STOPPING CRITERIA FOR RATIONAL MATRIX FUNCTIONS OF HERMITIAN AND SYMMETRIC MATRICES

ANDREAS FROMMER* AND VALERIA SIMONCINI†

Abstract. Building upon earlier work by Golub, Meurant, Strakos and Tichy, we derive new *a posteriori* error bounds for Krylov subspace approximations to $f(A)b$, the action of a function f of a real symmetric or complex Hermitian matrix A on a vector b. To this purpose we assume that a rational function in partial fraction expansion form is used to approximate f, and the Krylov subspace approximations are obtained as linear combinations of Galerkin approximations to the individual terms in the partial fraction expansion. Our error estimates come at very low computational cost. In certain important special situations they can be shown to actually be lower bounds of the error. Our numerical results include experiments with the matrix exponential, as used in exponential integrators, and with the matrix sign function, as used in lattice QCD simulations, and demonstrate the accuracy of the estimates. The use of our error estimates within acceleration procedures is also discussed.

Key words. Matrix functions, partial fraction expansions, error estimates, error bounds, Lanczos method, Arnoldi method, CG iteration

AMS subject classifications. 65F30, 65F10, 65F50.

1. Introduction. Matrix functions arise in a large number of application problems, and over the last years efforts to enhance their effective numerical computation have significantly encouraged their use in discretization and approximation methods. These include exponential integrators, which require the computation of the matrix exponential $\exp(A)$ or of $\varphi(A)$ with $\varphi(t) = (\exp(t) - 1)/t$, and have recently emerged for numerically solving stiff or oscillatory systems of ordinary differential equations; see, e.g., [22], [15]. Another example arises when simulating chiral fermions in lattice QCD. Here, one has to solve linear systems of the form $(P + \mathrm{sign}(A))x = b$, where P is a permutation matrix and A is the Wilson fermion matrix; see [3].

In general, given a square matrix A, the matrix function $f(A)$ can be defined for a sufficiently smooth function f by means of an appropriate spectral decomposition of A; see, e.g., [24]. In both the examples cited above, as well as in many other applications, the matrix A can be very large. Then it is practically impossible to explicitly compute $f(A)$ by means of a spectral decomposition of A, since $f(A)$ will in general be full even if A is sparse. Fortunately, in these applications the knowledge of the action of the matrix function on a vector is usually all that is required, and not the matrix function itself. For example, when solving the system $(P + \mathrm{sign}(A))x = b$ with an iterative solver, each step will usually require the computation of $(P + \mathrm{sign}(A))p$ for a certain vector p which changes at each iteration.

In the case of some functions such as the exponential, the sign, the square-root and trigonometric functions, a particularly attractive approach for large matrices exploits the powerful rational function approximation

$$f(t) \approx g(t) = \frac{p_{s_1}(t)}{p_s(t)},$$

where $p_i(t)$ are polynomials of degree i; see, e.g., [37], [2]. The built-in Matlab ([27])

*Fachbereich Mathematik und Naturwissenschaften, Universität Wuppertal, Gauß-Straße 20, D-42097 Wuppertal, Germany (frommer@math.uni-wuppertal.de). Partly supported by DFG grant Fr755 17-2.

†Dipartimento di Matematica, Università di Bologna, Piazza di Porta S. Donato, 5, I-40127 Bologna, Italy and CIRSA, Ravenna, Italy (valeria@dm.unibo.it).

function for the matrix exponential, for example, uses a Padé rational approximation. Rational functions may be conveniently employed in a matrix context by using a partial fraction expansion, namely (assuming that multiple poles do not occur)

$$g(t) = \frac{p_{s_1}(t)}{p_s(t)} = p_{s_2}(t) + \sum_{i=1}^{s} \omega_i \frac{1}{t - \sigma_i}. \tag{1.1}$$

Since the computation of $p_{s_2}(A)b$ is trivial – although numerical accuracy may become an issue if $\|A\|$ is large –, we assume $p_{s_2} = 0$ and concentrate on the sum representing the fractional part. When applied to a matrix A, this gives

$$z = g(A)b = \sum_{i=1}^{s} \omega_i (A - \sigma_i I)^{-1} b = \sum_{i=1}^{s} \omega_i x_i. \tag{1.2}$$

For large dimensional problems, the solutions x_i of the shifted linear systems

$$(A - \sigma_i I)x_i = b, \qquad i = 1, \ldots, s, \tag{1.3}$$

are approximated by more cheaply computable vectors $\widetilde{x}_i$, so as to obtain $\widetilde{z} = \sum_{i=1}^{s} \omega_i \widetilde{x}_i$. This can be done by using a single approximation space for all shifted systems. More precisely, if the columns of the matrix V span such an approximation space, then each approximate system solution may be written as $\widetilde{x}_i = V y_i$ for some y_i, $i = 1, \ldots, s$, so that

$$\widetilde{z} = \sum_{i=1}^{s} \omega_i \widetilde{x}_i = V \sum_{i=1}^{s} \omega_i y_i. \tag{1.4}$$

This strategy is particularly welcomed in view of the fact that constructing the approximation space is commonly the most expensive step in the whole approximation process. It is important to realize that such a procedure corresponds to employing a projection technique for $g(A)b$; we refer to [9] for a discussion of this connection and for further references.

The aim of this paper is to obtain cheap as well as sufficiently accurate estimates of the Euclidean norm of the error

$$\|\widetilde{z} - g(A)b\|, \tag{1.5}$$

which can then serve as a stopping criterion for an iterative process computing a sequence of approximations $\widetilde{z}$. To this end, we build upon a large amount of work available in the literature, on estimating the energy-norm or Euclidean norm of the error when the Conjugate Gradient method (CG) is used to solve $Mx = b$ with M Hermitian and positive definite (hereafter Hpd); see, e.g., [4, 12, 13, 29, 31, 44, 45].

We recall at this point that ad-hoc (mostly a-priori) error estimates for projection-type approximations of the matrix exponential have been recently developed [7], [43], [21]. On the other hand, a-posteriori bounds have also been derived for the matrix sign function [46].

The estimates proposed in this paper directly aim at approximating the error norm (1.5) of the rational approximation and are therefore of a general nature. The overall approximation error is then a combination of the error of the rational function as an approximation to f, a quantity which is usually known *a priori*, and the error (1.5). Our estimates are particularly useful in two situations: the matrix A is Hpd

and all σ_i are negative or A is real symmetric with the σ_i being complex. We show experimentally that the new estimates can be very tight, and that they are lower bounds when $A-\sigma_i I$ is Hpd for all i. The stopping criterion associated with the bound can be cheaply included in Krylov subspace projection-type methods for solving the combined shifted linear systems, and we discuss the corresponding implementation within the Lanczos and the CG framework.

The paper is organized as follows: In Section 2 we review some important facts for Krylov subspace approximations to the solution of linear systems and we establish the connection between the projection approach and Galerkin approximations. Section 3 discusses in more detail the CG method as an iterative method producing Galerkin approximations. After briefly explaining the known fact that CG can be implemented very efficiently in the context of multiple shifted linear systems, we highlight the role of the computed CG coefficients as ingredients to error estimates for the individual systems. In Section 4 we shortly discuss alternative estimates and stopping criteria. Section 5 is the core section where we develop our new error estimates and prove some results on their accuracy. In Section 6 we discuss details of an efficient implementation. Numerical experiments are contained in Section 7 and, finally, Section 8 shows how our estimates can be extended to shift-and-invert acceleration procedures that were recently proposed in [23] and [32].

Exact arithmetic is assumed throughout. This is a crucial issue, since rounding errors may substantially affect computed quantities in our context. To ensure the practical usefulness of the error estimates and the related stopping criteria to be developed in this work one thus needs an additional error analysis with respect to floating point arithmetic. Such an error analysis is beyond the scope of the work presented here. Let us just mention that in the case of the CG iteration such an error analysis has been presented in [44]. The fact that the stopping criteria to be developed here build on those for CG which have been identified as stable in [44] may thus be taken as a first hint on their numerical reliability.

We close this introduction with a word on notation. As a rule, we will use superscript to denote the iteration number, and subscript to denote the associated system, so that $r_i^{(k)}$ is the residual after k iterations of the ith system. For certain matrices that are independent of the systems considered, we will prefer the simpler notation V_k (say) rather than $V^{(k)}$. We use the Euclidean 2-norm for vectors. Moreover, for a column vector $x \in \mathbb{C}^n$, we use $x^{\mathtt{T}} = [(x)_1, \ldots, (x)_n]$ and $x^{\mathtt{H}} = [(\overline{x})_1, \ldots, (\overline{x})_n]$, where $(v)_i$ denotes the ith component of the vector v, and $(\overline{x})_i$ the conjugate of the ith component of x.

2. Basic facts on Krylov subspace approximation. Given a linear system $Mx = b$ and an approximation space $\mathcal{K}$, an approximate solution $\widetilde{x} \in \mathcal{K}$ to x may be obtained by imposing some orthogonality condition on the corresponding residual $\widetilde{r} = b - M\widetilde{x}$. For M Hermitian or complex symmetric, a common strategy is to impose that the residual be orthogonal to the approximation space, and this is called the Galerkin condition. Such orthogonality condition may be given, for instance, in terms of the standard Euclidean inner product, that is,

$$v^{\mathtt{H}}\widetilde{r} := \sum_{i=1}^{n} (\overline{v})_i (\widetilde{r})_i = 0, \quad v \in \mathcal{K}. \tag{2.1}$$

A particularly convenient approximation space is given by the Krylov subspace $\mathcal{K}_m(M, b) = \mathrm{span}\{b, Mb, M^2b, \ldots, M^{m-1}b\}$. Starting with $v^{(0)}$, a normalized ver-

sion of b, a basis of this space can be constructed iteratively, one vector at a time, as $\{v^{(0)}, v^{(1)}, \ldots, v^{(m-1)}\}$, so that for each m, $\mathcal{K}_m \subseteq \mathcal{K}_{m+1}$. The following classical matrix relation, the Arnoldi recurrence, provides the overall procedure for generating a nested orthonormal basis for $\mathcal{K}_m(A, b)$ as the columns of the matrices $V_m = [v^{(0)}, v^{(1)}, \ldots, v^{(m-1)}]$,

$$MV_m = V_m T_m + v^{(m)} t_{m+1,m} e_m^{\mathtt{T}}, \qquad V_m = [v^{(0)}, v^{(1)}, \ldots, v^{(m-1)}]. \tag{2.2}$$

Here, T_m is an $m \times m$ upper Hessenberg matrix which reduces to a tridiagonal matrix if M is self-adjoint with respect to the inner product in use. The vector e_m is the mth vector of the canonical basis, and $t_{m+1,m}$ is a normalization factor [41]. Note that while the entries in V_m and T_m depend on the underlying inner product, the general recursive form (2.2) remains unchanged. In case of the standard Euclidean inner product, V_m will satisfy $V_m^{\mathtt{H}} V_m = I_m$, the identity matrix of dimension m, and if $M \in \mathbb{C}^{n\times n}$ is Hermitian, T_m is tridiagonal and the Arnoldi recurrence (2.2) simplifies accordingly (see below).

The choice of an appropriate inner product depends on the matrix properties. In our context, the use of a bilinear form different from the Euclidean inner product turns out to be very convenient in the case where $M \in \mathbb{C}^{n\times n}$ is complex symmetric ($M = M^{\mathtt{T}}$). In this case, M is self-adjoint with respect to the bilinear form $x^{\mathtt{T}} y = \sum_{i=1}^{n} (x)_i (y)_i$. To make the notation more uniform, we use $\langle \cdot, \cdot \rangle_*$ to denote either of

$$\langle x, y \rangle_{\mathtt{H}} := x^{\mathtt{H}} y = \sum_{i=1}^{n} (\overline{x})_i (y)_i, \qquad \langle x, y \rangle_{\mathtt{T}} := x^{\mathtt{T}} y = \sum_{i=1}^{n} (x)_i (y)_i.$$

Both forms will be called an 'inner product', the first one being definite, the second one indefinite. Hence, we assume that the basis generated with the Arnoldi process (2.2) is orthogonal with respect to the employed inner product $\langle \cdot, \cdot \rangle_*$, when no additional specification is needed. With a slight abuse of terminology, we will use the following definition from now on:

DEFINITION 2.1. *A shifted complex symmetric matrix is a matrix of the form $M = A - \sigma I$, with A real symmetric, and $\sigma \in \mathbb{C}$.*

Note that shifted complex symmetric matrices are a subclass of the complex symmetric matrices.

When M is Hermitian (resp. complex symmetric) and the basis vectors are orthogonal with respect to $\langle \cdot, \cdot \rangle_{\mathtt{H}}$ (resp. $\langle \cdot, \cdot \rangle_{\mathtt{T}}$), the upper Hessenberg matrix $T_m = V_m^{\mathtt{H}} M V_m$ is Hermitian (resp. $T_m = V_m^{\mathtt{T}} M V_m$ is complex symmetric) and thus tridiagonal. This allows one to derive the next basis vector $v^{(m)}$ in (2.2) by only using the previous two basis vectors. The resulting short-term recurrence is the well-known Lanczos procedure for generating the Krylov subspace associated with M and b. We will henceforth always employ $*$=H when M is Hpd, and $*$=T when M is shifted complex symmetric.

We next recall some key facts about Galerkin approximations in the Krylov subspace when the coefficient matrix has a shifted form; see, e.g., [35], [36].

PROPOSITION 2.2. *Let the system $Mx = b$ be given, with $M = A - \sigma I$ for some $\sigma \in \mathbb{C}$. Then*

1. *$\mathcal{K}_m(A - \sigma I, b) = \mathcal{K}_m(A, b)$ (invariance with respect to shift)*
2. *Let $\widetilde{x}(\sigma) \in \mathcal{K}_m(A, b)$ be the Galerkin approximation to $Mx = b$ with the given inner product, and $\widetilde{r}(\sigma) = b - M\widetilde{x}(\sigma)$. For any $\sigma \in \mathbb{C}$ we have*

$$\widetilde{r}(\sigma) = (-1)^m \rho(\sigma) v^{(m)},$$

where $v^{(m)}$ is the $(m+1)$st Krylov subspace basis vector from (2.2) and $|\rho(\sigma)| = \|\widetilde{r}(\sigma)\|$.

The first result shows that when solving systems that only differ for the shifting parameter σ, approximations can be carried out in a single approximation space. The second result says that the residuals associated with the shifted systems are all collinear to the next basis vector.

In terms of the Arnoldi relation, shift invariance results in

$$(A-\sigma I)V_m = V_m(T_m-\sigma I) + v^{(m)}t_{m+1,m}e_m^{\mathtt{T}}, \qquad V_m^*V_m = I,$$

where V_m and T_m are the same for all σ's. Denoting by $\widetilde{x}(\sigma) = V_m y(\sigma) \in \mathcal{K}_m(A,b)$ the Galerkin approximation for $x(\sigma)$, the solution of $(A-\sigma I)x = b$, we have

$$0 = V_m^*b - V_m^*(A-\sigma I)V_m y(\sigma) = e_1\beta - (T_m-\sigma I)y(\sigma), \quad \beta = \langle b,b\rangle_*^{1/2}.$$

Assuming $(T_m-\sigma I)$ nonsingular, it follows that $y(\sigma) = (T_m-\sigma I)^{-1}e_1\beta$ so that

$$\widetilde{x}(\sigma) = V_m y(\sigma) = V_m(T_m-\sigma I)^{-1}e_1\beta. \tag{2.3}$$

By substituting this quantity in the residual $\widetilde{r}(\sigma) = b - M\widetilde{x}(\sigma)$, it also follows that $\widetilde{r}(\sigma) = v^{(m)}e_m^{\mathtt{T}}y(\sigma)t_{m+1,m}$, which is related to Proposition 2.2, part 2. In particular, this relation shows that neither $\widetilde{x}(\sigma)$ nor $\widetilde{r}(\sigma)$ need to be explicitly computed to get

$$\langle \widetilde{r}(\sigma), \widetilde{r}(\sigma)\rangle_* = (e_m^{\mathtt{T}}y(\sigma)t_{m+1,m})^*(e_m^{\mathtt{T}}y(\sigma)t_{m+1,m}). \tag{2.4}$$

REMARK 2.3. If A in $M = A-\sigma I$ is real and b is also real, then V_m is real. Complex arithmetic for $\sigma \in \mathbb{C}$ only arises in the computation of the approximate solution $y(\sigma)$. Moreover, all residuals are complex multiples of a real vector. In this case, we also have a particularly simple relation between the iterates belonging to pairs of conjugate shifts. If $\widetilde{x}(\overline{\sigma}) = V_m y(\overline{\sigma})$ is the approximate solution to $(A-\overline{\sigma}I)x = b$ in (2.3), then

$$y(\overline{\sigma}) = (T_m-\overline{\sigma}I)^{-1}e_1\beta = \overline{(T_m-\sigma I)^{-1}e_1\beta} = \overline{y(\sigma)}.$$

This shows that $\widetilde{x}(\overline{\sigma})$ is identical to $\overline{\widetilde{x}(\sigma)}$, so that $\widetilde{x}(\overline{\sigma})$ needs not be computed explicitly. Clearly, one also has $\widetilde{r}(\overline{\sigma}) = \overline{\widetilde{r}(\sigma)}$.

3. CG-type approximations and their error estimates. In the previous section we recalled that when T_m is tridiagonal the Arnoldi process reduces to a short (three)-term recurrence. In fact, if in addition T_m can be factorized as a product of two bidiagonal matrices, then a coupled two-term recurrence can be obtained. In the case when M (and thus T_m) is Hpd, then $T_m = L_m L_m^{\mathtt{H}}$, with L_m lower bidiagonal, which gives rise to the classical Conjugate Gradient method. A possible implementation is reported in Figure 3.1 (take $* = \mathtt{H}$).

In the case when $M = A-\sigma I$ is shifted complex symmetric and b is real, the Lanczos procedure in the inner product $\langle\cdot,\cdot\rangle_{\mathtt{T}}$ yields the shifted complex symmetric tridiagonal matrix $T_m - \sigma I$. An implementation of the Conjugate Gradient method in this inner product has been proposed in [28]. It is also subsumed in Figure 3.1, now with $* = \mathtt{T}$.

We next focus on some specific properties of the CG algorithm that allow us to derive error estimates for the rational function approximation. In section 6 we show that these quantities are easily available in practical algorithms that realize the rational function approximation.

Choose $x^{(0)}$, set $r^{(0)} = b - Mx^{(0)}$, $p^{(0)} = r^{(0)}$
for $k = 1, 2, \ldots$ **do**
  $\gamma^{(k-1)} = \langle r^{(k-1)}, r^{(k-1)} \rangle_* / \langle p^{(k-1)}, Mp^{(k-1)} \rangle_*$
  $x^{(k)} = x^{(k-1)} + \gamma^{(k-1)} p^{(k-1)}$
  $r^{(k)} = r^{(k-1)} - \gamma^{(k-1)} Mp^{(k-1)}$
  $\delta^{(k)} = \langle r^{(k)}, r^{(k)} \rangle_* / \langle r^{(k-1)}, r^{(k-1)} \rangle_*$
  $p^{(k)} = r^{(k)} + \delta^{(k)} p^{(k-1)}$
end for

FIG. 3.1. *CG for the two inner products* ($* = \mathtt{H}$ *or* $* = \mathtt{T}$)

Before we move on to discussing error estimates obtained from the CG coefficients, we need to provide sufficient conditions for the CG recurrence to exist. If $M = A - \sigma I$ with A Hpd and $\sigma \leq 0$ real, then M as well as $T_m - \sigma I$ are also Hpd, therefore the Lanczos approximate solution can be computed by solving the equation (2.3). In addition, a Cholesky factorization of the form $T_m - \sigma I = L_m L_m^{\mathtt{H}}$ exists, which yields the CG recurrence, see [10, 41].

When A is Hermitian, and $\sigma \in \mathbb{C}$ is nonreal, the Lanczos recurrence still exists, T_m is Hermitian and $T_m - \sigma I$ is nonsingular, since T_m has only real eigenvalues. However, the existence of a Cholesky-type factorization of $T_m - \sigma I$ that formally produces the CG recurrence (in the inner product $\langle \cdot, \cdot \rangle_{\mathtt{T}}$), is not obvious. For a general complex symmetric matrix such a factorization does not necessarily exist unless additional hypotheses on the matrix are assumed, see [19]. Fortunately, for the shifted tridiagonal matrices we are dealing with, the existence of the Cholesky-type factorization is guaranteed by the following result. It thus ensures that the CG procedure does not break down, so that the associated coefficients can be recovered.

PROPOSITION 3.1. *Let $\widehat{T}$ be a Hermitian matrix and let σ be a complex number with non-zero imaginary part. Let $T = \widehat{T} - \sigma I$. Then the symmetric root-free Cholesky decomposition $T = LDL^{\mathtt{T}}$ with L lower triangular with unit diagonal, D diagonal, exists.*

Proof. We first show that the traditional LU-decomposition of T without pivoting exists. Therefore, L and U are lower and upper tridiagonal, respectively. By using [18, Theorem 9.2], this decomposition exists, and all diagonal entries of U are non-zero, if $\det(T_i)$ is non-zero for $i = 1, \ldots, k-1$. Here, T_i denotes the $i \times i$ principal submatrix of T. Now note that $\widehat{T}_i$ is Hermitian, since $\widehat{T}$ is. Therefore, if T_i were singular, σ would be an eigenvalue of $\widehat{T}_i$ which is impossible, since all eigenvalues of $\widehat{T}_i$ are real. To conclude, write $U = D\widetilde{L}^{\mathtt{T}}$ with D the diagonal matrix containing the diagonal elements of U. Then $T = LD\widetilde{L}^{\mathtt{T}}$ and $T = T^{\mathtt{T}} = \widetilde{L}DL^{\mathtt{T}}$. By the uniqueness of the LU-factorization we get $\widetilde{L} = L$. ∎

COROLLARY 3.2. *Let $A \in \mathbb{R}^{n \times n}$ be symmetric and let $\sigma \in \mathbb{C}$, $\Im(\sigma) \neq 0$. Then there is no breakdown in the CG algorithm with $\langle \cdot, \cdot \rangle_{\mathtt{T}}$ when applied to $M = A - \sigma I$ with b real.*

Proof. This algorithm updates iterates using the complex Cholesky factorization of the tridiagonal matrix $T_k = \widehat{T}_k - \sigma I$, where $\widehat{T}_k$ comes from the classical Lanczos procedure. Since $\widehat{T}_k = V_k^{\mathtt{T}} A V_k$ is real and symmetric, Proposition 3.1 guarantees the existence of the factorization. ∎

Our error estimates will make use of the following relations in the CG methods. For the standard Euclidean inner product, they date back to the original work of Hestenes and Stiefel [17] and have recently been highlighted in [44, formulas (1.4) -

(1.7)]. For the bilinear form $\langle x, y\rangle_{\mathtt{T}}$, they can be derived in a completely analogous manner so that we do not reproduce a proof here. For the expressions associated with $\langle e^{(k)}, e^{(k)}\rangle_*$, we refer to the corresponding relations in [44, formulas (11.1) - (11.3)]; see also [30] for related results.

LEMMA 3.3. *Let M be either Hpd, or shifted complex symmetric. Let $e^{(k)} = x - x^{(k)}$ denote the error at iteration k where $x^{(k)}$ is the iterate of the CG method for $Mx = b$ and $r^{(k)} = Me^{(k)}$ its residual. For $d \in \mathbb{N}$, using the CG coefficients $\gamma^{(k)}$, denote*

$$\eta^{(k,d)} := \sum_{i=0}^{d-1} \gamma^{(k+i)} \langle r^{(k+i)}, r^{(k+i)}\rangle_*$$

$$\varphi^{(k,d)} := \sum_{i=0}^{d} \frac{\langle p^{(k+i)}, p^{(k+i)}\rangle_*}{\langle p^{(k+i)}, Mp^{(k+i)}\rangle_*} \left(\langle r^{(k+i)}, e^{(k+i)}\rangle_* + \langle r^{(k+i+1)}, e^{(k+i+1)}\rangle_* \right).$$

The following relations hold at any iteration k and for any $d > 0$:

$$\langle r^{(k)}, e^{(k)}\rangle_* = \langle r^{(k+d)}, e^{(k+d)}\rangle_* + \eta^{(k,d)},$$
$$\langle e^{(k)}, e^{(k)}\rangle_* = \langle e^{(k+d)}, e^{(k+d)}\rangle_* + \varphi^{(k,d)}.$$

Note that $\langle r^{(k)}, e^{(k)}\rangle_* = \langle Me^{(k)}, e^{(k)}\rangle_*$. Therefore, in case M is Hpd (and $* = \mathtt{H}$), all inner products in the definition of $\eta^{(k,d)}$ and $\varphi^{(k,d)}$ are positive from which it follows that $\eta^{(k,d)}$ and $\varphi^{(k,d)}$ are both positive and

$$\langle r^{(k)}, e^{(k)}\rangle_{\mathtt{H}} \geq \eta^{(k,d)}, \qquad \langle e^{(k)}, e^{(k)}\rangle_{\mathtt{H}} \geq \varphi^{(k,d)}.$$

In the case that M is shifted complex symmetric, we can only write the estimates

$$\langle r^{(k)}, e^{(k)}\rangle_{\mathtt{T}} \approx \eta^{(k,d)}, \tag{3.1}$$
$$\langle e^{(k)}, e^{(k)}\rangle_{\mathtt{T}} \approx \varphi^{(k,d)}, \tag{3.2}$$

since these are in general complex quantities and $\langle \cdot, \cdot\rangle_{\mathtt{T}}$ is not definite.

In either case, the estimates (3.1) and (3.2) are close to equalities whenever the quantities $\langle r^{(k+d)}, e^{(k+d)}\rangle_*, \langle e^{(k+d)}, e^{(k+d)}\rangle_*$ at iteration $k+d$ are small (but not necessarily very small) compared to those at iteration k, see [44]. Despite the lack of any minimization property in the bilinear form with $* = \mathtt{T}$, it can be shown that the convergence behavior of CG applied to the shifted complex symmetric matrix M is driven by the spectral properties of M in a way somehow similar to the Hpd case; see [26]. We thus expect at least linear convergence for sufficiently large k, and in particular, we expect that the error $e^{(k)}$ decreases with k.

The term $\eta^{(k,d)}$ is very easily computable from the scalar quantities of the CG iteration. For $\varphi^{(k,d)}$ this is not so, since its computation involves the unavailable quantities $\langle r^{(k+i)}, e^{(k+i)}\rangle_*$. Estimating those as in (3.1), we get an estimate $\tau^{(k,d)}$ for $\varphi^{(k,d)}$ with

$$\tau^{(k,d)} = \sum_{i=0}^{d} \frac{\langle p^{(k+i)}, p^{(k+i)}\rangle_*}{\langle p^{(k+i)}, Mp^{(k+i)}\rangle_*} (\eta^{(k+i,d)} + \eta^{(k+i+1,d)}). \tag{3.3}$$

Using this instead of $\varphi^{(k,d)}$ in (3.2) we obtain the estimate

$$\langle e^{(k)}, e^{(k)}\rangle_* \approx \tau^{(k,d)}, \tag{3.4}$$

which is now again very cheaply computable from the CG coefficients. Therefore, after $k+d$ iterations it is possible to compute the estimates $\eta^{(k,d)}$, $\tau^{(k,d)}$ associated with the CG error at iteration k.

For M Hpd and $* =$ "H", we can summarize this discussion as follows; see [44].

LEMMA 3.4. *If M is Hpd, then for $d \in \mathbb{N}$, at any iteration k it holds*

$$\begin{aligned}
&\langle r^{(k)}, e^{(k)} \rangle_{\mathtt{H}} \geq \eta^{(k,d)}, \\
&\langle e^{(k)}, e^{(k)} \rangle_{\mathtt{H}} \geq \varphi^{(k,d)}, \\
&\langle e^{(k)}, e^{(k)} \rangle_{\mathtt{H}} \geq \langle e^{(k+d)}, e^{(k+d)} \rangle_{\mathtt{H}} + \tau^{(k,d)} \ \geq \ \tau^{(k,d)}.
\end{aligned}$$

When M is Hpd and has a shifted form, we now show that the norm of the residual of the Galerkin approximation gets smaller as the shift moves the matrix spectrum away from the origin. We will need this result later in Section 5.

LEMMA 3.5. *Assume that A is Hermitian and positive definite and consider the two linear systems $(A-\sigma_i I)x = b$, $\sigma_i \in \mathbb{R}$, $i = 1, 2$ with $0 \geq \sigma_1 > \sigma_2$. Let $v^{(k)}$ denote the $(k+1)$st Lanczos vector. Let $r_i^{(k)} = (-1)^k \rho_i^{(k)} v^{(k)}$, with $\rho_i^{(k)} \in \mathbb{R}, i = 1, 2$, be the residuals associated with a Galerkin approximation in $\mathcal{K}_k(A, b)$. Then for any $k \geq 0$ with $v^{(k)} \neq 0$, we have $\rho_i^{(k)} > 0$, $i = 1, 2$, and, in addition, for $d \geq 0$ with $v^{(k+d)} \neq 0$*

$$\frac{\rho_2^{(k)}}{\rho_1^{(k)}} \geq \frac{\rho_2^{(k+d)}}{\rho_1^{(k+d)}} \geq 0. \tag{3.5}$$

Proof. We use the expressions (see [35],[46])

$$\rho_1^{(k)} = \rho_0^{(k)} \prod_{\nu=1}^{k} \frac{1}{1-\sigma_1/\Theta_\nu^{(k)}}, \qquad \rho_2^{(k)} = \rho_0^{(k)} \prod_{\nu=1}^{k} \frac{1}{1-\sigma_2/\Theta_\nu^{(k)}}.$$

Here, $\rho_0^{(k)} \geq 0$ is the norm of the residual of the Galerkin approximation to $Ax = b$ in $K_k(A, b)$ and the $\Theta_\nu^{(k)}$'s denote the Ritz values of A in $\mathcal{K}_k(A, b)$, i.e. the (all positive) eigenvalues of the Hpd matrix T_k from the Lanczos process. This shows that $\rho_i^{(k)} \geq 0$, $i = 1, 2$. We have

$$\frac{\rho_2^{(k)}}{\rho_1^{(k)}} = \prod_{\nu=1}^{k} \frac{1-\sigma_1/\Theta_\nu^{(k)}}{1-\sigma_2/\Theta_\nu^{(k)}}\,. \tag{3.6}$$

Since the Ritz values for two consecutive values of k interlace, we can order them in such a manner that $\Theta_\nu^{(k+d)} \leq \Theta_\nu^{(k)}$ for $\nu = 1, \ldots, k$. Because $\sigma_2 < \sigma_1 \leq 0$, the fraction $\frac{1-\sigma_1/t}{1-\sigma_2/t} = 1 + \frac{\sigma_2-\sigma_1}{t-\sigma_2}$ is a positive and monotonically increasing function of $t \in [0, \infty)$. Applying this to each factor in (3.6) we obtain

$$\frac{\rho_2^{(k)}}{\rho_1^{(k)}} \geq \prod_{\nu=1}^{k} \frac{1-\sigma_1/\Theta_\nu^{(k+d)}}{1-\sigma_2/\Theta_\nu^{(k+d)}}.$$

Since $0 < \frac{1-\sigma_1/\Theta_\nu^{(k+d)}}{1-\sigma_2/\Theta_\nu^{(k+d)}} \leq 1$ for $\nu = k+1, \ldots, k+d$, we can multiply the expression to the right with these factors to obtain

$$\frac{\rho_2^{(k)}}{\rho_1^{(k)}} \geq \prod_{\nu=1}^{k+d} \frac{1-\sigma_1/\Theta_\nu^{(k+d)}}{1-\sigma_2/\Theta_\nu^{(k+d)}} = \frac{\rho_2^{(k+d)}}{\rho_1^{(k+d)}},$$

which is the desired inequality. ∎

4. Standard estimates. Given the vector $z = g(A)b$ and its approximation z_k obtained after k steps of an iterative process, a simple stopping criterion consists in monitoring the change of these iterates. Since (see (1.4)) $z_k = \sum_{i=1}^{s} \omega_i V_k y_i^{(k)}$, within the Lanczos method we can cheaply compute

$$\Delta_{k,d} = \|z_k - z_{k+d}\| = \|\sum_{i=1}^{s} \omega_i (\begin{pmatrix} y_i^{(k)} \\ 0_d \end{pmatrix} - y_i^{(k+d)})\|, \tag{4.1}$$

where 0_d is the zero vector of d components. Note that the second equality only holds if $V_{k+d}^H V_{k+d} = I$ which is the case in both our situations: A real symmetric and all shifts complex ($* = \mathtt{T}$ and V_{k+d} is real) as well as A Hermitian and all shifts real ($* = \mathtt{H}$). In the case of the CG algorithm for instance, and for $d = 1$, $\Delta_{k,1}$ may be computed as $\Delta_{k,1} = \|\sum_{i=1}^{s} \omega_i \gamma_i^{(k-1)} p_i^{(k-1)}\|$. A criterion based on $\Delta_{k,d}$ is commonly employed in linear and nonlinear iterative solvers [25]. Its effectiveness strongly depends on the convergence behavior of the iterative process, and in fact $\Delta_{k,d}$ may be small due to temporary stagnation, rather than to a satisfactory convergence of the approximate solution. In our context, convergence is in many cases linear (with a good rate) or even superlinear (cf., e.g., [21] for superlinear convergence results for $\exp(x)$), except possibly for the first phase, where almost complete stagnation may occur. Our numerical experience fully confirms these phenomena, showing that (4.1) is very reliable when convergence enters the second phase (good linear convergence), whereas it may fail completely in the case of initial stagnation; cf. section 7.

Another stopping criterion may be derived by trying to generalize the concept of residual stemming from the linear system setting, by resorting to the partial fraction expansion of g in (1.2). Let $x_i^{(k)}$ be the Galerkin approximation to $(A - \sigma_i I)^{-1} b$ in $\mathcal{K}_k(A, b)$, $r_i^{(k)}$ the associated residual and assume that $(A - \sigma_i I)$ is not highly ill-conditioned for any i. If the quantity $r^{(k)} = \sum_{i=1}^{s} \omega_i r_i^{(k)}$ is small, then one may reasonably argue that $z^{(k)} = \sum_{i=1}^{s} \omega_i x_i^{(k)}$ is a good approximation to z. This provides an argument for using $\|r^{(k)}\|$ as stopping criterion [40]. In particular, since for each i we have $r_i^{(k)} = (-1)^k \rho_i^{(k)} v^{(k)}$ and $\|v^{(k)}\| = 1$, we can define

$$\varrho^{(k)} := \|v^{(k)} \sum_{i=1}^{s} \omega_i (-1)^k \rho_i^{(k)}\| = |\sum_{i=1}^{s} \omega_i \rho_i^{(k)}|, \tag{4.2}$$

which can be cheaply computed at each iteration; cf. [26] for a related discussion. Experience in the context of evaluating the exponential (cf., e.g., [23, 26, 38, 40, 42]) has shown that $\varrho^{(k)}$ may dramatically underestimate the actual error in the early convergence phase, until good information on the spectrum of A is generated in the Krylov subspace. After the stagnation phase terminates, we have observed a reasonable agreement, at least in terms of slope, with respect to the true error; cf. section 7. However, $\varrho^{(k)}$ can still differ from the exact error by some orders of magnitude. The approach just discussed can be refined. Assume that we know bounds ℓ_1, ℓ_2 such that $\mathrm{spec}(A) \subseteq [\ell_1, \ell_2]$. Starting from

$$\begin{aligned} g(A)b - z^{(k)} &= \sum_{i=1}^{s} \omega_i \left((A - \sigma_i I)^{-1} b - x_i^{(k)} \right) = \sum_{i=1}^{s} \omega_i (A - \sigma_i I)^{-1} r_i^{(k)} \\ &= \sum_{i=1}^{s} (-1)^k \rho_i^{(k)} \omega_i (A - \sigma_i I)^{-1} v^{(k)}, \end{aligned}$$

with $\|v^{(k)}\| = 1$, we see that

$$\|g(A)b - z^{(k)}\| \;\le\; \|\sum_{i=1}^{s} \rho_i^{(k)} \omega_i (A - \sigma_i I)^{-1}\| \;\le\; \max_{t\in[\ell_1,\ell_2]} |h^{(k)}(t)|, \tag{4.3}$$

where

$$h^{(k)}(t) = \sum_{i=1}^{s} \frac{\rho_i^{(k)} \omega_i}{t - \sigma_i}.$$

Similarly,

$$\|g(A)b - z^{(k)}\| \ge \min_{t\in[\ell_1,\ell_2]} |h^{(k)}(t)|. \tag{4.4}$$

We can obtain upper and lower bounds for the error if we can bound $h^{(k)}$ in $[\ell_1, \ell_2]$ from above and from below, respectively. In some situations, for example for rational approximations to the sign function or the square root of an Hpd matrix, the function $h^{(k)}$ is monotone, so that the maximum and the minimum can be read off from its values at ℓ_1 or ℓ_2. In other situations, for example for the exponential, the extremal values of $h^{(k)}$ are attained in the interior of the interval $[\ell_1, \ell_2]$, and we can use standard optimization methods to find them. Of course, this process should be comparably cheap. In experiments not reported here, we used S. M. Rump's `intlab` toolbox [39], and we obtained bounds of the extrema by just subdividing the interval $[\ell_1, \ell_2]$ into several subintervals and computing an image interval containing the range of $|h^{(k)}|$ on each of these subintervals using standard interval arithmetic. Bounds for the maximum and the minimum are then obtained from the endpoints of the image intervals.

The previous discussion allows us to better interpret the classical estimate $\varrho^{(k)}$ from (4.2), at least in the case where A is Hpd, $0 \ge \sigma_1 > \sigma_2 \ldots > \sigma_s$ and all ω_i are positive. By Lemma 3.5, we then know that $\rho_i^{(k)} \ge 0$ for all i. Thus we have

$$\frac{\varrho^{(k)}}{\ell_2 - \sigma_s} \le h^{(k)}(t) \le \frac{\varrho^{(k)}}{\ell_1 - \sigma_1}, \qquad t \in [\ell_1, \ell_2].$$

We finish this section by referring to some earlier work on Gaussian quadrature and iterative methods. For A Hpd, in [11] Golub and Meurant developed an elegant theory on moments and Gaussian quadrature with respect to discrete measures allowing to obtain methods for computing lower and upper bounds to quantities of the form $v^H f(A)v$; see also [13]. This theory was then elaborated to be cheaply included in conjugate gradient type algorithms, leading to important practical developments, on which the results of this paper are based; see, e.g., [12], [29], [30], [44].

5. New error estimates for rational approximations. Our goal is to develop estimates for the 2-norm of the error when the approximation $z^{(k)}$ is determined as $z^{(k)} = \sum_{i=1}^{s} \omega_i x_i^{(k)}$ and each $x_i^{(k)}$ is obtained by a Galerkin procedure. For simplicity of notation we assume $\|b\| = 1$ from now on. Then

$$\|g(A)b - z^{(k)}\| = \|\sum_{i=1}^{s} \omega_i (A - \sigma_i I)^{-1} b - \sum_{i=1}^{s} \omega_i V_k (T_k - \sigma_i I)^{-1} e_1\|.$$

Our estimates only rely on the quantities available from the CG processes for each of the s systems $(A-\sigma_i I)x = b$, as described in section 3. In particular, using the estimates from (3.1) and (3.4) we now prove the following result.

THEOREM 5.1. *In the systems $(A-\sigma_i I)x_i = b$, assume that either of the following assumptions holds:*

1. *A is Hpd and σ_i real, $i = 1, \ldots, s$,*
2. *A is real symmetric and b is real.*

For each i, let $x_i^{(k)}$ be the Galerkin approximation to x_i with respect to the kth Krylov subspace (obtained via CG or Lanczos) which is assumed to exist. This is the case, e.g., if $\sigma_i \leq 0$ in case 1 and if $\Im(\sigma_i) \neq 0$ in case 2. Let $r_i^{(k)} = (-1)^k \rho_i^{(k)} v^{(k)}$ be the associated residual. With the definitions in Lemma 3.3 and in (3.3), it holds

$$\|g(A)b - \sum_{i=1}^{s} \omega_i x_i^{(k)}\|_2^2 \approx \boldsymbol{\eta}^{(k,d)} + \boldsymbol{\tau}^{(k,d)} \tag{5.1}$$

where

$$\boldsymbol{\eta}^{(k,d)} = \sum_{i,j=1,\bar{\sigma}_i \neq \sigma_j}^{s} \frac{\overline{\omega}_i \omega_j}{\overline{\sigma}_i - \sigma_j} \left(\frac{\rho_j^{(k)}}{\overline{\rho}_i^{(k)}} \overline{\eta}_i^{(k,d)} - \frac{\overline{\rho}_i^{(k)}}{\rho_j^{(k)}} \eta_j^{(k,d)} \right),$$

$$\boldsymbol{\tau}^{(k,d)} = \sum_{i,j=1,\overline{\sigma}_i = \sigma_j}^{s} \bar{\omega}_i \omega_j \tau_j^{(k,d)}.$$

Proof. We have

$$\begin{aligned} \|g(A)b - \sum_{i=1}^{s} \omega_i x_i^{(k)}\|_2^2 &= \sum_{i,j=1}^{s} \overline{\omega}_i \omega_j \left((A-\sigma_i I)^{-1} r_i^{(k)} \right)^{\mathtt{H}} (A-\sigma_j I)^{-1} r_j^{(k)} \\ &= \sum_{i,j=1}^{s} \overline{\omega}_i \omega_j (e_i^{(k)})^{\mathtt{H}} e_j^{(k)}. \end{aligned} \tag{5.2}$$

We discuss each summand in (5.2) depending on whether $\overline{\sigma}_i = \sigma_j$ or not. First, let $\overline{\sigma}_i \neq \sigma_j$. Using

$$\frac{1}{(t-\overline{\sigma}_i)(t-\sigma_j)} = \frac{1}{\overline{\sigma}_i - \sigma_j} \cdot \left(\frac{1}{t-\overline{\sigma}_i} - \frac{1}{t-\sigma_j} \right), \tag{5.3}$$

we see that in this case

$$\begin{aligned} &\left((A-\sigma_i I)^{-1} r_i^{(k)} \right)^{\mathtt{H}} (A-\sigma_j I)^{-1} r_j^{(k)} \\ &= \frac{1}{\overline{\sigma}_i - \sigma_j} \left((r_i^{(k)})^{\mathtt{H}} (A-\overline{\sigma}_i I)^{-1} r_j^{(k)} - (r_i^{(k)})^{\mathtt{H}} (A-\sigma_j I)^{-1} r_j^{(k)} \right) \\ &= \frac{1}{\overline{\sigma}_i - \sigma_j} \left((e_i^{(k)})^{\mathtt{H}} r_j^{(k)} - (r_i^{(k)})^{\mathtt{H}} e_j^{(k)} \right). \end{aligned}$$

Recall that $r_i^{(k)} = (-1)^k \rho_i^{(k)} v^{(k)}$. If A and b are real and A is symmetric, then $v^{(k)}$ is a real vector and

$$\begin{aligned} (e_i^{(k)})^{\mathtt{H}} r_j^{(k)} &= (-1)^k \rho_j^{(k)} (e_i^{(k)})^{\mathtt{H}} v^{(k)} \;=\; (-1)^k \rho_j^{(k)} \overline{(v^{(k)})^T e_i^{(k)}} \\ &= \frac{\rho_j^{(k)}}{\overline{\rho}_i^{(k)}} \overline{(r_i^{(k)})^T e_i^{(k)}} \;=\; \frac{\rho_j^{(k)}}{\overline{\rho}_i^{(k)}} \overline{\langle r_i^{(k)}, e_i^{(k)} \rangle_{\mathtt{T}}}. \end{aligned}$$

Analogously, $(r_i^{(k)})^{\mathtt{H}} e_j^{(k)} = \frac{\overline{\rho}_i^{(k)}}{\rho_j^{(k)}} \langle r_j^{(k)}, e_j^{(k)} \rangle_{\mathtt{T}}$.

If A is Hermitian and all poles σ_i are real, then $\rho_i^{(k)}$ is real, and also

$$(e_i^{(k)})^{\mathtt{H}} r_j^{(k)} = \frac{\rho_j^{(k)}}{\rho_i^{(k)}} (e_i^{(k)})^{\mathtt{H}} r_i^{(k)} = \frac{\rho_j^{(k)}}{\rho_i^{(k)}} \overline{\langle r_i^{(k)}, e_i^{(k)} \rangle_{\mathtt{H}}},$$

as well as, analogously, $(r_i^{(k)})^{\mathtt{H}} e_j^{(k)} = \frac{\rho_i^{(k)}}{\rho_j^{(k)}} \langle r_j^{(k)}, e_j^{(k)} \rangle_{\mathtt{H}}$. With our general notation $\langle \cdot, \cdot \rangle_*$, we can subsume both cases in writing

$$\left((A - \sigma_i I)^{-1} r_i^{(k)} \right)^{\mathtt{H}} (A - \sigma_j I)^{-1} r_j^{(k)} = \frac{1}{\overline{\sigma}_i - \sigma_j} \left(\frac{\rho_j^{(k)}}{\overline{\rho}_i^{(k)}} \overline{\langle r_i^{(k)}, e_i^{(k)} \rangle_*} - \frac{\overline{\rho}_i^{(k)}}{\rho_j^{(k)}} \langle r_j^{(k)}, e_j^{(k)} \rangle_* \right).$$

Now, let $\overline{\sigma}_i = \sigma_j$. If A is real symmetric and b is real, then $x_i^{(k)} = \overline{x}_j^{(k)}$ and $r_i^{(k)} = \overline{r}_j^{(k)}, e_i^{(k)} = \overline{e}_j^{(k)}$, see Remark 2.3. Therefore,

$$(e_i^{(k)})^{\mathtt{H}} e_j^{(k)} = (e_j^{(k)})^{\mathtt{T}} e_j^{(k)} = \langle e_j^{(k)}, e_j^{(k)} \rangle_{\mathtt{T}},$$

while for A Hermitian and $\sigma_i = \sigma_j$ real, we have $(e_i^{(k)})^{\mathtt{H}} e_j^{(k)} = \langle e_j^{(k)}, e_j^{(k)} \rangle_{\mathtt{H}}$. Therefore, (5.2) can be written as

$$\begin{aligned} \|g(A)b - \sum_{i=1}^{s} \omega_i x_i^{(k)}\|_2^2 = & \sum_{i,j=1, \overline{\sigma}_i \neq \sigma_j}^{s} \frac{\overline{\omega}_i \omega_j}{\overline{\sigma}_i - \sigma_j} \left(\frac{\rho_j^{(k)}}{\overline{\rho}_i^{(k)}} \overline{\langle r_i^{(k)}, e_i^{(k)} \rangle_*} - \frac{\overline{\rho}_i^{(k)}}{\rho_j^{(k)}} \langle r_j^{(k)}, e_j^{(k)} \rangle_* \right) \\ & + \sum_{i,j=1, \bar{\sigma}_i = \sigma_j}^{s} \overline{\omega}_i \omega_j \langle e_j^{(k)}, e_j^{(k)} \rangle_*. \end{aligned} \tag{5.4}$$

By using the estimates from (3.1) and (3.4), the result follows. □

Except for the $\rho_i^{(k)}$'s – which can be updated very cheaply; see section 6 – all quantities used in estimate (5.1) are directly available from the respective CG processes. More precisely, the quantities $\eta_i^{(k,d)}$ and $\tau_j^{(k,d)}$ needed in $\boldsymbol{\eta}^{(k,d)}$ and $\boldsymbol{\tau}^{(k,d)}$ are built up from quantities available in iterations $k, \ldots, k+d$ and $k, \ldots, k+2d$, resp. Therefore, after $k+2d$ iterations of the CG or Lanczos methods, it is possible to estimate the 2-norm of the error at iteration k. At convergence, the overall estimation procedure will have required only $2d$ additional iterations to get often very accurate estimates throughout the convergence history. The parameter d can be fixed a-priori or adjusted dynamically; in many cases a small constant value, $d = 2, \ldots, 5$, is satisfactory.

REMARK 5.2. Since $\boldsymbol{\tau}^{(k,d)}$ requires the quantities $\eta_i^{(k+j)}$ for $j = 0, \ldots, 2d$, we can even use, at the same computational cost, the improved estimate

$$\|g(A)b - \sum_{i=1}^{s} \omega_i x_i^{(k)}\|_2^2 \approx \boldsymbol{\eta}^{(k,2d)} + \boldsymbol{\tau}^{(k,d)}. \tag{5.5}$$

REMARK 5.3. For any $d > 0$ the quantities $\boldsymbol{\eta}^{(k,d)}$ and $\boldsymbol{\tau}^{(k,d)}$ are real in the following two cases of interest:

1. A is (complex) Hermitian and all shifts are real;

2. A is real symmetric, b is real, the non-real shifts σ_i come in conjugate pairs and $\overline{\omega}_i = \omega_j$ whenever $\overline{\sigma}_i = \sigma_j$.

Indeed, in the case 1), every single summand in $\boldsymbol{\eta}^{(k,d)}$ and $\boldsymbol{\tau}^{(k,d)}$ is real. In the case 2), each summand in $\boldsymbol{\eta}^{(k,d)}$ and $\boldsymbol{\tau}^{(k,d)}$ corresponding to an index pair (i,j) has a complex conjugate summand corresponding to the index pair (j,i).

We now proceed to further investigate the error estimate (5.1) in the case of A Hpd and negative poles σ_i. This is so, for instance, for rational approximations to the sign function as well as to the inverse square root. We show that the estimates we derived are then *lower* bounds for the true error.

THEOREM 5.4. *Let A be Hermitian and positive definite and $b \in \mathbb{C}^n$. Let $g(t) = \sum_{i=1}^s \omega_i (t-\sigma_i)^{-1}$. Assume that $\omega_i \in \mathbb{R}$, $\omega_i > 0$, $i = 1, \ldots, s$ and that the poles σ_i are real and satisfy $\sigma_i < \sigma_j < 0$ for $i < j$. Let $\boldsymbol{\eta}^{(k,d)}$, $\boldsymbol{\tau}^{(k,d)}$ be defined as in Theorem 5.1. Then, for any $d \geq 0$,*

a) $\|g(A)b - \sum_{i=1}^s \omega_i x_i^{(k)}\|^2 \geq \boldsymbol{\eta}^{(k,d)} + \boldsymbol{\tau}^{(k,d)}$;

b) $\|g(A)b - \sum_{i=1}^s \omega_i x_i^{(k)}\|^2 \geq \boldsymbol{\eta}^{(k,2d)} + \boldsymbol{\tau}^{(k,d)}$.

Proof. Note first that the condition $\overline{\sigma}_i \neq \sigma_j$ is now equivalent to $i \neq j$. Since $\sigma_j > \sigma_i$ for $i < j$, Lemma 3.5 yields

$$\rho_i^{(k)}/\rho_j^{(k)} \geq \rho_i^{(k+d)}/\rho_j^{(k+d)}, \tag{5.6}$$

where all quantities are real. In the following, we use the symbol $\widehat{\ }$ to denote quantities related to iteration $k+d$. Define $\widehat{w}_i = \widehat{v}^{\mathtt{H}}(A-\sigma_i I)^{-1}\widehat{v} \in \mathbb{R}$ with $\widehat{v}$ the (complex) Lanczos vector at stage $k+d$. We have $\widehat{w}_i > 0$ for all i and $(r_i^{(k+d)})^{\mathtt{H}} e_i^{(k+d)} = \widehat{\rho}_i^2 \widehat{w}_i$. Therefore, using (3.1), the part that was neglected when passing from the first sum in (5.4) to its estimate $\boldsymbol{\eta}^{(k,d)}$ can be bounded as follows:

$$\begin{aligned}
&\sum_{i,j=1, i\neq j}^{s} \frac{\omega_i\omega_j}{\sigma_i-\sigma_j}\left(\frac{\rho_j^{(k)}}{\rho_i^{(k)}}(r_i^{(k+d)})^{\mathtt{H}} e_i^{(k+d)} - \frac{\rho_i^{(k)}}{\rho_j^{(k)}}(r_j^{(k+d)})^{\mathtt{H}} e_j^{(k+d)}\right) \\
&\qquad = 2\sum_{i,j=1, i<j}^{s} \frac{\omega_i\omega_j}{\sigma_i-\sigma_j}\left(\frac{\rho_j^{(k)}}{\rho_i^{(k)}}\widehat{\rho}_i^2\widehat{w}_i - \frac{\rho_i^{(k)}}{\rho_j^{(k)}}\widehat{\rho}_j^2\widehat{w}_j\right) \\
&\qquad \geq 2\sum_{i,j=1, i<j}^{s} \frac{\omega_i\omega_j}{\sigma_i-\sigma_j}\left(\widehat{\rho}_j\widehat{\rho}_i\widehat{w}_i - \widehat{\rho}_j\widehat{\rho}_i\widehat{w}_j\right).
\end{aligned}$$

In the last line we have used that $\sigma_i - \sigma_j < 0$ for $i < j$ and that, according to (5.6),

$$\frac{\rho_j^{(k)}}{\rho_i^{(k)}}\widehat{\rho}_i^2 \leq \widehat{\rho}_i\widehat{\rho}_j, \qquad \text{and} \quad \frac{\rho_i^{(k)}}{\rho_j^{(k)}}\widehat{\rho}_j^2 \geq \widehat{\rho}_j\widehat{\rho}_i.$$

Now, $\widehat{\rho}_j\widehat{\rho}_i\widehat{w}_i = (r_j^{(k+d)})^H (A-\sigma_i I)^{-1} r_i^{(k+d)}$, which together with (5.3) shows that

$$\frac{1}{\sigma_i-\sigma_j}\left(\widehat{\rho}_j\widehat{\rho}_i\widehat{w}_i - \widehat{\rho}_j\widehat{\rho}_i\widehat{w}_j\right) = (e_j^{(k+d)})^{\mathtt{H}} e_i^{(k+d)}.$$

This gives us

$$\sum_{i,j=1,i\neq j}^{s} \frac{\omega_i\omega_j}{\sigma_i-\sigma_j}\left(\frac{\rho_j^{(k)}}{\rho_i^{(k)}}(r_i^{(k+d)})^{\mathtt{H}}e_i^{(k+d)} - \frac{\rho_i^{(k)}}{\rho_j^{(k)}}(r_j^{(k+d)})^{\mathtt{H}}e_j^{(k+d)}\right)$$
$$\geq 2\sum_{i,j=1,i<j}^{s} \omega_i\omega_j(e_i^{(k+d)})^{\mathtt{H}}e_j^{(k+d)}$$
$$= \sum_{i,j=1,i\neq j}^{s} \omega_i\omega_j(e_i^{(k+d)})^{\mathtt{H}}e_j^{(k+d)}. \tag{5.7}$$

According to Lemma 3.4, the part that was neglected when passing from the second sum in (5.4) to its estimate $\boldsymbol{\tau}^{(k,d)}$ is not smaller than $\sum_{j=1}^{s}\omega_j^2(e_j^{(k+d)})^{\mathtt{H}}e_j^{(k+d)}$. Putting things together, we obtain

$$\begin{aligned}
&\|g(A)b-\sum_{i=1}^{s}\omega_i x_i^{(k)}\|_2^2\\
&\geq \boldsymbol{\eta}^{(k,d)}+\boldsymbol{\tau}^{(k,d)}+2\sum_{i,j=1,i<j}^{s}\omega_i\omega_j(e_i^{(k+d)})^{\mathtt{H}}e_j^{(k+d)}+\sum_{j=1}^{s}\omega_j^2(e_j^{(k+d)})^{\mathtt{H}}e_j^{(k+d)}\\
&= \boldsymbol{\eta}^{(k,d)}+\boldsymbol{\tau}^{(k,d)}+\|g(A)b-\sum_{i=1}^{s}\omega_i x_i^{(k+d)}\|_2^2.
\end{aligned}$$

This proves part a). For b), we obtain in a similar manner as before

$$\begin{aligned}
&\|g(A)b-\sum_{i=1}^{s}\omega_i x_i^{(k)}\|_2^2\\
&\geq \boldsymbol{\eta}^{(k,2d)}+\boldsymbol{\tau}^{(k,d)}+\sum_{i,j=1,i\neq j}^{s}\frac{\omega_i\omega_j\rho_j^{(k)}}{(\sigma_i-\sigma_j)\rho_i^{(k)}}(r_i^{(k+2d)})^{\mathtt{H}}e_i^{(k+2d)}\\
&\quad-\sum_{i,j=1,i\neq j}^{s}\frac{\omega_i\omega_j\rho_i^{(k)}}{(\sigma_i-\sigma_j)\rho_j^{(k)}}(r_j^{(k+2d)})^{\mathtt{H}}e_j^{(k+2d)}+\sum_{j=1}^{s}\omega_j^2(e_j^{(k+d)})^{\mathtt{H}}e_j^{(k+d)}\\
&= \boldsymbol{\eta}^{(k,2d)}+\boldsymbol{\tau}^{(k,d)}+2\cdot\left(\sum_{i,j=1,i<j}^{s}\frac{\omega_i\omega_j\rho_j^{(k)}}{(\sigma_i-\sigma_j)\rho_i^{(k)}}(r_i^{(k+2d)})^{\mathtt{H}}e_i^{(k+2d)}\right.\\
&\quad\left.-\frac{\omega_i\omega_j\rho_i^{(k)}}{(\sigma_i-\sigma_j)\rho_j^{(k)}}(r_j^{(k+2d)})^{\mathtt{H}}e_j^{(k+2d)}\right)+\sum_{j=1}^{s}\omega_j^2(e_j^{(k+d)})^{\mathtt{H}}e_j^{(k+d)}.
\end{aligned}\tag{5.8}$$

The same steps that lead to (5.7) yield

$$\sum_{i,j=1,i<j}^{s}\left(\frac{\omega_i\omega_j\rho_j^{(k)}}{(\sigma_i-\sigma_j)\rho_i^{(k)}}(r_i^{(k+2d)})^{\mathtt{H}}e_i^{(k+2d)}-\frac{\omega_i\omega_j\rho_i^{(k)}}{(\sigma_i-\sigma_j)\rho_j^{(k)}}(r_j^{(k+2d)})^{\mathtt{H}}e_j^{(k+2d)}\right)$$
$$\geq \sum_{i,j=1,i<j}^{s}\omega_i\omega_j(e_i^{(k+2d)})^{\mathtt{H}}e_j^{(k+2d)}$$
$$= \sum_{i,j=1,i<j}^{s}\omega_i\omega_j\widehat{\rho}_i\widehat{\rho}_j\widehat{v}^{\mathtt{H}}(A-\sigma_i I)^{-1}(A-\sigma_j I)^{-1}\widehat{v}_j\geq 0,$$

where the last inequality follows from the positive definiteness of the matrix $(A-\sigma_i I)^{-1}(A-\sigma_j I)^{-1}$ and the fact that $\widehat{\rho}_i\widehat{\rho}_j$ is positive.
Noticing that $\sum_{j=1}^{s}\omega_j^2(e_j^{(k+d)})^{\mathtt{H}}e_j^{(k+d)}\geq 0$, the result in b) follows. □

Note that the proof of this theorem also shows that we have a monotone behavior, $\|g(A)b-\sum_{i=1}^{p}\omega_i x_i^{(k)}\|\geq\|g(A)b-\sum_{i=1}^{p}\omega_i x_i^{(k+1)}\|$, as soon as $\boldsymbol{\eta}^{(k,1)}+\boldsymbol{\tau}^{(k,1)}\geq 0$. Although it is easy to see that the individual terms $\eta_i^{(k,d)}$ and $\tau_j^{(k,d)}$ are nonnegative (see [44]), we did not succeed in proving that the latter condition always holds. Numerical evidence strongly suggests that this is true.

6. Implementation aspects. The bounds obtained in Theorem 5.1 and Theorem 5.4 allow us to derive a computable and rather cheap stopping criterion. For a fixed small value of d, after k iterations of the iterative process, with $k>2d$, it is possible to derive an error estimate at iteration $k-2d$. As an example, for $d=3$, after 20 iterations it is possible to give an estimate of the error norm at iteration 14.

Starting with $k=1$, a sketch of the procedure is as follows

1. Expand Krylov subspace from $k-1$ to k
2. Update projected approximate solution $y(\sigma)$ for all σ's
3. If $k>2d$, compute estimate of error at step $k-2d$
4. If not converged, set $k=k+1$ and goto 1.

We suggest to stop the iteration as soon as the following criterion is satisfied

$$|\boldsymbol{\eta}^{(k,2d)}+\boldsymbol{\tau}^{(k,d)}|^{\frac{1}{2}}\leq \mathtt{tol},$$

where `tol` is a user-selected tolerance. We next discuss some additional details associated with the actual implementation of the overall approach.

The computation of $\boldsymbol{\eta}^{(k,2d)},\boldsymbol{\tau}^{(k,d)}$ requires knowledge of the CG coefficients $\gamma_i^{(k)}$ and $\delta_i^{(k)}$. When using the Lanczos recurrence (2.2), the CG coefficients are related to the entries of the tridiagonal matrix $T_k=(t_{ij})$ as follows (cf., e.g., [41]):

$$t_{k+1,k+1}-\sigma_i=\frac{1}{\gamma_i^{(k)}}+\frac{\delta_i^{(k-1)}}{\gamma_i^{(k-1)}},\qquad t_{k,k+1}=\frac{\sqrt{\delta_i^{(k-1)}}}{\gamma_i^{(k-1)}},\tag{6.1}$$

while $t_{k+1,k}=t_{k,k+1}$ and $t_{k+1,k}=\overline{t}_{k,k+1}$ for $*=$ "$\mathtt{T}$" and $*=$"$\mathtt{H}$", respectively. Moreover, $\boldsymbol{\tau}^{(k,d)}$ requires the computation of $\tau_i^{k,d}$, which for $i=1,\ldots,s$, contains the factor

$$\hat{\pi}_{i,k}=\frac{\langle p_i^{(k)},p_i^{(k)}\rangle_*}{\langle p_i^{(k)},(A-\sigma_i I)p_i^{(k)}\rangle_*}.$$

Once again, this quantity is not directly available in case the Lanczos process is employed. However, a short-term recurrence can be used for its update. We first observe that

$$\hat{\pi}_{i,k}=\frac{\langle p_i^{(k)},p_i^{(k)}\rangle_*}{\langle p_i^{(k)},(A-\sigma_i I)p_i^{(k)}\rangle_*}=\gamma_i^{(k)}\frac{\langle p_i^{(k)},p_i^{(k)}\rangle_*}{\langle r_i^{(k)},r_i^{(k)}\rangle_*}.$$

Using the CG recurrence and the orthogonality between $p_i^{(k-1)}$ and $r_i^{(k)}$, we have

$$\langle p_i^{(k)},p_i^{(k)}\rangle_*=\langle r_i^{(k)},r_i^{(k)}\rangle_*+(\delta_i^{(k)})^2\langle p_i^{(k-1)},p_i^{(k-1)}\rangle_*.$$

Therefore, starting with $\delta_i^{(0)} = 0, \pi_{i,0} = \langle p_i^{(0)}, p_i^{(0)} \rangle_* = \langle r_i^{(0)}, r_i^{(0)} \rangle_*$, $i = 1, \ldots, s$, we have

$$\pi_{i,k} = \langle r_i^{(k)}, r_i^{(k)} \rangle_* + (\delta_i^{(k)})^2 \pi_{i,k-1}, \;\; \hat{\pi}_{i,k} = \gamma_i^{(k)} \frac{\pi_{i,k}}{\langle r_i^{(k)}, r_i^{(k)} \rangle_*}, \;\; k = 1, 2, \ldots . \quad (6.2)$$

In the following we report a possible implementation[1] of the whole procedure with the new stopping criterion. The Lanczos recurrence is employed as underlying Krylov space method. We assume that A and b are given as well as the coefficients ω_i and the poles σ_i of the rational function $g(t) = \sum_{i=1}^{s} \omega_i \frac{1}{t - \sigma_i}$. Without loss of generality, we scale b so that $\|b\| = 1$.

Algorithm PFE-Lanczos.

Choose d ,tol, maxit {for error estimate and stopping}
$\mathcal{I} = \{1, 2, \ldots, s\}$ {non-converged systems}
for $i, j = 1, , \ldots, s$ **do** {various initializations}
if $\overline{\sigma}_i \neq \sigma_j$ **then** $\xi_{i,j} = \overline{\omega}_i \omega_j / (\overline{\sigma}_i - \sigma_j)$ **else** $\xi_{i,j} = 0$, **end**
end for
$\beta = 0, v_0 = b, v_{-1} = 0, \pi_{i,0} = 1, \gamma_i^{(-1)} = 1, \delta_i^{(0)} = 0, \rho_i^{(0)} = 1, \pi_i^{(0)} = 1 \; (i \in \mathcal{I})$
for $k = 1, \ldots$,maxit **do** {iteration}
$q = A v_{k-1} - \beta v_{k-2}, \;\; \alpha = v_{k-1}^* q, \;\; t_{k,k} = \alpha$ {Lanczos coefficients and vectors}
$\widetilde{v} = q - \alpha v_{k-1}$
$\beta = (\widetilde{v}^* \widetilde{v})^{1/2}, \;\; v_k = \widetilde{v}/\beta, \;\; t_{k+1,k} = \beta, \;\; t_{k,k+1} = \beta, \;\; v_k = \widetilde{v}/\beta$
$\gamma_i^{(k-1)} = 1/(\alpha - \sigma_i - \delta_i^{(k-1)}/\gamma_i^{(k-2)}), \;\; i \in \mathcal{I}$ {CG coefficients γ_i using (6.1)}
$\pi_i^{(k)} = (\rho_i^{(k-1)})^2 + (\delta_i^{(k-1)})^2 \pi_i^{(k-1)}, \;\; i \in \mathcal{I}$ {update factors in (6.2)}
$\hat{\pi}_i^{(k)} = (\pi_i^{(k)}/\rho_i^{(k-1)})(\gamma_i^{(k-1)}/\rho_i^{(k-1)}), \quad i \in \mathcal{I}$
$\delta_i^{(k)} = \beta^2 (\gamma_i^{(k-1)})^2, \; i \in \mathcal{I}$ {CG coefficients δ_i using (6.1)}
if $k - 1 \geq 2d$ **then** {compute error estimate}

$$\ell_i = \sum_{j=k-d}^{k} \hat{\pi}_i^{(j)} \left(\gamma_i^{(j-1)} (\rho_i^{(j-1)})^2 + 2 \sum_{m=j-d}^{k} \gamma_i^{(m-1)} (\rho_i^{(m-1)})^2 \right), \quad i \in \mathcal{I}$$

$$\boldsymbol{\tau}^{k,d} = \sum_{i \in \mathcal{I}} \omega_i^2 \ell_i$$

$$\hat{\ell}_i = \sum_{m=k-d}^{k} \gamma_i^{(m-1)} (\rho_j^{(m-1)})^2, \quad i \in \mathcal{I}$$

$$\boldsymbol{\eta}^{k,2d} = \sum_{j \in \mathcal{I}} \sum_{m \in \mathcal{I}} \left(\frac{\overline{\hat{\ell}_j}}{\overline{\rho}_j^{(k-2d)}} \xi_{j,m} \rho_m^{(k-2d)} - \frac{\hat{\ell}_m}{\rho_m^{(k-2d)}} \xi_{j,m} \overline{\rho}_j^{(k-2d)} \right)$$

est $= |\boldsymbol{\tau}^{k,d} + \boldsymbol{\eta}^{k,2d}|^{\frac{1}{2}}$
end if
$y_i = (T_k - \sigma_i I_k)^{-1} e_1, \;\; i \in \mathcal{I}$ {get projected solutions}
$\rho_i^{(k)} = e_k^{\mathsf{T}} y_i t_{k+1,k}, \quad i \in \mathcal{I}$ {factors for residuals}
if est $<$ tol **then** {iteration converged}

$$w_k = \sum_{i=1}^{s} \omega_i y_i, \;\; z_k = \sum_{i=0}^{k-1} (w_k)_{i+1} v_i, \; \textbf{stop}$$ {z_k approx. solution as in (1.4)}

else

[1] For the sake of readability we do not address the possible optimizations in the presence of complex conjugate shifts, see Remark 2.3.

remove i with $|\rho_i^{(k)}| \approx \epsilon_{mach}$ from $\mathcal{I}$ {remove converged systems}
end if
end for

Note that in this implementation, the whole set of Lanczos basis vectors is only needed when retrieving the approximate solution upon convergence. So, during the iteration phase, older vectors may be stored in secondary memory and accessed again only at convergence. An alternative is a two-sweep strategy, where the basis vectors are recomputed a second time. On the other hand, the number of basis vectors involved in the approximation procedure does not depend on the number of shifted systems to be solved. Therefore, the storage requirements do not increase significantly with the number of shifted systems, and the additional cost for getting a solution for the s shifted tridiagonal systems $(T_k - \sigma_i I_k)y_i = e_1$ is just $\mathcal{O}(sk)$ in iteration k. The computational cost for computing the various coefficients is $\mathcal{O}(s^2)$ per iteration. This shows that the dominant cost per iteration are caused by the computation of the next Lanczos basis vector where we have to invest one matrix-vector multiplication and $\mathcal{O}(n)$ additional work for the vector updates.

An analogous implementation relying on the CG method for multiple shifted systems can also be derived, see [47]. There, again, the Krylov subspace is expanded only once for all shifted systems at a time. In exact arithmetic, at each CG iteration the approximate solutions and the associated residuals are the same as those obtained in the Lanczos procedure, see (2.3). The algorithmic difference is that they are updated at each iteration, so that the whole subspace basis needs not be stored. As opposed to the Lanczos procedure, however, the direction vectors $p^{(k)}$ depend on the shift σ in $M = A - \sigma I$. As was shown in [47], it is possible to retrieve the CG coefficients $\gamma^{(k)}$ and $\delta^{(k)}$ for a shifted system from those of the non-shifted system at little constant cost. For each additional shift, one direction vector and one solution iterate need to be stored so that the memory requirements increase linearly with the number of shifts. This is in contrast to the Lanczos based approach, where memory requirements increase linearly with the number of iterations performed. The computational cost of CG for each additional shifted system is $\mathcal{O}(n)$ per iteration for the updates of the iterates and of the direction vectors, plus $\mathcal{O}(1)$ for retrieving the shifted CG coefficients. Summarizing, we see that the computational cost of the Lanczos and the CG methods for solving several shifted linear systems are slightly in favor of the former, which does not require s direction vector updates. However, if memory requirements become crucial, the multiple CG approach is preferable.

7. Numerical experiments. In this section we report on our numerical experience with the discussed stopping criteria. In particular, we compare our new estimate (5.5) for a very low value of d, usually $d \leq 5$, with the estimates using $\Delta_{k,d}$ from (4.1), for the same value of d, and the classical estimate $\varrho^{(k)}$ from (4.2). For completeness, in one case with A Hpd we also give the bounds obtained via the operator norm bounds in (4.3) and (4.4).

We consider four different rational functions, each approximating a function of interest in applications: $\text{sign}(x)$, $x^{-\frac{1}{2}}$, $\exp(x)$ and $\cos(x)$. In our experiments we use both matrices from real application problems, as well as "academic" examples with diagonal matrices. In the latter case, this is not a major restriction since the performance of the methods in exact arithmetic only depends on the spectrum and the initial vector, but not on the sparsity of the matrix.

EXAMPLE 7.1. The Wilson fermion matrix Q results as a discretization of the theory of Quantum Chromodynamics (QCD) which explains the strong interaction

between the quarks as constituents of matter. Recent developments that for the first time respect the physically important 'chiral symmetry' (see, e.g., [3]), produce models in which systems of the form $P + \mathrm{sign}(Q)$, P a simple permutation matrix, have to be solved repeatedly. This is done with a Krylov subspace method, so that in each step one has to compute $\mathrm{sign}(Q)b$ for some vector b. The matrix Q is Hermitian and indefinite. Here we report on numerical results obtained with the matrix D available in the QCD collection of the matrix market [33] as configuration `conf5.4-00l8x8-2000.mtx` with $\kappa_c = 0.15717$. Q is then given as $Q = P(I - \frac{4}{3}\kappa_c D)$, with P the permutation

$$P = I_3 \otimes \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \otimes I_{\frac{n}{12}}.$$

The dimension of the system is $n = 12 \cdot 8^4 \approx 50\,000$. We compute $\mathrm{sign}(Q)b$ for a randomly generated vector b. To this purpose, we first compute two numbers $0 < a_1 < a_2$ such that $\mathrm{spec}(Q) \subset [-a_2, -a_1] \cup [a_1, a_2]$. We then approximate $\mathrm{sign}(t)$ on $[-a_2, -a_1] \cup [a_1, a_2]$ with the Zolotarev rational approximation, see [37]. It has the form $\widehat{g}(t) = \sum_{i=1}^{s} \omega_i \frac{t}{t^2+\alpha_i}$, $\omega_i, \alpha_i > 0$ and it is an ℓ_∞ best approximation. The number of poles s was chosen such that the ℓ_∞-error was less than 10^{-7}, that is $s = 11$. To compute $\widehat{g}(Q)b$, we actually computed $g(Q) \cdot (Qb)$ with

$$g(t) = \sum_{i=1}^{s} \omega_i \frac{1}{t^2 + \alpha_i}. \tag{7.1}$$

Since Q^2 is Hpd, Theorem 5.4 applies so that our estimates will be lower bounds of the true error. We also used a deflation technique which has become common in realistic QCD computations: Since the density of the eigenvalues of Q close to 0 is relatively small, it pays out to compute q eigenvalues of Q which are smallest in modulus, $\lambda_1, \ldots, \lambda_q$, say, beforehand using a Lanczos procedure for Q^2. With Π denoting the orthogonal projector along the space spanned by the corresponding eigenvectors $w_i, i = 1, \ldots, q$, we then work with the matrix $\Pi Q \Pi$ and the vector Πb. 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 to be solved converge more rapidly. The vector $\mathrm{sign}(Q)b$ can be retrieved as $\mathrm{sign}(\Pi Q \Pi)\Pi b + \mathrm{sign}(\mathrm{diag}(\lambda_1, \ldots, \lambda_q)) \cdot (I - \Pi)b$. In QCD practice, this approach results in a major speedup, since $\mathrm{sign}(Q)b$ must usually be computed repeatedly for various vectors b.

The convergence plot in Figure 7.1 shows that the error is monotonically decreasing. Our new estimate (with d=5) is quite close to the true error, and it is a lower bound in accordance with Theorem 5.4. The plot also gives upper and lower bounds as obtained by (4.3) and (4.4), as well as the classical estimate $\varrho^{(k)}$ from (4.2), and the estimate $\Delta_{k,d}$ from (4.1), with $d = 5$. Our new estimate is the most precise of all.

Let us mention that [46] gives an alternative way for obtaining an upper estimate for the error with the Zolotarev approximation. This estimate assumes that we first compute the Galerkin approximation for $g(Q)b$ and then post-multiply by Q, whereas here we use just the opposite order, i.e. we first multiply b by Q.

Example 7.2. We consider the Zolotarev rational function approximation to the inverse square root function, $\widetilde{g}(A)b \approx A^{-1/2}b$, see [37, Chapter 4]. This is directly related to the sign function as $\widetilde{g}(t) = g(t^{1/2})$ with g from (7.1). So, again, our new

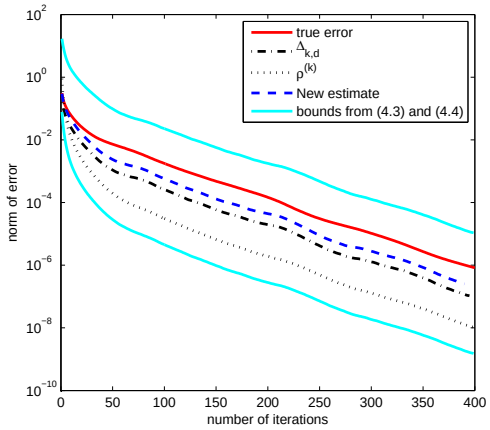

FIG. 7.1. *Results for the Zolotarev rational approximation of sign(Q) for the QCD matrix Q. The 30 smallest eigenvalues are deflated. We plot the convergence history of the Krylov subspace approximation and various error estimates for $s = 11$ poles. Both the new estimate and $\Delta_{k,d}$ use $d = 5$.*

estimates will be lower bounds of the error if A is Hpd. In both experiments below, we employ $d = 2$. We use a rational function with denominator of degree $s = 12$ (and numerator of degree $s - 1 = 11$), and a 200×200 diagonal matrix A with uniformly distributed values in the interval $[1, 1000]$; b is the normalized vector of all ones. With these parameters, the accuracy of the Zolotarev approximation turns out to be of the order of 10^{-7}. The convergence history of our Krylov subspace method together with the error estimates are reported in the left plot of Figure 7.2. The figure shows the good agreement of the new estimate with the true error. The norm $\Delta_{k,d}$ (for $d = 2$) between different approximate solutions is also a good estimate, since complete stagnation of the process is never observed. The estimate $\varrho^{(k)}$ is not sharp, losing at least two orders of magnitude. In the right plot A was taken a diagonal matrix of size 3000×3000 with clustered eigenvalues. There are 300 clusters the centers of which are uniformly distributed in $[1, 10\,000]$. Each cluster contains 10 eigenvalues obtained by perturbing the center with a random relative change uniformly distributed in $[0, 10^{-4}]$. The analysis in [34] explains that we are now to expect stagnation phases as those observed in the plot. The error estimates become less tight – and they are too small – when stagnation occurs, whereas they are quite good when there is progress in the approximation. The new error estimate is again slightly better than $\Delta_{k,d}$ ($d = 2$ in both estimates). As in Example 7.1 we know that the new estimate represents a lower bound by Theorem 5.4.

EXAMPLE 7.3. In this example we consider the Chebyshev rational approximation $g(A)b$ to the exponential function $\exp(-A)b$. The coefficients of the two polynomials of the same degree appearing in g have been tabulated in [5] for several different degrees. It is known that the error associated with this approximation is $\max_{t>0} |\exp(-t) - g(t)| = \mathcal{O}(10^{-s})$, where s is the degree of the polynomials in the rational function. In this case, the poles σ_i and the coefficients ω_i in the partial fraction expansion are complex, and appear in conjugate pairs. Therefore, the code with $* = \mathtt{T}$ is employed. In both experiments below we use $d = 2$. We first consider the 900×900 matrix $A = 0.1\,\widetilde{A}$, with $\widetilde{A}$ stemming from the finite difference discretization of the two-dimensional

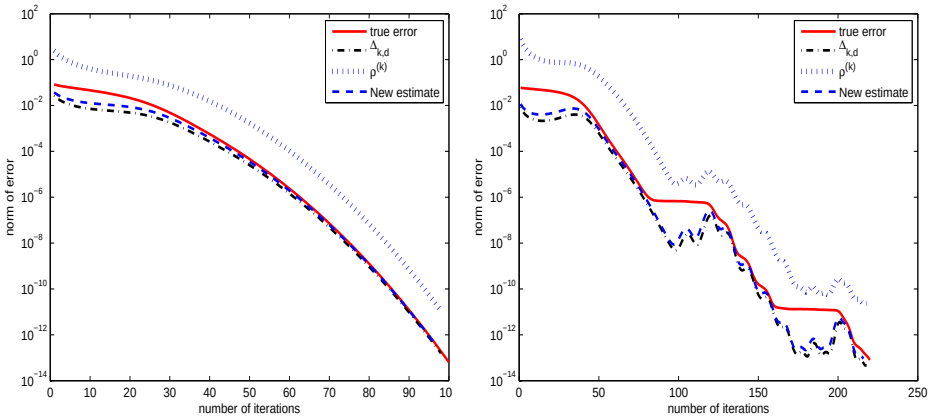

FIG. 7.2. *Results for the (11,12) Zolotarev rational function approximation of* $1/\sqrt{x}$. *Convergence history of the Krylov subspace approximation with various error estimates. Left: A is a* 200×200 *diagonal matrix with uniformly distributed values in* $[1, 1000]$. *Right: A is a* 3000×3000 *diagonal matrix with clustered values in* $[1, 10\,000]$. *In both examples, we used* $d = 2$ *for the new estimate and for* $\Delta_{k,d}$.

Laplace operator in the unit square and homogeneous boundary conditions, and thus A is real symmetric and positive definite; b is taken as the scaled vector of all ones. The convergence in the approximation to the vector $g(A)b$, with g of degree $s = 14$ is reported in the left plot of Figure 7.3, together with the considered error estimates. Both the new bound and the difference $\Delta_{k,d}$, for the same value of d, are able to closely follow the true convergence behavior of the error. The classical bound $\varrho^{(k)}$ stays almost two orders of magnitude above the actual error, although the slope is very similar.

We next consider an example in which the convergence history shows an initial stagnation phase, and the picture changes significantly. To this end, we let A be a diagonal 400×400 matrix with uniformly distributed values in the interval $[0, 2000]$, and b, $g(t)$ as above. The results are reported in the right plot of Figure 7.3. Both the classical estimates $\varrho^{(k)}$ and those relying on $\Delta_{k,d}$ are completely unreliable throughout the stagnation stage, whereas our new estimate provides a useful estimate. After that phase, all curves behave as in the previous example, with the new estimate being the sharpest one. We refer to [40] for an improved, possibly more reliable stopping criterion than that based on $\varrho^{(k)}$, in the case of the exponential function. In this context, we also mention that a stopping criterion based on a variant of $\Delta_{k,d}$ is also used in [23, sec. 4].

EXAMPLE 7.4. Finally, we consider the $(8, 8)$ Padé rational approximation to the cosine function, as described in [20, formula (4.1)]; see also [16]. We consider a diagonal 1000×1000 matrix A with diagonal real elements uniformly distributed in $[0, \pi]$, and b equal to a normalized vector of uniformly distributed random values in [0,1]. The moderate norm of A ensures an accuracy of the Padé approximation of the order of 10^{-8}. The poles and coefficients in the partial fraction expansion of the Padé function arise in complex conjugates, yielding a shifted complex symmetric matrix M.

The results for the Krylov subspace iteration are displayed in Figure 7.4. They confirm the good accuracy of the new estimate as those based on $\Delta_{k,d}$ to the point

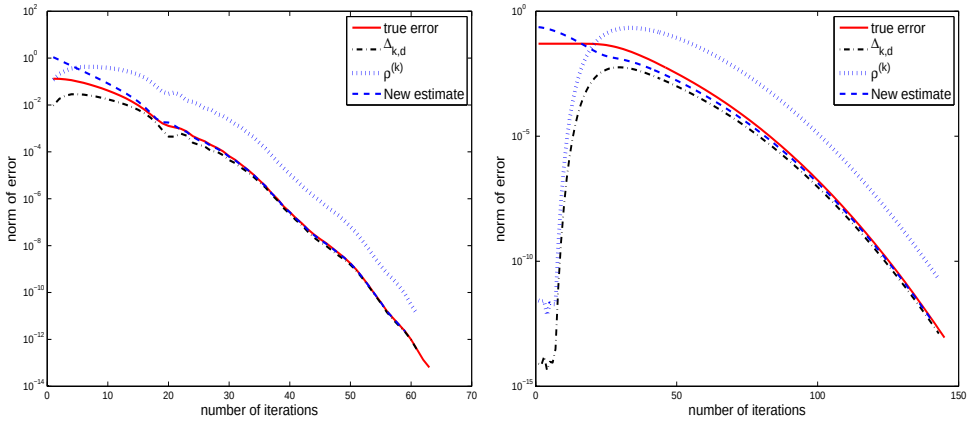

FIG. 7.3. *Results for the Chebyshev rational function approximation of the exponential. Convergence history of the Krylov subspace approximation with various error estimates. Left: A stems from the 2D Laplace operator on a 30×30 grid. Right: A is a diagonal matrix of dimension 400 with uniformly distributed eigenvalues in $[0, 2000]$. In both examples, we used $d = 2$ for the new estimate and for $\Delta_{k,d}$.*

that they fully overlap. We used $d = 2$ for both the new estimate and $\Delta_{k,d}$. The classical estimate $\varrho^{(k)}$ has a dramatic oscillating behavior, which makes the estimate completely unreliable. We also observe that the true error stagnates at the level $\approx 10^{-13}$), which for this problem appears to be the method's final attainable accuracy. This phenomenon deserves a deeper analysis, in view of similar discussions in the case of iterative system solvers for linear systems; see, e.g., [14].

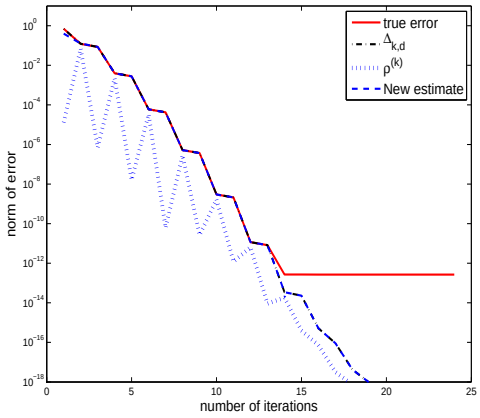

FIG. 7.4. *Results for the (8,8) Padé rational function approximation to the cosine. Convergence history of the Krylov subspace approximation with various error estimates. A is a 1000×1000 diagonal matrix with uniformly distributed values in $[-\pi, \pi]$. We used $d = 2$ for the new estimate and for $\Delta_{k,d}$.*

8. Acceleration procedures. Acceleration procedures have been proposed to enhance the convergence of the approximation to matrix functions; see, e.g., [23], [32], [6], [8], [1]. In [23] and [32] a procedure based on shift-invert Lanczos is proposed to accelerate the Krylov subspace approximation of the action of matrix functions to a

vector, $f(A)b$, so that fewer iterations are required to reach the desired accuracy. For a given real parameter $\mu > 0$, the procedure first constructs the Krylov subspace $\mathcal{K}_k((I-\mu A)^{-1}, b)$, by means of the Arnoldi recurrence $(I-\mu A)^{-1}V_k = V_k T_k + t_{k+1,k} v_{k+1} e_k^{\mathsf{T}}$, then determines an approximation to $f(A)b$ as $V_k f((I - T_k^{-1})/\mu)e_1$. For f a rational function, it was shown in [38, Proposition 3.1] that the shift-invert Lanczos procedure amounts to approximating each system solution $(A-\sigma_i I)^{-1}b$ in $\mathcal{K}_k((I-\mu A)^{-1}, b)$ by imposing the Galerkin condition. To be specific, put $\widehat{A} = (I-\mu A)^{-1}$ and $\widehat{\sigma}_i = \frac{1}{\sigma_i \mu - 1}$. Then

$$\widehat{A} \cdot (A - \sigma_i I) = \frac{1-\mu\sigma_i}{\mu}\left(\widehat{A} - \widehat{\sigma}_i I\right),$$

and the shift-invert Lanczos procedure amounts to solve, for $i = 1, \ldots, p$, the systems

$$\left(\widehat{A} - \widehat{\sigma}_i I\right)\widehat{x} = \widehat{b}, \quad \text{with} \quad \widehat{x} = \frac{1-\mu\sigma_i}{\mu}x,\ \widehat{b} = \widehat{A}b. \tag{8.1}$$

The linear systems in (8.1) have precisely the same shifted structure as those in the previous sections. Let $\widehat{x}_i^{(k)}$ be the Galerkin solution to system i in $\mathcal{K}_k(\widehat{A}, \widehat{b})$; note that the generation of this subspace requires solving a system with $(I-\mu A)$ at each iteration. The acceleration procedure is therefore effective only if one can solve these systems efficiently, e.g. using a multigrid method. Let $x_i^{(k)} = \frac{\mu}{1-\mu\sigma_i}\widehat{x}_i^{(k)}$ be the corresponding approximate solution to the original system $(A-\sigma_i I)x = b$, see (8.1). Then

$$\begin{aligned} g(A)b - \sum_{i=1}^{s}\omega_i x_i^{(k)} &= \sum_{i=1}^{s}\omega_i(x_i - x_i^{(k)}) \\ &= \sum_{i=1}^{s}\frac{\mu\omega_i}{1-\mu\sigma_i}(\widehat{x}_i - \widehat{x}_i^{(k)}) \equiv \sum_{i=1}^{s}\widehat{\omega}_i(\widehat{x}_i - \widehat{x}_i^{(k)}). \end{aligned}$$

The results of Theorems 5.1 and 5.4 can now be used to estimate the error $\sum_{i=1}^{s}\widehat{\omega}_i(\widehat{x}_i - \widehat{x}_i^{(k)})$. We show the behavior of the shift-invert acceleration procedure in the next two examples.

EXAMPLE 8.1. We consider the approximation of the operation $\exp(-tA)b$ where $t = 0.1$, b is the scaled unit vector, and A is the $10\,000 \times 10\,000$ matrix stemming from the five point finite difference discretization of the operator

$$\mathcal{L}(u) = (a(x,y)u_x)_x + (b(x,y)u_y)_y, \qquad a(x,y) = 1+y-x, \quad b(x,y) = 1+x+x^2,$$

in $[0,1]^2$; see [23]. The standard Lanczos approach is extremely inefficient on this problem, whereas the shift-invert acceleration strategy is very competitive [38]. In Figure 8.1 we report the performance of the procedure for $s = 14$, $d = 2$ and the acceleration parameter μ taken as $\mu = 1/\max_i|\sigma_i|$; cf. [38] for a justification of this choice. The results fully confirm the effectiveness of the error estimate even in the acceleration context, and highlight its sharpness whenever convergence is fast. Note that the simple estimate $\Delta_{k,d}$ is equally good, owing to the fast convergence rate, whereas the classical residual-based estimate $\varrho^{(k)}$ is unable to capture the true order of magnitude of the error.

EXAMPLE 8.2. We conclude with the use of the shift-invert Lanczos procedure for accelerating the approximation of $A^{-1/2}b$ in Example 7.2 with A of size 3000×3000. For this example, after some tuning we set the acceleration parameter to be equal to

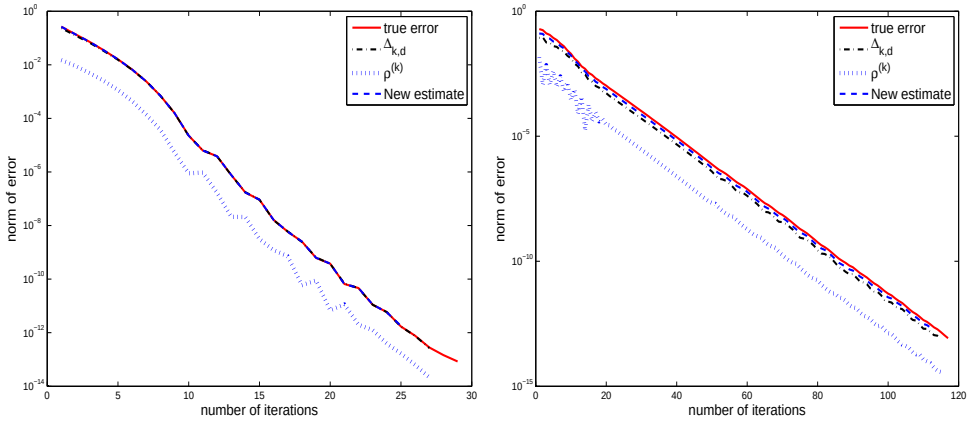

FIG. 8.1. *Convergence of the acceleration procedure. Left: approximation to the Chebyshev rational function approximating the exponential. A is the discretization of an elliptic operator; see Example 8.1. Right: approximation to the (11,12) Zolotarev rational function approximation of the function* $1/\sqrt{x}$*; see Example 8.2. In both cases, we used* $d = 2$ *for the new estimate and for* $\Delta_{k,d}$*.*

$\mu = 1/(100 + \min_i |\sigma_i|)$ and we considered $d = 2$. With the same data as in Example 7.2, the results are displayed in the right plot of Figure 8.1 and once again report an optimal behavior of our new error estimate compared to $\varrho^{(k)}$.

9. Conclusions. In this paper we have shown that a sharp error estimate may be obtained for the approximation by Krylov subspace methods of the action of rational matrix functions to a vector. Our results are sufficiently general to be applicable to a large class of rational functions, commonly employed to approximate not necessarily analytic functions such as the exponential, the sign, the square-root functions and trigonometric functions. Our estimates rely on known error estimates for Hermitian positive definite systems, however we apply them to a wider class of matrices and to the far more general context of rational functions. Under certain hypotheses, we were able to prove that our estimates are true lower bounds of the *Euclidean norm* of the rational function error. We have also discussed practical implementation issues, showing that our estimates can be cheaply included in a Lanczos or CG procedure. We also showed that a classical measure of the error, the difference $\Delta_{k,d}$ between two iterates, may be a good indicator of the actual convergence history, unless complete stagnation takes place.

Acknowledgement. We would like to thank Zdenek Strakoš for his helpful comments on an earlier version of this paper and for suggesting the second matrix in Example 7.2.

REFERENCES

[1] O. AXELSSON AND A. KUCHEROV, *Real valued iterative methods for solving complex symmetric linear systems*, Numer. Linear Algebra Appl., 7 (2000), pp. 197–218.

[2] G. A. BAKER AND P. GRAVES-MORRIS, *Padé Approximants*, Encyclopedia of Mathematics and its Applications, Cambridge University Press, Cambridge, 1996.

[3] A. BORICI, A. FROMMER, B. JOÓ, A. KENNEDY, AND B. PENDLETON, *QCD and Numerical Analysis III. Proceedings of the third international workshop on numerical analysis and lattice QCD, Edinburgh, UK, June 30 – July 4, 2003*, Lecture Notes in Computational Science and Engineering 47. Berlin: Springer, 2005.

[4] D. Calvetti, S. Morigi, L. Reichel, and F. Sgallari, *Computable error bounds and estimates for the conjugate gradient method*, Numer. Algorithms, 25 (2000), pp. 79–88.
[5] A. J. Carpenter, A. Ruttan, and R. S. Varga, *Extended numerical computations on the 1/9 conjecture in rational approximation theory*, in Rational Approximation and Interpolation, P. R. Graves-Morris, E. B. Saff, and R. S. Varga, eds., vol. 1105 of Lecture Notes in Mathematics, Springer-Verlag, Berlin, 1984, pp. 383–411.
[6] P. Castillo and Y. Saad, *Preconditioning the matrix exponential operator with applications*, J. Scientific Computing, 13 (1999), pp. 225–302.
[7] V. Druskin and L. Knizhnerman, *Two polynomial methods of calculating functions of symmetric matrices*, U.S.S.R. Comput. Math. Math. Phys., 29 (1989), pp. 112–121.
[8] ———, *Extended Krylov subspaces: approximation of the matrix square root and related functions*, SIAM J. Matrix Analysis and Appl., 19 (1998), pp. 755–771.
[9] A. Frommer and V. Simoncini, *Matrix functions*, in Model Order Reduction: Theory, Research Aspects and Applications, W. H. A. Schilders and H. A. van der Vorst, eds., Mathematics in Industry, Springer, Heidelberg, To appear 2007.
[10] G. Golub and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins Univ. Press, Baltimore, 3rd ed., 1996.
[11] G. H. Golub and G. A. Meurant, *Matrices, moments and quadrature*, in Numerical Analysis 1993, D. Griffiths and G. Watson, eds., vol. 303 of Pitman research Notes in Mathematics, Longman Sci., 1994, pp. 105–156.
[12] ———, *Matrices, moments and quadrature II; how to compute the norm of the error in iterative methods*, BIT, 37 (1997), pp. 687–705.
[13] G. H. Golub and Z. Strakos, *Estimates in quadratic formulas*, Numerical Algorithms, 8 (1994), pp. 241–268.
[14] A. Greenbaum, *Estimating the attainable accuracy of recursively computed residual methods*, SIAM J. Matrix Analysis and Appl., 18 (1997), pp. 535–551.
[15] E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration. Structure-preserving algorithms for ordinary differential equations*, vol. 31 of Springer Series in Computational Mathematics, Springer, Berlin, 2002.
[16] G. I. Hargreaves and N. J. Higham, *Efficient algorithms for the matrix cosine and sine*, Numerical Algorithms, 40 (2005), pp. 383–400.
[17] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, Journal of Research of the National Bureau of Standards, 49 (1952), pp. 409–436.
[18] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1996.
[19] N. J. Higham, *Factorizing complex symmetric matrices with positive definite real and imaginary parts*, Math. Comput., 67 (1998), pp. 1591–1599.
[20] N. J. Higham and M. I. Smith, *Computing the matrix cosine*, Numerical Algorithms, 34 (2003), pp. 13–26.
[21] M. Hochbruck and C. Lubich, *On Krylov subspace approximations to the matrix exponential operator*, SIAM J. Numer. Anal., 34 (1997), pp. 1911–1925.
[22] M. Hochbruck and A. Ostermann, *Exponential Runge-Kutta methods for parabolic problems*, Applied Numer. Math., 53 (2005), pp. 323–339.
[23] M. Hochbruck and J. van den Eshof, *Preconditioning Lanczos approximations to the matrix exponential*, SIAM J. Sci. Comput., 27 (2006), pp. 1438–1457.
[24] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge: Cambridge University Press, 1994.
[25] C. T. Kelley, *Iterative Methods for Linear and Nonlinear Equations*, SIAM, Philadelphia, 1995.
[26] L. Lopez and V. Simoncini, *Analysis of projection methods for rational function approximation to the matrix exponential*, SIAM J. Numer. Anal., 44 (2006), pp. 613 – 635.
[27] The MathWorks, Inc., *MATLAB 7*, September 2004.
[28] J. A. Meijerink and H. A. van der Vorst, *An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix*, Math. Comp., 31 (January 1977), pp. 148–162.
[29] G. Meurant, *The computation of bounds for the norm of the error in the conjugate gradient algorithm*, Numerical Algorithms, 16 (1997), pp. 77–87.
[30] ———, *Estimates of the l_2 norm of the error in the conjugate gradient algorithm*, Numerical Algorithms, 40 (2005), pp. 157–169.
[31] G. Meurant and Z. Strakoš, *The Lanczos and conjugate gradient algorithms in finite precision arithmetic*, in Acta Numerica, vol. 15, Cambridge University Press, 2006, pp. 471–542.
[32] I. Moret and P. Novati, *RD-rational approximations of the matrix exponential*, BIT, Nu-

merical Mathematics, 44 (2004), pp. 595–615.
[33] National Institute of Standards and Technology, *Matrix Market*.
[34] D. P. O'Leary, Z. Strakoš, and P. Tichý, *On sensitivity of Gauß-Christoffel quadrature*, Numer. Math., 107 (2007), pp. 147–174.
[35] C. C. Paige, B. N. Parlett, and H. A. van der Vorst, *Approximate solutions and eigenvalue bounds from Krylov subspaces*, Numerical Linear Algebra with Applications, 2 (1995), pp. 115–134.
[36] B. N. Parlett, *A new look at the Lanczos algorithm for solving symmetric systems of linear equations*, Lin. Alg. Appl., 29 (1980), pp. 323–346.
[37] P. P. Petrushev and V. A. Popov, *Rational Approximation of Real Functions*, Cambridge University Press, Cambridge, 1987.
[38] M. Popolizio and V. Simoncini, *Refined acceleration techniques for approximating the matrix exponential operator*, tech. rep., Dipartimento di Matematica, Università di Bologna, 2006. Submitted.
[39] S. M. Rump, *INTLAB—INTerval LABoratory*, in Developments in Reliable Computing, T. Csendes, ed., Kluwer Academic Publishers, Dordrecht, NL, 1999, pp. 77–104.
[40] Y. Saad, *Analysis of some Krylov subspace approximations to the matrix exponential operator*, SIAM J. Numer. Anal., 29 (1992), pp. 209–228.
[41] ———, *Iterative Methods for Sparse Linear Systems*, The PWS Publishing Company, Boston, 1996. Second edition, SIAM, Philadelphia, 2003.
[42] R. B. Sidje, *Expokit: A Software Package for Computing Matrix Exponentials*, ACM Transactions on Math. Software, 24 (1998), pp. 130–156.
[43] D. E. Stewart and T. S. Leyk, *Error estimates for Krylov subspace approximations of matrix exponentials*, J. Computational and Applied Mathematics, 72 (1996), pp. 359–369.
[44] Z. Strakoš and P. Tichý, *On error estimation in the conjugate gradient method and why it works in finite precision computations*, Electronic Transactions on Numerical Analysis, 13 (2002), pp. 56–80.
[45] Z. Strakoš and P. Tichý, *Error estimation in preconditioned conjugate gradients*, BIT Numerical Mathematics, 45 (2005), pp. 789–817.
[46] J. van den Eshof, A. Frommer, T. Lippert, K. Schilling, and H. van der Vorst, *Numerical methods for the QCD overlap operator. I: Sign-function and error bounds.*, Comput. Phys. Commun., 146 (2002), pp. 203–224.
[47] J. van den Eshof and G. L. Sleijpen, *Accurate conjugate gradient methods for families of shifted systems*, Appl. Numer. Math., 49 (2004), pp. 17–37.