MATRIX FUNCTIONS

ANDREAS FROMMER* AND VALERIA SIMONCINI†

1. Introduction. In this chapter, we give an overview on methods to compute functions of a (usually square) matrix A with particular emphasis on the matrix exponential and the matrix sign function. We will distinguish between methods which indeed compute the entire matrix function, i.e. they compute a matrix, and those which compute the action of the matrix function on a vector. The latter task is particularly important in the case where we have to deal with a very large (and possibly sparse) matrix A or in situations, where A is not available as a matrix but just as a function which returns Ax for any input vector x. Computing the action of a matrix function on a vector is a typical model reduction problem, since the resulting techniques usually rely on approximations from small-dimensional subspaces.

This chapter is organized as follows: In section 2 we introduce the concept of a matrix function $f(A)$ in detail, essentially following [37] and [27]. Section 3 gives an assessment of various general computational approaches for either obtaining the whole matrix $f(A)$ or its action $f(A)v$ on a vector v. Sections 4 and 5 then give much more details for two specific functions, the exponential and the sign functions, which, as we will show, are particularly important in many areas like control theory, simulation of physical systems and other application fields involving the solution of certain ordinary or partial differential equations. The applicability of matrix functions in general, and of the exponential and the sign functions in particular, is vast. However, we will limit our discussion to characterizations and to application problems that are mostly related to Model Order Reduction.

2. Matrix Functions. In this section we address the following general question: Given a function $f : \mathbb{C} \to \mathbb{C}$, is there a canonical way to extend this function to square matrices, i.e. to extend f to a mapping from $\mathbb{C}^{n\times n}$ to $\mathbb{C}^{n\times n}$? If f is a polynomial p of degree d, $f(z) = p(z) = \sum_{k=0}^{d} a_k z^k$, the canonical extension is certainly given by

$$p : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}, \quad p(A) = \sum_{k=0}^{d} a_k A^k.$$

If $f(z)$ can be expressed by a power series, $f(z) = \sum_{k=0}^{\infty} a_k z^k$, a natural next step is to put

$$f(A) = \sum_{k=0}^{\infty} a_k A^k, \tag{2.1}$$

but for (2.1) to make sense we must now discuss convergence issues. The main result is given in the following theorem, the proof of which gives us valuable further information on matrix functions. Recall that the spectrum $\mathrm{spec}(A)$ is the set of all eigenvalues of A.

*Fachbereich Mathematik und Naturwissenschaften, Universität Wuppertal, D-42097 Wuppertal, frommer@math.uni-wuppertal.de

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

THEOREM 2.1. *Assume that the power series* $f(z) = \sum_{k=0}^{\infty} a_k z^k$ *is convergent for* $|z| < \rho$ *with* $\rho > 0$ *and assume that* $\operatorname{spec}(A) \subset \{z \in \mathbb{C} : |z| < \rho\}$. *Then the series (2.1) converges.*

Proof. Let T be the transformation matrix occuring in the Jordan decomposition

$$A = TJT^{-1}, \tag{2.2}$$

with

$$J = \begin{bmatrix} J_{m_1}(\lambda_1) & & 0 \\ & \ddots & \\ 0 & & J_{m_\ell}(\lambda_\ell) \end{bmatrix} =: \operatorname{diag}(\, J_{m_1}(\lambda_1), \ldots, J_{m_\ell}(\lambda_\ell)\,).$$

Here, $\lambda_1, \ldots, \lambda_\ell$ are the (not necessarily distinct) eigenvalues of A and m_j is the size of the jth Jordan block associated with λ_j, i.e.

$$J_{m_j}(\lambda_j) = \begin{pmatrix} \lambda_j & 1 & 0 & \cdots & 0 \\ 0 & \lambda_j & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & 1 \\ 0 & \cdots & \cdots & 0 & \lambda_j \end{pmatrix} =: \lambda_j I + S_{m_j} \in \mathbb{C}^{m_j \times m_j},$$

and $\sum_{j=1}^{\ell} m_j = n$. For each λ_j, the powers of $J_{m_j}(\lambda_j)$ are given by

$$J_m(\lambda_j)^k = \sum_{\nu=0}^{k} \binom{k}{\nu} \lambda_j^{k-\nu} \cdot S_{m_j}^{\nu}.$$

Note that $S_{m_j}^{\nu}$ has zero entries everywhere except for the ν-th upper diagonal, whose entries are equal to 1. In particular, $S_{m_j}^{\nu} = 0$ for $\nu \geq m_j$. Therefore,

$$f(J_{m_j}(\lambda_j)) = \sum_{k=0}^{\infty} a_k \sum_{\nu=0}^{k} \binom{k}{\nu} \lambda_j^{k-\nu} \cdot S_{m_j}^{\nu},$$

and for ν and j fixed we have

$$\sum_{k=0}^{\infty} a_k \binom{k}{\nu} \lambda_j^{k-\nu} = \sum_{k=0}^{\infty} \frac{1}{\nu!} \cdot a_k \cdot (k \cdot \ldots \cdot (k-\nu+1)) \lambda_j^{k-\nu} = \frac{1}{\nu!} f^{(\nu)}(\lambda_j).$$

Note that the last equality holds in the sense of absolute convergence because λ_j lies within the convergence disk of the series. This shows that the series $f(J_{m_j}(\lambda_j))$ converges. Plugging these expressions into the series from (2.1) we obtain the value of the original (now convergent) series,

$$\begin{aligned} f(A) &= T \operatorname{diag}\left(f(J_{m_1}(\lambda_1)), \ldots, f(J_{m_\ell}(\lambda_\ell))\right) T^{-1} \\ &= T \operatorname{diag}\left(\sum_{\nu=0}^{m_1-1} \frac{1}{\nu!} f^{(\nu)}(\lambda_1) \cdot S_{m_1}^{\nu}, \ldots, \sum_{\nu=0}^{m_\ell-1} \frac{1}{\nu!} f^{(\nu)}(\lambda_\ell) \cdot S_{m_\ell}^{\nu} \right) T^{-1}. \end{aligned} \tag{2.3}$$

□

It may happen that a function f cannot be expressed by a series converging in a large enough disk. If f is sufficiently often differentiable at the eigenvalues of A, then the right-hand side of (2.3) is still defined. We make it the basis of our final definition of a matrix function.

DEFINITION 2.2. *Let $A \in \mathbb{C}^{n\times n}$ be a matrix with* $\operatorname{spec}(A) = \{\lambda_1, \ldots, \lambda_\ell\}$ *and Jordan normal form*

$$J = T^{-1}AT = \operatorname{diag}(\, J_{m_1}(\lambda_1), \ldots, J_{m_\ell}(\lambda_\ell)\,).$$

Assume that the function $f : \mathbb{C} \to \mathbb{C}$ is $m_j - 1$ times differentiable at λ_j for $j = 1, \ldots, \ell$. Then the matrix function $f(A)$ is defined as $f(A) = Tf(J)T^{-1}$ where

$$f(J) = \operatorname{diag}(f(J_{m_1}(\lambda_1)), \ldots, f(J_{m_\ell}(\lambda_\ell))) \quad \text{with } f(J_{m_j}(\lambda_j)) = \sum_{\nu=0}^{m_j-1} \frac{1}{\nu!} f^{(\nu)}(\lambda_j) \cdot S_{m_j}^{\nu}.$$

This definition makes explicit use of the Jordan canonical form and of the associated transformation matrix T. Neither T nor J are unique, but it can be shown – as is already motivated by (2.1) – that $f(A)$ as introduced in Definition 2.2 does not depend on the particular choice of T or J.

As a first consequence of Definition 2.2 we note the following important property.

PROPOSITION 2.3. *With the notation above, it holds $f(A) = p(A)$, where p is the polynomial of degree not greater than $n-1$ which interpolates the eigenvalues λ_j of A in the Hermite sense (i.e. $f^{(\nu)}(\lambda_j) = p^{(\nu)}(\lambda_j)$ for all relevant ν's and j's).*

The polynomial p in Proposition 2.3 will not only depend on f, but also on A or, more precisely, on the minimal polynomial of A (of which the multiplicity of an eigenvalue λ determines the maximal block size m_j for the Jordan blocks corresponding to this eigenvalue). When A is normal, T is an orthogonal matrix and all Jordan blocks have size one, i.e. we have

$$J = \operatorname{diag}(\lambda_1, \ldots, \lambda_n).$$

So, in this particular case, we do not need any differentiability assumption on f.

A further representation of $f(A)$ can be derived in the case when f is analytic in a simply connected region Ω containing $\operatorname{spec}(A)$. Let γ be a curve in Ω with winding number $+1$ w.r.t. a point $z \in \Omega$. The Residue Theorem tells us

$$\frac{f^{(\nu)}(z)}{\nu!} = \frac{1}{2\pi i} \oint_\gamma \frac{f(t)}{(t-z)^{\nu+1}} dt. \tag{2.4}$$

Let $J_{m_j}(\lambda_j)$ be a Jordan block associated with λ_j and let $z \neq \lambda_j$. Then

$$(zI - J_{m_j})^{-1} = ((z-\lambda_j)I - S_{m_j})^{-1} = \frac{1}{z-\lambda_j} \cdot \sum_{\nu=0}^{m_j-1} \left(\frac{1}{z-\lambda_j} \cdot S_{m_j} \right)^{\nu},$$

from which we get

$$\begin{aligned} \frac{1}{2\pi i} \oint_\gamma f(z)(zI - J_{m_j})^{-1} &= \sum_{\nu=0}^{m_j-1} \frac{1}{2\pi i} \oint_\gamma \frac{f(z)}{(z-\lambda_j)^{\nu+1}} S_{m_j}^{\nu} dt \\ &= \sum_{\nu=0}^{m_j-1} \frac{f^{(\nu)}(\lambda_j)}{\nu!} \cdot S_{m_j}^{\nu}, \end{aligned}$$

the second line holding due to (2.4). Using this for each Jordan block in Definition 2.2 and recombining terms we obtain the following integral representation of $f(A)$,

$$f(A) = \frac{1}{2\pi i} \oint_\gamma f(t)(tI - A)^{-1} dt. \tag{2.5}$$

3. Computational aspects. It is not necessarily a good idea to stick to one of the definitions of matrix function given in the previous section when it comes to numerically compute a matrix function $f(A)$. In this section we will discuss such computational issues, describing several numerical approaches having their advantages in different situations, basically depending on spectral properties of A, on the dimension and sparsity of A and on whether we really want to obtain the matrix $f(A)$ rather than "just" its action $f(A)v$ on a vector v.

3.1. Normal matrices. A matrix $A \in \mathbb{C}^{n\times n}$ is said to be *normal* if it commutes with its adjoint, $AA^H = A^H A$. Normal matrices may also be characterized as being unitarily diagonalizable, i.e. we have the representation

$$A = Q\Lambda Q^H \quad \text{with } Q^{-1} = Q^H,\ \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n),\ \operatorname{spec}(A) = \{\lambda_1, \ldots, \lambda_n\}.$$

This representation is also the Jordan decomposition of A from (2.2), so that

$$f(A) = Qf(\Lambda)Q^H, \qquad f(\Lambda) = \operatorname{diag}(f(\lambda_1), \ldots, f(\lambda_n)). \tag{3.1}$$

Normal matrices have the very attractive property that their eigenvalues λ_i and the corresponding invariant subspaces are well conditioned (see [16], for example), i.e. small changes in A yield only small changes in Λ and Q. Therefore, if we use a numerically (backward) stable algorithm to compute Λ and Q, like, for example, the standard Householder reduction to upper Hessenberg form followed by the QR-iteration, we may safely use the so computed Λ and Q to finally compute $f(A)$ via (3.1). The computational cost of this approach is $\mathcal{O}(n^3)$ due to the various matrix-matrix multiplications and to the cost for computing the eigendecomposition.

If A is not normal, its eigenvalues are not necessarily well conditioned, the condition number being related to $\|T\|_2 \cdot \|T^{-1}\|_2$ with T from the Jordan decomposition (2.2). It is also important to realize that the size of the Jordan blocks may widely vary under infinitesimal perturbations in A. Therefore, if A is not normal, Definition 2.2 does not provide a numerically stable means for computing $f(A)$.

3.2. Quadrature rules. Assume that f is analytic in Ω and that γ and Ω are as in (2.5) so that we have

$$f(A) = \frac{1}{2\pi i} \oint_\gamma f(t)(tI - A)^{-1} dt.$$

We apply a quadrature rule with m nodes $t_j \in \gamma$ and weights ω_j to the right-hand side to get

$$\frac{1}{2\pi i} \oint_\gamma \frac{f(t)}{t - z} dt = \sum_{j=1}^{m} \omega_j \frac{f(t_j)}{t_j - z} + r.$$

This shows that we can approximate

$$f(A) \approx \sum_{j=1}^{m} \omega_j f(t_j) \cdot (t_j I - A)^{-1}. \tag{3.2}$$

For such quadrature rules, the approximation error r can be expressed or bounded using higher derivatives of f. Actually, since we integrate over a closed curve, taking the right nodes the quadrature error is usually much smaller than what one would expect from quadrature formulas over finite (real) intervals, and the accuracy often increases exponentially with the number of nodes, see [14],[15]. In principle, this can then be used to obtain bounds on the approximation error in (3.2), but to do so we usually need some knowledge about the norms of T and T^{-1} in (2.2), as well as on the size of the eigenvalues of A. See also section 3.6.

For specific functions, other integral representations may be used. For example, for $z \in \mathbb{C}$, z not on the non-positive real line, we have (see [14])

$$\log(z) = \int_0^1 (z-1)[t(z-1)+1]^{-1} dt,$$

so that using a quadrature rule for the interval $[0,1]$, we can use the approximation

$$\log(A) \approx \sum_{j=1}^{m} \omega_j \cdot (A-I)[t_j(A-I)+I]^{-1}.$$

As another example, for $z > 0$ we can write

$$z^{-1/2} = \frac{2}{\pi} \cdot \int_0^\infty \frac{1}{t^2+z}\, dt,$$

and use a quadrature rule on $[0,\infty]$ to approximate $A^{-1/2}$ when $\mathrm{spec}(A) \subset (0,\infty]$.

Similar approaches have been proposed for various other functions like the p-th root or the sign function, see [6], [57], for example.

Within this quadrature framework, the major computational cost will usually be due to the inversion of several matrices. As is explained in [14], this cost can often be reduced if we first compute a unitary reduction to upper Hessenberg form (which can be done in a numerically stable manner using Householder transformations), i.e.

$$A = QHQ^H, \;\; Q \text{ unitary }, \; H \text{ zero below the first subdiagonal.}$$

Then, for example,

$$(t_j I - A)^{-1} = Q \cdot (t_j I - H)^{-1} \cdot Q^H \text{ for all } j,$$

with the inversion of the matrix $t_j I - H$ having cost $\mathcal{O}(n^2)$ rather than $\mathcal{O}(n^3)$.

3.3. Matrix iterations. Sometimes, it is convenient to regard $f(z)$ as the solution of a fixed point equation $g_z(f) = f$ with g_z being contractive in a neighbourhood of the fixed point $f(z)$. The method of successive approximations

$$f_{k+1} = g_z(f_k) \tag{3.3}$$

can then be turned into a corresponding matrix iteration

$$F_{k+1} = g_A(F_k). \tag{3.4}$$

Approaches of this kind have, for example, been proposed for the matrix square root [31], [32], where Newton's method

$$f_{k+1} = \frac{1}{2} \cdot \left(f_k + \frac{z}{f_k} \right) \tag{3.5}$$

to compute $\sqrt{z}$ results in the iteration

$$F_{k+1} = \frac{1}{2} \cdot \left(F_k + A \cdot F_k^{-1} \right). \tag{3.6}$$

Similar other iterations, not always necessarily derived from Newton's method, have been proposed for the matrix p-th root [6] or for the matrix sign function [40]. A major catch with these approaches is that numerical stability of the matrix iteration (3.4) is not always guaranteed, even when the scalar iteration (3.3) is perfectly stable. Then, some quite subtle modifications, like e.g. the coupled two-term iteration for the square root analyzed in [31] must be used in order to achieve numerical stability. The iteration (3.4) is usually also quite costly. For example, (3.6) requires the inversion of F_k at every step, so that each step has complexity $\mathcal{O}(n^3)$. Therefore, for these methods to be efficient, convergence should be fast, at least superlinear.

3.4. Rational approximations. Polynomial approximations for a function f often require a quite high degree of the approximating polynomial in order to achieve a reasonable quality of approximation. *Rational* approximations typically obtain the same quality with substantially fewer degrees of freedom.

Assume that we have the rational approximation

$$f(z) \approx \frac{\mathcal{N}_{\mu\nu}(z)}{\mathcal{D}_{\mu\nu}(z)},$$

where $\mathcal{N}_{\mu\nu}, \mathcal{D}_{\mu\nu}$ are polynomials of degree μ and ν, respectively. (The use of the two indices μ and ν in both polynomials may appear abusive at this point, but it will be very convenient when discussing Padé approximations to the exponential in section 4.2). Then

$$f(A) \approx \mathcal{N}_{\mu\nu}(A) \cdot \left(\mathcal{D}_{\mu\nu}(A) \right)^{-1}.$$

Assume that A is diagonalizable. If we know

$$\left| f(z) - \frac{\mathcal{N}_{\mu\nu}(z)}{\mathcal{D}_{\mu\nu}(z)} \right| \leq \epsilon \text{ for } z \in \operatorname{spec}(A),$$

we get

$$\| f(A) - \mathcal{N}_{\mu\nu}(A) \cdot \left(\mathcal{D}_{\mu\nu}(A) \right)^{-1} \|_2 \leq \epsilon \cdot \|T\|_2 \cdot \|T^{-1}\|_2$$

which further simplifies when A is normal, since then T is unitary so that $\|T\|_2 \cdot \|T^{-1}\|_2 = 1$. Rational functions can be expressed as partial fraction expansions. Simplifying our discussion to the case of single poles, this means that we can expand

$$\frac{\mathcal{N}_{\mu\nu}(z)}{\mathcal{D}_{\mu\nu}(z)} = p(z) + \sum_{j=1}^{\nu} \frac{\omega_j}{z - \tau_j},$$

with $p(z)$ being a polynomial of degree $\mu - \nu$ if $\mu \geq \nu$ and $p \equiv 0$ if $\mu < \nu$. This representation is particularly useful if we are interested only in $f(A)v$ for some vector v, as we will discuss later in section 3.6. Note also that the quadrature rules from (3.2) immediately give a partial fraction expansion, so that the two approaches are very closely related. For a recent investigation, see [65].

3.5. Krylov subspace approaches. When A has large dimension, the action of $f(A)$ on a vector v, namely $f(A)v$, may be effectively approximated by projecting the problem onto a subspace of possibly much smaller dimension. The Krylov subspace

$$K_k(A, v) = \text{span}\{v, Av, \ldots, A^{k-1}v\}$$

has been extensively used to this purpose, due to its favourable computational and approximation properties, see, e.g., van der Vorst [67], [68] for a discussion for general f. Let V_k be a full column rank $n \times k$ matrix whose columns span $K_k(A, v)$, and assume the following Arnoldi type recurrence holds for V_k,

$$AV_k = V_{k+1}H_{k+1,k} = V_kH_k + h_{k+1,k}v_{k+1}e_k^T. \tag{3.7}$$

An approximation to $x = f(A)v$ may be obtained as

$$x_k = V_k f(H_k)e_1\|v\|. \tag{3.8}$$

The procedure amounts to projecting the matrix onto the much smaller subspace $K_k(A, v)$, by means of the representation matrix H_k and $v = V_k e_1\|v\|$. If V_k has orthonormal columns then $H_k = V_k^T AV_k$. If in addition A is Hermitian, the iteration (3.7) reduces to the Lanczos three-term recurrence, in which case H_k is tridiagonal and Hermitian.

The functional evaluation is carried out within this reduced space, and the obtained solution is expanded back to the original large space. Assume now that $k = n$ iterations can be carried out, so that the square matrix V_n is orthogonal. Then (3.7) gives $AV_n = V_nH_n$ and thus $A = V_nH_nV_n^T$. Using this relation, for $k < n$, the approximation in $K_k(A, v)$ may be viewed as a problem order reduction to the first k columns of V_n and corresponding portion of H_n as

$$x = f(A)v = V_n f(H_n)V_n^T v \approx V_k f(H_k)V_k^T v.$$

For k small compared to n, the quality of the approximation strongly depends on the spectral properties of A and on the capability of $K_k(A, v)$ to capture them. A first characterization in this sense is given by the following result, which can be deduced from Proposition 2.3 applied to the matrix H_k and the fact that $p(A)v = V_k p(H_k)v$ for all polyomials of degree less than or equal to $k-1$; see [60, Proposition 6.3]. This is a generalization of [59, Theorem 3.3].

PROPOSITION 3.1. *Let the columns of V_k, with $V_k^T V_k = I_k$ span $K_k(A, v)$ and let $H_k = V_k^T AV_k$. Then, the approximation $V_k f(H_k)e_1\|v\|$ represents a polynomial approximation $p(A)v$ to $f(A)v$, in which the polynomial p of degree $k-1$ interpolates the function f in the Hermite sense on the set of eigenvalues of H_k.*

Other polynomial approximations have been explored, see, e.g., [18]; approaches that interpolate over different sets have been proposed for the exponential function [52]. Note that the projection nature of the approach allows to derive estimates for $\|f(A)\|$ as $\|f(A)\| \approx \|f(H_k)\|$ which may be accurate even for small k when A is Hermitian.

All these results assume exact precision arithmetic. We refer to [17] for an analysis of finite precision computation of matrix functions with Krylov subspace methods when A is Hermitian.

It should be mentioned that the projection onto a Krylov subspace does not require A to be stored explicitly, but it only necessitates a function that given v,

returns the action of A, namely $y = Av$. This operational feature is of paramount importance in applications where, for instance, A is the (dense) product or other combination of sparse matrices, so that the operation $y = Av$ may be carried out by a careful application of the given matrix combination.

Another practical aspect concerns the situation where k, the dimension of the Krylov subspace, becomes large. Computing $f(H_k)$ with one of the methods presented in the previous sections can then become non-negligible. Moreover, we may run into memory problems, since approximating $f(A)v$ via (3.8) requires the whole matrix V_k to be stored. This is needed even when, for istance, A is Hermitian, in which case (3.7) is the Lanczos recurrence and H_k is tridiagonal. In such a situation, however, we can resort to a "two–pass" procedure which crucially reduces the amount of memory needed: In the first pass, we run the short-term recurrence Lanczos process. Here, older columns from V_k can be discarded, yet the whole (tridiagonal) matrix H_k can be built column by column. Once $f(H_k)$ has been generated, we compute $y_k = f(H_k)e_1 \cdot \|v\|$. Then we run the short-term recurrence modification of the Lanczos process once again to recompute the columns of V_k and use them one at a time to sum up $V_k f(H_k)e_1 = V_k y_k$. Of course, this two-stage approach essentially doubles the computational work.

For a general matrix A the Arnoldi process cannot be turned into a short-term recurrence, so one must search for alternatives in the case that k gets too large. Recently, Eiermann and Ernst [20] have developed an interesting scheme that allows to *restart* Krylov subspace methods for computing $f(A)v$, in the same flavour as with linear system solvers; in fact, the two approaches are tightly related; see [46]. Having computed a not yet sufficiently good approximation x_k via (3.8), the idea is to start again a Krylov subspace approximation based on the error $x_k - f(A)v$ which is expressed as a *new* matrix function of A. The algorithmic formulation is non-trivial, particularly since special care has to be taken with regard to numerical stability, see [20].

Other alternatives include acceleration procedures, that aim at improving the convergence rate of the approximation as the Krylov subspace dimension increases. Promising approaches have been recently proposed in the Hermitian case by Druskin and Knizhnerman [19], by Moret and Novati [51] and by Hochbruck and van den Eshof [36].

3.6. Krylov subspaces and rational approximations. As a last contribution to this section, let us turn back to rational approximations for f which we assume to be given in the form of a partial fraction expansion (no multiple poles for simplicity)

$$f(z) \approx p(z) + \sum_{j=1}^{\nu} \frac{\omega_j}{z - \tau_j}.$$

Then $f(A)v$ can be approximated as

$$f(A)v \approx p(A)v + \sum_{j=1}^{\nu} \omega_j (A - \tau_j I)^{-1} v. \tag{3.9}$$

Since evaluating $p(A)v$ is straightforward, let us assume $p \equiv 0$ in the sequel.

The computation of $(A - \tau_j I)^{-1}v$ means that we have to solve a linear system for each j, where all linear systems have the same right-hand side, while the coefficient matrix only differs for the shift. In general, shifts may be complex even for real and

symmetric A, although they appear in conjugate pairs. Interestingly, the particular "shifted" structure of these systems can be exploited in practical computation. If we solve each system iteratively using a Krylov subspace method with initial zero guess for all j, the kth iterate for each system lies in $K_k(A - \tau_j I, v)$ which is identical to $K_k(A, v)$. The fact that Krylov subspaces are invariant with respect to shifts can now be exploited in various Krylov subspace solvers like CG, BiCG, FOM and QMR (and also with modifications in BiCGStab and restarted GMRES) to yield very efficient procedures which require only *one* matrix-vector multiplication with A, and possibly with A^T, in order to update the iterates for *all* m systems simultaneously; see [62] for a survey of these methods for shifted systems and also [21], [22], [23], [24]. Denote by $x_k^{(j)}$ the iterate of the Krylov solver at step k for system j. Then the linear combination

$$x_k = \sum_{j=1}^{\nu} \omega_j x_k^{(j)} \in K_k(A, v) \tag{3.10}$$

is an approximation to $f(A)v$. In fact, it is an approximation to the action of the rational function approximating $f(A)$. Therefore, what we obtained in (3.10) is an approximation to $f(A)v$ in $K_k(A, v)$, which is different from (3.8) presented before. A special case is when f is itself a rational function. In such a situation, the two approaches may coincide if, for instance, a Galerkin method is used to obtain the approximate solutions $x_k^{(j)}$. Indeed, for $f = \mathcal{R}_{\mu\nu} = \mathcal{N}_{\mu\nu}/\mathcal{D}_{\mu\nu}$,

$$\begin{aligned} f(A)v = \mathcal{N}_{\mu\nu}(A)(\mathcal{D}_{\mu\nu}(A))^{-1}v &= \sum_{j=1}^{\nu} \omega_j (A - \tau_j I)^{-1} v \qquad (3.11) \\ &\approx \sum_{j=1}^{\nu} \omega_j V_k (H_k - \tau_j I)^{-1} e_1 \|v\| = V_k f(H_k) e_1 \|v\|. \end{aligned}$$

The approach outlined above has several attractive features for a general function f. Firstly, if we have a bound for the error between $x_k^{(j)}$ and the solution $(A - \tau_j)^{-1}v$ for each j, we can combine these bounds with the approximation error of the rational approximation to get an overall a posteriori bound for $\|f(A)v - x^{(k)}\|$. Sometimes, such bounds might be obtained quite easily. For example, if A is Hermitian and positive definite and all shifts τ_j are real and negative, the norm of the inverse $(A - \tau_j I)^{-1}$ is bounded by $1/|\tau_j|$. Since the residuals $r_k^{(j)} = (A - \tau_j I)x_k^{(j)} - v$ are usually available in the Krylov solver in use, we can use the bound

$$\|x_k^{(j)} - (A - \tau_j I)^{-1} v\|_2 \leq \frac{1}{|\tau_j|} \|r_k^{(j)}\|_2.$$

Similar bounds that require estimates of the spectrum of A may be obtained also for complex poles τ_j, see [46].

Secondly, in the Hermitian case, the memory requirements of this approach only depend on m, the number of poles in the rational approximation, but not on k, the dimension of the Krylov subspace. Indeed, the symmetry of the problem can be exploited to devise a short-term recurrence which dynamically updates the solution x_k without storing the whole Krylov subspace basis. So even if k has to be sensibly large in order to get a good approximation, we will not run into memory problems. This is in contrast to the approach from section 3.5, although the two approaches are

strictly related. Indeed, using x_k in (3.10), by the triangle inequality we have

$$\begin{aligned} |\,\|f(A)v - x_k\| - \|f(A)v - V_k f(H_k)e_1\|v\|\,\|\,| &\le \|V_k f(H_k)e_1\|v\| - x_k\| \\ &= \|\,(f(H_k) - \mathcal{R}_{\mu\nu}(H_k))\,e_1\|\|v\|\,. \end{aligned}$$

Therefore, whenever the chosen rational function $\mathcal{R}_{\mu\nu}$ accurately approximates f, the two approaches evolve similarly as the Krylov subspace dimension increases.

4. The exponential function. We next focus our attention on methods specifically designed to approximate the matrix exponential, $\exp(A)$, and its action on a vector v. We start by briefly discussing the role of this function within Model Order Reduction applications. Depending on the setting, we shall use either of the two equivalent notations $\exp(A)$ and e^A. Finally, we explicitly observe that Definition 2.2 ensures that $\exp(A)$ is nonsingular for any matrix A.

4.1. The exponential matrix in model order reduction applications. In this section we briefly review some application problems whose numerical solution benefits from the approximate computation of the exponential.

Numerical solution of time-dependent differential equations. The numerical solution of ordinary and time-dependent partial differential equations (ODEs and PDEs, respectively) may involve methods that effectively employ the matrix exponential. Recent developments in the efficient approximation of $\exp(A)v$ have increased the use of numerical "exponential-based" (or just "exponential") techniques that allow one to take larger time steps. More precisely, consider the system of ODEs of the form

$$u'(t) = Au(t) + b(t), \qquad u(0) = u_0,$$

where A is a negative semidefinite matrix. The analytic solution is given by

$$u(t) = e^{tA}u_0 + \int_0^t e^{(\tau - t)A}b(\tau)d\tau,$$

Whenever a good approximation to the propagation operator e^{sA} is available, it is possible to approximate the analytic solution by simply approximating the integral above with convenient quadrature formulas, leading to stable solution approximations. The generalization of this approach to the numerical solution of partial differential equations can be obtained, for instance, by employing a semidiscretization (in space) of the given problem. Consider the following self-adjoint parabolic equation

$$\frac{\partial u(x,t)}{\partial t} = \mathrm{div}(a(x)\nabla u(x,t)) - b(x)u(x,t) + c(x),$$

with $x \in \Omega$, Dirichlet boundary conditions and $b(x) \ge 0$, $a(x) > 0$ in Ω, with a, b, c sufficiently regular functions. A continuous time-discrete space discretization leads to the ordinary differential equation

$$E\frac{d\mathbf{u}(t)}{dt} = -A\mathbf{u}(t) + \mathbf{c}, \quad t \ge 0,$$

where A, E are positive definite Hermitian matrices, so that the procedure discussed above can be applied; see, e.g., [11], [25], [50], [64], [69]. Further attempts to generalize this procedure to non-selfadjoint PDEs can be found in [25, section 6.2], although the

theory behind the numerical behavior of the ODE solver in this case is not completely understood yet.

The use of exponential integrators is particularly effective in the case of certain stiff systems of nonlinear equations. Consider, e.g., the initial value problem

$$\frac{du(t)}{dt} = f(u), \qquad u(t_0) = u_0.$$

If the problem is stiff, standard integrators perform very poorly. A simple example of an exponential method for this system is the *exponentially fitted Euler* scheme, given by

$$u_1 = u_0 + h\phi(hA)f(u_0),$$

where h is the step size, $\phi(z) = \frac{e^z-1}{z}$, and $A = f'(u_0)$. The recurrence $\{u_k\}_{k=0,1,\ldots}$ requires the evaluation of $\phi(hA)v$ at each iteration, for some vector v; see, e.g., [35].

An application that has witnessed a dramatic increase in the use of the matrix exponential is Geometric Integration. This research area includes the derivation of numerical methods for differential equations whose solutions are constrained to belong to certain manifolds equipped with a group structure. One such example is given by linear Hamiltonian problems of the form

$$\begin{cases} \dot{Y}(t) = \mathcal{J}A(t)Y(t), \\ Y(t_0) = Y_0, \end{cases}$$

where $\mathcal{J}$ is the matrix $[0, I; -I, 0]$, A is a continuous, bounded, symmetric matrix function, and $Y_0 \in \mathbb{R}^{N\times p}$ is symplectic, that is it satisfies $Y_0^T \mathcal{J} Y_0 = \mathcal{J}$. The solution $Y(t)$ is symplectic for any $t \geq t_0$. Using the fact that $\mathcal{J}A$ is Hamiltonian, it can be shown that $\exp(\mathcal{J}A(t))$ is symplectic as well. Numerical methods that aim at approximating $Y(t)$ should also preserve its symplecticity property. This is achieved for instance by the numerical scheme $Y_{k+1} = \exp(h\mathcal{J}A(t_k))Y_k$, $t_{k+1} = t_k + h$, $k = 0, 1, \ldots$. Structure preserving methods associated with small dimensional problems have received considerable attention, see, e.g., [10], [29], [38], [70] and references therein. For large problems where order reduction is mandatory, approximations obtained by specific variants of Krylov subspace methods can be shown to maintain these geometric properties; see, e.g., [47].

Analysis of dynamical systems. The exponential operator has a significant role in the analysis of linear time-invariant systems of the form

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t) \\ y(t) = Cx(t) \end{cases} \tag{4.1}$$

where A, B and C are real matrices of size $n \times n$, $n \times m$ and $p \times n$, respectively. In the following we assume that A is stable, that is its eigenvalues are in the left half plane $\mathbb{C}^-$, and that the system is controllable and observable; see, e.g., [1].

The matrix of the states of the system for impulsive inputs is $x(t) = e^{tA}B$, whereas in general, for an initial state x_0 at time t_0, the resulting state at time $t \geq t_0$ is given by

$$x(t) = e^{(t-t_0)A}x_0 + \int_{t_0}^{t} e^{(t-\tau)A}Bu(\tau)d\tau.$$

Therefore, an approximation to the state involves the approximation of the matrix exponential. Moreover, the state function is used to define the first of the following two matrices which are called the controllability and the observability Gramians, respectively,

$$P = \int_0^\infty e^{tA} B B^T e^{tA^T} dt, \quad Q = \int_0^\infty e^{tA^T} C^T C e^{tA} dt. \tag{4.2}$$

The following result shows that these are solutions to Lyapunov equations.

THEOREM 4.1. *Given the linear time-invariant system (4.1), let P, Q be as defined in (4.2). Then they satisfy*

$$AP + PA^T + BB^T = 0, \qquad A^T Q + QA + C^T C = 0.$$

Proof. The proof follows from substituting the definition of P and Q into the corresponding expressions $AP + PA^T$, $A^T Q + QA$. By using the fact that $e^{tA} A = \frac{d}{dt}(e^{tA})$ and integrating, we obtain, e.g., for Q,

$$\begin{aligned} QA + A^T Q &= \int_0^\infty \left(e^{tA^T} C^T C e^{tA} A + A^T e^{tA^T} C^T C e^{tA} \right) dt \\ &= \int_0^\infty \left(e^{tA^T} C^T C \frac{de^{tA}}{dt} + \frac{de^{tA^T}}{dt} C^T C e^{tA} \right) dt \\ &= \int_0^\infty \frac{d(e^{tA^T} C^T C e^{tA})}{dt} dt = \lim_{\tau \to \infty} (e^{tA^T} C^T C e^{tA}) \Big|_0^\tau = -C^T C. \quad \square \end{aligned}$$

It can also be shown that the solution to each Lyapunov equation is unique. In a more general setting, the matrix $M := -(A^T Q + QA)$ is not commonly given in factored form. In this case, if it can be shown that M is positive semidefinite and that the pair (A, M) is observable, then Q is positive definite (a corresponding result holds for P); see, e.g., [4], [1], [13].

The Lyapunov equation may be used to compute estimates for $\|e^{tA}\|$, which in turn provides information on the stability of the original system in the case of $C^T C$ full rank; see, e.g., [13, Th. 3.2.2] for a proof.

THEOREM 4.2. *Let A be stable and $C^T C$ full rank. Then the unique solution Q to the Lyapunov equation $A^T Q + QA + C^T C = 0$ satisfies*

$$\|e^{tA}\| \leq \left(\frac{\lambda_{\max}(Q)}{\lambda_{\min}(Q)} \right)^{\frac{1}{2}} e^{-\alpha t},$$

where $\alpha = \lambda_{\min}(Q^{-1} C^T C)/2 > 0$.

For large problems, other devices can be used to directly approximate $\|e^{tA}\|$ without first resorting to the solution of a Lyapunov equation; cf. section 3.5. We also refer to [45] for a general discussion on the norm $\|e^{tA}\|$ and some of its bounds.

4.2. Computing the exponential of a matrix. Over the years, several methods have been devised and tested for the computation of the matrix exponential; we refer to [49] for a recent survey of several approaches and for a more complete bibliographic account. The algorithmic characteristics may be very different depending on whether the matrix has small or large dimension, or whether it is dense or sparse; the structural and symmetry properties also play a crucial role; see, e.g., the discussion in [61]. In this section we discuss the case of small matrices. When A is normal, the

spectral decomposition discussed in section 3.1 can be employed, namely $A = TJT^H$ with T unitary. This gives $\exp(A) = T\exp(J)T^H$, once the decomposition of A is computed.

In the non-normal case, one method has emerged in the last decade, for its robustness and efficiency: Padé approximation with scaling and squaring. The basic method employs a rational function approximation to the exponential function as

$$\exp(\lambda) \approx \mathcal{R}_{\mu\nu}(\lambda) = \frac{\mathcal{N}_{\mu\nu}(\lambda)}{\mathcal{D}_{\mu\nu}(\lambda)},$$

where $\mathcal{N}_{\mu\nu}, \mathcal{D}_{\mu\nu}$ are polynomials of degree μ and ν, respectively. One attractive feature of the $[\mu/\nu]$ Padé approximation is that the coefficients of the two polynomials are explicitly known, that is

$$\mathcal{N}_{\mu\nu}(\lambda) = \sum_{j=0}^{\mu} \frac{(\mu+\nu-j)!\mu!}{(\mu+\nu)!(\mu-j)!j!}\lambda^j, \qquad \mathcal{D}_{\mu\nu}(\lambda) = \sum_{j=0}^{\nu} \frac{(\mu+\nu-j)!\nu!}{(\mu+\nu)!(\nu-j)!j!}(-\lambda)^j.$$

These two polynomials have a rich structure. For example, one has the relation $\mathcal{N}_{\mu\nu}(\lambda) = \mathcal{D}_{\nu\mu}(-\lambda)$ as well as several other important properties which can be found, e.g., in [26, section 5.2].

Diagonal Padé approximation ($\mu = \nu$), is usually preferred because computing $\mathcal{R}_{\mu\nu}$ with say, $\mu > \nu$, is not cheaper than computing the more accurate $\mathcal{R}_{\nu_*\nu_*}$ where $\nu_* = \max\{\mu, \nu\}$. Nonetheless, because of their stability properties, Padé $[\nu+1/\nu]$ approximations are used, together with $[\nu/\nu]$ approximations, in the numerical solution of initial value problems with one-step methods. Another attractive property of the diagonal Padé approximation is that if A has eigenvalues with negative real part, then the spectral radius of $\mathcal{R}_{\nu\nu}(A)$ is less than one, for any ν. In the following, diagonal rational approximation will be denoted by $\mathcal{R}_{\nu\nu} = \mathcal{R}_\nu$. The accuracy of the approximation can be established by using the following result.

THEOREM 4.3. [26, Theorem 5.5.1] *Let the previous notation hold. Then*

$$e^\lambda - \mathcal{R}_{\mu\nu}(\lambda) = (-1)^\nu \frac{\mu!\,\nu!}{(\mu+\nu)!\,(\mu+\nu+1)!}\lambda^{\mu+\nu+1} + O(\lambda^{\mu+\nu+2}).$$

This error estimate shows that the approximation degrades as λ gets away from the origin. This serious limitation motivated the introduction of the scaling and squaring procedure. By exploiting the property $e^A = (e^{A/k})^k$, for any square matrix A and scalar k, the idea is to determine k so that the scaled matrix A/k has norm close to one, and then employ the approximation

$$e^{A/k} \approx \mathcal{R}_\nu(A/k).$$

The approximation to the original matrix e^A is thus recovered as $e^A \approx \mathcal{R}_\nu(A/k)^k$. The use of powers of two in the scaling factor is particularly appealing. Indeed, by writing $k = 2^s$, the final approximation $\mathcal{R}_\nu(A/2^s)^{2^s}$ is obtained by repeated squaring. The scalar s is determined by requiring that $\|A\|_\infty/2^s$ is bounded by some small constant, say $1/2$. In fact, this constant could be allowed to be significantly larger with no loss in stability and accuracy; see [33]. The approach oulined here is used in Matlab 7.1. [48].

A rational function that is commonly used in the case of symmetric negative semidefinite matrices, is given by the Chebychev rational function. The Chebychev

approximation $\mathcal{R}^\star_{\mu\nu}$ determines the best rational function approximation in $[0,+\infty)$ to $e^{-\lambda}$ by solving the problem

$$\min_{\mathcal{R}_{\mu\nu}} \max_{\lambda\in[0,+\infty)} \left|e^{-\lambda} - \mathcal{R}_{\mu\nu}(\lambda)\right|,$$

where the minimum is taken over all rational functions. In particular, the cases $\mu = 0$ and $\mu = \nu$ have been investigated in greater detail, and the coefficients of the polynomials of $\mathcal{R}^\star_\nu$ have been tabulated first by Cody, Meinardus and Varga in [12] for $\nu \le 14$ and then in [9] for degree up to 30. Setting $\mathcal{E}_\nu = \max_{\lambda\in[0,+\infty)} \left|e^{-\lambda} - \mathcal{R}^\star_\nu(\lambda)\right|$, great efforts in the approximation theory community have been devoted to show the following elegant result on the error asymptotic behavior,

$$\lim_{\nu\to\infty} \mathcal{E}_\nu^{1/\nu} = \frac{1}{9.28903...},$$

disproving the so-called "1/9" conjecture. From the result above it follows that $\sup_{\lambda\in[0,+\infty)} \left|e^{-\lambda} - \mathcal{R}_\nu(\lambda)\right| \approx 10^{-\nu}$.

Other rational function approximations that have recently received renewed interest are given by rational functions with real poles, such as $\mathcal{R}_{\mu\nu}(\lambda) = \mathcal{N}_\mu(\lambda)/(1+h\lambda)^\nu$; see, e.g., [7], [51], [54]. An advantage of these functions is that they avoid dealing with *complex* conjugate poles.

4.3. Reduction methods for large matrices. In many application problems where A is large, the action of $\exp(A)v$ is required, rather than $\exp(A)$ itself, so that the methods of section 3.5 and of section 3.6 can be used. We first discuss some general convergence properties, and then show the role of the Krylov subspace approximation to $\exp(A)v$ in various circumstances. Note that time dependence can, in principle, be easily acommodated in the Krylov approximation as, for instance, $\exp(tA)v \approx V_k \exp(tH_k)e_1\|v\|$. In the following, we shall assume that A already incorporates time dependence. In particular, estimates involving spectral information on the matrix will be affected by possible large values of t.

An analysis of the Krylov subspace approximation $V_k \exp(H_k)e_1\|v\|$ to $\exp(A)v$ was given by Saad [59], where the easily computable quantity

$$h_{k+1,k} \cdot \left|e_k^T \exp(H_k)e_1\|v\|\right|$$

was proposed as stopping criterion for the iterative Arnoldi process; a higher order estimate was also introduced in [59]. Further study showed that the convergence rate of the approximation is often superlinear. In the Hermitian negative semidefinite case, a complete characterization of this superlinearity behavior can be derived using the following bounds. We refer to [18], [64] for qualitatively similar, although asymptotic bounds.

THEOREM 4.4 (see Hochbruck and Lubich [34]). *Let A be a Hermitian negative semidefinite matrix with eigenvalues in the interval $[-4\rho, 0]$, with $\rho > 0$. Then the error in the approximation (3.8) of $\exp(A)v$ is bounded as follows:*

$$\|\exp(A)v - V_k \exp(H_k)e_1\| \le 10e^{-k^2/(5\rho)}, \qquad \sqrt{4\rho} \le k \le 2\rho,$$
$$\|\exp(A)v - V_k \exp(H_k)e_1\| \le \frac{10}{\rho} e^{-\rho} \left(\frac{e\rho}{k}\right)^k, \qquad k \ge 2\rho.$$

Other bounds that emphasize the superlinear character of the approximation have also been proposed in [63], and earlier in [25]. Similar results also hold in the case when A is skew-symmetric, or when A is non-symmetric, under certain hypotheses on the location of its spectrum, see [18], [34].

A typical convergence curve of the error together with the bounds of Theorem 4.4 (called HL bound) are shown in Figure 4.1, for a diagonal 1001×1001 matrix A with entries uniformly distributed in $[-40, 0]$ and a random vector v with uniformly distributed values in $[0, 1]$ and unit norm; this example is taken from [34].

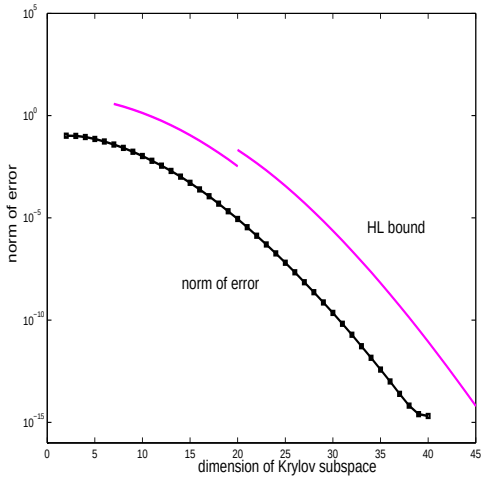

FIG. 4.1. *Typical convergence of Krylov subspace approximation to* $\exp(A)v$ *and upper bounds of Theorem 4.4.*

Rational approximation and partial fraction expansion. When A is large, the methods of the previous sections can be employed to approximate $\exp(A)v$. In particular, using diagonal Padé or Chebyshev rational functions and their partial fraction expansion, one gets

$$\exp(A)v \approx \mathcal{N}_\nu(A)\mathcal{D}_\nu(A)^{-1}v = \omega_0 v + \sum_{j=1}^{\nu} \omega_j (A - \tau_j I)^{-1} v,$$

where the coefficients and the poles are pairwise complex conjugates. A recent discussion on the main properties of this approximation and on related references can be found in [46].

Application to the solution of Lyapunov matrix equations. The approximate solution of the Lyapunov matrix equation has been addressed in a large body of literature, in which both the sign function and the exponential function have played a leading role over the years. Here we focus on low-rank approximations, assuming that B has few columns and that it is full column rank.

The dependence of P (and Q) on the exponential has classically motivated the use of quadrature formulas for the approximation of the integral defining P (cf. (4.2)), together with the use of low degree polynomial or rational function approximations

to $\exp(t\lambda)$. More precisely, setting $X(t) = \exp(tA)B$, P could be approximated as

$$P(\tau) = \int_0^\tau X(t)X(t)^T dt,$$

for some $\tau \geq 0$. Then, by using a quadrature formula with discretization points t_i and weights δ_i, the integral in $[0, \tau]$ is approximated as $P(\tau) \approx \sum_{i=1}^k X(t_i)\delta_i X(t_i)^T$. The practical effectiveness of the approach depends on the value of τ, but most notably on the quadrature formula used, under the constraint that all δ_i be positive, to ensure that $P(\tau)$ is positive semidefinite; see [58] for some experiments using different quadrature formulas.

The procedure above determines a low-rank approximation to the corresponding Gramian since the number of columns of B is small. An alternative approach that bypasses the integral formulation within the low-rank framework, is obtained by reducing the problem dimension. If an approximation to $\exp(tA)B$ is available as $x_k = V_k \exp(tH_k)E$, where E and V_k are defined so that $B = V_k E$, then

$$P_k = V_k \int_0^\infty \exp(tH_k)EE^T \exp(tH_k^T)dt\, V_k^T =: V_k G_k V_k^T.$$

If H_k is stable, Theorem 4.1 ensures that G_k is the solution to the following small dimensional Lyapunov equation:

$$H_k G + G H_k^T + EE^T = 0. \tag{4.3}$$

This derivation highlights the theoretical role of the exponential in the approximation procedure. However, one can obtain G_k by directly solving the small matrix equation, by means of methods that exploit matrix factorizations [3], [30].

The following result sheds light onto the reduction process performed by this approximation; see, e.g., [39], [58].

PROPOSITION 4.5. *Let the columns of V_k, with $V_k^T V_k = I_k$, span $K_k(A, B) = \text{span}\{B, AB, \ldots, A^{k-1}B\}$. The approximate solution $P_k = V_k G_k V_k^T$ where G_k solves (4.3) is the result of a Galerkin process onto the space $K_k(A, B)$.*

Proof. Let V_k be a matrix whose orthonormal columns span $K_k(A, B)$. Let $R_k = AP_k + P_k A^T + BB^T$ be the residual associated with $P_k = V_k G_k V_k^T$ for some G_k and let $H_k = V_k^T A V_k$. A Galerkin process imposes the following orthogonality condition on the residual[1]

$$V_k^T R_k V_k = 0.$$

Expanding R_k and using $V_k^T V_k = I$, we obtain

$$\begin{aligned} V_k^T A V_k G_k V_k^T + G_k V_k^T A^T V_k + V_k^T BB^T V_k = 0 \\ H_k G_k + G_k H_k + V_k^T BB^T V_k = 0. \end{aligned}$$

Recalling that $B = V_k E$, the result follows. □

Other methods have been proposed to approximately solve large-scale Lyapunov equations; see [28],[44],[55] and references therein.

[1] This "two-sided" condition can be derived by first defining the matrix inner product $\langle X, Y \rangle = \text{tr}(XY^T)$ and then imposing $\langle R_k, P_k \rangle = 0$ for any $P_k = V_k G V_k^T$ with $G \in \mathbb{R}^{k \times k}$.

5. The matrix sign function. In this section we discuss methods for the matrix sign function, with the sign function on $\mathbb{C}$ defined as

$$\operatorname{sign}(z) = \begin{cases} +1 & \text{if } \Re(z) > 0, \\ -1 & \text{if } \Re(z) < 0. \end{cases}$$

We do not define $\operatorname{sign}(z)$ on the imaginary axis where, anyway, it is not continuous. Outside the imaginary axis, sign is infinitely often differentiable, so that $\operatorname{sign}(A)$ is defined as long as the matrix $A \in \mathbb{C}^{n\times n}$ has no eigenvalues on the imaginary axis. We first recall a few application problems where the sign function is commonly employed.

5.1. Motivation. The *algebraic Riccati equation* arises in control theory as a very fundamental system to be solved in order to compute, for example, certain observers or stabilizers, see [4], [43]. It is a quadratic matrix equation of the form

$$G + A^T X + XA - XFX = 0, \tag{5.1}$$

where $A, F, G \in \mathbb{R}^{n\times n}$ and F and G are symmetric and positive definite. One aims at finding a symmetric positive definite and stabilizing solution X, i.e. the spectrum of $A - FX$ should lie in $\mathbb{C}^-$. The quadratic equation (5.1) can be linearized by turning it into a system of doubled size as

$$K := \begin{bmatrix} A^T & G \\ F & -A \end{bmatrix} = \begin{bmatrix} X & -I \\ I & 0 \end{bmatrix} \cdot \begin{bmatrix} -(A-FX) & -F \\ 0 & (A-FX)^T \end{bmatrix} \cdot \begin{bmatrix} X & -I \\ I & 0 \end{bmatrix}^{-1}.$$

If we assume that X is a stabilizing solution (such a solution exists under mild conditions, see [4],[43]), a standard approach is to use the matrix sign function to compute X. We have $\operatorname{sign}(-(A-FX)) = I$ and $\operatorname{sign}(A-FX) = -I$. Therefore,

$$\operatorname{sign}\begin{bmatrix} -(A-FX) & -F \\ 0 & (A-FX)^T \end{bmatrix} = \begin{bmatrix} I & Z \\ 0 & -I \end{bmatrix}, \quad Z \in \mathbb{R}^{n\times n}$$

and we see that

$$\operatorname{sign}(K) - I = \begin{bmatrix} X & -I \\ I & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & Z \\ 0 & -2I \end{bmatrix} \cdot \begin{bmatrix} X & -I \\ I & 0 \end{bmatrix}^{-1}. \tag{5.2}$$

Split $\operatorname{sign}(K) - I$ vertically in its middle as $[M|N]$, move the inverse matrix to the left-hand side in (5.2) and then equate the first halves, to get

$$MX = N.$$

This is an overdetermined, but consistent linear system for X, and by working with the second half blocks, it can be shown that M is full column rank. Therefore, the procedure above outlines a method to derive a stabilizing solution X to (5.1) by means of the sign function.

As discussed in section 4.3, the Lyapunov equation

$$A^T X + XA + C^T C = 0, \text{ where } A, C \in \mathbb{R}^{n\times n}$$

also arises in control theory. It was already shown in [57] that the (Hermitian) solution X is the (2,1) block of the sign-function of a matrix of twice the dimension, that is

$$\begin{bmatrix} 0 & 0 \\ X & I \end{bmatrix} = \frac{1}{2}\left(I + \operatorname{sign}\left(\begin{bmatrix} A & 0 \\ C^T C & -A^T \end{bmatrix}\right)\right).$$

This follows in a way similar to what we presented for the algebraic Riccati equation; see also [5] for a generalization.

The matrix sign function also appears in the modelling (and subsequent simulation) of complex physical systems. One example is given by the so-called overlap fermions of lattice quantum chromodynamics [53], where one has to solve linear systems of the form

$$(I + \Gamma_5 \mathrm{sign}(Q))x = b. \tag{5.3}$$

Here Γ_5 is a simple permutation matrix and Q is a huge, sparse, complex Hermitian matrix representing a nearest neighbour coupling on a regular 4-dimensional grid with 12 variables per grid point. Note that here we may situate ourselves in an order reduction context, since if we solve (5.3) with some iterative method, the basic operation will be to compute matrix-vector products, i.e. we need the action $\mathrm{sign}(Q)v$ rather than $\mathrm{sign}(Q)$ itself.

5.2. Matrix methods. A detailed survey on methods to compute the whole matrix $\mathrm{sign}(A)$ is given in [42], see also [2]. We shortly describe the most important ones.

The Newton iteration to solve $z^2 - 1 = 0$ converges to $+1$ for all starting values in the right half plane, and to -1 for all those from $\mathbb{C}^-$. According to (3.5), the corresponding matrix iteration reads

$$S_{k+1} = \frac{1}{2}\left(S_k + S_k^{-1}\right), \quad \text{where } S_0 = A. \tag{5.4}$$

Although the convergence is global and asymptotically quadratic, it can be quite slow in the presence of large eigenvalues or eigenvalues with a small real part. Therefore, several accelerating scaling strategies have been proposed [41], for example by using the determinant [8], i.e.

$$S_{k+1} = \frac{1}{2}\left((c_k S_k) + \frac{1}{c_k} S_k^{-1}\right), \quad \text{with } c_k = \det(S_k).$$

Note that $\det(S_k)$ is easily available if S_k is inverted using the LU-factorization. An alternative which avoids the computation of inverses is the Schulz iteration, obtained as Newton's method for $z^{-2} - 1 = 0$, which yields

$$S_{k+1} = \frac{1}{2} \cdot S_k \cdot \left(3I - S_k^2\right), \;\; S_0 = A.$$

This iteration is guaranteed to converge only if $\|I - A^2\| < 1$ (in an arbitrary operator norm).

In [40], several other iterations were derived, based on Padé approximations of the function $(1 - z)^{-1/2}$. They have the form

$$S_{k+1} = S_k \cdot \mathcal{N}_{\mu\nu}(S_k^2) \cdot \mathcal{D}_{\mu\nu}(S_k^2)^{-1}, \;\; S_0 = A.$$

For $\mu = 2p, \nu = 2p - 1$, an alternative representation is

$$S_{k+1} = \left((I + S_k)^{2p} + (I - S_k)^{2p}\right) \cdot \left((I + S_k)^{2p} - (I - S_k)^{2p}\right)^{-1}. \tag{5.5}$$

In this case, the coefficients of the partial fraction expansion are explicitly known, giving the equivalent representation

$$S_{k+1} = \frac{1}{p} \cdot S_k \cdot \sum_{i=1}^{p} \frac{1}{\xi_i} \left(S_k^2 + \alpha_i I\right)^{-1}, \quad S_0 = A, \tag{5.6}$$
$$\text{with } \xi_i = \frac{1}{2}\left(1 + \cos\frac{(2i-1)\pi}{2p}\right), \quad \alpha_i^2 = \frac{1}{\xi_i} - 1, \quad i = 1, \ldots, p.$$

Interestingly, ℓ steps of the iteration for parameter p are equivalent to one step with parameter $p\ell$. The following global convergence result on these iterations was proved in [40].

THEOREM 5.1. *If A has no eigenvalues on the imaginary axis, the iteration (5.5) converges to* $\operatorname{sign}(A)$. *Moreover, one has*

$$(\operatorname{sign}(A) - S_k)(\operatorname{sign}(A) + S_k)^{-1} = \left((\operatorname{sign}(A) - A)(\operatorname{sign}(A) + A)^{-1}\right)^{(2p)^k},$$

which, in the case that A is diagonalizable, gives

$$\|(\operatorname{sign}(A) - S_k)(\operatorname{sign}(A) + S_k)^{-1}\| \leq \|T\| \cdot \|T^{-1}\| \cdot \left(\max_{\lambda \in \operatorname{spec}(A)} \frac{\operatorname{sign}(\lambda) - \lambda}{\operatorname{sign}(\lambda) + \lambda}\right)^{2p^k}. \tag{5.7}$$

5.3. Krylov subspace approximations. We now look at Krylov subspace approximations for

$$\operatorname{sign}(A)v, \quad v \in \mathbb{C}^n$$

with special emphasis on A Hermitian. The Krylov subspace projection approach from (3.8) gives

$$\operatorname{sign}(A)v \approx V_k \operatorname{sign}(H_k) e_1 \cdot \|v\|. \tag{5.8}$$

If one monitors the approximation error in this approach as a function of k, the dimension of the Krylov subspace, one usually observes a non-monotone, jig-saw like behaviour. This is particularly so for Hermitian indefinite matrices, where the real eigenvalues lie to the left and to the right of the origin. This can be explained by the fact, formulated in Proposition 3.1, that the Krylov subspace approximation is given as $p_{k-1}(A)v$ where p_{k-1} is the degree $k-1$ polynomial interpolating at the Ritz values. But the Ritz values can get arbitrarily close to 0 (or even vanish), even though the spectrum of A may be well separated from 0, then producing a (relatively) large error in the computed approximation. A Ritz value close to 0 is likely to occur if $k-1$ is odd, so the approximation has a tendency to degrade every other step. A remedy to this phenomenon is to use the polynomial that interpolates A at the *harmonic* Ritz values, since these can be shown to be as well separated from zero as $\operatorname{spec}(A)$. Computationally, this can be done using the same Arnoldi recurrence (3.7) as before, but applying a simple rank-one modification to H_k before computing its sign function. Details are given in [66].

An alternative is to use the identity

$$\operatorname{sign}(z) = z \cdot (z^2)^{-\frac{1}{2}},$$

then use the Krylov subspace approach on the squared matrix A^2 to approximate

$$\left.\begin{array}{l} \left(A^2\right)^{-\frac{1}{2}} v \approx V_k (H_k)^{-\frac{1}{2}} e_1 \cdot \|v\| =: y_k, \\ \mathrm{sign}(A) v \approx x_k = A y_k. \end{array}\right\} \tag{5.9}$$

Note that in (5.9) the matrix H_k represents the projection of A^2 (not of A !), onto the Krylov subspace $K_k(A^2, b)$. Interestingly, this is one of the special cases when explicitly generating the space $K_k(A^2, b)$ is more effective than using $K_k(A, b)$; see [67] for a general analysis of cases when using the latter is computationally more advantageous.

It is also remarkable that, in case that A is Hermitian, it is possible to give *a posteriori* error bounds on the quality of the approximation x_k as formulated in the following theorem taken from [66].

THEOREM 5.2. *Let A be Hermitian and non-singular. Then x_k from (5.9) satisfies*

$$\|\mathrm{sign}(A) v - x_k\|_2 \leq \|r_k\|_2 \leq 2\kappa \left(\frac{\kappa - 1}{\kappa + 1}\right)^k \cdot \|v\|_2, \tag{5.10}$$

where $\kappa \equiv \|A\|_2 \|A^{-1}\|_2$ and r_k is the residual in the k-th step of the CG method applied to the system $A^2 x = v$ (with initial residual v, i.e. initial zero guess).

The residual norms $\|r_k\|$ need not be computed via the CG method since they can be obtained at almost no cost from the entries of the matrix H_k in the Lanczos recursion (3.7). This can be seen as follows: Since A^2 is Hermitian, H_k is Hermitian and tridiagonal. If p_{k-1} is the degree $k-1$ polynomial expressing the k-th Lanczos vector v_k as $v_k = p_{k-1}(A^2) v$, the Lanczos recursion gives $h_{k+1,k} p_k(z) = (z - h_{k,k}) \cdot p_{k-1}(z) - h_{k-1,k} p_{k-2}(z)$. On the other hand, it can be shown that $r_k = \sigma v_{k+1}$ with $\|v_{k+1}\| = 1$ for some scalar σ (see [60, Proposition 6.20]), and since $r_k = q_k(A^2) v$ for some polynomial q_k of degree k satisfying $q_k(0) = 1$, it must be $q_k = p_k / p_k(0)$, so that $r_k = p_k(A^2) v / p_k(0) = v_{k+1} / p_k(0)$. Therefore, along with the Lanczos process we just have to evaluate the recursion $p_k(0) = -[h_{k,k} \cdot p_{k-1}(0) + h_{k-1,k} p_{k-2}(0)] / h_{k+1,k}$ to obtain $\|r_k\| = 1 / |p_k(0)|$.

We do not know of comparable error bounds for the other two approaches outlined earlier ((5.8) and its variant using harmonic Ritz values). Note also that x_k from (5.9) satisfies $x_k = A p_{k-1}(A^2) v$, where $q(z) = z \cdot p_{k-1}(z^2)$ is an odd polynomial, that is $q(-z) = -q(z)$, of degree $2k-1$ in z. This is a restriction as compared to the other two approaches where we do not enforce any symmetry on the interpolating polynomials. However, this restriction will most probably have an effect only if the spectrum of A is very unsymmetric with respect to the origin. Our computational experience in simulations from lattice QCD indicates that x_k is actually the best of the three approximations discussed so far. Since, in addition, x_k comes with a bound of the true error norm, we definitely favor this approach.

5.4. Partial fraction expansions. If we have sufficient information on the eigensystem of A available, we can use (5.7) to estimate a number $p \in \mathbb{N}$ such that the first iterate from (5.6) already gives a sufficiently good approximation to the sign function. As discussed in section 3.6, we can then approximate

$$\mathrm{sign}(A) v \approx A \cdot \sum_{i=1}^{p} \frac{1}{p \xi_i} \tilde{x}^{(j)},$$

with $\tilde{x}^{(j)}$ an approximate solution of the linear system

$$(A^2 + \alpha_j I)x^{(j)} = v, \qquad j = 1, \ldots, p.$$

We already discussed in section 3.6 how we can make efficient use of the shifted nature of these systems when solving them with standard Krylov subspace methods.

A particularly important situation arises when A is Hermitian and the intervals $[-b, -a] \cup [c, d]$, with $0 < a \leq b, 0 < c \leq d$, containing the eigenvalues of A are available. Under these hypotheses, Zolotarev explicitly derived the best rational approximation in the Chebyshev sense; see [56]. The next theorem states this result for $[-b, -a] = -[c, d]$. The key point is that for fixed $\mu = 2p - 1, \nu = 2p$, finding the optimal rational approximation $\mathcal{R}_{\mu\nu}(z) = \mathcal{N}_{\mu\nu}(z)/\mathcal{D}_{\mu\nu}(z)$ to the sign function on $[-b, -a] \cup [a, b]$ is equivalent to finding the best such rational approximation $\mathcal{S}_{p-1,p}(z) = \mathcal{N}_{p-1,p}(z)/\mathcal{D}_{p-1,p}(z)$ in *relative sense* to the inverse square root on $[1, (b/a)^2]$. The two functions are then related via $\mathcal{R}_{2p-1,2p}(z) = az \cdot \mathcal{S}_{p-1,p}(az)$.

PROPOSITION 5.3. *Let $\mathcal{R}_{2p-1,2p}(z) = \mathcal{N}_{2p-1,2p}(z)/\mathcal{D}_{2p-1,2p}(z)$ be the Chebyshev best approximation to* sign(z) *on the set $[-b, -a] \cup [a, b]$, i.e. the function which minimizes*

$$\max_{a<|z|<b} |\text{sign}(z) - \widetilde{\mathcal{R}}_{2p-1,2p}(z)|$$

over all rational functions $\widetilde{\mathcal{R}}_{2p-1,2p}(z) = \widetilde{\mathcal{N}}_{2p-1,2p}(z)/\widetilde{\mathcal{D}}_{2p-1,2p}(z)$. Then the factored form of $\mathcal{R}_{2p-1,2p}$ is given by

$$\mathcal{R}_{2p-1,2p}(z) = az \cdot \mathcal{S}_{p-1,p}((az)^2) \qquad \textit{with} \qquad \mathcal{S}_{p-1,p}(z) = D\frac{\prod_{i=1}^{p-1}(z + c_{2i})}{\prod_{i=1}^{p}(z + c_{2i-1})},$$

where

$$c_i = \frac{\text{sn}^2\left(iK/(2p); \sqrt{1 - (b/a)^2}\right)}{1 - \text{sn}^2\left(iK/(2p); \sqrt{1 - (b/a)^2}\right)},$$

K is the complete elliptic integral, sn *is the Jacobi elliptic function, and D is uniquely determined by the condition*

$$\max_{z\in[1,(b/a)^2]} \left(1 - \sqrt{z}\mathcal{S}_{p-1,p}(z)\right) = -\min_{z\in[1,(b/a)^2]} \left(1 - \sqrt{z}\mathcal{S}_{p-1,p}(z)\right).$$

For a given number of poles, the Zolotarev approximation is much more accurate than that of the rational approximation (5.6) and is therefore to be preferred. This is illustrated in Table 5.1, taken form [66]. However, the use of the Zolotarev approximation is restricted to Hermitian matrices for which lower and upper bounds (a and b, resp.) on the moduli of the eigenvalues are known.

As a final point, let us again assume that A is Hermitian and that we approximate sign(A) by some rational approximation $\mathcal{R}(A)$ with $\mathcal{R}$ having a partial fraction expansion of the form

$$\mathcal{R}(z) = \sum_{j=1}^{p} \omega_j \frac{z}{z^2 + \alpha_j}, \quad \omega_j \geq 0, \quad \alpha_j \geq 0, \quad j = 1, \ldots, p.$$

TABLE 5.1
Number of poles necessary to achieve accuracy of 0.01

b/a	(5.6)	Zolotarev
200	19	5
1000	42	6

Note that this is the case for the Zolotarev approximation from Proposition 5.3 as well as for the Padé approximations from (5.6). In order to compute $\mathcal{R}(A)v$, let us assume that we use the (shifted) CG-method to simultaneously solve $(A^2 + \alpha_j I)x^{(j)} = v$ for all j of interest, so that we get CG-iterates $x_k^{(j)}$ with residual $r_k^{(j)} = v-(A^2+\alpha_j I)x_k^{(j)}$. Then the following estimate holds [66].

PROPOSITION 5.4. *Let $g_j > 0$ be such that $\sum_{j=1}^{p} g_j = 1$ and $\varepsilon > 0$. If the CG iteration for system j is stopped at step k_j in which the residual satisfies*

$$\|r_{k_j}^{(j)}\|_2 \leq \varepsilon g_j \frac{\sqrt{\alpha_j}}{\omega_j},$$

then

$$\| \mathcal{R}(A)v - \sum_{j=1}^{p} \omega_j x_{k_j}^{(j)} \|_2 \leq \varepsilon.$$

This proposition formulates a computationally feasible stopping criterion. If we also know the approximation accuracy of the rational approximation, i.e. if we have an information of the kind

$$\max_{z \in \mathrm{spec}(A)} |\mathcal{R}(z) - \mathrm{sign}(z)| \leq \varepsilon_2,$$

then we know that

$$\| \mathrm{sign}(A)v - \sum_{j=1}^{p} \omega_j x_{k_j}^{(j)} \|_2 \leq \varepsilon + \varepsilon_2.$$

This fact is in agreement with the discussion on rational approximation of section 3.6.

Acknowledgement. We thank Daniel Kressner, Marlis Hochbruch and Henk A. van der Vorst for helpful comments on an earlier version of this chapter.

REFERENCES

[1] A. C. Antoulas. *Approximation of large-scale Dynamical Systems*. Advances in Design and Control. SIAM, Philadelphia, 2005.
[2] Zhaojun Bai and James Demmel. Using the matrix sign function to compute invariant subspaces. *SIAM J. Matrix Anal. Appl.*, 19(1):205–225, 1998.
[3] R. H. Bartels and G. W. Stewart. Algorithm 432: Solution of the Matrix Equation AX+XB=C. *Comm. of the ACM*, 15(9):820–826, 1972.
[4] P. Benner. Control Theory. Technical report, http://www.tu-chemnitz.de/~benner, 2006. to appear in 'Handbook of Linear Algebra'.
[5] P. Benner and E.S. Quintana-Orti. Solving stable generalized Lyapunov equations with the matrix sign function. *Numer. Algorithms*, 20:75–100, 1999.
[6] D. A. Bini, N. J. Higham, and B. Meini. Algorithms for the matrix pth root. *Numerical Algorithms*, 39(4):349–378, 2005.

[7] P. B. Borwein. Rational approximations with real poles to e^{-x} and x^n. *Journal of Approximation Theory*, 38:279–283, 1983.
[8] R. Byers. Solving the algebraic Riccati equation with the matrix sign function. *Linear Algebra Appl.*, 85:267–279, 1987.
[9] A. J. Carpenter, A. Ruttan, and R. S. Varga. Extended numerical computations on the "1/9" conjecture in rational approximation theory. In P. R. Graves-Morris, E. B. Saff, and R. S. Varga, editors, *Rational approximation and interpolation. Proceedings of the United Kingdom-United States conference held at Tampa, Florida, December 12-16, 1983*, pages 503–511, Berlin, 1990. Springer-Verlag, Lecture Notes Math.
[10] E. Celledoni and A. Iserles. Approximating the exponential from a Lie algebra to a Lie group. *Mathematics of Computation*, 69(232):1457–1480, 2000.
[11] E. Celledoni and I. Moret. A Krylov projection method for systems of ODEs. *Applied Num. Math.*, 24:365–378, 1997.
[12] W. J. Cody, G. Meinardus, and R. S. Varga. Chebyshev rational approximations to e^{-x} in $[0, +\infty)$ and applications to heat-conduction problems. *J. Approx. Theory*, 2(1):50–65, March 1969.
[13] M. J. Corless and A. E. Frazho. *Linear systems and control - An operator perspective*. Pure and Applied Mathematics. Marcel Dekker, New York - Basel, 2003.
[14] Ph. I. Davies and N. J. Higham. Computing $f(A)b$ for matrix functions f. In *QCD and numerical analysis III*, volume 47 of *Lect. Notes Comput. Sci. Eng.*, pages 15–24. Springer, Berlin, 2005.
[15] P. J. Davis and P. Rabinowitz. *Methods of Numerical Integration*. Academic Press, London, 2nd edition edition, 1984.
[16] J. W. Demmel. *Applied numerical linear algebra*. SIAM, Philadelphia, 1997.
[17] 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.
[18] 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.
[19] V. Druskin and L. Knizhnerman. Extended Krylov subspaces: approximation of the matrix square root and related functions. *SIAM J. Matrix Analysis and Appl.*, 19(3):755–771, 1998.
[20] M. Eiermann and O. Ernst. A restarted Krylov subspace method for the evaluation of matrix functions. Technical report, Institut für Numerische Mathematik und Optimierung, Technische Universität Bergakademie Freiberg, 2005.
[21] R. W. Freund. Solution of shifted linear systems by quasi-minimal residual iterations. In L. Reichel et al., editor, *Numerical linear algebra. Proceedings of the conference in numerical linear algebra and scientific computation, Kent, OH, USA, March 13-14, 1992*, pages 101–121. Walter de Gruyter, Berlin, 1993.
[22] A. Frommer. BiCGStab(ℓ) for families of shifted linear systems. *Computing*, 70:87–109, 2003.
[23] A. Frommer and U. Glässner. Restarted GMRES for shifted linear systems. *SIAM J. Sci. Comput.*, 19(1):15–26, 1998.
[24] A. Frommer and P. Maass. Fast CG-based methods for Tikhonov–Phillips regularization. *SIAM J. Sci. Comput.*, 20(5):1831–1850, 1999.
[25] E. Gallopoulos and Y. Saad. Efficient solution of parabolic equations by Krylov approximation methods. *SIAM J. Sci. Stat. Comput.*, 13(5):1236–1264, 1992.
[26] W. Gautschi. *Numerical Analysis. An Introduction*. Birkhäuser, Boston, 1997.
[27] G. Golub and C. F. Van Loan. *Matrix computations*. The Johns Hopkins Univ. Press, Baltimore, 3rd edition, 1996.
[28] S. Gugercin, D. C. Sorensen, and A. C. Antoulas. A modified low-rank Smith method for large-scale Lyapunov equations. *Numer. Algorithms*, 32:27–55, 2003.
[29] E. Hairer, C. Lubich, and G. Wanner. *Geometric numerical integration. Structure-preserving algorithms for ordinary differential equations*, volume 31 of *Springer Series in Computational Mathematics*. Springer, Berlin, 2002.
[30] S. J. Hammarling. Numerical solution of the stable, non-negative definite Lyapunov equation. *IMA J. Numer. Anal.*, 2:303–323, 1982.
[31] N. J. Higham. Newton's method for the matrix square root. *Math. Comp.*, 46(174):537–549, April 1986.
[32] N. J. Higham. Stable iterations for the matrix square root. *Numer. Algorithms*, 15(2):227–242, 1997.
[33] N. J. Higham. The Scaling and Squaring Method for the Matrix Exponential Revisited. *SIAM J. Matrix Analysis and Appl.*, 26(4):1179–1193, 2005.
[34] M. Hochbruck and C. Lubich. On Krylov subspace approximations to the matrix exponential

operator. *SIAM J. Numer. Anal.*, 34(5):1911–1925, 1997.
[35] M. Hochbruck, C. Lubich, and H. Selhofer. Exponential integrators for large systems of differential equations. *SIAM J. Sci. Comput.*, 19(5):1552–1574, 1998.
[36] M. Hochbruck and J. van den Eshof. Preconditioning Lanczos approximations to the matrix exponential. *SIAM J. Sci. Comput.*, 27(4):1438–1457, 2006.
[37] R. A. Horn and C. R. Johnson. *Topics in matrix analysis. 1st paperback ed. with corrections.* Cambridge University Press, Cambridge, 1994.
[38] A. Iserles, H. Z. Munthe-Kaas, S. P. Nørsett, and A. Zanna. Lie-group methods. *Acta Numerica*, 9:215–365, 2000.
[39] I. M. Jaimoukha and E. M. Kasenally. Krylov subspace methods for solving large Lyapunov equations. *SIAM J. Numer. Anal.*, 31(1):227–251, Feb. 1994.
[40] C. Kenney and A. J. Laub. Rational iterative methods for the matrix sign function. *SIAM J. Matrix Anal. Appl.*, 12(2):273–291, 1991.
[41] C. Kenney and A. J. Laub. On scaling Newton's method for polar decomposition and the matrix sign function. *SIAM J. Matrix Anal. Appl.*, 13(3):688–706, 1992.
[42] C. S. Kenney and A. J. Laub. The matrix sign function. *IEEE Trans. Autom. Control*, 40(8):1330–1348, 1995.
[43] P. Lancaster and L. Rodman. *The Algebraic Riccati Equation*. Oxford University Press, Oxford, 2nd edition edition, 1995.
[44] J.-R. Li and J. White. Low-Rank solutions of Lyapunov equations. *SIAM Review*, 46(4):693–713, 2004.
[45] Charles Van Loan. The sensitivity of the matrix exponential. *SIAM J. Numer. Anal.*, 14(6):971–981, 1977.
[46] 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.
[47] L. Lopez and V. Simoncini. Preserving geometric properties of the exponential matrix by block Krylov subspace methods. Technical report, Dipartimento di Matematica, Università di Bologna, December 2005.
[48] The MathWorks, Inc. *MATLAB 7*, September 2004.
[49] C. Moler and C. Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review*, 45(1):3–49, 2003.
[50] I. Moret and P. Novati. An interpolatory approximation of the matrix exponential based on Faber polynomials. *J. Comput. and Applied Math.*, 131:361–380, 2001.
[51] I. Moret and P. Novati. RD-rational approximations of the matrix exponential. *BIT, Numerical Mathematics*, 44(3):595–615, 2004.
[52] I. Moret and P. Novati. Interpolating functions of matrices on zeros of quasi-kernel polynomials. *Numer. Linear Algebra Appl.*, 11(4):337–353, 2005.
[53] H. Neuberger. The overlap Dirac operator. In *Numerical Challenges in Lattice Quantum Chromodynamics*, volume 15 of *Lect. Notes Comput. Sci. Eng.*, pages 1–17. Springer, Berlin, 2000.
[54] S. P. Nørsett. Restricted Padé approximations to the exponential function. *SIAM J. Numer. Anal.*, 15(5):1008–1029, Oct. 1978.
[55] T. Penzl. A cyclic low-rank Smith method for large sparse Lyapunov equations. *SIAM J. Sci. Comput.*, 21(4):1401–1418, 2000.
[56] P. P. Petrushev and V. A. Popov. *Rational approximation of real functions*. Cambridge University Press, Cambridge, 1987.
[57] J. D. Roberts. Linear model reduction and solution of the algebraic Riccati equation by use of the sign function. *Int. J. Control*, 32:677–687, 1980.
[58] Y. Saad. Numerical solution of large Lyapunov equations. In M. A. Kaashoek, J. H. van Schuppen, and A. C. Ran, editors, *Signal Processing, Scattering, Operator Theory, and Numerical Methods. Proceedings of the international symposium MTNS-89, vol III*, pages 503–511, Boston, 1990. Birkhauser.
[59] Y. Saad. Analysis of some Krylov subspace approximations to the matrix exponential operator. *SIAM J. Numer. Anal.*, 29:209–228, 1992.
[60] Y. Saad. *Iterative Methods for Sparse Linear Systems*. The PWS Publishing Company, Boston, 1996. Second edition, SIAM, Philadelphia, 2003.
[61] R. B. Sidje. Expokit: A Software Package for Computing Matrix Exponentials. *ACM Transactions on Math. Software*, 24(1):130–156, 1998.
[62] V. Simoncini and D. B. Szyld. Recent developments in Krylov subspace methods for linear systems. Technical report, Dipartimento di Matematica, Università di Bologna, September 2005. Submitted.
[63] D. E. Stewart and T. S. Leyk. Error estimates for Krylov subspace approximations of matrix

exponentials. *J. Comput. and Applied Math.*, 72:359–369, 1996.

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

[65] L. N. Trefethen, J. A. C. Weideman, and T. Schmelzer. Talbot Quadratures and Rational Approximations. *BIT*, to appear.

[66] 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.

[67] H.A. van der Vorst. An iterative solution method for solving $f(A)x = b$, using Krylov subspace information obtained for the symmetric positive definite matrix A. *J. Comput. Appl. Math.*, 18:249–263, 1987.

[68] Henk A. van der Vorst. Solution of $f(A)x = b$ with projection methods for the matrix A. In *Numerical Challenges in Lattice Quantum Chromodynamics*, volume 15 of *Lect. Notes Comput. Sci. Eng.*, pages 18–28. Springer, Berlin, 2000.

[69] R. S. Varga. On higher order stable implicit methods for solving parabolic partial differential equations. *J. of Mathematics and Physics*, XL:220–231, 1961.

[70] A. Zanna and H. Z. Munthe-Kaas. Generalized polar decompositions for the approximation of the matrix exponential. *SIAM J. Matrix Analysis and Appl.*, 23(3):840–862, 2002.