator from the exact solution is given by

$$M^{-1} - M_E^{-1} = M^{-1} \times \left(\mathbb{I} - \frac{1}{N}\sum_{j=1}^{N}|\eta\rangle\langle\eta|\right). \tag{2.3}$$

From Eq. (2.1) it is clear that the more stochastic sources are used, the smaller the stochastic error becomes. In fact, from Eq.(2.1) and (2.3) we learn that the errors decrease as $O\left(\frac{1}{\sqrt{N}}\right)$, as expected.

Since we also have to deal with the gauge error, we would like to minimize the statistical error by increasing the number of stochastic sources N until we reach the gauge noise. For some cases, this may result in a large value of N, an expensive choice.

2.2 The Truncated Solver Method

The Truncated Solver Method (TSM) [2] increases N at a reduced cost by aiming at a low precision (LP) estimation of the inverse $\left|s_j\right\rangle_{LP} = \left(M^{-1}\right)_{LP}\left|\eta_j\right\rangle$, where the inverter is truncated at reduced accuracy. The truncation criterium can be a large residual or equivalently a fixed number of iterations. This way we can increase the number of sources N_{LP} cheaply, but we are introducing a bias in our estimate due to the truncation. We correct the bias stochastically, by inverting a number of sources to high and low precision and taking the difference:

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

which requires N_{HP} high precision inversions and $N_{HP}+N_{LP}$ low precision inversions. If the convergence of the solver is fast, we only need a few high precision inversions to estimate properly the correction, and then the error falls as $O\left(\sqrt{1/N_{LP}}\right)$. Therefore we want to ensure a good convergence for the solver; in our case this is ensured by the twisted mass regularization, which introduces a lower bound for the eigenvalues of the dirac operator.

The TSM needs tuning of its parameters, namely the precision of the LP inversions and N_{HP}/N_{LP} ratio, to get a safe result with maximum performance. For the first parameter we chose values already used in the literature, i.e., the residual $\rho_{LP} \sim 10^{-2}$ [3]. The tuning of the second parameter was performed empirically: we took a disconnected diagram we expected to yield a large stochastic error, and we optimized N_{HP} and N_{LP} so as to get the minimum error at the lowest computer cost. As shown in Fig. 1, the error decreases as the number of HP or LP increases. A good compromise for this particular diagram is $HP = 12$ and $LP = 300$ as the cheapest point that saturates to the gauge noise. Since the tuning depends on the diagram to be computed, we decided to take the more conservative number of 24 for the number of HP sources.

2.3 The one-end trick

The properties of the twisted mass action provide a powerful method to reduce the variance of the disconnected diagrams. The standard way to compute the disconnected diagrams is to generate N stochastic sources η_r, invert them, and compute the diagrams corresponding to operator X as $\frac{1}{N}\sum_r^N \left\langle \eta_r^\dagger X s_r\right\rangle \approx \mathrm{Tr}\left(M^{-1}X\right)$, where the operator X is expressed in the twisted basis. However, if the operator X involves an isovector combination in the twisted basis, one can resort to the identity $M_u - M_d = 2i\mu a\gamma_5$, which becomes $M_u^{-1} - M_d^{-1} = -2i\mu a M_d^{-1}\gamma_5 M_u^{-1}$ for the propagators:

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

As a result of this substitution, the fluctuations are reduced by the small μ factor. Most important is the implicit sum of V terms in the product $M_d^{-1}\gamma_5 M_u^{-1}$. The difference of propagators exhibits

a signal-to-noise ratio of $1/\sqrt{V}$, but in the product it becomes $V/\sqrt{V^2}$. In fact, a comparison between the two methods reveals a large reduction in the errors at the same computer cost [4–6]. The drawback of this technique is its inapplicability to operators lacking a τ_3 flavour matrix in the twisted basis. A generalized version of the trick can be developed from the identity $M_u + M_d = 2D_W$, with D_W the Dirac-Wilson operator without a twisted mass term. After some algebra,

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

but the lack of the μ suppresion factor introduces a considerable penalty in the signal-to-noise ratio.

3. Simulation details

In order to test these methods, we analyzed 4698 configurations of the *B*55 ensemble of the ETMC collaboration. This ensemble is a $32^3 \times 64$ lattice and was generated with $2+1+1$ dynamical fermions, at pion mass $m_\pi \approx 360$ MeV and strange and charm quark masses fixed at about their physical values. The resulting lattice spacing is $a = 0.086(1)$ fm determined from the nucleon mass resulting in $m_\pi L \sim 5$. The disconnected diagrams were computed by making intensive use of a modified version of the QUDA library [7, 8], which implemented new code and kernels to do the required inversions and contractions on the GPUs. For the Fourier transform we used the CUFFT library.

The QUDA library allowed for multi-GPU calculations, so 2 GPUs worked in parallel per configuration. As seen in the right graph of Fig. 1, the scaling for a few GPUs is very good, with a $\sim 90\%$ increase in performace when adding the second GPU. This result holds up to 8 GPUs, where there is a drop, beyond that the advantages of adding new GPUs are only useful in the case of lack of memory. It is remarkable however that we can reach TFlop sustained performance with just a few GPUs.

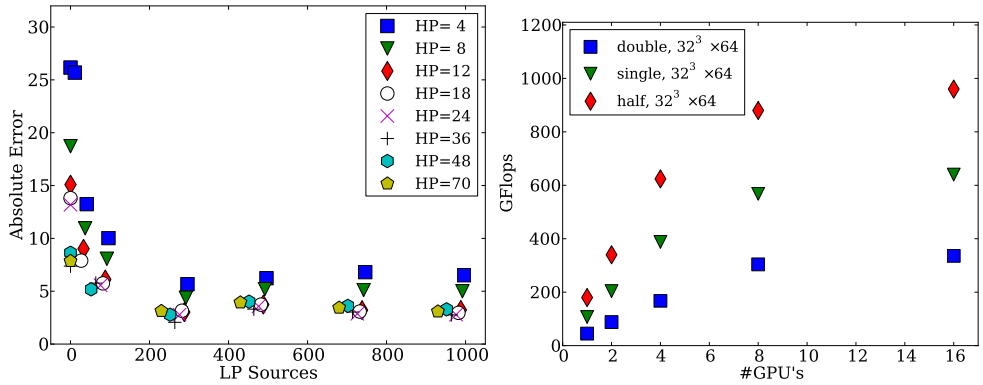

Figure 1: Left: Tuning of the number of HP and LP stochastic noise vectors for the TSM using 50 configurations of the B55.32 ensemble for the the traceless version of the operator $i\bar{\psi}\gamma_3 D_3\psi$ at a given value of the insertion time $t_i = 8$ and sink time $t_s = 16$. The error is shown versus N_{LP} for different values of N_{HP} marked by the different ploting symbols given in the legend. Right: Strong scaling of the multi-GPU code for this ensemble.

The computations were performed on GPU clusters with NVidia fermi GPUs, mainly Tesla M2070 with 6Gb of memory, but also Tesla M2090 and M2050. The noise sources were generated on-the-fly, and the propagators were not stored, in order to save storage and I/O time.

4. The analysis with the summation method

One of the advantages of the one-end trick for twisted mass fermions is the fact that, since the noise sources must be on all sites, we obtained results for all the possible insertions for free. This feature enables us to use the summation method to perform the ratio analysis.

The method is known since a long time [9 – 11], and requires the knowledge of the three point function for all possible insertion times. The advantage is the reduction of the noise due to the excited states by an exponential of the sink time, e^{-kt_s}, as opposed to the standard decrease with the insertion time e^{-Kt_i}. In this method we sum, for every value of t_s, the ratios from $t_i = 0$ up to $t_i = t_s$, $R_{Sum}(t_s) = \sum_{t_i=0}^{t_i=t_s} R_{Plateau}(t_i, t_s)$. Thence the dependence of the ratio on t_i dissappears. The ratio $R_{Plateau}$, computed as the quotient between the three-point function and the two-point function, can be written as $R_{Plateau}(t_i, t_s) = R_{GS} + O(e^{-Kt_i}) + O(e^{-K't_s})$, where R_{GS} is the uncontaminated ratio, and the other contributions are the undesired excited states. After performing the sum in t_i, we get the ratio as a slope $R_{Sum}(t_s) = t_s R_{GS} + c(K, K') + O(e^{-Kt_s}) + O(e^{-K't_s})$, and the contributions of the excited states become a geometrical series in t_i whose sum decays as t_s. Therefore we expect a better suppression of the excited states at the same t_s. The drawback is that we now need to fit to a straight line with two fitting parameters instead of one.

5. Results

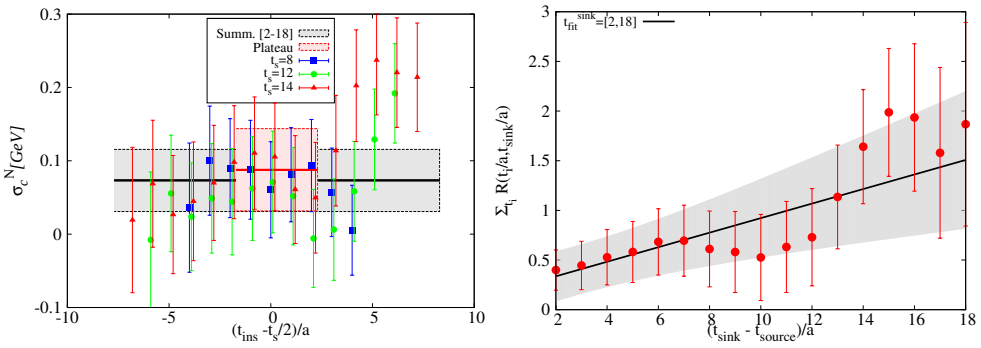

Figure 2: Left: Charm content for the nucleon, from $R_{Plateau}(t_i, t_s)$. The grey band is the value obtained from the summation method (right).

We combined the GPU-computed diagrams with nucleon 2-point functions in order to get the ratios for g_A and σ_c^N. Each disconnected diagram was combined with a set of 5 2-point functions, with randomized positions for each one of the 2912 configurations, where the 2-point functions were computed for proton and neutron, propagating backwards and forwards. In this manner we produced 20 measurements per gauge configuration. The slope obtained in the summation method changes as the sink-source separation increases and fitting too early would yield a wrong result.

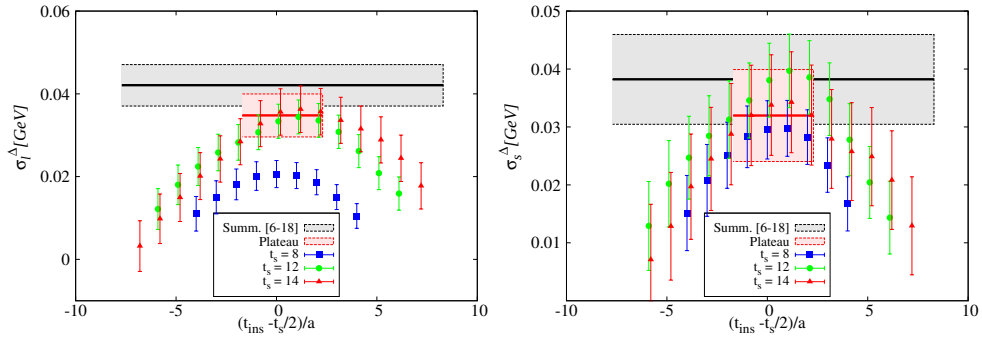

Figure 3: Left: Disconnected contribution to the light σ-term of the Δ from $R_{Plateau}(t_i, t_f)$. Right: The strange σ-term of the Δ.

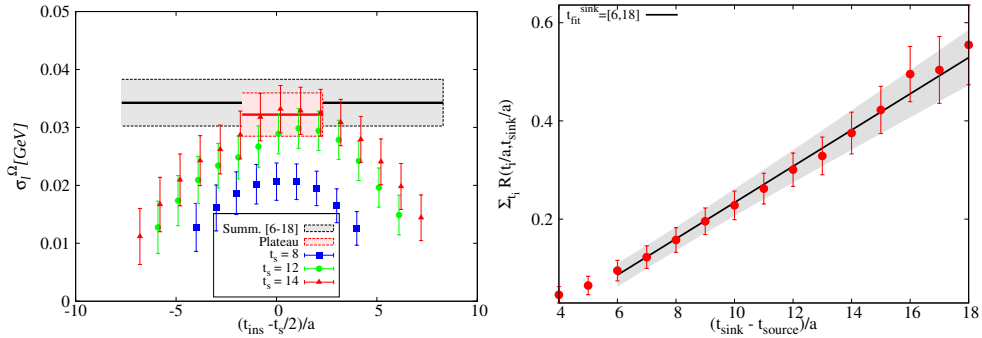

Figure 4: Left: Light σ-term of the Ω from $R_{Plateau}(t_i, t_s)$. The grey band is the value obtained from the summation method by fitting the slope shown on the right.

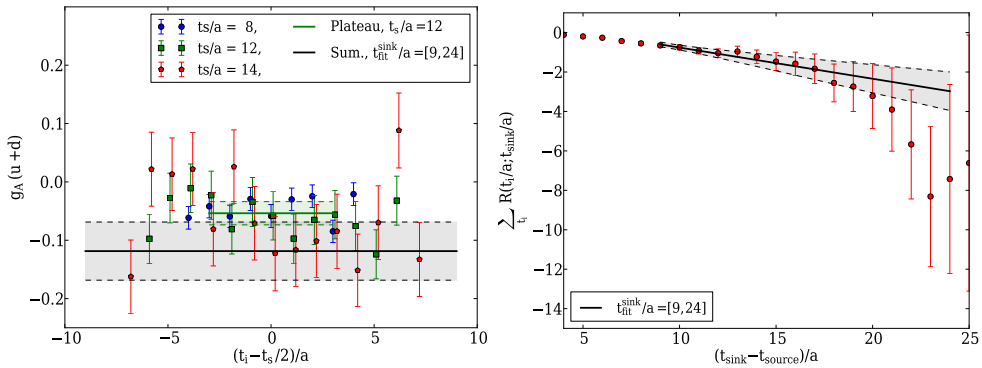

Figure 5: Disconnected contribution to the isoscalar g_A using the generalized one-end trick of Eq.(2.6). The results are noisier than those obtained for operators calculated using the standard one-end trick of Eq.(2.5).

The two methods give consistent results, and therefore combining both one can ensure that we have a large enough sink-source separation for excited states to be neglected.

For the σ_c^N (Fig.2), more statistics are needed to understand the change in slope in the summation method. In the summation of the Δ we observed a similar behaviour, but it was quite reduced

and the results agree with the plateaus (Fig.3), even when our statistics were smaller: 4643 configurations combined with 4 different 2-point functions propagating forwards and backwards (8 measurements). In contrast, the Ω (Fig.4) yields a strong signal with the same statistics.

The generalized version of the one-end trick as expected is more noisy. Our results for the isoscalar nucleon axial charge, g_A^{is} are shown in Fig. 5 and are in agreement with recent evaluation using Clover fermions [12].

6. Conclusions

The computation of disconnected contributions for flavour singlet quantities has become feasible, due to the improvement in the algorithms and to the increase in computational resources. In this work we show that we can get reliable results for disconnected contributions to the σ-terms and the isoscalar axial charge. GPUs are particularly efficient for the evaluation of disconnected diagrams using the TSM, yielding a huge improvement in the computation of LP inversions and contractions. In additon, the one-end trick allows a reduction of the variance at the same computational cost, as well as getting the fermion loops for all the possible insertion times for free. This property, together with the application of the plateau and the summation methods, as well as the generalized one-end trick, allowed us to compute nucleon observables wiere disconnected diagrams play an important role.

Acknowledgments

A. Vaquero is supported by the Research Promotion Foundation (RPF) of Cyprus under grant ΠΡΟΣΕΛΚΥΣΗ/ΝΕΟΣ/0609/16. Computations are performed on GPUs on Cy-Tera (Cyprus) supported by RPF under the grant NEA ΥΠΟΔΟΜΗ/ΣΤΡΑΤΗ/0308/31, Judge at Jülich Forschungzentrum (JSC) (Germany), Forge at NCSA Illinois (USA) and Minotauro at BSC (Spain) through PRACE. Forward propagators were computed on Jugene at JSC through PRACE Tier-0 access.

References

[1] K. Bitar *et al.*, Nucl. Phys. **B**313, (1989) 348.

[2] G. Bali, S. Collins and A. Schäffer, PoSLaT**2007**, 141.

[3] C. Alexandrou *et al.*, Comput. Phys. Commun. **183**, 1215 (2012) arXiv:1108.2473.

[4] Ph. Boucaud, *et al.* (ETM Collaboration), Comput. Phys. Commun. **179** (2008), 695.

[5] Chris Michael and Carsten Urbach (ETM Collaboration), PoSLAT**2007**, 122.

[6] S. Dinter *et al.* (ETM Collaboration), JHEP **1208**, 037 (2012) arXiv:1202.1480.

[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] L. Maiani, G. Martinelli, M. L. Paciello and B. Taglienti, Nucl. Phys. **B**293, 420 (1987).

[10] S. Güsken, arXiv:hep-lat/9906034v1.

[11] S. Capitani, B. Knippschild, M. Della Morte and H. Wittig, PoSLaT**2010**, 147.

[12] G. Bali *et al.*(QCDSF Collaboration), Phys. Rev. Lett. **108**, 222001 (2012).