Computer Physics Communications 185 (2014) 1370–1382

Contents lists available at ScienceDirect

Computer Physics Communications

journal homepage: www.elsevier.com/locate/cpc

Evaluation of disconnected quark loops for hadron structure using GPUs

C. Alexandrou [a,b,*], M. Constantinou [a], V. Drach [c], K. Hadjiyiannakou [a], K. Jansen [a,c], G. Koutsou [b], A. Strelchenko [b,1], A. Vaquero [b]

[a] Department of Physics, University of Cyprus, P.O. Box 20537, 1678 Nicosia, Cyprus

[b] Computation-based Science and Technology Research Center, The Cyprus Institute, 20 Kavafi Str., Nicosia 2121, Cyprus

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

ARTICLE INFO

Article history:
Received 14 September 2013
Received in revised form
18 December 2013
Accepted 17 January 2014
Available online 25 January 2014

Keywords:
Hadron physics
Lattice QCD

ABSTRACT

A number of stochastic methods developed for the calculation of fermion loops are investigated and compared, in particular with respect to their efficiency when implemented on Graphics Processing Units (GPUs). We assess the performance of the various methods by studying the convergence and statistical accuracy obtained for observables that require a large number of stochastic noise vectors, such as the isoscalar nucleon axial charge. The various methods are also examined for the evaluation of sigma-terms where noise reduction techniques specific to the twisted mass formulation can be utilized thus reducing the required number of stochastic noise vectors.

1. Introduction

The evaluation of disconnected quark loops is of paramount importance in order to eliminate a systematic error inherent in the determination of hadron matrix elements in lattice QCD. For flavor singlet quantities, these contributions, even though smaller in magnitude as compared to the connected contributions that are computationally easier to evaluate, are substantial and cannot be neglected. The explanation of why these quark loop contributions are large for flavor singlet quantities is the fact that, in a flavor singlet, the disconnected contributions coming from different flavors add up, and hence there is no *a priori* reason to neglect them. Naive perturbative calculations of some of these flavor singlet contributions differ from their experimental value, which suggests that flavor singlet phenomena are inherently linked with non-perturbative properties of the vacuum. A good example to support this point is the axial anomaly in the case of the η' mass, which is connected to the topological properties and non-perturbative nature of QCD.

The computation of disconnected quark loops within the lattice QCD formulation requires the calculation of all-to-all or time-slice-to-all propagators, which are impractical to compute exactly, and for which the computational resources required to estimate them with, e.g. stochastic methods, are much larger than those required for the corresponding connected contributions. Therefore, in most hadron studies up to now the disconnected contributions were neglected introducing an uncontrolled systematic uncertainty.

Recent progress in algorithms, however, combined with the increase in computational power, have made such calculations feasible. On the algorithmic side, a number of improvements like the one-end trick [1–3], dilution [4–8], the Truncated Solver Method (TSM) [8–10] and the Hopping Parameter Expansion (HPE) [1,11] have led to a significant reduction in both stochastic and gauge noise associated with the evaluation of disconnected quark loops. Moreover, using special properties of the twisted mass fermion Lagrangian, one can further enhance the signal-to-noise ratio by taking the appropriate combination of flavors. On the hardware side, graphics cards (GPUs) can provide a large speed-up in the evaluation of quark propagators and contractions. In particular, for the TSM, which relies on a large number of inversions of the Dirac matrix in single or half precision, GPUs provide an optimal platform.

In this paper, our aim is to assess recently developed methods and examine how reliably one can compute disconnected contributions to flavor singlet quantities by combining the algorithmic advances with the numerical power of GPUs. We will describe the various improvements using one ensemble of twisted mass fermion (TMF) gauge field configurations. The ensemble is generated with two light degenerate quarks and a strange and charm quark with masses fixed to their physical values, referred to

* Corresponding author at: Department of Physics, University of Cyprus, P.O. Box 20537, 1678 Nicosia, Cyprus.

E-mail address: alexand@ucy.ac.cy (C. Alexandrou).

1 Present address: Scientific Computing Division, Fermilab, Batavia, IL 60510-5011, USA.

0010-4655/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.cpc.2014.01.009

as $N_f = 2+1+1$ simulations. The lattice size is $32^3 \times 64$, the lattice spacing extracted from the nucleon mass [12] $a = 0.082(1)(4)$ and pion mass about 370 MeV. This ensemble will be hereafter referred to as the B55.32 ensemble. This paper intends to describe the methodology and identify the efficiency of the various methods with respect to the observable under investigation, rather than to arrive at precise physical results. The latter we reserve for a follow-up publication. Although we will use the nucleon to test our methodology the conclusions apply to any hadron. The paper is organized as follows: in Section 2 we present the algorithms and variance reduction techniques we will employ. In Section 3 we explain our particular formulation, including information on the gauge configurations used, as well as details on the GPU implementation of our methods. Section 4 explains our analysis to extract the desired matrix elements, followed by Section 5 in which we summarize the comparisons between the different methods employed. In Section 6 we give our conclusions and outlook.

2. Methods for disconnected calculations

2.1. Stochastic estimate

The exact computation of all-to-all (time-slice-to-all) propagators on a lattice volume of physical interest is outside our current computer power, since this would require volume (spatial volume) times inversions of the Dirac matrix, whose size ranges from $\sim 10^7$ for a $24^3 \times 48$ lattice to $\sim 10^9$ for the largest volumes of $96^3 \times 192$ considered nowadays. The typical way around this problem is to compute an unbiased stochastic estimate of the all-to-all propagator [13]. The method consists of generating a set of N_r sources $|\eta_r\rangle$ randomly, by filling each component of the source with random numbers drawn from a particular representation of the $\mathbb{Z}_2$ or $\mathbb{Z}_4$ groups (more exactly $\{1, -1\}$ for $\mathbb{Z}_2$ and $\{1, i, -1, -i\}$ for $\mathbb{Z}_4$), or from a representation of $\mathbb{Z}_2 \otimes i\mathbb{Z}_2$. Other noise sets may be used, however it has been shown that $\mathbb{Z}_N$-noise has smaller variance than e.g. Gaussian noise [14]. The $\mathbb{Z}_N$-noise sources have the following properties:

$$\frac{1}{N_r}\sum_{r=1}^{N_r} |\eta_r\rangle = |0\rangle + \mathcal{O}\left(\frac{1}{\sqrt{N_r}}\right), \tag{1}$$

$$\frac{1}{N_r}\sum_{r=1}^{N_r} |\eta_r\rangle\langle\eta_r| = \mathbb{I} + \mathcal{O}\left(\frac{1}{\sqrt{N_r}}\right). \tag{2}$$

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

$$M\,|s_r\rangle = |\eta_r\rangle \tag{3}$$

and calculating

$$M_E^{-1} := \frac{1}{N_r}\sum_{r=1}^{N_r} |s_r\rangle\langle\eta_r| \approx M^{-1}. \tag{4}$$

Since in general the number of noise vectors N_r required is much smaller than the lattice volume V, the computation becomes feasible, although it can still be very expensive depending on the value of N_r required to achieve a good estimate of M^{-1} in Eq. (4).

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

$$M^{-1} - M_E^{-1} = M^{-1} \times \left(\mathbb{I} - \frac{1}{N_r}\sum_{r=1}^{N_r} |\eta\rangle\langle\eta|\right), \tag{5}$$

so as N_r increases the introduced stochastic error decreases, as Eq. (2) clearly shows. In fact, from Eqs. (2), (5) we see that the errors decrease as $\mathcal{O}\left(\frac{1}{\sqrt{N_r}}\right)$, as expected from the properties of these noise sources.

Since we have to deal with gauge error, i.e. the error coming from the fact that we analyze a representative set of gauge configurations, the number of stochastic noise sources should be taken so that the stochastic error is comparable to the gauge error. This criterion ideally determines the number of stochastic sources N_r, which can differ for each observable. Since we will be interested in evaluating a range of observables we will choose N_r that can yield good results for the most demanding among these observables.

2.2. The Truncated Solver Method

The Truncated Solver Method (TSM) [8–10] is a way to increase N_r at a reduced computational cost. The idea behind the method is the following: instead of inverting to high precision the stochastic sources in Eq. (3), we can aim at a low precision (LP) estimate

$$|s_r\rangle_{LP} = \left(M^{-1}\right)_{LP} |\eta_r\rangle\,, \tag{6}$$

where the inverter, which is a Conjugate Gradient (CG) solver in this work, is truncated. The truncation criterion can be a low precision stop condition for the residual (for instance, $|\hat{r}| < 10^{-2}$, with $\hat{r}$ the residual vector in the CG algorithm), or a fixed number of iterations, roughly around 1/10 or 1/20 of what would be needed to obtain a high precision (HP) solution. This way we can increase the number of stochastic sources $N_{\rm LP}$ at a very small cost. Using the low precision sources our estimate of the inverse matrix given by Eq. (4) is not unbiased, so we are introducing new errors in the computation of the all-to-all propagator.

In order to correct for the bias introduced using low precision, we estimate the correction C_E to this bias stochastically by inverting a number of sources to high and low precision, and calculating the difference,

$$C_E := \frac{1}{N_{\rm HP}}\sum_{r=1}^{N_{\rm HP}} \left[|s_r\rangle_{\rm HP} - |s_r\rangle_{\rm LP}\right]\langle\eta_r|\,, \tag{7}$$

where the $|s_r\rangle_{\rm HP}$ are calculated by solving Eq. (3) up to high precision, so our final estimate becomes

$$M_{E_{\rm TSM}}^{-1} := \frac{1}{N_{\rm HP}}\sum_{r=1}^{N_{\rm HP}} \left[|s_r\rangle_{\rm HP} - |s_r\rangle_{\rm LP}\right]\langle\eta_r| + \frac{1}{N_{\rm LP}}\sum_{j=N_{\rm HP}+1}^{N_{\rm HP}+N_{\rm LP}} |s_r\rangle_{\rm LP}\langle\eta_r|\,, \tag{8}$$

which requires $N_{\rm HP}$ high precision inversions and $N_{\rm HP} + N_{\rm LP}$ low precision inversions. Following the discussion in Ref. [15], one expects the error of this improved estimate of the fermion loop to scale as:

$$e\sqrt{2(1-r_c) + \frac{1}{n_{\rm LP}}}, \tag{9}$$

where the unimproved error e scales as $\propto 1/\sqrt{N_{\rm HP}}$ and $n_{\rm LP} = N_{\rm LP}/N_{\rm HP}$. r_c is the correlation between the $N_{\rm HP}$ quark propagators in low and high precision, which is expected to be close to unity (with the optimal being one) and depends on the criterion for the LP inversions and on how well-conditioned the Dirac fermion matrix is. In this work, we use the twisted mass formulation for the fermion action, hence the smallest eigenvalues depend on the value of the twisted mass parameter μ, and our matrix is protected from near-zero eigenvalues.

In the TSM one needs to tune the precision of the LP inversions as well as the $n_{\rm LP}$ ratio, with the goal of choosing as large a ratio as possible while still ensuring that the final result is unbiased and that $r_c \simeq 1$. In the next subsection we give details on how we optimized the TSM parameters with this criterion in mind.