Domain Wall Fermion Simulations with the Exact One-Flavor Algorithm

David Murphy
Department of Physics, Columbia University, New York, NY 10027, USA
E-mail: dmurphy@phys.columbia.edu

As algorithmic developments have driven down the cost of simulating degenerate light quark flavors the relative cost of simulating single quark flavors with the Rational Hybrid Monte Carlo (RHMC) algorithm has become more expensive. TWQCD has proposed an exact one-flavor algorithm (EOFA) that allows for HMC simulations of a single quark flavor without taking a square root of the fermion determinant. We have independently implemented EOFA in the Columbia Physics System (CPS) and BAGEL Fermion Sparse-Matrix Library (BFM) for Shamir and Möbius domain wall fermions, and begun to optimize and test our implementation against RHMC. In this talk we discuss the derivation of the EOFA action, our tests of its equivalence to RHMC, and the current state of our implementation and optimization. We find, after introducing a novel preconditioning technique for the EOFA Dirac operator, that EOFA is a factor of 2.4 times faster than RHMC per molecular dynamics trajectory for the strange quark determinant on an $N_f = 2 + 1$ Möbius DWF ensemble with physical quark masses and a $24^3 \times 64 \times 24$ volume. We expect that further improvement is possible by retuning the integrator parameters for EOFA and by continuing to optimize our code.
1. Introduction

A number of recent developments in the hybrid Monte Carlo (HMC) algorithm used by the RBC/UKQCD collaboration have driven down the cost of simulating degenerate quark flavor pairs. These developments include: extensive force tuning via Hasenbush mass preconditioning [1], the zMöbius domain wall fermion action [2], reduced $L_s$ approximations to the light quark determinant [3], and the use of mixed-precision methods in the conjugate gradient (CG) algorithm. On a recent production run of a large volume, physical quark mass $N_f = 2 + 1 + 1$ ensemble we observed that the strange and charm quark determinants were collectively more expensive than the light quark determinant. To address this, we have turned to exploring TWQCD’s exact one-flavor algorithm (EOFA) [4] as an alternative to the rational HMC (RHMC) algorithm for single quark flavors. This effort is further motivated by our $I = 0$ $K \to \pi\pi$ simulations with G-parity boundary conditions, where $D^\dagger D$ describes four flavors and RHMC is needed for the light quark pair as well [5].

2. The Exact One-Flavor Algorithm

The exact one-flavor algorithm was introduced by TWQCD for efficient simulations of single quark flavors on GPU clusters. In a series of papers the authors first demonstrate how to construct a positive-definite pseudofermion action describing a single quark flavor for Wilson and domain wall fermions [6]. They then benchmark EOFA against RHMC, finding a 20% speed-up and a substantially reduced memory footprint for the case of EOFA [4, 7]. Their construction uses block manipulations in spin space to factorize a ratio of fermion determinants as

$$\det \left[ \frac{D(m_1)}{D(m_2)} \right] = \frac{1}{\det (\mathcal{M}_L)} \cdot \frac{1}{\det (\mathcal{M}_R)},$$

with $\mathcal{M}_L$ and $\mathcal{M}_R$ manifestly Hermitian and positive-definite. Contrast this with RHMC, where we instead compute

$$\det \left[ \frac{D(m_1)}{D(m_2)} \right] = \left\{ \det \left[ \frac{D^\dagger D(m_1)}{D^\dagger D(m_2)} \right] \right\}^{1/2}$$

using a rational approximation to the square root. Both algorithms are equivalent in the sense that they compute the same determinant ratio, but EOFA has the advantage that it avoids the need for computing an overall fractional power of the fermion determinant.

TWQCD’s construction in Ref. [4] begins with a factorization of the domain wall fermion Dirac operator ($D_{\text{DWF}}$). For our purposes we consider the Möbius kernel, and use the following notation: $\alpha$ denotes the Möbius scale, $c = \alpha/2$ and $d = 1/2$ are the weights along the fifth dimension, $D_w$ is the Wilson-Dirac operator, $L_{ss'}$ contains the 5D hopping terms, and $(R_5)_{ss'} = \delta_{s,L_s} - 1 - s'.$ is the 5D reflection operator. Factoring out the terms multiplying $D_w$ in $D_{\text{DWF}}$ results in

$$(D_{\text{DWF}})_{xx',ss'} = (c + d) D_w + \mathbb{1})_{xx'} \delta_{ss'} + ((c - d) D_w - \mathbb{1})_{xx'} L_{ss'}$$

$$= \{ (D_w)_{xx'} \delta_{ss'} + \delta_{xx'} D_{\perp} \} \cdot \left\{ d (1 - L) + c (1 + L) \right\}_{ss'}$$

$$\equiv (D_{\text{EOFA}})_{xx',ss'} \cdot \tilde{D}_{ss'}.$$
One can show analytically that

$$\det(\tilde{D}) = \left( (c + d)^{L_s} + m_f (c - d)^{L_s} \right)^{12V} \quad (2.4)$$

where $m_f$ is the fermion mass, $V$ is the 4D lattice volume, and $L_s$ is the number of $s$ sites. Since this has no dependence on the gauge field, $D_{\text{DWF}}$ can be replaced with $D_{\text{EOFA}}$ in the path integral without modifying the physics. This formalism has the advantage that $H \equiv \gamma_5 R_S D_{\text{EOFA}}$ is Hermitian even for Möbius DWF, but comes at the cost of evaluating the dense 5D operator $D^\perp_{ss}$. The authors then show, using block manipulations in spin space and the Schur determinant identity, that the factorization (2.1) holds with $D = D_{\text{EOFA}}$. Defining $\Delta_{\pm} \equiv R_5 (D^\perp_{\pm}(m_2) - D^\perp_{\pm}(m_1))$, and observing that $\Delta_{\pm}$ factorizes as $\Delta_{\pm} = k\Omega_\pm \Omega_\pm^\dagger$, the authors further demonstrate that this same determinant ratio can be written as a pseudofermion path integral in terms of the action $S_f = \phi^\dagger \mathcal{M}_{\text{EOFA}} \phi$, with

$$\mathcal{M}_{\text{EOFA}} = 1 - kP_+ \Omega_+^\dagger [H(m_1)]^{-1} \Omega_- P_- + kP_- \Omega_-^\dagger [H(m_2) - \Delta_+ P_+]^{-1} \Omega_+ P_+. \quad (2.5)$$

This is the final form of the EOFA action explored in this work.

In the following sections we perform tests of EOFA using two $N_f = 2+1$ RBC/UKQCD domain wall fermion ensembles. The properties of these ensembles are summarized in Table 1.

<table>
<thead>
<tr>
<th>Ensemble</th>
<th>Action</th>
<th>$\beta$</th>
<th>$L^3 \times T \times L_s$</th>
<th>Möbius scale</th>
<th>$am_l$</th>
<th>$am_h$</th>
<th>$m_T$ (MeV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>16I [8]</td>
<td>DWF + I</td>
<td>2.13</td>
<td>$16^3 \times 32 \times 16$</td>
<td>—</td>
<td>0.01</td>
<td>0.032</td>
<td>400(11)</td>
</tr>
<tr>
<td>24ID [9]</td>
<td>MDWF + ID</td>
<td>1.633</td>
<td>$24^3 \times 64 \times 24$</td>
<td>4.0</td>
<td>0.00107</td>
<td>0.0850</td>
<td>137.1(5)</td>
</tr>
</tbody>
</table>

Table 1: Summary of the ensembles used in this work. (M)DWF denotes (Möbius) domain wall fermions, and I(D) denotes the Iwasaki gauge action (+ DSDR term) with coupling $\beta$.
we plot distributions of the magnitude of the PoS(LATTICE2016)272 quire three multishift CG inversions. In Figure 3.3 Pseudofermion Force interval, allowing us to accurately compute the RHMC heatbath. In practice, we observe that the eigenvalues of ∆ algorithm since the operators ∆±P± are singular, making the EOFA heatbath more expensive than the RHMC heatbath. In practice, we observe that the eigenvalues of ∆EOFA cover a relatively small interval, allowing us to accurately compute φ using a rational approximation with a modest number of poles and partially ameliorate this cost.

Table 2: Stochastic evaluations of − logdet(D_{DWF}(m_1)/D_{DWF}(m_2)) using RHMC and EOFA.

<table>
<thead>
<tr>
<th>Ensemble</th>
<th>N_{hits}</th>
<th>am_1</th>
<th>am_2</th>
<th>RHMC</th>
<th>EOFA</th>
</tr>
</thead>
<tbody>
<tr>
<td>16I</td>
<td>10</td>
<td>0.032</td>
<td>0.042</td>
<td>67.2(9)</td>
<td>67.1(1)</td>
</tr>
<tr>
<td>24ID</td>
<td>10</td>
<td>0.085</td>
<td>0.09</td>
<td>521.9(2.0)</td>
<td>520.0(2)</td>
</tr>
</tbody>
</table>

3.2 Heatbath
At the start of each HMC trajectory we draw a random pseudofermion field φ according to \( P(\phi) \propto \exp(-\phi^\dagger M_{\text{EOFA}} \phi) \). This is accomplished by generating a random Gaussian vector \( \eta \), and then computing \( \phi = M_{\text{EOFA}}^{-1/2} \eta \) using a rational approximation \( x^{-1/2} \approx \alpha_0 + \sum_{i=1}^{N_p} \alpha_i / (\beta_i + x) \). Defining \( \gamma = (1 + \beta_i)^{-1} \), one can show that the resulting rational approximation to \( M_{\text{EOFA}}^{-1/2} \) takes the form
\[
M_{\text{EOFA}}^{-1/2} \approx \alpha_0 \mathbb{I} + \sum_{i=1}^{N_p} \alpha_i \left\{ \mathbb{I} + k \gamma P \Omega^\dagger \left[ H(m_1) - \gamma \Delta P \right]^{-1} \Omega P - k \gamma P \Omega^\dagger \left[ H(m_2) - \gamma \beta \Delta P \right]^{-1} \Omega P \right\},
\]
requiring \( 2N_p \) CG inversions to compute \( \phi \). These inversions are not amenable to a multishift CG algorithm since the operators \( \Delta \pm P \) are singular, making the EOFA heatbath more expensive than the RHMC heatbath. In practice, we observe that the eigenvalues of \( M_{\text{EOFA}} \) cover a relatively small interval, allowing us to accurately compute \( \phi \) using a rational approximation with a modest number of poles and partially ameliorate this cost.

3.3 Pseudofermion Force
The EOFA pseudofermion force is derived by varying (2.5) with respect to the gauge field:
\[
\partial_{\alpha,\mu}S_f[U] = k x_{L}^\dagger \gamma S R\left( \partial_{\alpha,\mu}D_w \right) x_L - k x_{R}^\dagger \gamma S R\left( \partial_{\alpha,\mu}D_w \right) x_R,
\]
with \( x_L \equiv [H(m_1)]^{-1} \Omega P \phi \) and \( x_R \equiv [H(m_2) - \Delta_P]^{-1} \Omega P \phi \). This can be evaluated at the cost of two CG inversions, in contrast to the corresponding RHMC force evaluations, which require three multishift CG inversions. In Figure 1 we plot distributions of the magnitude of the pseudofermion force associated with each gauge link on a single configuration of the 16I ensemble, and confirm TWQCD’s observation that the average EOFA force is somewhat smaller in magnitude than the average RHMC force. While TWQCD has reported a speed-up by using a Sexton-Weingarten integration scheme to exploit the asymmetry in the size of the left-handed and right-handed EOFA force contributions [7], we have yet to explore this direction in our work.

4. Reproduction of the 16I Ensemble Using EOFA
As a final test of EOFA we run two, parallel streams to reproduce the 16I ensemble using the parameters in Ref. [8]. On one stream the RHMC action was used to evolve the strange quark, and on the other stream the EOFA action was used to evolve the strange quark; otherwise the evolutions are identical. We generated 1500 trajectories of each stream, and then compared a number of observables — including the plaquette, topological susceptibility, and pion and kaon masses and decay constants — measured every ten trajectories from 500 to 1500. We find complete agreement within the computed statistical errors. We also observe similar rates of topological tunneling, with the caveat that a proper study of the autocorrelation time would require a longer run.
5. Performance Improvements

In this section we discuss various algorithmic refinements to the basic EOFA formalism, and benchmark EOFA against RHMC on the 24ID ensemble. Each MD trajectory of the 24ID ensemble consists of 12 steps of a nested force gradient QPQPQ integrator, with the strange quark on the outermost time step. Timings are reported for a 256-node BG/Q partition.

5.1 Accelerating EOFA Inversions

Since the majority of the computational effort in an HMC evolution lies in repeatedly inverting the Dirac operator, techniques to accelerate these inversions can lead to substantial increases in the overall efficiency of the evolution. In the context of EOFA, the linear system we invert takes the general form

\[ (H + \alpha \Delta \sigma P \pm ) \psi = \phi. \]  

(5.1)

We have introduced a number of such refinements, including: even-odd preconditioning, Cayley-form preconditioning\(^1\), and the use of mixed-precision CG. In the left panel of Figure 2 we show the successive improvements in inversion time as each of these techniques is introduced for a single inversion of (5.1) at the strange quark mass on the 24ID ensemble.

\(^1\)This is a novel technique specific to EOFA, which exploits the relationship \(D_{\text{DWF}} = D_{\text{EOFA}} \cdot \tilde{D}\) from Eqn. (2.3) to right-precondition Eqn. (5.1), resulting in an equivalent linear system in terms of \(D_{\text{DWF}}\) rather than \(D_{\text{EOFA}}\). Since the 5D structure of \(D_{\text{EOFA}}\) is dense, whereas the 5D structure of \(D_{\text{DWF}}\) has a tridiagonal Cayley form, this results in a preconditioned system that is substantially cheaper to solve.
5.2 Forecasted Solutions for the Heatbath Step

The EOFA heatbath requires (5.1) to be simultaneously solved for $2N_p$ values of $\alpha_l$, arising from the rational approximation to $\mathcal{M}^{-1/2}_{\text{EOFA}}$. Like TWQCD, we use the chronological inversion technique introduced by Brower et al. [10] to forecast solutions for a given $\alpha_l$ from previous solutions for other $\{\alpha_l\}$. We observe that by the tenth pole the iteration count has been approximately halved relative to using zero or the solution for the previous $\alpha_l$ as the initial CG guess.

Figure 2: Left: wall clock inversion time for a single EOFA strange quark solve on the 24ID ensemble as the acceleration techniques described in Section 5.1 are introduced. We find an overall $31.8 \times$ speed-up, and a $3.5 \times$ speed-up relative to an even-odd preconditioned multishift inversion of $D_{\text{DWF}}^{1/2} \psi = \phi$ at the strange quark mass. Right: CG iteration counts for each solve in the EOFA heatbath comparing the different schemes for the initial guesses described in Section 5.2.

5.3 HMC Timing Benchmarks

In Table 4 we compare timings for a single MD trajectory using RHMC and EOFA with and without Cayley-form preconditioning. The details of the ensemble parameters and force gradient integrator are identical except for the choice of strange quark action. We find that without this technique RHMC and EOFA break even: the expensive heatbath and cost of inverting $D_{\text{EOFA}}$ negate the expected gain from the simpler form of the energy and force evaluations. Once Cayley-form preconditioning is introduced, we observe a $2.4 \times$ speed-up over RHMC. We expect that by optimizing our code and retuning the details of the force gradient integrator for EOFA a speed-up of $3 \times$ or more should be possible.

6. Conclusion

We have independently implemented and tested TWQCD’s exact one-flavor algorithm. We find, after optimizing, that the HMC evolution of the strange quark is $2.4$ times faster per trajectory with EOFA for a $24^3 \times 64 \times 24$ physical mass Möbius domain wall fermion ensemble. The key to this improvement is a preconditioning technique that relates inversions of $D_{\text{EOFA}}$ to cheaper inversions of $D_{\text{DWF}}$. We expect that further improvements are possible, and are working to implement EOFA with G-parity boundary conditions for our ongoing $I = 0 \to \pi \pi$ calculation [5]. We will elaborate on the details of our EOFA implementation in a forthcoming publication [11].
## EOFA Domain Wall Fermion Simulations

David Murphy

<table>
<thead>
<tr>
<th>Step</th>
<th>RHMC Time (s)</th>
<th>RHMC %</th>
<th>EOFA (dense) Time (s)</th>
<th>EOFA (dense) %</th>
<th>EOFA (Cayley precond.) Time (s)</th>
<th>EOFA (Cayley precond.) %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heatbath</td>
<td>42.9</td>
<td>2.0</td>
<td>340.6</td>
<td>15.1</td>
<td>160.1</td>
<td>18.4</td>
</tr>
<tr>
<td>Force gradient integration (total)</td>
<td>1865.2</td>
<td>88.9</td>
<td>1840.6</td>
<td>81.8</td>
<td>684.0</td>
<td>78.7</td>
</tr>
<tr>
<td>Final Hamiltonian evaluation</td>
<td>189.4</td>
<td>9.0</td>
<td>68.8</td>
<td>3.0</td>
<td>25.0</td>
<td>2.9</td>
</tr>
<tr>
<td>Total</td>
<td>2097.5</td>
<td></td>
<td>2250.0</td>
<td></td>
<td>869.1</td>
<td></td>
</tr>
<tr>
<td>(Total RHMC) / Total</td>
<td>1.00</td>
<td></td>
<td>0.93</td>
<td></td>
<td>2.41</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Strange quark timings for a single 24ID MD trajectory on a 256-node BG/Q partition.

## 7. Acknowledgments

The authors would like to thank members of the RBC/UKQCD collaboration for helpful discussions. This work was supported in part by U.S. DOE grant #DE-SC0011941. Calculations were performed on the BG/Q computers at Brookhaven National Lab.

## References


