# A NEW INVESTIGATION OF THE EXTENDED KRYLOV SUBSPACE METHOD FOR MATRIX FUNCTION EVALUATIONS *

L. KNIZHNERMAN† AND V. SIMONCINI‡

**Abstract.** For large square matrices $A$ and functions $f$, the numerical approximation of the action of $f(A)$ to a vector $v$ has received considerable attention in the last two decades. In this paper we investigate the *Extended Krylov subspace method*, a technique that was recently proposed to approximate $f(A)v$ for $A$ symmetric. We provide a new theoretical analysis of the method, which improves the original result for $A$ symmetric, and gives a new estimate for $A$ nonsymmetric. Numerical experiments confirm that the new error estimates correctly capture the linear asymptotic convergence rate of the approximation. By using recent algorithmic improvements, we also show that the method is computationally competitive with respect to other enhancement techniques.

**1. Introduction.** Given a large matrix $A \in \mathbb{R}^{n\times n}$ (or $\mathbb{C}^{n\times n}$ when explicitly stated) and a vector $v \in \mathbb{R}^n$, we are interested in the approximation of

$$u = f(A)v, \tag{1.1}$$

where $f$ is a function that is sufficiently regular so that $f(A)$ is well defined. More precise classes of functions will be introduced later. Without loss of generality we also assume that $\|v\| = 1$. We stress that the problem of approximating the action of $f(A)$ to a vector is significantly different from that of approximating $f(A)$. In particular, for large-scale problems the computation of $f(A)$ is not feasible since, although $A$ may be a sparse matrix, $f(A)$ is dense in general. We refer to the recent monograph of Higham ([27]) for a detailed account of the algorithmic and theoretical progress in the computation of $f(A)$.

The development of numerical procedures for approximating the action of $f(A)$ to a vector has received considerable attention in the last two decades. This is possibly related to the significant increase of methods for the numerical solution of partial differential equations that either directly approximate the exact solution (see, e.g., [20], [47], [24], [23]), or employ matrix functional integrators; see, e.g., [30], [28], [29]. In addition, large-scale advanced scientific applications often require function evaluations of matrices; see, e.g., [5], [43], [55], [17].

For large $A$, a now standard way of approximating (1.1) consists of projecting the original problem onto a subspace of much smaller dimension, which for conveniency reasons is taken to be the Krylov subspace associated with $A$ and $v$ [13], [15], [34], [48]. For particularly challenging problems, however, an unacceptably large approximation space may be required to obtain a sastisfactory approximation. This difficulty has lead to the study of enhancement techniques that aim either at enriching the approximation space or at making the overall procedure less expensive [4], [16], [18], [22], [33], [39], [40], [41], [46], [54].

In this paper we investigate the *Extended Krylov subspace method*, a technique that was proposed in [16]. By using recent algorithmic improvements developed in [50], we show that the method is competitive with respect to other enhancements techniques. Moreover, we provide a new theoretical analysis of the method, which

*Version of 12 January 2009.

†Central Geophysical Expedition, house 38, building 3, Narodnogo opolcheniya St., Moscow, 123298 Russia (mmd@cge.ru)

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

improves the result in [16] for $A$ symmetric, and gives a new estimate for $A$ nonsymmetric. We show with numerical experiments that the new error estimates correctly capture the linear asymptotic convergence rate of the approximation.

Our theoretical analysis considers the class of (Markov) functions $f$ that can be written as

$$f(z) = \int_{-\infty}^{0} \frac{d\mu(\zeta)}{z-\zeta}, \qquad z \in \mathbb{C}\backslash ]-\infty, 0], \tag{1.2}$$

where $\mu$ is a (possibly signed) measure such that the absolute convergence of the integral takes place. The results can also be applied to a linear combination of such integrals with monomials $\lambda^\ell$, $\ell \in \mathbb{Z}$ as coefficients. This is the case, for instance, for functions such as $\lambda^\alpha$ with $\alpha \in \mathbb{R}\backslash\mathbb{Z}$, $\exp(-\sqrt{\lambda})$, $\tanh(\sqrt{\lambda})/\sqrt{\lambda}$ [13]; see also [21], [33] for a similar framework and for further examples. Nonetheless, we stress that the algorithm is also effective when used with other functions that do not belong to this class, as is the case, for instance, for $f(\lambda) = \exp(-\lambda)$; see, e.g., [45], [1].

A synopsis of the paper is as follows. In section 2 we describe the method with some algorithmic details. In section 3 we provide a new convergence analysis of the Extended Krylov subspace method; in section 3.1 and in section 3.2 we specialize our theory to the case of $A$ symmetric and nonsymmetric, and we provide numerical evidence of the accuracy of our asymptotic bounds. In section 4 we compare our method with the Standard Krylov subspace, and with another acceleration method, the shift-invert Lanczos.

We end this section with some notation. In the following, $\|x\|$ is the standard Euclidean norm for vectors, induced by the usual inner product. With $x^T$ ($x^*$) we denote the tranpose (conjugate transpose) of a vector $x$. With $W(A)$ we denote the field of values, or numerical range, of an $n \times n$ matrix $A$, that is $W(A) = \{x^*Ax,\ x \in \mathbb{C}^n, \|x\| = 1\}$.

**2. The method.** Given a space $\mathcal{K}$ and a matrix $V$ such that $\mathrm{range}(V) = \mathcal{K}$, an approximation to (1.1) may be obtained as $f(A)v \approx Vf(V^TAV)V^Tv$. A method that has been used since the 1980s employs the Krylov subspace $K_k(A,v) = \mathrm{span}\{v, Av, \ldots, A^{k-1}v\}$ as approximation space $\mathcal{K}$ [42], [13], [53]; note that these spaces are nested, that is $K_k(A,v) \subseteq K_{k+1}(A,v)$, so that $K_{k+1}(A,v)$ may be obtained from $K_k(A,v)$ by adding a single vector to the basis.

Let $V_k = [v_1, \ldots, v_k]$, where the set $\{v_1, \ldots, v_k\}$ is an orthonormal basis of $K_k(A,v)$ generated by a Gram–Schmidt process. We thus obtain

$$f(A)v \approx V_k f(H_k)e_1, \tag{2.1}$$

where $H_k = V_k^TAV_k$ is an upper Hessenberg matrix, and $V_k^Tv = e_1$, in which $e_1$ is the first column of the identity matrix whose dimension is clear from the context. We shall refer to the approximation in (2.1) as the *Standard Krylov method*, and it reduces to the Spectral *Lanczos* decomposition method (or SLDM) when $A$ is symmetric (see, e.g., [13], [48]), so that $H_k$ is also symmetric and thus tridiagonal.

The general description above suggests that the approximation space $\mathcal{K}$ could contain more information than that of the regular Krylov subspace $\mathcal{K}_k(A,v)$. To this end, it was proposed in [16] for $A$ symmetric to generate a sequence of *extended* approximation spaces that contain information on both $A$ and $A^{-1}$, that is

$$\mathrm{span}\{v, A^{-1}v, Av, A^{-2}v, A^2v, A^{-3}v, \ldots\}.$$

The algorithm [16] proceeds by first running $k$ steps of a Lanczos process with $A^{-1}$, and then continuing with $m$ iterations of a Lanczos process with $A$, while maintaining orthogonalization among all generated vectors in the sequence. Then an approximation to (1.1) is obtained by projection and restriction in the generated space of dimension $k + m - 1$.

In [50] an alternative implementation of this method was proposed. Starting with the pair $\{v, A^{-1}v\}$, in [50] a block Arnoldi-type method is used to generate an orthonormal basis of the extended subspace, by adding two vectors at the time, one multiplied by $A$, and one by $A^{-1}$. The new procedure is very general and it also applies for $A$ nonsymmetric. Let the columns of $\mathcal{V}_m$ span the extended subspace of dimension $2m$, and let $\mathcal{T}_m = \mathcal{V}_m^T A \mathcal{V}_m$. Once the approximation space is built, the approximate solution may be obtained as

$$u_m = \mathcal{V}_m f(\mathcal{T}_m) e_1.$$

It is important to realize that the matrix $\mathcal{T}_m$ may be constructed iteratively, without additional multiplications by $A$; see [50, Proposition 3.2]. The iterative update of $\mathcal{T}_m$ makes the algorithm much more efficient, compared with the original implementation in [16], where the number of applications of $A^{-1}$ was fixed a-priori and presumably small. We remark that such an implementation was proposed in [50] as an acceleration procedure for solving the Lyapunov equation, therefore its use in the context of matrix functions was not explored.

An outline of the Extended Krylov Subspace Method (EKSM) is given next. Here `gram_sh` implements the Gram–Schmidt procedure to orthogonalize the columns of the given matrix.

Given $v$, $A$, set $\mathbf{V}_1 = \texttt{gram\_sh}([v, A^{-1}v])$, $\mathcal{V}_0 = \emptyset$.
For $m = 1, 2, \ldots,$
1. $\mathcal{V}_m = [\mathcal{V}_{m-1}, \mathbf{V}_m]$
2. Set $\mathcal{T}_m = \mathcal{V}_m^T A \mathcal{V}_m$
3. Compute $y_m = f(\mathcal{T}_m) e_1$
4. If converged then $u_m = \mathcal{V}_m y_m$ and stop
5. $\mathbf{V}'_{m+1} = [A\mathbf{V}_m e_1, A^{-1}\mathbf{V}_m e_2]$
6. $\widehat{\mathbf{V}}_{m+1} \leftarrow$ orthogonalize $\mathbf{V}'_{m+1}$ w.r.to $\mathcal{V}_m$
7. $\mathbf{V}_{m+1} = \texttt{gram\_sh}(\widehat{\mathbf{V}}_{m+1})$

At each iteration of this process, two new vectors are added to the space. Unless breakdown occurs (cf. Proposition 2.1), at the $m$th iteration the method has constructed an orthonormal basis of dimension $2m$, given by the columns of the matrix $\mathcal{V}_m = [\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_m]$, $\mathbf{V}_i \in \mathbb{R}^{n\times 2}$, and $\mathbf{K}_m := \text{Range}(\mathcal{V}_m)$. The orthogonalization is performed first with respect to the previous basis vectors, and then within the new block of 2 vectors. It is also interesting that the matrices $\mathbf{V}_k$, $k = 1, 2, \ldots$ formally satisfy the following Arnoldi-like recurrence,

$$A\mathcal{V}_m = \mathcal{V}_m \mathcal{T}_m + \mathbf{V}_{m+1} \boldsymbol{\tau}_{m+1,m} E_m^T, \tag{2.2}$$

where $E_m$ contains the last 2 columns of the identity matrix of dimension $2m$, and $\boldsymbol{\tau}_{m+1,m} = \mathbf{V}_{m+1}^T A \mathbf{V}_m$; see [16] for a similar relation. At each iteration (cf. step 3), the function of a small $2m \times 2m$ matrix is computed with methods for dense matrices. In the absence of a method specifically developed for these purposes, and if $\mathcal{T}_m$ is diagonalizable, the eigenvalue decomposition of the matrix $\mathcal{T}_m$ is used to compute

$f(\mathcal{T}_m)$; see, e.g., [27], [32]. Only at convergence (step 4), the approximate solution is constructed as

$$u_m = \mathcal{V}_m y_m. \tag{2.3}$$

Next Proposition shows that in exact arithmetic, the full rank of the costructed basis is ensured as long as the associated space keeps growing.

PROPOSITION 2.1. *Let* $\mathbf{V}_j = [v_j^{(1)}, v_j^{(2)}]$, $j = 1, \ldots, m$. *With the previous notation, if*

$$\dim(\mathrm{span}\{A^{-m}v, \ldots, A^{m-1}v\}) = 2m,$$

*then for* $1 \le j \le m$

$$v_j^{(1)} = p_{j-1}(A)v + q_{j-1}(A^{-1})v, \quad v_j^{(2)} = r_{j-1}(A^{-1})A^{-1}v + s_{j-1}(A^{-1})v,$$

*where* $\deg p_{j-1} = \deg r_{j-1} = j-1$, $\deg q_{j-1} \le j-1$, $\deg s_{j-1} \le j-1$, *and no premature termination takes place.*

*Proof.* In the following we use the notation $\mathrm{l.c.}\{A^k v, \ldots, A^j v\}$ with $k \le j$ to denote any linear combination of the given vectors with (possibly negative) powers of $A$ from $k$ to $j$. Analogously, we use $\mathrm{l.c.}\{v_1, \ldots, v_k\}$ to denote a linear combination of $k$ given vectors.

We proceed by induction on $m$. For $m = 1$ we have $v_1^{(1)} = v = p_0(A)v$ and $v_1^{(2)} = cA^{-1}v + \mathrm{l.c.}\{A^0 v\} = r_0(A^{-1})A^{-1}v + s_0(A)v$, with $c \neq 0$. This proves the claim for $m = 1$.

We proceed with $m+1$. We have

$$Av_m^{(1)} = Ap_{m-1}(A)v + Aq_{m-1}(A^{-1})v = cA^m v + \mathrm{l.c.}\{A^{-m+2}v, \ldots, A^{m-1}v\},$$

with $c \neq 0$ since $\deg p_{m-1} = m-1$. Orthogonalization with respect to the previous vectors $v_j^{(1)}, v_j^{(2)}$, $j = 1, \ldots, m$ gives

$$\begin{aligned} cA^m v + \mathrm{l.c.}\{A^{-m+2}v, \ldots, A^{m-1}v\} + \mathrm{l.c.}\{v_1^{(1)}, \ldots, v_m^{(2)}\} \\ = cA^m v + \mathrm{l.c.}\{A^{-m}v, \ldots, A^{m-1}v\} \neq 0, \end{aligned}$$

owing to linear independence. Thus $v_{m+1}^{(1)} = p_m(A)v + q_m(A^{-1})v$, with $\deg p_m = m$.

Analogously, we obtain

$$\begin{aligned} A^{-1}v_m^{(2)} &= A^{-1}r_{m-1}(A^{-1})A^{-1}v + A^{-1}s_{m-1}(A)v \\ &= cA^{-m-1}v + \mathrm{l.c.}\{A^{-m}v, \ldots, A^{m-2}v\},\ c \neq 0. \end{aligned}$$

Orthogonalization with respect to $v_j^{(1)}, v_j^{(2)}$, $j = 1, \ldots, m$ and to $v_{m+1}^{(1)}$ gives

$$\begin{aligned} cA^{-m-1}v + \mathrm{l.c.}\{A^{-m}v, \ldots, A^{m-1}v\} + \mathrm{l.c.}\{p_m(A)v + q_m(A^{-1})v\} \\ = cA^{-m-1}v + \mathrm{l.c.}\{A^{-m}v, \ldots, A^{m}v\} \neq 0, \end{aligned}$$

owing to linear independence, again. Therefore, $v_{m+1}^{(2)} = r_m(A^{-1})A^{-1}v + s_m(A)v$, with $\deg r_m = m$. ∎

We remark that the proof of Proposition 2.1 also implies that if breakdown occurs, then it is a happy breakdown, that is, an invariant subspace of $A$ associated with $v$ is found, containing the exact solution of the given problem.

It remains to decide how convergence is detected. This is discussed in the next section.

**2.1. Stopping criteria.** Following the classical proposal in [48], a possible simple stopping criterion is given by the quantity:

$$\|\boldsymbol{\tau}_{m+1,m} E_m^T f(\mathcal{T}_m) e_1\|, \tag{2.4}$$

which can be cheaply computed during the iteration since all the quantities are available. A natural motivation for this quantity stems from a few application problems. For istance, assume the following boundary value problem is given:

$$Au - u_{zz} = 0, \quad u(0) = v, \quad u(+\infty) = 0,$$

with $A$ positive definite. It may be readily shown that $u(z) = f(A)v = \exp(-z\sqrt{A})v$ is the solution to this problem. Moreover, writing $u_m(z) = \mathcal{V}_m f(\mathcal{T}_m) e_1$ and using (2.2), we obtain

$$Au_m - (u_m)_{zz} = \mathbf{V}_{m+1} \boldsymbol{\tau}_{m+1,m} E_m^T f(\mathcal{T}_m) e_1.$$

Due to the orthogonality of the basis, it follows that the quantity in (2.4) is the residual norm of the differential equation; similar reasonings were used for instance in [12].

Another possibility, first suggested for the exponential function in [31], is given by the following estimate

$$\frac{\|u - u_m\|}{\|u_m\|} \approx \frac{\delta_{m+j}}{1 - \delta_{m+j}}, \tag{2.5}$$

where $\delta_{m+j} = \|u_{m+j} - u_m\|/\|u_m\|$, namely the norm of the difference between two computed approximations, so that after $m + j$ iterations it is possible to compute an estimate of the error at step $m$. We next give an algebraic proof for this approximation.

PROPOSITION 2.2. *Assume that $m + j$ iterations of the Extended Krylov method have been taken. With the notation above, let $u_{m+j} = \mathcal{V}_{m+j} f(\mathcal{T}_{m+j}) e_1$. Then*

$$\frac{\|u - u_m\|}{\|u\|} \leq \frac{\delta_{m+j} + \frac{\varepsilon_{m+j}}{\|u_m\|}}{1 - \delta_{m+j} - \frac{\varepsilon_{m+j}}{\|u_m\|}},$$

*where $\varepsilon_{m+j} = \|u - u_{m+j}\|$. If $\varepsilon_{m+j} \ll \varepsilon_m$ and $\|u\| \approx \|u_m\|$ then (2.5) follows.*

*Proof.* We write $\|u_m\| \leq \|u_m - u_{m+j}\| + \|u - u_{m+j}\| + \|u\|$, so that $\|u\| \geq \|u_m\| - \varepsilon_{m+j} - \|u_{m+j} - u_m\|$. Therefore

$$\frac{\|u - u_m\|}{\|u\|} \leq \frac{\|u_{m+j} - u_m\| + \|u - u_{m+j}\|}{\|u_m\| - \varepsilon_{m+j} - \|u_{m+j} - u_m\|}.$$

Collecting $\|u_m\|$ at the numerator and denominator of the right-hand side, we obtain the inequality. The approximation readily follows. ∎

We note that for $A$ symmetric and positive definite, it holds that $\|u_m\| \leq \|u\|$ for all $m$ (cf. [21]; see an analogous earlier result for the exponential in [11]), so that only the hypothesis on $\varepsilon_m$ remains.

A third option fully exploits the theoretical information on the a-priori rate of convergence, namely that the rate is $aq^m$ for some $q$ and $a$ to be estimated; see, e.g., [14], for a similar approach. We use the ideal equalities

$$\|u_{m+j} - u_m\| = cq^m, \qquad \|u_{m+2j} - u_{m+j}\| = cq^{m+j}.$$

We set $\chi_m := \log \|u_{m+j} - u_m\|$, $\chi_{m+j} := \log \|u_{m+2j} - u_{m+j}\|$, from which we obtain

$$q = \exp\left(-\frac{1}{j}(\chi_m - \chi_{m+j})\right), \qquad c = \exp\left(\frac{j+m}{j}\chi_m - \frac{m}{j}\chi_{m+j}\right). \tag{2.6}$$

Our albeit limited experience showed that the last two estimates are quite reliable. Given a fixed tolerance, we have used them as stopping criterion for the considered methods, which for (2.5) reads as follows:

$$\texttt{if } \tfrac{\delta_{m+j}}{1-\delta_{m+j}} \leq \texttt{tol then stop} \tag{2.7}$$

In the case of EKSM the two criteria provided graphically undistinguishable convergence curves.

Both stopping criteria require saving the last $j$ (resp. $2j$) solution vectors for (2.5) (resp. (2.6)). However, we noticed that a very small value of $j$ ($j = 2$ and $j = 1$ respectively) was sufficient for the analysis. Moreover, since the basis is orthogonal, we only needed to store much shorter vectors (e.g., vector $y_m$ in (2.3)) to compute $\|u_{m+j} - u_m\| = \|y_{m+j} - \hat{y}_m\|$, where $\hat{y}_m$ is the vector $y_m$ padded with $j$ zeros at the bottom.

**3. Convergence theory.** In this section we derive a new estimate for the convergence rate of the Extended Krylov subspace method. The estimate is based on a very general approximation result, which can be applied to symmetric and nonsymmetric matrices, and to a quite large class of functions. We stress that the derivations of these results are new, and that these derivations significantly differ from those used in [16] in the symmetric case. In particular, the new approach allows us to improve the original convergence estimate even in the symmetric case.

We consider the class of Markov type functions, which can be written in integral form as in (1.2) or as linear combination with monomials. For $a > 0$, we split the integral in (1.2) as

$$f = f_1 + f_2, \qquad f_1(z) = \int_{-\infty}^{-a} \frac{d\mu(\zeta)}{z-\zeta}, \qquad f_2(z) = \int_{-a}^{0} \frac{d\mu(\zeta)}{z-\zeta}. \tag{3.1}$$

In our analysis the integrals $f_1$ and $f_2$ are conveniently written as Faber series expansions and then the error of this series truncation of these series are evaluated. This idea was originally proposed in [16] for the symmetric case. Technically speaking, the main difference in this new approximation result is the way the expansion errors for $f_1$ and $f_2$ are derived. The tools are very general, so that the proof applies as is to nonsymmetric matrices.

We define $W_1 := W(A)$, and[1] $W_2 = (W(A))^{-1} := \{z^{-1} | z \in W_1\}$, and we assume that both $W_j$ are symmetric with respect to the real axis $\mathbb{R}$ and (strictly) lie in the right half-plane. Let $D$ denote the closed unit circle, and let $\Psi_j : \bar{\mathbb{C}} \backslash D \to \bar{\mathbb{C}} \backslash W_j$, $\Phi_j = \Psi_j^{-1}$ be the direct and inverse Riemann mappings for $W_j$, for $j = 1, 2$. Moreover, $F_{j,k}$, $k \in \mathbb{N}$, $j = 1, 2$ denote the corresponding Faber polynomials of degree $k$; see, e.g., Suetin [52].

We start with an approximation result.

[1] We remark that $W_2$ is not necessarily the same as $W(A^{-1})$. This definition will be crucial in the proof of Theorem 3.4.

LEMMA 3.1. *Let $f$ be defined by (1.2) and satisfy (3.1) for some $a > 0$. With the notation above, for any $m \in \mathbb{N}$, $m > 1$ it holds*

$$\left| f_1(z) - \sum_{k=0}^{m-1} \gamma_{1,k} F_{1,k}(z) \right| \le c_1 |\Phi_1(-a)|^{-m}, \qquad z \in W_1$$
$$\left| f_2(z) - \sum_{k=0}^{m} \gamma_{2,k} F_{2,k}(z^{-1}) \right| \le c_2 |\Phi_2(-a^{-1})|^{-m}, \qquad z \in W_1,$$

*where $c_1, c_2$ are positive real constants independent of $m$, $n$, and where $\gamma_{j,k}$, $j = 1, 2$ are some real numbers.*

*Proof.* By using the splitting in (3.1), we next approximate $f_1(A)$ as a function of $A$ with a polynomial in $A$, which corresponds to a truncated Faber series expansion of $f_1$ built on $W_1$. Morever, we approximate $f_2\left[(A^{-1})^{-1}\right]$ as a function of $A^{-1}$ with a truncation of a Faber series in $A^{-1}$ built on $W_1^{-1}$.

The generating relation for Faber polynomials is (see, e.g., [52, Ch. 2, § 2, formula (1)])

$$\frac{1}{z-\zeta} = -\frac{1}{\Psi'[\Phi(\zeta)]} \sum_{k=0}^{\infty} \frac{F_k(z)}{\Phi(\zeta)^{k+1}}, \qquad z \in W, \quad \zeta \notin W. \tag{3.2}$$

Exploiting (3.2) for $W = W_1$ and definition (3.1), we derive

$$\begin{aligned} f_1(z) &= -\int_{-\infty}^{-a} \frac{1}{\Psi_1'[\Phi_1(\zeta)]} \sum_{k=0}^{\infty} \Phi_1(\zeta)^{-k-1} F_{1,k}(z) d\mu(\zeta) \\ &= -\sum_{k=0}^{\infty} F_{1,k}(z) \int_{-\infty}^{-a} \frac{d\mu(\zeta)}{\Psi_1'[\Phi_1(\zeta)]\Phi_1(\zeta)^{k+1}}. \end{aligned}$$

The following properties are valid for any continuum $W$ symmetric with respect to $\mathbb{R}$ and lying in the right half-plane (due to the symmetry and bijectivity, $\Psi$ monotonically maps $]-\infty, -1[$ onto $]-\infty, \min \mathbb{R} \cap W[$):

$$\begin{aligned} |\Phi(\zeta)| &\ge c_3 |\zeta|, \qquad \zeta \in ]-\infty, 0], \\ |\Phi(\zeta)| &\ge |\Phi(-b)|, \qquad \zeta \in ]-\infty, -b], \quad b > 0, \\ |\Psi'[\Phi(\zeta)]| &\ge c_4, \qquad \zeta \in ]-\infty, 0], \end{aligned} \tag{3.3}$$

where $c_i$ are positive constants depending on $W$.

Utilizing (3.3) with $W = W_1$ and $b = a$, we obtain

$$\left| -\int_{-\infty}^{-a} \frac{d\mu(\zeta)}{\Psi_1'[\Phi_1(\zeta)]\Phi_1(\zeta)^{k+1}} \right| \le \int_{-\infty}^{-a} \frac{|d\mu(\zeta)|}{c_4 c_3 |\zeta|\, |\Phi_1(-a)|^k} \le c_5 |\Phi_1(-a)|^{-k},$$

where $c_5 > 0$ depends on $W_1$ and $a$. Thus, since $\max_{z \in W_1} |F_{1,k}(z)| \le 2$ (see [52, Ch. IX, § 3, theorem 10]),

$$\begin{aligned} \left| f_1(z) + \sum_{k=0}^{m-1} F_{1,k}(z) \int_{-\infty}^{-a} \frac{d\mu(\zeta)}{\Psi_1'[\Phi_1(\zeta)]\Phi_1(\zeta)^{k+1}} \right| &\le c_5 \sum_{k=m}^{\infty} |\Phi_1(-a)|^{-k} |F_{1,k}(z)| \\ &\le c_1 |\Phi_1(-a)|^{-m}, \qquad z \in W_1. \end{aligned}$$

As to $f_2$, we consider the change of variables $y = z^{-1}$. Analogously to what was done for $f_1$, we have

$$\begin{aligned} f_2(z) = f_2\left(y^{-1}\right) &= \int_{-a}^{0} \frac{d\mu(\zeta)}{y^{-1}-\zeta} = \int_{-a}^{0} \frac{y\, d\mu(\zeta)}{1-\zeta y} = y\int_{-a}^{0} \frac{d\mu(\zeta)}{\zeta\left(\zeta^{-1}-y\right)} \\ &= y\int_{-a}^{0} \frac{d\mu(\zeta)}{\zeta\Psi_2'\left[\Phi_2\left(\zeta^{-1}\right)\right]} \sum_{k=0}^{\infty} \frac{F_{2,k}(y)}{\Phi_2(\zeta^{-1})^{k+1}} \\ &= y\sum_{k=0}^{\infty} F_{2,k}(y) \int_{-a}^{0} \frac{d\mu(\zeta)}{\zeta\Psi_2'\left[\Phi_2\left(\zeta^{-1}\right)\right]\Phi_2(\zeta^{-1})^{k+1}} \end{aligned}$$

and

$$\begin{aligned} \left|\int_{-a}^{0} \frac{d\mu(\zeta)}{\zeta\Psi_2'\left[\Phi_2\left(\zeta^{-1}\right)\right]\Phi_2(\zeta^{-1})^{k+1}}\right| &\le \int_{-a}^{0} \frac{|d\mu(\zeta)|}{|\zeta|\cdot c_6\cdot|\zeta^{-1}|\cdot|\Phi_2(-a^{-1})|^k} \\ &\le c_7|\Phi_2(-a^{-1})|^{-k}. \end{aligned}$$

Since the conformal mapping $z \mapsto z^{-1}$ preserves angles and preserves the order of arc lengths on $\partial W_1$, $\partial W_2 = \partial W_1^{-1}$ is a finite rotation (Radon) curve. Therefore, Theorem 11 from [52, Ch. IX, § 3] ensures that $\max_{y\in W_1^{-1}} |F_{2,k}(y)| \le c_8$. Hence, for $y \in W_1^{-1}$,

$$\left|f_2\left(y^{-1}\right) - y\sum_{k=0}^{m-1} F_{2,k}(y)\int_{-a}^{0} \frac{d\mu(\zeta)}{\zeta\Psi_2'\left[\Phi_2\left(\zeta^{-1}\right)\right]\Phi_2(\zeta^{-1})^{k+1}}\right| \le c_2|\Phi_2(-a^{-1})|^{-m}. \qquad \square$$

Remark 3.2. The proof strongly relies on the splitting $f = f_1 + f_2$, and the effectiveness of the bound depends on the value of $a$, which drives the integral splitting. In particular, it becomes readily apparent that the expansion of $f_1$ allows one to derive an error bound for the Standard Arnoldi method (for $a = 0$), whereas the error bound for the expansion of $f_2$ has connections with the convergence analysis of Shift-Invert Lanczos method ($a = \infty$); see also [39].

Next lemma ensures that polynomials in $A$ and in $A^{-1}$ up to a certain degree are exactly represented in the extended Krylov space; see [16] for a proof for the symmetric case which is valid for the nonsymmetric case as well.

Lemma 3.3. *Matrix polynomials in $A$ (resp. in $A^{-1}$) of degree $k \le m-1$ (resp. of degree $k \le m$) are exact in $\mathbf{K}_m$. In particular, $p_k(A)b = \mathcal{V}_m p_k(\mathcal{T}_m)e_1 \in \mathbf{K}_m$, $k \le m-1$, and $p_k(A^{-1})b = \mathcal{V}_m p_k(\mathcal{T}_m^{-1})e_1 \in \mathbf{K}_m$, $k \le m$.*

Next theorem provides an error estimate for the approximation with the extended Krylov subspace method.

Theorem 3.4. *Let $A \in \mathbb{R}^{n\times n}$ be nonsingular with $W(A) \subset \mathbb{C}^+$, and let $f$ satisfy (1.2). There exists $a > 0$ such that it holds*

$$\|f(A)v - \mathcal{V}_m f(\mathcal{T}_m)e_1\| \le \frac{c_9}{|\Phi_1(-a)|^m}, \tag{3.4}$$

*where $c_9$ is a positive constant depending on $W(A)$ and on the measure $\mu$ but independent of $n$ and $m$.*

*Proof.* Define the functions

$$g(z) = f_1(z) - \sum_{k=0}^{m-1} \gamma_{1,k} F_{1,k}(z), \qquad h(z) = f_2(z) - \sum_{k=0}^{m} \gamma_{2,k} F_{2,k}(z^{-1}). \tag{3.5}$$

Using the decomposition in (3.1) and Lemma 3.3, we have

$$\begin{aligned}
\|f(A)v - \mathcal{V}_m f(\mathcal{T}_m)e_1\| = \Big\| & f_1(A)v - \sum_{k=0}^{m-1} \gamma_{1,k} F_{1.k}(A)v - \mathcal{V}_m f_1(\mathcal{T}_m)e_1 \\
& + \mathcal{V}_m \sum_{k=0}^{m-1} \gamma_{1,k} F_{1,k}(\mathcal{T}_m)e_1 + f_2(A)v - \sum_{k=0}^{m} \gamma_{2,k} F_{2,k}(A^{-1})v \\
& - \mathcal{V}_m f_2(\mathcal{T}_m)e_1 + \mathcal{V}_m \sum_{k=0}^{m} \gamma_{2,k} F_{2,k}(\mathcal{T}_m^{-1})e_1 \Big\| \\
= \|g(A)v - \mathcal{V}_m g(\mathcal{T}_m)e_1 + h(A)v - \mathcal{V}_m h(\mathcal{T}_m)e_1\| \\
\le \|g(A)\| + \|g(\mathcal{T}_m)\| + \|h(A)\| + \|h(\mathcal{T}_m)\|.
\end{aligned} \tag{3.6}$$

Since both functions in (3.5) are analytic in $W_1$ and since $W(\mathcal{T}_m) \subseteq W_1$, we deduce from [8, Theorem 2 and formula (1)]:

$$\max\{\|g(A)\|, \|g(\mathcal{T}_m)\|\} \le 11.08 \max_{z \in W_1} |g(z)|, \tag{3.7}$$

$$\max\{\|h(A)\|, \|h(\mathcal{T}_m)\|\} \le 11.08 \max_{z \in W_1} |h(z)|. \tag{3.8}$$

These inequalities combined with (3.6) and Lemma 3.1, give (3.4). Here $a$ is chosen so that

$$|\Phi_1(-a)| = |\Phi_2(-a^{-1})|\,. \tag{3.9}$$

□

Remark 3.5. We wish to stress that the use of (3.7) from [8] is crucial for our bound, without which we obtained a less sharp result in the nonsymmetric case; see also [2, 3] for more inspiring results in this direction.

**3.1. The symmetric case.** We derive the optimal value of $a$ in (3.9) by explicitly writing the conformal mappings.

Proposition 3.6. *Let the spectrum of the symmetric matrix $A$ be contained in $[\alpha, \beta] \subset \mathbb{R}^+$. Then for $a = \sqrt{\alpha\beta}$ it holds that*

$$\|f(A)v - \mathcal{V}_m f(\mathcal{T}_m)e_1\| \le \frac{c_{10}}{|\Phi_1(-a)|^m} \simeq \mathcal{O}\left[\exp\left(-2m\sqrt[4]{\alpha/\beta}\right)\right],$$

*where the last equality asymptotically holds for large $\beta/\alpha$.*

*Proof.* For $\sigma(A) \subset [\alpha, \beta] \subset \mathbb{R}^+$, the mappings are given in terms of the scaled inverse Zhukovsky function as

$$\Phi_1(z) = \frac{z-c}{d} + \sqrt{\left(\frac{z-c}{d}\right)^2 - 1}, \quad \Phi_2(z^{-1}) = \frac{z^{-1}-\hat{c}}{\hat{d}} + \sqrt{\left(\frac{z^{-1}-\hat{c}}{\hat{d}}\right)^2 - 1},$$

with $c = (\alpha+\beta)/2$, $d = (\beta-\alpha)/2$, $\hat{c} = (1/\alpha + 1/\beta)/2$, $\hat{d} = (1/\alpha - 1/\beta)/2$. Imposing (3.9) yields

$$\frac{-a-c}{d} = \frac{-1/a-\hat{c}}{\hat{d}},$$

from which we readily obtain

$$a = \sqrt{\alpha\beta}, \qquad |\Phi_1(-a)| = |\Phi_2(-1/a)| = \left|z + \sqrt{z^2-1}\right|, \quad \text{with} \quad z = \frac{\sqrt{\beta}+\sqrt{\alpha}}{\sqrt{\beta}-\sqrt{\alpha}}.$$

The result follows from using Theorem 3.4. □

The result of Proposition 3.6 should be compared with the estimate appearing in [16, Theorem 6], namely $\|f(A)v - \mathcal{V}_m f(\mathcal{T}_m)e_1\| = \mathcal{O}\left(m^2 \exp(-2m\sqrt[4]{\alpha/\beta})\right)$, containing the weakening factor $m^2$.

EXAMPLE 3.7. In Figure 3.1 we consider the convergence history (solid curve) of EKSM for $f(\lambda) = \lambda^{-1/2}$ and a real symmetric matrix $A$ of dimension $n = 400$ with eigenvalues uniformly distributed in $[0.01, 0.9]$ (left plot) and $[1, 50]$ (right plot). The dashed curve is the asymptotic convergence rate predicted by the estimate above, namely $1/|\Phi_1(-a)|^m$. The agreement of the estimate is very good in both cases.

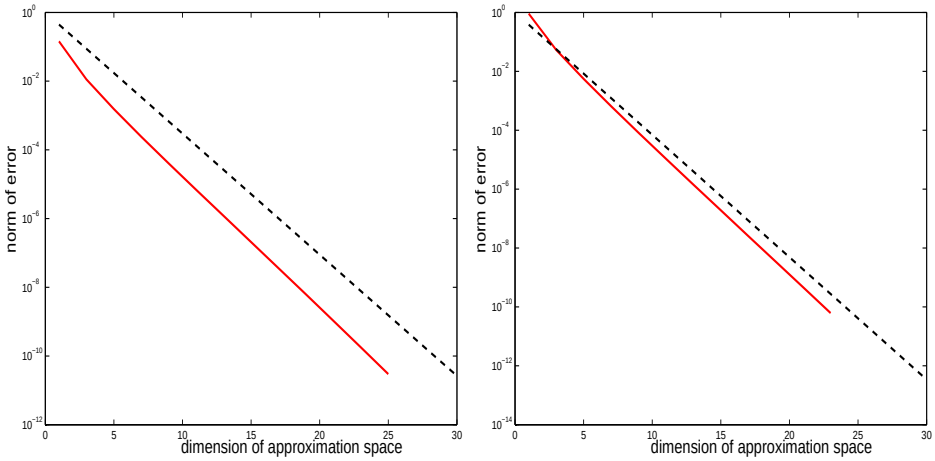


FIG. 3.1. *Example 3.7. True error (solid line) and its estimate (dashed line) for the function* $f(\lambda) = \lambda^{-1/2}$ *and a symmetric matrix with eigenvalues uniformly distributed in* $[0.01, 0.9]$ *(left plot) and* $[1, 50]$ *(right plot).*

**3.2. The nonsymmetric case.** In the nonsymmetric case, the situation is significantly complicated by the fact that even assuming that it is possible to analytically determine a simple curve containing $W(A)$, then little is known about the curve containing $(W(A))^{-1}$; see, e.g., [32], [25]. It should also be added that even when both mappings $\Phi_1$ and $\Phi_2$ can be explicitly derived, then the optimal value of $a$ satisfying (3.9) is hard to find analytically, and numerical procedures need be adopted for its estimation. In the following we consider examples where it is possible to enclose $W_1$ and $W_2$ within an explicit conformal mapping, and then we estimate $a$ by numerically solving the nonlinear equation $|\Phi_1(-a)| - |\Phi_2(-a^{-1})| = 0$. This allows us to derive

the asymptotic term of Theorem 3.4. A natural candidate as an approximation to $a$ is $a = \sqrt{\lambda_{\min}(H)\lambda_{\max}(H)}$, where $H$ is the symmetric part of $A$, $H = (A + A^T)/2$, since $W(H) = W(A) \cap \mathbb{R}$. However, this turns out to be in most cases too optimistic.

In all examples the field of values was approximated with the function `fv` in the Matrix Computation Toolbox by N. Higham [26]. The images of the conformal mappings were obtained with the Schwarz–Christoffel Mapping Toolbox by T. Driscoll [10].

EXAMPLE 3.8. For $f(\lambda) = \lambda^{-1/2}$, we consider the normal diagonal $400 \times 400$ matrix $A$ whose eigenvalues lie on the ellipse with semi-axes $a_1 = 0.097$, $a_2 = 0.01$ and center $c = 0.1$. The vector $v$ is taken as the normalized vector of ones. The true convergence rate of EKSM (solid line) and its estimate $1/|\Phi_1(-a)|^m$ (dashed line) are reported in Figure 3.2. The agreement is impressive. The optimal (numerically computed) value of the asymptotic parameter was $a = 0.02$, whereas $a = \sqrt{\ell_1 \ell_n} = \sqrt{\lambda_{\min}(H)\lambda_{\max}(H)} = 0.0243$.

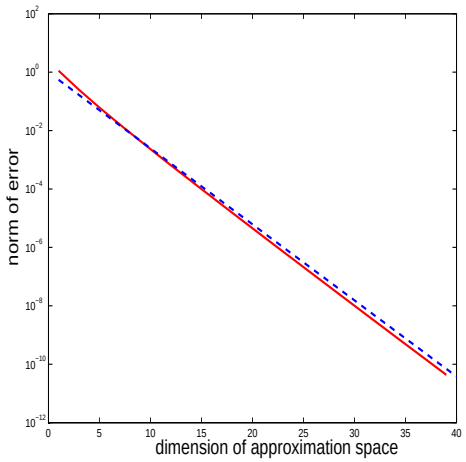


FIG. 3.2. *Example 3.8 with normal matrix. True convergence history of EKSM (solid line) and its estimate (dashed line).*

EXAMPLE 3.9. We consider the function $f(\lambda) = \sqrt{\lambda}$ and the $200 \times 200$ Jordan block matrix associated with the single eigenvalue $\lambda = 4$. The associated field of values is a disk centered at $\lambda$ of radius close to 1 (see [32, exercise 9 in § 1.3, exercise 29 in § 1.5] and [19]); $v$ is the normalized vector of all ones. In our experiments, $f(A)$ and $f(\mathcal{T}_m)$ are computed by Heron's method (see [27, formula (4.27)]). Figure 3.3 shows the convergence rate of the Extended Krylov Subspace Method, together with our asymptotic estimate, where the optimal parameter $a$ has been determined numerically. The new estimate perfectly matches the true asymptotic convergence rate.

EXAMPLE 3.10. We conclude with a diagonalizable $200 \times 200$ upper triangular matrix, with diagonal elements uniformly distributed in $[50, 400]$ and second upper diagonal elements equal to $a_{i,i+2} = -30$. The vector $v$ is chosen as $v = X[1, \ldots, 1]^T$, scaled so as to have unit norm, where $X$ is the matrix of eigenvectors of $A$. The considered function is the inverse square root. Figure 3.4 shows the convergence rate of the Extended Krylov Subspace Method, together with our asymptotic estimate, where the optimal parameter $a$ has once again been determined numerically. The agreement is sufficiently satisfactory, taking into account that on this problem adaptation of the method to the spectrum can be observed [35].

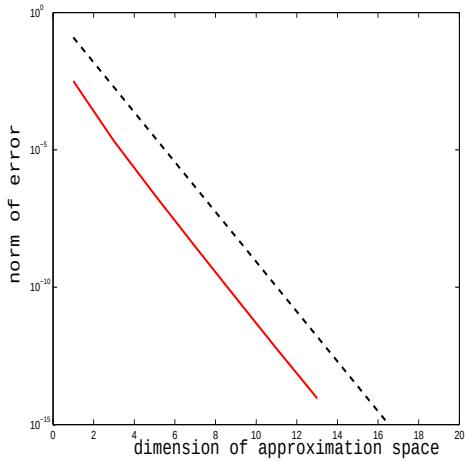


FIG. 3.3. *Example 3.9. True convergence history and its estimate from Theorem 3.4.*

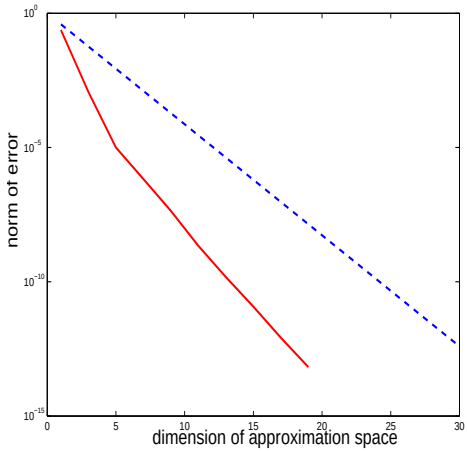


FIG. 3.4. *Example 3.10. True convergence history and its estimate from Theorem 3.4.*

**4. Numerical comparisons.** In this section we compare the new version of the Extended Krylov subspace method with other approaches that have been studied for the same purposes. In most cases, we compare EKSM with Standard Lanczos, and with the Shift-Invert Lanczos (SI-Lanczos) method. Given a real parameter $\gamma > 0$, the Shift-Invert Lanczos method constructs the Krylov subspace $K_m((I+\gamma A)^{-1}, v)$, computes the projection and restriction $T_m$ of the shifted and inverted matrix $(I+\gamma A)^{-1}$, and then computes an approximation as $Q_m f(\gamma^{-1}(T_m^{-1} - I))e_1$, where the columns of $Q_m$ form an orthonormal basis of $K_m((I+\gamma A)^{-1}, v)$. The method was analyzed in [40] for a class of functions, and in [31] for the exponential function. In [39] a study of the parameter $\gamma$ was performed, and in the symmetric non-singular case the value $\gamma = 1/\sqrt{\lambda_{\min}\lambda_{\max}}$ was obtained as a quasi-optimal estimate.

In the following we analyze methods by comparing the dimension of the approximation space required to achieve convergence, and in some cases by also comparing the CPU time. Other memory requirements may become significant, such as those

needed to store the matrix factors, for methods that require system solves. We should add, however, that unless stated otherwise, the factorization of $A$ was neither time nor memory consuming. All experiments were carried out in Matlab [37] on a a single CPU of a 2GB memory PC, running an Intel dual Core System at 2GHz. All methods were stopped when the error norm was sufficiently small. This forced us to compute the exact solution, by means of an eigenvalue decomposition, so that we could not test very large problems. Alternatively, one could consider adopting a (cheap) a-posteriori stopping criterion, which may vary depending on the method. Since the early contribution in [48] for the exponential function, stopping criteria have become an important ingredient of methods that approximate matrix functions, and it is therefore a current topic of active research; see, e.g., [9], [22], [31], [46]. A simple although not always reliable stopping criterion for EKSM, as well as for other methods based on an Arnoldi-like relation, is given by (2.4). Except for Example 4.6 where (2.7) was used, we decided to measure the true error for all methods to minimize the influence of the stopping criterion on our comparison.

The implementation of all discussed methods requires the generation of an orthonormal basis of the Krylov subspace (in the standard or accelerated versions). If $A$ is symmetric and if exact arithmetic is assumed, the next basis vector may be obtained with a short-term recurrence which involves only two previous (block) basis vectors. However, it is important to realize that in finite precision computation, orthogonality of the basis is soon lost, unless full reorthogonalization of the basis vectors is enforced as the subspace grows [12]. In our experiments we explicitly orthogonalized all vectors only in the nonsymmetric case, and we did not observe any performance degradation in the symmetric case when no reorthogonalization was performed.

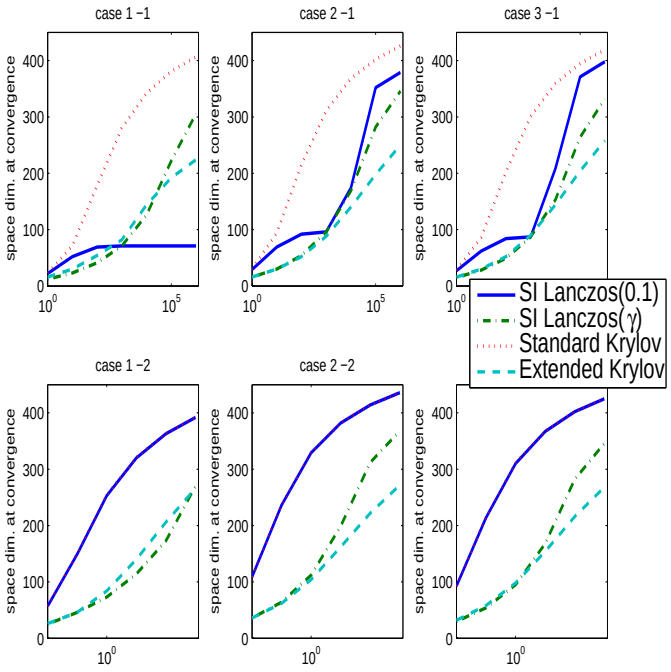


FIG. 4.1. *Example 4.1. Performance of all methods on the functions in (4.1). Upper plots: spectrum in* $[10^{-1}, 10^{k}]$*, with* $k = 0, \ldots, 6$*. Lower plots: spectrum in* $[10^{-4}, 10^{k}]$*, with* $k = -2, \ldots, 3$.

EXAMPLE 4.1. We compare the performance of EKSM, of SI-Lanczos and of the Standard Krylov method for the following three functions

$$f_{(1)}(\lambda) = \exp(-\sqrt{\lambda}), \quad f_{(2)}(\lambda) = \lambda^{-1/2}, \quad f_{(3)}(\lambda) = \lambda^{-1/4} \tag{4.1}$$

and a diagonal $500 \times 500$ real matrix whose eigenvalues are log-uniformly distributed in the interval $[10^{-1}, 10^k]$, $k = 0, \ldots, 6$ (case $\star - 1$) and $[10^{-4}, 10^k]$, $k = -2, \ldots 3$ (case $\star - 2$). The results are reported in the plots of Figure 4.1. On the abscissa is the order of magnitude of the largest eigenvalue, while on the y-axis is the dimension of the approximation space at convergence, for tol=$10^{-8}$. The first row of plots shows case $1 - \star$, while the second row reports case $2 - \star$, for the three test functions. Each column plot in the figure refers to the same function. The curves correspond to Standard Krylov, EKSM, and SI-Lanczos with the quasi-optimal value of $\gamma$ suggested in [39], and with $\gamma = 0.1$. Except for case 1 for $f_{(1)}(\lambda) = \exp(-\sqrt{\lambda})$, convergence deteriorates far less for EKSM than for the other methods, as the condition number of the matrix worsens, that is, the dimension of the final approximation space grows far less for EKSM than for the other methods. In case 2, the curve corresponding to SI-Lanczos with $\gamma = 0.1$ fully overlaps that of Standard Krylov, showing a disappointing behavior of SI-Lanczos for that choice of the parameter. Case 1 and $f(\lambda) = \exp(-\sqrt{\lambda})$ deserves a closer look, reporting a particularly good performance for SI-Lanczos with $\gamma = 0.1$. A detailed analysis of the SI-Lanczos($\gamma$) convergence history is described in Figure 4.2 for $\sigma(A) \subseteq [10^{-1}, 10^6]$. We also report the behavior of SI-Lanczos for other values of the parameter. Clearly, the behavior of the method is highly sensitive to the choice of the parameter, and the value of $\gamma$ suggested by the theory may be far from optimal. This phenomenon was also discussed in [39], where it is noticed that the inadequacy of that value of $\gamma$ is particularly apparent for large $\lambda_{\max}$ [39, Remark 6].

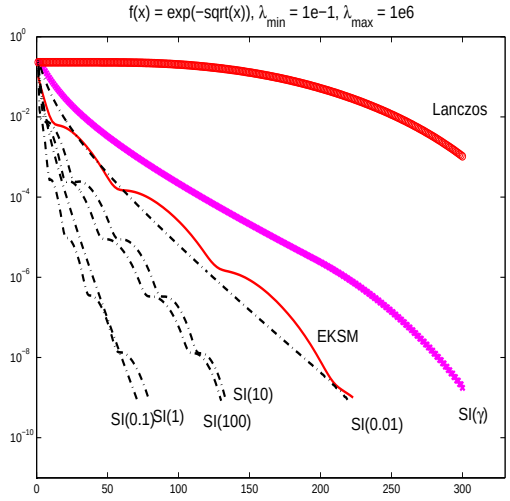


FIG. 4.2. *Example 4.1. Convergence history of all methods for* $\sigma(A) \subseteq [10^{-1}, 10^6]$ *and* $f(\lambda) = \exp(-\sqrt{\lambda})$ *(case 1-1 in Figure 4.1). "SI" stands for SI-Lanczos.*

EXAMPLE 4.2. We consider the matrix $A \in \mathbb{R}^{4900 \times 4900}$, discretization with centered finite differences of the 2D selfadjoint operator $\mathcal{L}(u) = -\frac{1}{10}u_{xx} - 100u_{yy}$, in $[0, 1]^2$ with homogeneous boundary conditions. The spectrum of $A$ is contained in

$[9.6 \cdot 10^2,\ 1.96 \cdot 10^6]$, and the considered function is $f(\lambda) = \lambda^{-1/2}$. CPU times obtained in Matlab ([37]) for various methods are reported in Table 4.1. We can readily observe that the Standard Krylov method is very slow, requiring a large approximation subspace to reach the requested accuracy of $10^{-8}$. It is important to realize that for a large approximation space, also the computation of $f$ on the reduced matrix becomes expensive, yielding an overall unacceptable timing. SI-Lanczos severely depends on the accuracy with which the optimal $\gamma$ (=2e-5) is approximated. The numbers show that capturing the order of magnitude of the parameter is not sufficient to achieve comparable performance. We emphasize that in general, the optimal $\gamma$ is not cheaply available. The Extended Krylov method performs very well on this problem, and it is free of parameters. We also report the performance when the Zolotarev rational approximation with 16 nodes of $f(\lambda) = \lambda^{-1/2}$ is used, and applied to $A$ by means of a partial fraction expansion [44, Chapter 4]. We recall that the Zolotarev approximation requires information on the spectral interval of $A$, which was exactly provided in this example. Although competitive with respect to the Standard Krylov method, the rational function approximation does not perform as well as the Extended Krylov method. Finally, we stress that all linear systems involved in the computation were solved with direct methods after reordering of the matrix entries to improve fill-in in the factorization. This cost was taken into account for the methods under consideration.

TABLE 4.1
*Example 4.2. Approximation of $A^{-1/2}v$ where $A \in \mathbb{R}^{4900\times 4900}$ is the discretization of a self-adjoint elliptic operator in $[0,1]^2$ with zero Dirichlet boundary conditions.*

| Method | space dim. | CPU Time |
|---|---|---|
| Standard Krylov | 185 | 16.02 |
| Rational (Zolotarev) | | 0.40 |
| SI-Lanczos(0.001) | 62 | 1.00 |
| SI-Lanczos (1e-5) | 49 | 0.60 |
| SI-Lanczos ($\gamma$=2.3e-5) | 33 | 0.32 |
| Extended Krylov | 32 | 0.20 |

EXAMPLE 4.3. We consider the computation of the following product

$$\mathrm{sign}(Q)v = (Q^2)^{-1/2}(Qv)$$

which arises, for instance, in the theory of Quantum Chromodynamics (QCD) in quarks strong interaction. We consider a matrix available in the QCD collection of the Matrix Market as `conf5.0-00l4x4-2600.mtx` [38]. The complex Hermitian matrix $Q$ is indefinite and has dimension $3072 \times 3072$; $v$ is chosen as $v = e_1$. The spectrum of $Q^2$ is in the interval $[1.8 \cdot 10^{-6}, 8.1]$. We refer to [6] for more details on the application and on the specific linear algebra problems associated with the computation of the sign function. To speed up convergence, it is customary to deflate the smallest eigenvalues in modulo, which are computed beforehand together with their eigenvectors. In Figure 4.3 we report the performance of the Standard Krylov method and of the Extended method for approximating $(Q^2)^{-1/2}(Qv)$, when deflation of the first 20 eigenvalues is carried out. The deflated spectrum of $Q^2$ is contained in the interval $[3.8 \cdot 10^{-3}, 8.1]$. Performance is measured in terms of error norm versus the dimension of the approximation subspace. The final stopping tolerance was set to $10^{-6}$. The plot clearly shows the superiority of the extended method

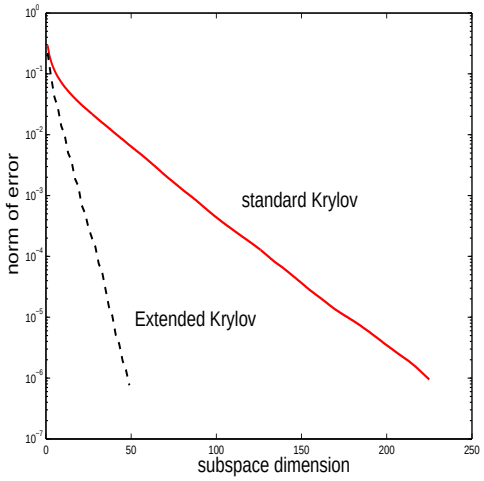


FIG. 4.3. *Example 4.3. Approximation of the sign function.*

TABLE 4.2
*Example 4.3. Approximation of the sign function. Left:Approximation of deflated problem;* $\gamma = 5.68$. *Right: Approximation without deflation;* $\gamma = 259$.

| Method Deflated $Q$ | CPU time | Space dim | Method No Deflation | CPU time | Space dim |
|---|---|---|---|---|---|
| SI-Lanczos(1) | 16.90 | 77 | SI-Lanczos(10) | 51.38 | 175 |
| SI-Lanczos(10) | 11.95 | 37 | SI-Lanczos(100) | 22.84 | 109 |
| SI-Lanczos($\gamma$) | 11.98 | 35 | SI-Lanczos($\gamma$) | 36.02 | 153 |
| EKSM | 11.19 | 56 | EKSM | 12.52 | 76 |
| Standard Krylov | 36.99 | 226 | Standard Krylov | - | 500 |

over the Standard method. A more detailed analysis is shown in Table 4.2, where the performance of SI-Lanczos for various values of the shift parameter is also reported. In there, CPU time and the dimension of the approximation space are shown, both with and without deflation of the smallest eigenvalues. We remark that the factorization of the (reordered) Hermitian positive definite matrix $Q^2$ and of its shifted version costs about 7 seconds, which is a significant portion of the computation when deflation is used. SI-Lanczos is sensitive to the choice of the shift parameter, in terms of dimension of the approximation space, reaching the best performance for the value $\gamma = 5.68$. When no deflation is carried out, so that the cost of pre-computing the eigenspace is avoided, performance degrades significantly for all methods except for EKSM, whose costs, especially in terms of CPU time, remain moderate. The theoretical analysis suggests a value of the shift parameter equal to $\gamma = 259$ for SI-Lanczos. The standard procedure is unable to reach convergence within the maximum allowed number of iterations, providing a final residual of the order of $10^{-2}$.

EXAMPLE 4.4. We next consider the approximation of $A^{1/2}v$, in which the matrix $A \in \mathbb{R}^{900\times 900}$ is the nonsymmetric matrix stemming from the centered finite difference discretization of the 2D operator $\mathcal{L}(u) = -100u_{xx} - u_{yy} + 10xu_x$ in $[0,1]^2$, with homogeneous Dirichlet boundary conditions. The matrix has real spectrum with $\lambda_{\min} \approx 9.2 \cdot 10^2$ and $\lambda_{\max} \approx 3.6 \cdot 10^5$. In Figure 4.4 we report the performance of

the extended Krylov and SI-Lanczos methods, as a function of the dimension of the approximation space. For SI-Lanczos we used various values of the shift parameter, since the optimal value, though now known ([3]), is hard to estimate in the general nonsymmetric case. The results show the much better performance of the extended method.

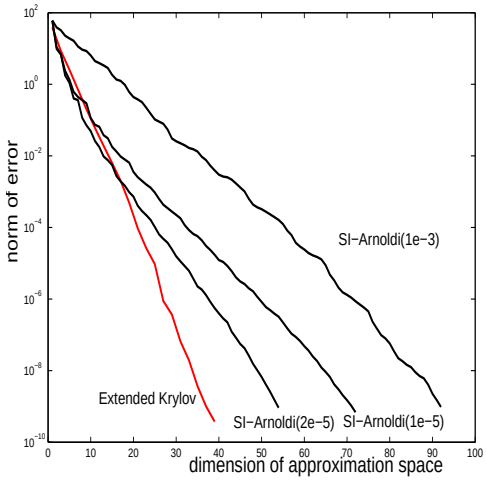


FIG. 4.4. *Example 4.4. Approximation of $A^{1/2}v$ where $A \in \mathbb{R}^{900\times 900}$ is the discretization of a non-selfadjoint elliptic operator in $[0,1]^2$ with zero Dirichlet boundary conditions.*

EXAMPLE 4.5. We report on an example where the inner systems with $A$ are solved iteratively, leading to an inner-outer method. We use GMRES with no preconditioning to solve these inner systems. Such a computation may be efficiently carried out by *relaxing* the accuracy with which GMRES solves the inner system at each outer iteration. More precisely, we use a dynamic inner stopping tolerance, that is inversely proportional to the error at the previous outer iteration as follows:

$$\epsilon_m^{(\text{inner})} = \frac{\texttt{tolin}}{\|u - u_{m-1}\|}.$$

We refer to [49] for an analysis of this relaxed criterion in the context of linear systems, and to [31] for a discussion of relaxation in the matrix function setting. To solely analyze the effect of the relaxed strategy, in the following tests we use the true error, however in practice some error estimate will substitute the error norm at the denominator. We report on two different values of `tolin`: a first value coincides with the outer tolerance, a second selection is more conservative. In the experiments we consider the nonsymmetric $1000 \times 1000$ matrix stemming from the centered finite difference discretization of the 3D operator $\mathcal{L}(u) = -u_{xx} - u_{yy} - u_{zz} + 50(x+y)u_x$ in the unit cube, with homogeneous boundary conditions. The spectrum of the matrix is complex. We use the function $f(\lambda) = \lambda^{-1/3}$. The outer tolerance was set equal to $10^{-10}$. We compare the performance of the Extended Krylov method when GMRES is used with fixed or with flexible stopping tolerance. Next table shows the final error; in parentheses are the final dimension of the outer approximation space, and the total number of inner iterations performed. Convergence curves are not reported as they visually fully overlap.

| tolin | fixed inner tol | relaxed inner tol |
|---|---|---|
| 1e-10 | 6.97e-11 (24/901) | 6.58e-11 (24/559) |
| 1e-12 | 6.48e-11 (24/1052) | 6.48e-11 (24/716) |

The digits show that the total number of inner iterations is considerably reduced with the flexible strategy, with no loss in the overall performance of the outer method. A complete theoretical justification of the flexible strategy in the computation of general matrix functions is substantially an open problem, although better understanding is available for rational functions [36] and for the exponential [31].

EXAMPLE 4.6. We conclude with a series of experiments with large-scale matrices, for which the inner systems must be solved with an iterative method; a fixed inner tolerance was used as stopping criterion. We consider the matrices stemming from the centered finite difference discretization of the following four non-selfadjoint elliptic operators

$$\mathcal{L}_1(u) = -100u_{x_1x_1} - u_{x_2x_2} + 10x_1u_{x_1}, \tag{4.2}$$

$$\mathcal{L}_2(u) = -100u_{x_1x_1} - u_{x_2x_2} - u_{x_3x_3} + 10x_1u_{x_1}, \tag{4.3}$$

$$\mathcal{L}_3(u) = -\exp(-x_1x_2)u_{x_1x_1} - \exp(x_1x_2)u_{x_2x_2} + 1/(x_1+x_2)u_{x_1}, \tag{4.4}$$

$$\mathcal{L}_4(u) = -\mathrm{div}(\exp(3x_1x_2)\mathrm{grad}u) + 1/(x_1+x_2)u_{x_1}, \tag{4.5}$$

on the unit square or cube, with Dirichlet homogeneous boundary conditions; operators $\mathcal{L}_1, \mathcal{L}_3$ and $\mathcal{L}_4$ are two-dimensional, whereas $\mathcal{L}_2$ is three-dimensional. Our results for SI-Arnoldi, EKSM and Standard Krylov methods are reported in Table 4.3 for various grid sizes (the matrix dimension $n$ is reported), where the functions $f(\lambda) = \lambda^{1/2}$ and $f(\lambda) = \lambda^{-1/3}$ are used. The table reports the CPU time to reach an accuracy of $10^{-7}$. The vector $v$ is the vector of all ones, normalized so as to have unit norm. In parentheses is the generated subspace dimension. For the sake of consistency, we used (2.7) (with $j = 2$) as stopping criterion for all methods. For Standard Krylov, a maximum approximation space of size 300 was allowed. An asterisk next to the CPU time means that no convergence was reached within the maximum subspace dimension allowed. Inner (shifted and scaled) systems in SI-Arnoldi were solved with IDR(4) ([51]) preconditioned by ILU with fill-in threshold equal to $10^{-4}$, requiring on average 4 iterations to converge. Higher thresholds were not as effective in terms of elapsed times. The same preconditioner was not efficient within EKSM to solve with $A$. In this latter case, GMRES with the Algebraic Multigrid Preconditioner HSL-MI20 was employed [7]; default values were used to build the preconditioner. On average, only about 5 GMRES iterations were necessary to reach a residual below $10^{-7}$. Our numerical experience showed that HSL-MI20 preconditioning within SI-Arnoldi was less competitive than ILU preconditioning. This justifies the use of different solvers for SI-Arnoldi and for EKSM.

We would like to stress that both forms of preconditioning are significantly cheaper than using a sparse direct solver. A full LU decomposition (Matlab function `lu`) of the matrix associated with $\mathcal{L}_3$ for $n = 160,000$ would deliver factors with about 43M nonzeros each, whereas only a few million nonzeros were necessary for the ILU factors.

The determination of an effective parameter value for SI-Arnoldi was rather difficult. For $f(\lambda) = \lambda^{1/2}$, we report in Figure 4.5 the number of SI-Arnoldi iterations as a function of the parameter value for the operator $\mathcal{L}_1$ with $n = 10\,000$ (solid line), and for the operator $\mathcal{L}_3$ with $n = 40\,000$ (dashed line). It can be readily noticed that performance deteriorates considerably for values that are close to the (computationally determined) optimal one. As a consequence, SI-Arnoldi timings shown in Table

TABLE 4.3

*Example 4.6. Approximation of $f(A)v$ for two functions and different matrix dimensions, for the operators in (4.2)-(4.5). In the SI-Arnoldi method the values $\gamma = 1.2 \cdot 10^{-5}$ and $\gamma = 1.8 \cdot 10^{-4}$, $\gamma = 3.5 \cdot 10^{-5}$, for problems $\mathcal{L}_{1,2}$, $\mathcal{L}_3$ and $\mathcal{L}_4$ respectively, were used. In parenthesis is the total space dimension. An asterisk indicates that the maximum number of iterations was reached with no convergence.*

| $f$ | Oper. | $n$ | SI-Arnoldi | EKSM | Std Krylov |
|---|---|---|---|---|---|
| $\lambda^{1/2}$ | $\mathcal{L}_1$ | 2500 | 0.9 (59) | 0.6 (48) | 7 (193) |
| | | 10000 | 4.0 (66) | 3.6 (68) | *46 (300) |
| | | 160000 | 642.9(246) | 219.7(122) | *458(300) |
| | $\mathcal{L}_2$ | 27000 | 10.8 (55) | 7.4 (40) | 6.7(119) |
| | | 125000 | 86.7 (60) | 65.3 (52) | 138.7(196) |
| | $\mathcal{L}_3$ | 40000 | 26.3 (75) | 21.1 (72) | *87 (300) |
| | | 160000 | 318.5(144) | 173.3 (96) | *442(300) |
| | $\mathcal{L}_4$ | 40000 | 41.1(117) | 25.4(106) | *89 (300) |
| | | 160000 | 580.2(442) | 231.2(144) | *461 (300) |
| $\lambda^{-1/3}$ | $\mathcal{L}_1$ | 2500 | 0.6 (43) | 0.4 (30) | 2.2(131) |
| | | 10000 | 2.6 (46) | 1.8 (38) | 26.2(252) |
| | | 160000 | 79.3 (48) | 99.7 (64) | *460(300) |
| | $\mathcal{L}_2$ | 27000 | 7.8 (41) | 4.8 (26) | 3.1 (82) |
| | | 125000 | 64.8 (45) | 38.9 (32) | 67.5(138) |
| | $\mathcal{L}_3$ | 40000 | 20.7 (61) | 13.7 (48) | *88 (300) |
| | | 160000 | 116.5 (62) | 105.2 (62) | *460 (300) |
| | $\mathcal{L}_4$ | 40000 | 35.8(104) | 14.2 (66) | *88 (300) |
| | | 160000 | 208.1(104) | 112.2 (84) | *461 (300) |

4.3 may provide an optimistic picture of the situation when the parameter cannot be determined with good accuracy.

Table 4.3 clearly shows that EKSM is competitive with respect to SI-Arnoldi, also when computing the same approximation space dimension, because it requires half the number of system solves. On the other hand, we observe that the performance of EKSM may be affected by the possible difficulty in solving systems with $A$, whereas the shift-and-scaling process in the SI-Arnoldi method may alleviate this difficulty.

**Acknowledgments**. We thank I. Moret for discussions on [39] and T. Driscoll for his help with the use of the SCToolbox [10]. We thank the referees for their comments, which lead us to the inclusion of Example 4.6, of section 2.1 and of Proposition 2.1.

REFERENCES

[1] F. O. Alpak, C. Torres-Verdín, K. Sepehrnoori, S. Fang, and L. Knizhnerman. An extended Krylov subspace method to simulate single-phase fluid flow phenomena in axisymmetric and anisotropic porous media. *J. of Petroleum Science and Engineering*, 40:121–144, 2003.

[2] B. Beckermann. Image numérique, GMRES et polynômes de Faber. *C. R. Acad. Sci. Paris, Ser. I*, 340:855–860, 2005.

[3] B. Beckermann and L. Reichel. Error Estimation and Evaluation of matrix functions via the Faber Transform, manuscript, December 2008.

[4] L. Bergamaschi and M. Vianello. Efficient computation of the exponential operator for large, sparse, symmetric matrices. *Numer. Linear Algebra Appl.*, 7(1):27–45, 2000.

[5] A. Borici. Computational methods for UV suppresses fermions. *J. Comp. Phys.*, 189:454–462,

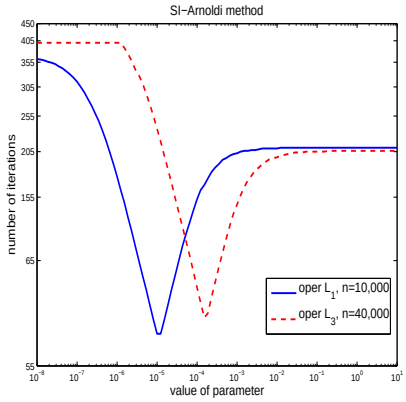


FIG. 4.5. *Example 4.6. Number of SI-Arnoldi iterations as a function of the parameter, for* $f(\lambda) = \lambda^{1/2}$, $\mathcal{L}_1$ *in (4.2) and* $n = 10\,000$, *and* $\mathcal{L}_3$ *in (4.4) with* $n = 40\,000$.

2003.

[6] 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. xii, 201 p., 2005.

[7] J. Boyle, M. D. Mihajlovic and J. A. Scott. *HSL_MI20: an efficient AMG preconditioner.* Tech. Rep. RAL-TR-2007-021, Rutherford Appleton Laboratory, Chilton, England, 2007.

[8] M. Crouzeix. Numerical range and numerical calculus in Hilbert space. *J. Functional Analysis*, 244:668–690, 2007.

[9] F. Diele, I. Moret, and S. Ragni. Error estimates for polynomial Krylov approximations to matrix functions. *SIAM J. Matrix Analysis and Appl.*, 30(4):1546-1565, 2008.

[10] T. Driscoll. Algorithm 756: A MATLAB Toolbox for Schwarz–Christoffel Mapping. *ACM Transactions on Mathematical Software*, 22(2):168–186, 1996.

[11] V.L. Druskin. On monotonicity of the Lanczos approximation to the matrix exponential. *Linear Algebra and its Applications*, 429(7):1679-1683, 2008.

[12] V. Druskin, A. Greenbaum, and L. Knizhnerman. Using nonorthogonal Lanczos vectors in the computation of matrix functions. *SIAM J. Sci. Comput.*, 19(1):38–54, 1998.

[13] V. Druskin and L. Knizhnerman. Two polynomial methods of calculating functions of symmetric matrices. *U.S.S.R. Comput. Math. Math. Phys.*, 29:112–121, 1989.

[14] V. Druskin and L. Knizhnerman. Spectral approach to solving three-dimensional Maxwell's diffusion equations in the time and frequency domains. *Radio Science*, v. 29(4):937-953, 1994.

[15] V. Druskin and L. Knizhnerman. Krylov subspace approximation of eigenpairs and matrix functions in exact and computer arithmetic. *Numerical Linear Algebra with Appl.*, 2(3):205–217, 1995.

[16] V. Druskin and L. Knizhnerman. Extended Krylov subspaces: approximation of the matrix square root and related functions. *SIAM J. Matrix Anal. Appl.*, 19(3):755–771, 1998.

[17] V. L. Druskin, L. A. Knizhnerman, and P. Lee. New spectral Lanczos decomposition method for induction modeling in arbitrary 3-D geometry. *Geophysics*, 64(3):701–706, 1999.

[18] M. Eiermann and O. Ernst. A restarted Krylov subspace method for the evaluation of matrix functions. *SIAM J. Numer. Anal.*, 44(6):2481–2504, 2006.

[19] V. Faber, A. Greenbaum, and D. E. Marshall. The polynomial numerical hulls of Jordan blocks and related matrices. *Lin. Alg. Appl.*, 374:231–246, 2003.

[20] R.A. Friesner, L.S. Tuckerman, B.C. Dornblaser, and T.V. Russo. A method for exponential propagation of large systems of stiff nonlinear differential equations. *J. Sci. Comput.*, 4:327–354, 1989.

[21] A. Frommer. Monotone Convergence of the Lanczos Approximations to Matrix Functions of Hermitian Matrices. Technical report, Fachbereich Universität, Wuppertal - D, 2008.

[22] A. Frommer and V. Simoncini. Stopping criteria for rational matrix functions of Hermitian and symmetric matrices. *SIAM J. Sci. Comput.*, 30(3):1387–1412, 2008.

[23] E. Gallopoulos and Y. Saad. A parallel block cyclic reduction algorithm for the fast solution

of elliptic equations. *Parallel Computing*, 10:143–159, 1989.
[24] E. Gallopoulos and Y. Saad. Efficient solution of parabolic equations by Krylov approximation methods. *SIAM J. Sci. Stat. Comput.*, 13(5):1236–1264, 1992.
[25] K. Gustafson and D. K. M. Rao. *Numerical Range: The Field of Values of Linear Operators and Matrices*. Universitext. Springer, New York, NY, USA, 1977.
[26] N. J. Higham. The Matrix Computation Toolbox. `http://www.ma.man.ac.uk/~higham/mctoolbox`.
[27] N. J. Higham. *Matrix Functions – Theory and Applications*. SIAM, Philadelphia, USA, 2008.
[28] M. Hochbruck and C. Lubich. Exponential integrators for quantum-classical molecular dynamics. *BIT*, 39(1):620–645, 1999.
[29] M. Hochbruck, C. Lubich, and H. Selhofer. Exponential integrators for large systems of differential equations. *SIAM J. Sci. Comput.*, 19(5):1552–1574, 1998.
[30] M. Hochbruck and A. Ostermann. Exponential Runge-Kutta methods for parabolic problems. *Applied Numer. Math.*, 53:323–339, 2005.
[31] J. van den Eshof and M. Hochbruck. Preconditioning Lanczos approximations to the matrix exponential. *SIAM J. Sci. Comput.*, 27(4):1438–1457, 2006.
[32] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, Cambridge, 1991.
[33] M. Ilic, D. P. Simpson, and I. W. Turner. A Restarted Lanczos Approximation to Functions of a Symmetric Matrix. Technical Report 8011, Queensland University of Technology, Australia, 2007.
[34] L. Knizhnerman. Calculation of functions of unsymmetric matrices using Arnoldi's method. *U.S.S.R. Comput. Math. Math. Phys.*, 31:1–9, 1991.
[35] L. Knizhnerman. Adaptation of the Lanczos and Arnoldi methods to the spectrum, or why the two Krylov subspace methods are powerful. *Chebyshev Digest (in Russian: Chebyshyovsky Sbornik)*, 3(2):141-164, 2002. Also available as: http://www.tspu.tula.ru/res/math/c_sbor/tom3/v2/knizhnerman.rar
[36] L. Lopez and V. Simoncini. Analysis of projection methods for rational function approximation to the matrix exponential. *SIAM J. Numer. Anal.*, 44(2):613–635, 2006.
[37] The MathWorks, Inc. *MATLAB 7*, September 2004.
[38] Matrix Market. A visual repository of test data for use in comparative studies of algorithms for numerical linear algebra, Mathematical and Computational Sciences Division, National Institute of Standards and Technology. Online at `http://math.nist.gov/MatrixMarket` .
[39] I. Moret. Rational Lanczos approximations to the matrix square root and related functions. Technical report, Dipartimento di Matematica e Informatica, Università degli Studi di Trieste, Trieste, Italy, 2005. To appear in *Numerical Linear Algebra with Appl.*
[40] I. Moret and P. Novati. An interpolatory approximation of the matrix exponential based on Faber polynomials. *Journal of Computational and Applied Mathematics*, 131:361–380, 2001.
[41] I. Moret and P. Novati. RD-rational approximations of the matrix exponential. *BIT*, 44(3):595–615, 2004.
[42] A. Nauts and R.E. Wyatt. New approach to many state quantum dynamics: The recursive residue generation method. *Phys. Rev. Lett.*, 51:2238–2241, 1983.
[43] P. Nettesheim and Ch. Schütte. Numerical integrators for Quantum-Classical Molecular Dynamics. In P. Deuflhard, J. Hermans, B. Leimkuhler, A. E. Marks, S. Reich, and R. D. Skeel, editors, *Computational Molecular Dynamics: Challenges, Methods, Ideas*, volume 4 of *Lecture Notes in Computational Science and Engineering*, pages 396–411. Springer Verlag, 1999.
[44] P. P. Petrushev and V. A. Popov. *Rational approximation of real functions*. Cambridge University Press, Cambridge, 1987.
[45] M. Popolizio. *Acceleration techniques for approximating the matrix exponential*. PhD thesis, Università degli Studi di Bari, March. 2008.
[46] M. Popolizio and V. Simoncini. Acceleration techniques for approximating the matrix exponential operator. *SIAM J. Matrix Anal. Appl.*, 30(2):657–683, 2008.
[47] I.V. Puzynin, A.V. Selin, and S.I. Vinitsky. Magnus-factorized method for numerical solving the time-dependent Schrödinger equation. *Comput. Phys. Commun.*, 126(1-2):158–161, 2000.
[48] Y. Saad. Analysis of some Krylov subspace approximations to the matrix exponential operator. *SIAM J. Numer. Anal.*, 29:209–228, 1992.
[49] V. Simoncini and D. B. Szyld. Theory of inexact Krylov subspace methods and applications to scientific computing. *SIAM J. Sci. Comput.*, 25(2):454–477, 2003.
[50] V. Simoncini. A new iterative method for solving large-scale Lyapunov matrix equations. *SIAM*

*J. Sci. Comput.*, 29(3):1268–1288, 2007.

[51] P. Sonneveld and M. B. van Gijzen. IDR(s): a family of simple and fast algorithms for solving large nonsymmetric linear systems. *SIAM J. Sci. Comput.*, 31(2):1035–1062, 2008.

[52] P. K. Suetin. *Series of Faber Polynomials*. Analytical Methods and Special Functions. Gordon and Breach Science, Amsterdam, 1998. Originally published in Russian in 1984 as Riady po mnogochlennam Fabera by Nauka, Moscow.

[53] H. Tal-Ezer. Spectral methods in time for parabolic problems. *SIAM J. Numer. Anal.*, 26:1–11, 1989.

[54] L. N. Trefethen, J. A. C. Weideman, and T. Schmelzer. Talbot quadratures and rational approximations. *BIT*, 46(3):653–670, 2006.

[55] J. van den Eshof, A. Frommer, Th. 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(2):203–224, 2002.