A comparative study of numerical methods for the overlap Dirac operator—a status report

J. van den Eshof, A. Frommer, Th. Lippert, K. Schilling and H. van der Vorst

Department of Mathematics, University of Utrecht, The Netherlands
Department of Mathematics, University of Wuppertal, Germany
Department of Physics, University of Wuppertal, Germany

Improvements of various methods to compute the sign function of the hermitian Wilson-Dirac matrix within the overlap operator are presented. An optimal partial fraction expansion (PFE) based on a theorem of Zolotarev is given. Benchmarks show that this PFE together with removal of converged systems within a multi-shift CG appears to approximate the sign function times a vector most efficiently. A posteriori error bounds are given.

1. INTRODUCTION

The overlap operator, $D = 1 + r\gamma_5 \text{sign}(Q)$, satisfies the Ginsparg-Wilson relation and thus exhibits chiral symmetry at finite lattice spacing $a$ (see [1] and references therein). However, due to the sign function of the hermitian Wilson-Dirac operator, $Q$, its numerical evaluation is extremely costly, with an overhead estimated to be at least a factor $O(100)$ compared to Wilson fermions.

In this status report of our ongoing interdisciplinary project, we demonstrate that well established methods to compute the sign function like Lanczos and multi-shift CG in combination with a partial fraction expansion (PFE/CG) can be improved substantially. We present benchmarks of Neuberger’s PFE/CG method [1], an optimal PFE/CG method with reduced number of poles, a PFE-improved version of Boriçi’s Lanczos process for $Q^2$ [3], as well as the standard Chebyshev approximation. It turns out that the PFE/CG method with removal of converged systems is most efficient.

Furthermore, for error monitoring and termination of iterations, we derive a posteriori error bounds of the approximation of the sign function for both Lanczos (in terms of the residual of a related CG-process) and PFE methods (in terms of the residuals in the multi-shift solver).

2. NUMERICAL PROBLEMS

Computations involving the overlap operator, $D$, are characterized by two nested iterations, (i) the outer iterative solution of

$$D x = (1 + r\gamma_5 \text{sign}(Q)) x = b \quad |r| \leq 1,$$

requiring (ii) an inner iteration for $s$

$$s = \text{sign}(Q)b$$

in each outer iteration step.

Despite the fact that nested schemes are suboptimal as information built-up for the sign function is discarded after each iteration, they might still be superior to alternatives from [1].

3. NUMERICAL METHODS FOR $Sx$

3.1. Polynomial approximations for $t^{-\frac{1}{2}}$

These methods determine polynomials $p_k$ which approximate $t^{-\frac{1}{2}}$ on $[a^2, b^2]$ with $a \leq |\lambda_{\text{min}}|$ and $b \geq |\lambda_{\text{max}}|$, the extremal eigenvalues of $Q$. The approximation to $s = \text{sign}(Q)b$ is then $s \approx Q p_k(Q^2)b$. Polynomials that have been used are Chebyshev polynomials with linear convergence (error $\propto (\frac{1}{\kappa+1})^k$, $\kappa$ being the condition number of $Q$), Legendre polynomials as applied in Ref. [4], Gegenbauer polynomials as introduced by Bunk (Ref. [5]) and Schulz polynomials (error...
\( \propto (v_{k+1}^T b)^+ \) which will be presented in a forthcoming publication of our collaboration.

3.2. Lanczos based methods

The Lanczos process in matrix form reads

\[
Q V_k = V_k T_k + \beta_{k+1} v_{k+1} e_k^T, \text{ with } V_k^T b = e_1. \tag{3}
\]

where we assume \( \|b\| = 1 \). We refer to eq. (3) as “Lanczos for \( Q \)”. Two ways have been proposed to approximate \( \text{sign}(Q)b \) by diagonalization of \( T_k \):

\[
\text{sign}(Q)b \approx Q V_k (T_k^2)^{-1/2} e_1. \tag{4}
\]

\[
\text{sign}(Q)b \approx V_k \text{sign}(T_k)e_1. \tag{6}
\]

The errors of both methods are highly oscillating as a function of \( k \). For the second one, the peaks are bounded, however. In order to avoid such oscillations Boriç has introduced an alternative based on a Lanczos process on \( Q^2 \) \( \#3 \):

\[
Q^2 V_k = V_k T_k + \beta_{k+1} v_{k+1} e_k^T, \text{ with } V_k^T b = e_1. \tag{6}
\]

\[
\text{sign}(Q)b \approx Q V_k T_k^{-1/2} e_1. \tag{7}
\]

The latter method (“Lanczos for \( Q^2 \)”) shows a smoother convergence rate as well as a potentially smaller projected system \( T_k \). However, in any case the spectral decomposition of \( T_k \) is computationally very costly.

3.3. Partial Fraction Expansion and multi-shift CG (PFE/CG)

The elegant idea to use a fixed number of vectors by means of partial fractions expansions has been proposed by Neuberger \( \#1 \):

\[
\text{sign}(Q)b \approx x^{\text{PFE}} = \sum_{i=1}^{m} \frac{\omega_i Q}{Q^2 + \tau_i} b = \sum_{i=1}^{m} \omega_i Q x_k^i. \tag{8}
\]

The \( m \) vectors \( x_k^i \) are computed in step \( k \) of the multi-shift CG method \( \# \) for the shifts \( \tau_i \).

Two rational approximations so far have been applied in the context of the overlap operator:

**Neuberger’s proposal (\( \#1 \))**:

The coefficients are defined by

\[
\tau_i = \tan^2 \left( \frac{\pi}{2m} \left( i - \frac{1}{2} \right) \right) \quad \omega_i = \frac{1}{m} \cos^2 \left( \frac{\pi}{2m} \left( i - \frac{1}{2} \right) \right). \tag{9}
\]

In general, a large number \( m \) of poles \( \tau_i \) is required to achieve practical precisions.

**Remez algorithm** (Edwards et al. \#):

By use of the Remez algorithm, an optimal approximation \( g(x) \) to \( x^{-1} \) in \( \| \cdotp \|_{\infty} \equiv [\lambda_\text{min}, \lambda_\text{max}] \) is constructed, resulting in a substantially smaller number of poles. However, the sign function is approximated as \( x g(x^2) \) which is not the \( \| \cdotp \|_{\infty} \)-optimal approximation to \( \text{sign}(x) \) in \( [-|\lambda_\text{max}|, -|\lambda_\text{min}|] \cup [|\lambda_\text{min}|, |\lambda_\text{max}|] \).

4. IMPROVEMENTS

4.1. Lanczos procedures

As mentioned, the Lanczos approach might be slow since a diagonalization of \( T_k \) is required. As far as the “Lanczos on \( Q^2 \)” approach is concerned we propose to use a PFE, as detailed next, to compute a first approximation to the inverse square root of the full matrix \( T_k \). Based on this approximation, the Lanczos procedure is repeated to yield the final approximation to \( (Q^2)^{-1/2} b \).

4.2. PFE/CG

The vector updates in PFE/CG play a significant role for large numbers of poles for practical implementations. Therefore, we seek for a reduction of the number of poles to improve PFE/CG. In contrast to Ref. \# we try to find a rational function \( f(x) \) that minimizes

\[
\| 1 - \sqrt{x} f(x) \|_{\infty} = \| [\lambda_\text{min}, \lambda_\text{max}] \|.
\]

Then \( x f(x) \) is the \( \| \cdotp \|_{\infty} \)-optimal rational approximation of the sign function on \( [-|\lambda_\text{max}|, -|\lambda_\text{min}|] \cup [|\lambda_\text{min}|, |\lambda_\text{max}|] \). By means of Zolotarev’s theorem \( \# \) \( f(x) \) can be given in analytic form:

\[
f(x) = D \prod_{l=1}^{k-1} \frac{x + c_{2l}}{x + c_{2l-1}}, \tag{11}
\]

where the coefficients can be expressed in terms of Jacobian elliptic functions:

\[
c_l = \frac{\sin^2 \left( l K/2k; \kappa \right)}{1 - \sin^2 \left( l K/2k; \kappa \right)}, \quad l = 1, \ldots, 2k - 1, \tag{12}
\]

with \( \sqrt{1 - \kappa^2} = |\lambda_\text{max}| \) and \( D \) being uniquely determined by the condition

\[
\max_{c \in [1, (\lambda_\text{min})^2]} \left| 1 - \sqrt{x} f(x) \right| = \min_{c \in [1, (\lambda_\text{max})^2]} \left| 1 - \sqrt{x} f(x) \right|.
\]
Table 1 shows that the method drastically reduces the number of poles, in particular for large condition numbers.

Table 1
Number of poles for precision 0.01.

| $\lambda_{\text{max}}$ | Neuberger | Remez | Zolotarev |
|------------------------|-----------|-------|-----------|
| 200                    | 19        | 7     | 5         |
| 1000                   | 42        | 12    | 6         |
| 100000                 | > 500     | ?     | 10        |

Another interesting idea is to remove converged systems from the multi-shift process early, as residuals for shifted matrices with large shifts $\tau_i$ reduce more quickly. Under some restrictions on the quality of PFE we can show that for a total error of at most $\epsilon$ we can stop updating system $j$ after step $k$ as soon as

$$||r^j_k|| \leq \frac{\epsilon \sqrt{\lambda_i^j}}{m \omega_j}.$$  (13)

5. A POSTERIORI ERROR BOUNDS

The error in the PFE/CG method is composed of 2 parts. For a total error of at most $\epsilon$ we demand:

I. $|\text{sign}(\lambda) - \sum_{i=1}^{m} \omega_i \frac{\lambda}{\lambda^2 + \tau_i}| \leq \epsilon/2,$  (14)

II. $||x^{\text{PFE}} - x_k|| \leq \epsilon/2.$  (15)

One can prove that the total error $\leq \epsilon$, if the CG residual for the smallest shift satisfies

$$||r^j_k|| \leq \frac{\epsilon}{2 + \epsilon}.$$  (16)

For “Lanczos for $Q^{2n}$ it is worth noting that we also managed to get a posteriori error bounds which will be presented in a forthcoming paper.

6. NUMERICAL EXPERIMENTS

Our tests have been carried out on quenched $16^4$ configurations at $\beta = 6.0$ and $m = -1.6$ with the error for the approximation of the sign function being $< 10^{-10}$. The timings are from 16 nodes of the Wuppertal cluster computer ALiCE.

Table 2
Benchmarks.

| confs | 1 | 2 | 3 | 4 | 5 |
|-------|---|---|---|---|---|
| $\lambda_{\text{min}} \cdot 10^4$ | 0.455 | 1.39 | 1.17 | 2.23 | 3.02 |
| $\lambda_{\text{max}}$ | 2.48 | 2.48 | 2.48 | 2.48 | 2.48 |
| poles Neub. | 143 | 82 | 89 | 65 | 56 |
| poles Zolo. | 21 | 18 | 19 | 17 | 16 |

7. OUTLOOK

The benchmark results demonstrate that the PFE/CG/Zolotarev procedure with removing of converged systems turns out to be most effective.

As a next step, we will tune the accuracy of the sign approximation within the solution of the outer problem, $Dx = b$ (eq. (1)). Furthermore, we will investigate the effect of projecting out some low eigenvalues of $Q$ onto our findings.

REFERENCES

1. H. Neuberger. In Frommer et al. [10], p. 1.
2. A. Boriçi. Phys. Lett., B453:46, 1999.
3. A. Boriçi. In Frommer et al. [10], p. 40.
4. P. Hernandez et al. Nucl. Phys., B552:363, 1999.
5. B. Bunk. Nucl. Phys. Proc. Suppl., B63:952, 1998.
6. H. A. van der Vorst. In [11], p. 18.
7. U. Glässner et al. Int. J. Mod. Phys., C7:635, 1996.
8. R. G. Edwards et al. Nucl. Phys., B540:457, 1999.
9. D. Ingerman et al. Comm. Pure Appl. Math., 53(8):1039, 2000.
10. A. Frommer et al. (eds.) *Numerical Challenges in Lattice Quantum Chromodynamics*, Springer Verlag, Heidelberg, 2000.