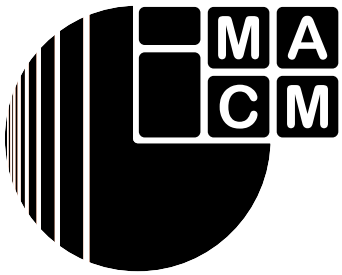

Bergische Universität Wuppertal

Fachbereich Mathematik und Naturwissenschaften

Institute of Mathematical Modelling, Analysis and Computational Mathematics (IMACM)

Preprint BUW-IMACM 11/30

A. Frommer, K. Kahl, Th. Lippert, H. Rittich

Error Bounds for the Sign Function

December 2011

http://www.imacm.uni-wuppertal.de
http://www-ai.math.uni-wuppertal.de/SciComp/

Error Bounds for the Sign Function*

Andreas Frommer† Karsten Kahl‡ Thomas Lippert§
H. Rittich¶

January 2, 2012

Abstract

The Overlap operator fulfills the Ginsparg-Wilson relation exactly and therefore represents an optimal discretization of the QCD Dirac operator with respect to chiral symmetry. When computing propagators or in HMC simulations, where one has to invert the overlap operator using some iterative solver, one has to approxomate the action of the sign function of the (symmetrized) Wilson fermion matrix Q on a vector b in each iteration. This is usually done iteratively using a 'primary' Lanczos iteration. In this process, it is very important to have good stopping criteria which allow to reliably assess the quality of the approximation to the action of the sign function computed so far. In this work we show how to cheaply recover a secondary Lanczos process, starting at an arbitrary Lanczos vector of the primary process and how to use this secondary process to efficiently obtain computable error estimates and error bounds for the Lanczos approximations to $\mathrm{sign}(Q)b$, where the sign function is approximated by the Zolotarev rational approximation.

1 Introduction

Overlap fermions as a lattice formulation of QCD respecting chiral symmetry have been proposed in [7] and been investigated since by many authors. The overlap operator still represents the discrete Dirac operator which most neatly deals with chiral symmetry, fulfilling the Ginsparg-Wilson relation on the lattice

*This work was supported by Deutsche Forschungsgemeinschaft through the Collaborative Research Centre SFB-TR 55 "Hadron Physics from Lattice QCD"

†Fachbereich C, Mathematik und Naturwissenschaften, Bergische Universität Wuppertal, D-42097 Wuppertal, Germany, frommer@math.uni-wuppertal.de

‡Fachbereich C, Mathematik und Naturwissenschaften, Bergische Universität Wuppertal, D-42097 Wuppertal, Germany, kahl@math.uni-wuppertal.de

§Jülich Supercomputing Centre, Forschungszentrum Jülich GmbH, D-52425 Jülich, Germany, th.lippert@fz-juelich.de

¶Fachbereich C, Mathematik und Naturwissenschaften, Bergische Universität Wuppertal, D-42097 Wuppertal, Germany, rittich@math.uni-wuppertal.de

exactly. If D_W describes the hopping part of the standard Wilson fermion matrix and κ_c its critical hopping parameter, the overlap operator is given as

$$D_O = I + \rho\gamma_5 \text{sign}(Q) \text{ with } Q = \gamma_5(I - \frac{4\kappa_c}{3} D_W).$$

Herein, ρ is a mass parameter which is close to 1.

A direct computation of sign(Q) is not feasible, since Q is large and sparse, whereas sign(Q) would be full. Therefore, numerical algorithms which invert systems with the matrix D_O have to follow an inner-outer paradigm: One performs an outer Krylov subspace method where each iteration requires the computation of a matrix-vector product involving sign(Q). Each such product is computed through another, inner iteration using matrix-vector multiplications with Q. In this context, it is very important to be able to assess the accuracy of the computed approximation to sign$(Q)b$ from the inner method, since one can steer the outer method so as to require less and less accurate computations of sign$(Q)b$, resulting in substantial savings in computational work, see [1].

In this work we precisely consider the task of obtaining reliable error estimates and bounds when computing approximations for sign$(Q)b$. Most preferably, we would like to have a precise *upper bound*, so that a stopping criterion based on that upper bound will guarantee that the exact error is below this bound. Actually, we will consider the case where the sign function sign(t) is approximated by a rational function $g(t)$, the Zolotarev approximation. This approach has established itself as the method of choice, since the multishift cg method allows for an efficient update of the iterates, involving only short recurrencies and thus few memory [9].

Usually, one fixes the rational Zolotarev approximation $g(Q)b$ such that the error w.r.t. the sign function is less than ϵ_1 on the spectrum of Q. An error bound ϵ_2 for the approximation of $g(Q)b$ then results in an overall error bound $\epsilon_1 + \epsilon_2$ w.r.t. sign$(Q)b$.

2 Lanczos process and Lanczos approximations

Assuming that $v_1 \in \mathbb{C}^n$ is normalized to $\|v_1\|_2 = 1$, the Lanczos process computes orthonormal vectors $v_1, v_2, \ldots$ such that $v_1, \ldots, v_m$ form an orthonormal basis of the nested sequence of Krylov subspaces $K_m(Q, v_1)$, $m = 1, 2, \ldots$. It is given here as Algorithm 2.1.

The Lanczos process can be summarized via the *Lanczos relation*

$$AV_m = V_{m+1}T_{m+1,m} = V_m T_m + \beta_m \cdot e_m^* v_{m+1}, \tag{1}$$

where $V_m = [v_1|\ldots|v_m] \in \mathbb{C}^{n\times m}$ is the matrix containing the Lanczos vectors,

Algorithm 2.1: Lanczos process with maztrix A and starting vector v_1

choose v_1 such that $\|v_1\| = 1$
let $\beta_0 := 0$, $v_0 := 0$
for $j = 1, \ldots, m$ **do**
  $w_j = Av_j - \beta_{j-1} v_{j-1}$
  $\alpha_j = v_j^* w_j$
  $w_j = w_j - \alpha_j v_j$
  $\beta_j = \|w_j\|_2$
  if $\beta_j = 0$ **then** stop
  $v_{j+1} = (1/\beta_j) \cdot w_j$
end

$e_m = (0, \ldots, 0, 1)^* \in \mathbb{C}^m$ and

$$T_{m+1,m} = \begin{bmatrix} \alpha_1 & \beta_1 & & \\ \beta_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_{m-1} \\ & & \beta_{m-1} & \alpha_m \\ & & & \beta_m \end{bmatrix} = \begin{bmatrix} T_m \\ \beta_m \cdot e_m^* \end{bmatrix} \in \mathbb{R}^{(m+1)\times m}$$

with T_m a (real) symmetric tridiagonal matrix.

Let $g(t) = \sum_{i=1}^{p} \frac{\omega_i}{t-\sigma_i}$ be the Zolotarev approximation to $t^{-1/2}$. We get the m-th Lanczos approximation to $g(Q^2)(Qb)$, which in turn approximates $\mathrm{sign}(Q)b$, by running a multishift cg method, based on the Lanczos process, for the p systems $(Q^2 - \sigma_i I)x_i = Qb$. This is summarized as Algorithm 2.2, where $A = Q^2, c = Qb$. Herein, the factors $\rho_m^{(i)}$ are the scaling factors between the Lanczos vector and the residuals, see [8]:

$$r_m^{(i)} = Qb - Q^2 x_m^{(i)} = \rho_m^{(i)} v_{m+1}, \quad \text{and } \rho_m^{(i)} = (-1)^m \|r_m^{(i)}\|_2. \tag{2}$$

For the error e_m of the m-th approximation x_m we obtain

$$\underbrace{\sum_{i=1}^{p} \omega_i (Q^2 - \sigma_i I)^{-1}(Qb)}_{:=x_*} - x_m = \sum_{i=1}^{p} \omega_i (Q^2 - \sigma_i I)^{-1} r_m^{(i)} = \sum_{i=1}^{p} \rho_m^{(i)} \omega_i (Q^2 - \sigma_i I)^{-1} v_{m+1},$$

so we can express $\|e_m\|^2$ as

$$\|e_m\|^2 = \|g_m(Q^2) v_{m+1}\|^2 = v_{m+1}^* g_m^2(Q^2) v_{m+1}, \text{ where } g_m(t) = \sum_{i=1}^{p} \frac{\rho_m^{(i)} \omega_i}{t - \sigma_i}. \tag{3}$$

The elegant theory of moments and quadrature developed in [5, 6] allows to bound this quantity, and more generally quantities of the form $v^* h(A) v$, from below and from above by performing some steps of the Lanczos process for Q^2 with starting vector v_{m+1}. The precise results is as follows:

Algorithm 2.2: Multishift cg

set $x_{-1} = 0$, $\rho_0^{(i)} = \|c\|_2$, $\tau_0^{(i)} = 1$, $v_1 = (1/\|c\|)c$
for $j = 0, 1, \ldots$ **do**
  compute α_{j+1}, β_{j+1}, v_{j+2} using the Lanczos process for A **for** $i = 1, \ldots, p$ **do**
    if $j > 0$ **then**
      $\tau_j^{(i)} = \left[1 - \frac{\alpha_j - \sigma_i}{\alpha_{j+1} - \sigma_i} \left(\frac{\rho_j^{(i)}}{\rho_{j-1}^{(i)}}\right)^2 \frac{1}{\tau_{j-1}^{(i)}}\right]^{-1}$
    end
    $\rho_{j+1}^{(i)} = -\tau_j^{(i)} \rho_j^{(i)} \frac{\beta_{j+1}}{\alpha_{j+1} - \sigma_i}$
    $x_{j+1}^{(i)} = \tau_j^{(i)}(x_j^{(i)} + \frac{1}{\alpha_{j+1}} r_j^{(i)}) + (1 - \tau_j^{(i)}) x_{j-1}^{(i)}$
    $r_{j+1}^{(i)} = \rho_{j+1}^{(i)} v_{j+2}$
  end
  $x_m = \sum_{i=1}^{p} x_m^{(i)}$;
end

Theorem 1 *Let $\hat{T}_k$ denote the tridiagonal matrix in the Lanczos relation (1) arising after k steps of the Lanczos process with starting vector $v, \|v\| = 1$. Assume that $h : \mathbb{R} \to \mathbb{R}$ is at least $2k + 2$ times continuously differentiable on an open set containing $[a, b]$, where* $\mathrm{spec}(A) \subseteq [a, b]$.

(i) Approximating $v^* h(A) v$ with the Gauss quadrature rule using k nodes $t_j \in (a, b)$ gives

$$v^* h(A) v = e_1^* h(T_k^{\mathrm{G}}) e_1 + R_k^{\mathrm{G}}[h], \text{ where } T_k^{\mathrm{G}} = \hat{T}_k,$$

with the error $R_k^{\mathrm{G}}[h]$ given as

$$R_k^{\mathrm{G}}[h] = \frac{h^{(2k)}(\xi)}{(2k)!} \int_a^b \left[\prod_{j=1}^{k} (t - t_j)\right]^2 d\gamma(t), \quad a < \xi < b\ . \tag{4}$$

(ii) Approximating $v^* h(A) v$ with the Gauss-Radau quadrature rule using $k-1$ nodes $t_j \in (a, b)$ with one additional node fixed at a gives

$$v^* h(A) v = e_1^* h(T_k^{\mathrm{GR}}) e_1 + R_k^{\mathrm{GR}}[h].$$

Here, the tridiagonal matrix T_k^{GR} differs from $\hat{T}_k$ in that its (k, k) entry α_k is replaced by $\widetilde{\alpha}_k = a + \delta_{k-1}$, where δ_{k-1} is the last entry of the vector δ with $(\hat{T}_{k-1} - aI)\delta = \beta_{k-1}^2 e_{k-1}$. The error $R_k^{\mathrm{GR}}[h]$ is given as

$$R_k^{\mathrm{GR}}[h] = \frac{h^{(2k-1)}(\xi)}{(2k-1)!} \int_a^b (t - a) \left[\prod_{j=1}^{k-1} (t - t_j)\right]^2 d\gamma(t), \quad a < \xi < b\ . \tag{5}$$

(iii) Approximating $v^*h(A)v$ with the Gauss-Lobatto quadrature rule using $k-2$ nodes $t_j \in (a,b)$ and two additional nodes, one fixed at a and one fixed at b, gives

$$v^*h(A)v = e_1^*h(T_k^{\mathrm{GL}})e_1 + R_k^{\mathrm{GL}}[h].$$

Here, the tridiagonal matrix T_k^{GL} differs from $\hat{T}_k$ in its last column and row. With δ and μ the solutions of the system $(\hat{T}_{k-1} - aI)\delta = e_{k-1}$, $(\hat{T}_{k-1} - bI)\mu = e_{k-1}$ and $\widetilde{\alpha}_k, \widetilde{\beta}_{k-1}^2$ the solution of the linear system

$$\begin{bmatrix} 1 & -\delta_k \\ 1 & -\mu_k \end{bmatrix} \begin{bmatrix} \widetilde{\alpha}_k \\ \widetilde{\beta}_{k-1}^2 \end{bmatrix} = \begin{bmatrix} a \\ b \end{bmatrix},$$

the tridiagonal matrix T_k^{GL} is obtained from $\hat{T}_k$ by replacing α_k by $\widetilde{\alpha}_k$ and β_{k-1} by $\widetilde{\beta}_{k-1}$. The error $R_k^{\mathrm{GL}}[h]$ is given as

$$R_k^{\mathrm{GL}}[h] = \frac{h^{(2k-2)}(\xi)}{(2k-2)!} \int_a^b (t-a)(t-b) \left[\prod_{j=1}^{k-2}(t-t_j)\right]^2 d\gamma(t), \quad a < \xi < b\ . \tag{6}$$

W apply Theorem 1 to the rational functions $h = g_m^2$ representing the error in (3) Inspecting the terms $R_k^{\mathrm{G}}[h]$, $R_k^{\mathrm{GR}}[h]$ and $R_k^{\mathrm{GL}}[h]$ and noticing that $h^{(\ell)}(t) < 0$ (> 0) for $t \in [0,\infty)$ if ℓ is odd (even), we get the following corollary.

Corollary 1 *In the case $h(t) = g_m(t)^2$ with g_m from (3), the estimates $e_1^*h(T_k^{\mathrm{G}})e_1$ and $e_1^*h(T_k^{\mathrm{GL}})e_1$ from Theorem 1 (i), (iii) represent lower bounds, the estimate $e_1^*h(T_k^{\mathrm{GR}})e_1$ from (ii) represents an upper bound for the (square of the) error $\|x_m - x_*\|^2$.*

3 Lanczos restart recovery

To avoid ambiguities, let us call *primary* Lanczos process the one of the multishift cg method, i.e. the Lanczos process through which we obtain the approximations x_m. The straightforward way to obtain the error estimates from Theorem 1 would be to perform k steps of a new, *restarted* Lanczos process which takes the current Lanczos vector v_{m+1} of the primary process as its starting vector. This results in the restarted Lanczos relation

$$AV_k^{\mathrm{r}} = V_{k+1}^{\mathrm{r}} T_{k+1,k}^{\mathrm{r}}, \tag{7}$$

and we can now apply the theorem using the tridiagonal matrix T_k^{r} arising from the restarted process. This is, however, far too costly in practice: computing the error estimate would require k multiplications with A—approximately the same amount of work that we would need to advance the primary iteration from step m to $m+k$.

Fortunately, it is possible to cheaply retrieve the matrix T_k^{r} of the secondary Lanczos process from the matrix T_{m+1+k} of the primary Lanczos process. This *Lanczos restart recovery* opens the way to efficiently obtain all the error estimates from Theorem 1 in a retrospective manner: At iteration $m+k$ we get the estimates for the error at iteration m without using any matrix-vector multiplications with A and with cost $\mathcal{O}(k^2)$, independently of the system size n.

For $m = 0, 1, \ldots$, we define the tridiagonal matrix $T^{(m+1,k)}$ as the diagonal block of T_{m+1+k} ranging from rows and columns $\max\{1, m+1-k\}$ to $m+1+k$. So $T^{(m+1,k)}$ is a $(2k+1) \times (2k+1)$ matrix, except for $m+1 \leq k$, where its size is $(m+1+k) \times (m+1+k)$.

The following theorem, see [3], shows that for Lanczos restart recovery we basically have to run the Lanczos process for the tridiagonal matrix $T^{(m+1,k)}$, starting with the $k+1$st unit vector $e_{k+1} \in \mathbb{C}^{2k+1}$.

Theorem 2 *Let the Lanczos relation for k steps of the Lanczos process for $T^{(m+1,k)}$ with starting vector $e_{k+1} \in \mathbb{C}^{2k+1}$ ($e_{m+1} \in \mathbb{C}^{m+1+k}$ if $m+1 \leq k$) be given as*

$$T^{(m+1,k)} \widetilde{V}_k = \widetilde{V}_{k+1,k} \widetilde{T}_{k+1,k}. \tag{8}$$

Then the matrix $T_{k+1,k}^{\mathrm{r}}$ of the restarted Lanczos relation (7) is given as

$$T_{k+1,k}^{\mathrm{r}} = \widetilde{T}_{k+1,k}. \tag{9}$$

The above theorem shows that we can retrieve $T_{k+1,k}^{\mathrm{r}}$ from $T_{m+k+1,m+k}$ by performing k steps of the Lanczos process for the $(2k+1) \times (2k+1)$ tridiagonal matrix $T^{(m+1,k)}$. Herein, each step has work $\mathcal{O}(k)$, so that the overall cost for computing $T_{k+1,k}^{\mathrm{r}}$ is $\mathcal{O}(k^2)$. So we conclude that the total cost for computing the error estimates from Theorem 1 is also $\mathcal{O}(k^2)$.

Algorithm 3.1 shows how we suggest to use the results exposed so far. It computes the Lanczos approximations x_m for $g(A)b$ with $g(t) = \sum_{i=1}^{p} \frac{\omega_i}{t-\sigma_i}$ and bounds ℓ_{m-k}, u_{m-k} for the error at iteration m based on the Gauss and the Gauss-Radau rule. The Algorithm can be modified to also obtain error estimates or bounds based on the Gauss-Lobatto rule and to get bounds for the A-norm in case we deal with a linear system.

4 Numerical results

In this section we report the results of several numerical experiments with relatively small lattices of size 8^4 to 16^4. In our computations we used the common deflation technique as described, e.g. in [9]: We precompute the first, $\lambda_1, \ldots, \lambda_q$ say, eigenpairs of smallest modulus. With Π denoting the orthogonal projection onto the space spanned by the corresponding eigenvectors, we then have $\mathrm{sign}(Q)b = \mathrm{sign}(Q(I-\Pi)b) + \mathrm{sign}(Q\Pi b)$. Herein, we know $\mathrm{sign}(Q\Pi b)$ explicitly, so that we now just have to approximate $\mathrm{sign}(Q(I-\Pi)b)$. In this manner, we effectively shrink the eigenvalue intervals for Q, so that we need fewer poles for an accurate Zolotarev approximation and, in addition, the linear systems

Algorithm 3.1: Lanczos approximation for Zolotarev function with error bounds

set $x_{-1} = 0$, $\rho_0 = \|b\|_2$, $\tau_0 = 1$
choose k
for $m = 0, 1, \ldots$ **do**
  compute α_{m+1}, β_{m+1}, v_{m+2} using the Lanczos process for A
  for $i = 1, \ldots, p$ **do** `/* loop over poles */`
    if $m > 0$ **then**
      $\tau_m^{(i)} = \left[1 - \frac{\alpha_m - \sigma_i}{\alpha_{m+1} - \sigma_i} \left(\frac{\rho_m^{(i)}}{\rho_{m-1}^{(i)}}\right)^2 \frac{1}{\tau_{m-1}^{(i)}}\right]^{-1}$
    end
    $\rho_{m+1}^{(i)} = -\tau_m^{(i)} \rho_m^{(i)} \frac{\beta_{m+1}}{\alpha_{m+1} - \sigma_i}$
    $x_{m+1}^{(i)} = \tau_m^{(i)} \left(x_m^{(i)} + \frac{\rho_m^{(i)}}{\alpha_{m+1} - \sigma_i} v_{m+1}\right) + \left(1 - \tau_j^{(i)}\right) x_{m-1}^{(i)}$
  end
  $x_{m+1} = \sum_{i=1}^{p} \omega_i x_{m+1}^{(i)}$
  if $m > k$ **then**
    perform k steps of the Lanczos process for $T^{(m-k,k)}$
    this yields the tridiagonal matrix $\hat{T}_k \in \mathbb{C}^{k \times k}$
    $\ell_{m-k} = \|g_m(\hat{T}_k) e_1\|_2$
    $u_{m-k} = \|g_m(\hat{T}^{\mathrm{GR}}) e_1\|_2$ `/* ` $\hat{T}_k, \hat{T}^{\mathrm{GR}}$ `given in Theorem 1(ii) */`
  end
end

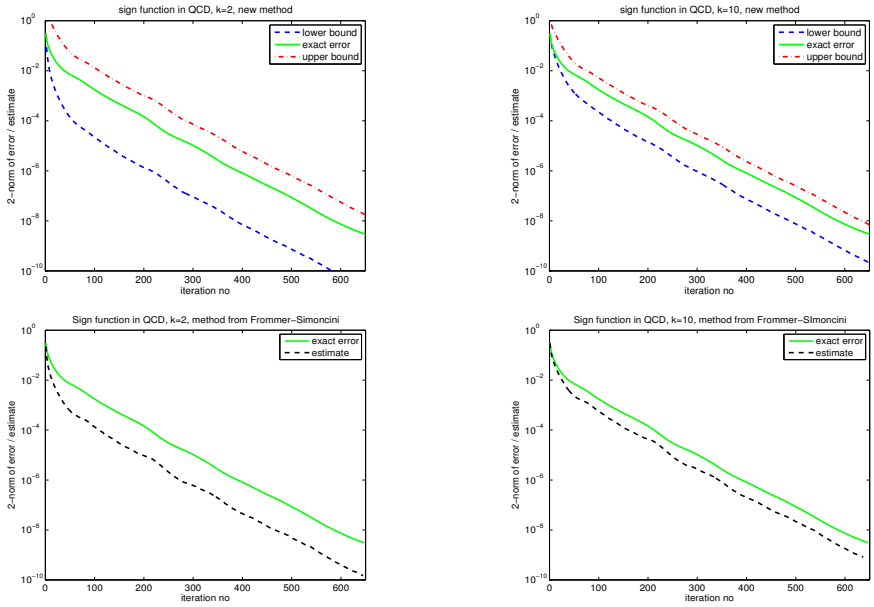

Figure 1: Error bounds and exact error for Zolotarev approximation for sign(Q), 8^4 lattice. Left column: $k = 2$, right column: $k = 10$. Top row: Algorithm 3.1, bottom row: method from [4].

to be solved converge more rapidly. Within an iterative solver for the overlap operator this approach results in a major speedup, since sign$(Q)b$ must usually be computed repeatedly for various vectors b. For Algorithm 3.1 it has the additional advantage that we immediately have a very good value for a, the lower bound on the smallest eigenvalue of Q^2 for which we can take λ_q^2. In all our computations we deflated the smallest 30 eigenvalues, and we chose the Zolotarev approximation to have error less than 10^{-9}.

Figure 1 shows results for the 8^4 configuration available in the matrix group `QCD` at the UFL sparse matrix collection [2] as matrix `conf5.4-0018x8-2000.mtx`. This is a dynamicyally generated configuration at $\beta = 5.4$. The (effective) condition number of the (deflated) matrix Q^2 is approximately $4.5 \cdot 10^4$. The left column of the figure reports upper and lower bounds from Algorithm 3.1 whereas the right columns gives the estimates from earlier work [4] which are known know to be lower bounds. The top row takes $k = 2$ in Algorithm 3.1 (and a similar parameter in the method from [4]), and the bottom row refers to $k = 10$. We see that going from $k = 2$ to 10 results in a significant gain in accuracy and that for $k = 10$ the upper and lower bounds just differ by a factor of 10.

Figure 2 gives the results for Algorithm 3.1 with $k = 10$ for a configuration on a 16^4 lattice. The configuration was the result of a quenched simulation.

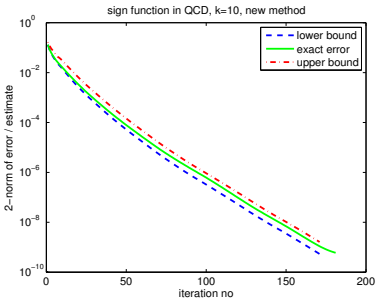

Figure 2: Error bounds and exact error for Zolotarev approximation for sign(Q), 16^4 lattice, Algorithm 3.1.

The condition number of the deflated matrix Q^2 is now 64^2, i.e. less than for the 8^4 lattice. Therefore, the convergence speed as well as the quality of the bounds are better than for the 8^4 lattice.

References

[1] N. Cundy, J. van den Eshof, A. Frommer, S. Krieg, and K. Schäfer. Numerical methods for the QCD overlap operator. III: Nested iterations. *Comput. Phys. Commun.*, 165:221–242, 2005.

[2] T. A. Davis and Y. F. Hu. The University of Florida sparse matrix collection. `http://www.cise.ufl.edu/research/sparse/matrices/`.

[3] A. Frommer, K. Kahl, T. Lippert, and H. Rittich. 2-norm error bounds and estimates for Lanczos approximations to linear systems and rational matrix functions, in preparation.

[4] A. Frommer and V. Simoncini. Stopping criteria for rational matrix functions of hermitian and symmetric matrices. *SIAM J. Sci. Comp.*, 30:1387–1412, 2008.

[5] G. H. Golub and G. Meurant. Matrices, moments and quadrature. In D. Griffiths and G. W. Eds., editors, *Numerical Analysis 1993*, volume 303 of *Pitman Research Notes in Mathematics Series*, pages 105–156. Longman Scientific & Technical, Harlow, 1994.

[6] G. H. Golub and G. Meurant. Matrices, moments and quadrature. II. How to compute the norm of the error in iterative methods. *BIT*, 37(3):687–705, 1997.

[7] R. Narayanan and H. Neuberger. An alternative to domain wall fermions. *Phys. Rev.*, D62:074504, 2000.

[8] C. C. Paige, B. N. Parlett, and H. A. van der Vorst. Approximate solutions and eigenvalue bounds from Krylov subspaces. *Numer. Linear Algebra Appl.*, 2:115–134, 1995.

[9] J. van den Eshof, A. Frommer, T. Lippert, K. Schilling, and H. A. van der Vorst. Numerical methods for the QCD overlap operator. I: Sign-function and error bounds. *Comput. Phys. Commun.*, 146:203–224, 2002.