SIAM J. MATRIX ANAL. APPL.
Vol. 42, No. 1, pp. 83–107

A COMPARISON OF LIMITED-MEMORY KRYLOV METHODS FOR STIELTJES FUNCTIONS OF HERMITIAN MATRICES*

STEFAN GÜTTEL† AND MARCEL SCHWEITZER‡

Abstract. Given a limited amount of memory and a target accuracy, we propose and compare several polynomial Krylov methods for the approximation of $f(A)\boldsymbol{b}$, the action of a Stieltjes matrix function of a large Hermitian matrix on a vector. Using new error bounds and estimates, as well as existing results, we derive predictions of the practical performance of the methods and rank them accordingly. As byproducts, we derive new results on inexact Krylov iterations for matrix functions in order to allow for a fair comparison of rational Krylov methods with polynomial inner solves.

Key words. matrix function, Krylov method, shift-and-invert method, restarted method, Stieltjes function, inexact Krylov method, outer-inner iteration

AMS subject classifications. 65F60, 65F50, 65F10, 65F30

DOI. 10.1137/20M1351072

1. Introduction.

In recent years considerable progress has been made in the development of numerical methods for the efficient approximation of $f(A)\boldsymbol{b}$, the action of a matrix function $f(A)$ on a vector $\boldsymbol{b}$. In applications, the matrix $A \in \mathbb{C}^{N\times N}$ is typically large and sparse, and the computation of the generally dense matrix $f(A)$ is infeasible. One therefore seeks to approximate $f(A)\boldsymbol{b}$ directly by means of some iterative method. By far the most popular methods for this task are polynomial [15, 42] or rational [16, 26, 27, 49] Krylov methods. In the latter class of methods is, in particular, the popular extended Krylov subspace method [16, 36] which utilizes matrix-vector products and linear system solves with the matrix A. In cases where the matrix size is such that direct solution methods for (shifted) linear systems with A are feasible, or in cases where a good preconditioner is available to solve such problems iteratively, rational Krylov methods can significantly outperform polynomial methods. On the other hand, even if applicable, rational Krylov methods can be somewhat more difficult to tune as generally more parameters need to be chosen to obtain fast convergence. There are also variants of rational Krylov methods for $f(A)\boldsymbol{b}$ that choose their shift parameters automatically based on some heuristics (see, e.g., [17, 27]), but then there is very little theory that governs their convergence.

In this work we investigate which polynomial Krylov methods are best suited *when A is Hermitian and the only feasible operations involving this matrix are matrix-vector products.* This might be the case, e.g., when direct solvers are inefficient due to A's sparsity structure or if A is only implicitly available through a routine that returns the result of a matrix-vector product. We further assume that memory is limited so that only a predefined number $m_{\max}$ of vectors of size N can be stored. This situation

*Received by the editors July 7, 2020; accepted for publication (in revised form) by J. Liesen November 2, 2020; published electronically January 21, 2021.
https://doi.org/10.1137/20M1351072
Funding: The work of the first author was supported by the Alan Turing Institute under the EPSRC grant EP/N510129/1.
†Department of Mathematics, the University of Manchester, M13 9PL Manchester, United Kingdom (stefan.guettel@manchester.ac.uk).
‡Mathematisch-Naturwissenschaftliche Fakultät, Heinrich-Heine-Universität Düsseldorf, Universitätsstraße 1, 40225 Düsseldorf, Germany (marcel.schweitzer@hhu.de).

arises in many applications, including lattice quantum chromodynamics [9, 11, 19] and statistical sampling [32, 33, 47].

The basis of polynomial Krylov methods for Hermitian matrices is the *Lanczos method* [37] (which corresponds to the *Arnoldi method* [2] in the non-Hermitian case). Applying m Lanczos iterations with A and $\boldsymbol{b}$ yields the *Lanczos relation*

$$AV_m = V_m T_m + \beta_{m+1} \boldsymbol{v}_{m+1} \boldsymbol{e}_m^T \tag{1.1}$$

with $V_m = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m] \in \mathbb{C}^{N\times m}$ containing an orthonormal basis of the Krylov space $\mathcal{K}_m(A, \boldsymbol{b}) := \operatorname{span}\{\boldsymbol{b}, A\boldsymbol{b}, A^2\boldsymbol{b}, \ldots, A^{m-1}\boldsymbol{b}\}$, a symmetric tridiagonal matrix

$$T_m = \begin{bmatrix} \eta_1 & \beta_2 & & & \\ \beta_2 & \eta_2 & \beta_3 & & \\ & \ddots & \ddots & \ddots & \\ & & \beta_{m-1} & \eta_{m-1} & \beta_m \\ & & & \beta_m & \eta_m \end{bmatrix} \in \mathbb{R}^{m\times m},$$

and $\boldsymbol{e}_m$ denoting the mth canonical unit vector in $\mathbb{R}^m$. The *Lanczos approximation* $\boldsymbol{f}_m \approx f(A)\boldsymbol{b}$ is obtained by projecting the original problem onto the Krylov space,

$$\boldsymbol{f}_m := V_m f(V_m^H A V_m) V_m^H \boldsymbol{b} = \|\boldsymbol{b}\| V_m f(T_m) \boldsymbol{e}_1, \tag{1.2}$$

where $\|\cdot\|$ denotes the Euclidean vector norm. The evaluation of (1.2) requires the storage of the full Lanczos basis V_m, i.e., m vectors of length N. In the situation described above, the value $m_{\max}$ which limits the number of vectors that can be stored therefore limits the maximum number of iterations that can be performed and thus also the attainable accuracy of the Lanczos approximation. This is different from the situation for Hermitian linear systems, where the short recurrence for the Krylov basis vectors translates into a short recurrence for the iterates, resulting in the famous conjugate gradient method [31].

There are several approaches for overcoming the memory problem, including

- the two-pass Lanczos method (see, e.g., [10, 24]) which overcomes memory limitations but roughly doubles the computational effort;
- the multi-shift CG method [19, 23], which replaces f by a rational approximation and then simultaneously solves the resulting linear systems using the short recurrence conjugate gradient method;
- restarted Krylov methods [1, 18, 21, 22, 33, 45, 48] which, similar to restarted methods for linear systems, construct a series of Krylov iterates in such a way that each "cycle" of the method only requires a fixed amount of storage.

We also refer the reader to the recent survey [28] covering limited-memory polynomial methods for the $f(A)\boldsymbol{b}$ problem.

Another approach, which has not been considered in this context in the literature so far, is to use rational Krylov methods [8, 26, 27] combined with an iterative short recurrence solver for the associated linear systems. It appears to be generally thought that using a polynomial Krylov solver inside a rational Krylov method is not sensible, because the approximation computed is then polynomial and hence could also be computed with a polynomial Krylov method alone. While this is true in theory, an outer-inner rational-polynomial Krylov method may still be interesting in our setting: if the overall number of outer iterations is small, the number of vectors to be stored is small (as the inner iteration uses a short recurrence), and hence the method could be a viable alternative to restarting strategies. Thus, we pose the following question:

Given a limited amount of memory (storage of at most $m_{\max}$ vectors of length N) and a target accuracy ε, what is an efficient way to extract an accurate approximation to $f(A)b$ from a polynomial Krylov space?

Of course, "efficient" can have several meanings like, e.g., "small number of matrix-vector products and inner products" or "low overall computation time." The latter criterion is highly dependent on the specific implementation of each method and also the hardware environment (e.g., parallel/distributed computing) and hence difficult to asses given only information on $f, A, \boldsymbol{b}$; see also [12] for a discussion of such difficulties. We take a more general, implementation-independent approach here by exploiting the considerable progress that has recently been made in the understanding of (restarted) Krylov methods for the $f(A)\boldsymbol{b}$ problem [21, 22, 45]. Together with the very well-understood convergence behavior of the conjugate gradient method (see, e.g., [43]), this opens up the possibility to assess and compare the efficiency of different algorithmic variants using upper error bounds. We hope that this theoretical work will serve as a starting point for a more practice-oriented comparison of the different algorithms for

(a) investigating the potential for efficient parallelization and tuning and
(b) comparing our theoretical estimates of iteration numbers to the real numbers occurring when solving different real-world problems.

Most of the available theoretical results mentioned above apply to the class of *Stieltjes functions*

$$f(z) = \int_0^\infty \frac{1}{t+z}\,\mathrm{d}\mu(t), \tag{1.3}$$

where $\mu(t)$ is a monotonically increasing and nonnegative measure on $[0,\infty)$ and $\int_0^\infty (t+1)^{-1}\,\mathrm{d}\mu(t) < \infty$; see, e.g., [6, 7, 30]. The latter condition ensures that $f(z)$ is finite for all $z > 0$. Important examples of Stieltjes functions include $f(z) = z^{-\alpha}$ with $\alpha \in (0,1)$ and $f(z) = \log(1+z)/z$. Stieltjes matrix functions are closely related to shifted linear systems (see, e.g., [21]), which allows us to transfer many theoretical results well-known for linear systems, like the classical CG convergence bound. In particular, the following theorem is central to many developments in this paper. The A-norm used in the statement of the theorem is defined as $\|\boldsymbol{x}\|_A := \sqrt{\boldsymbol{x}^H A \boldsymbol{x}}$.

THEOREM 1.1 (see, e.g., [43]). *Let $A \in \mathbb{C}^{N\times N}$ be Hermitian positive definite and $\boldsymbol{x}_0, \boldsymbol{b} \in \mathbb{C}^N$. Further, let $\boldsymbol{x}^*$ denote the solution of the linear system $A\boldsymbol{x} = \boldsymbol{b}$, and let $\boldsymbol{x}_m$ be the mth CG iterate with initial guess $\boldsymbol{x}_0$. Let $\lambda_{\min}$ and $\lambda_{\max}$ denote the smallest and largest eigenvalue of A, respectively, and let $\kappa = \frac{\lambda_{\max}}{\lambda_{\min}}$ denote the condition number of A. Define*

$$c = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \quad \text{and} \quad \alpha_m = \frac{1}{\cosh(m \ln c)}$$

(where we set $\alpha_m = 0$ if $\kappa = 1$). Then the error in the CG method satisfies

$$\|\boldsymbol{x}^* - \boldsymbol{x}_m\|_A \le \alpha_m \|\boldsymbol{x}^* - \boldsymbol{x}_0\|_A.$$

The remainder of this paper is structured as follows. In section 2 we give a survey of the different established polynomial methods for approximating $f(A)\boldsymbol{b}$, together with their (worst-case) convergence bounds. Sections 3 and 4 introduce new combinations of outer-inner rational-polynomial Krylov methods, namely, an inexact shift-and-invert method [38, 40, 49] and an inexact extended Krylov method [16, 36]

with polynomial Krylov solvers. We provide a new convergence analysis for the shift-and-invert method and discuss ways to relax the inner iterations of the proposed methods. In section 5 we use the obtained convergence results to estimate the total arithmetic cost of each of the considered methods and discuss general advantages, disadvantages, and prerequisites of each of the methods. The theoretical estimates are compared to real iteration counts for some (artificial) test problems. Concluding remarks and topics for future research are given in section 6.

2. Polynomial limited-memory Krylov methods. As the problem of approximating functions of very large, sparse matrices arises frequently in applications, several different strategies have been developed to overcome the problem of scarce memory. We briefly describe three established Krylov methods, together with theoretical results on their convergence behavior.

2.1. Two-pass Lanczos. The two-pass Lanczos method [10, 24] is a very simple approach that solves the scarce memory problem by applying the Lanczos process twice. Of course, this doubles the number of matrix-vector products and inner products that need to be computed.

In the first pass of the Lanczos method one computes the compressed matrix T_m and discards the basis vectors in V_m as soon as they are no longer needed to compute the next basis vector (i.e., only storing the last three basis vectors). Then, the coefficient vector $\boldsymbol{y}_m := \|\boldsymbol{b}\| f(T_m)\boldsymbol{e}_1$ is computed. In the second pass, the Lanczos approximation is computed as

$$\boldsymbol{f}_m = \sum_{i=1}^{m} [\boldsymbol{y}_m]_i \boldsymbol{v}_i,$$

which can be updated from one iteration to the next and thus also allows us to discard old basis vectors. This approach produces the same iterates as the standard Lanczos process but requires twice the number of matrix-vector products. When f is a Stieltjes function, the convergence of the two-pass Lanczos method is thus characterized by the following theorem from [21]. It is a special case of a more general result for the restarted Lanczos method; cf. Theorem 2.6 below.

THEOREM 2.1 (see [21, Corollary 4.4]). *Let $A \in \mathbb{C}^{N\times N}$ be Hermitian positive definite, $\boldsymbol{b} \in \mathbb{C}^N$, f a Stieltjes function* (1.3)*, and let $\boldsymbol{f}_m$ be the approximation to $f(A)\boldsymbol{b}$ after m iterations of the Lanczos method. Further, let*

$$\kappa(t) = \frac{\lambda_{\max}+t}{\lambda_{\min}+t},\ c(t) = \frac{\sqrt{\kappa(t)}-1}{\sqrt{\kappa(t)}+1},\ \text{ and } \alpha_m(t) = \frac{1}{\cosh(m \ln c(t))}, \tag{2.1}$$

and let $t_0 \geq 0$ be the left endpoint of the support of μ. Then

$$\|f(A)\boldsymbol{b} - \boldsymbol{f}_m\|_A \leq C\alpha_m(t_0), \tag{2.2}$$

where

$$C = \|\boldsymbol{b}\|\sqrt{\lambda_{\max}} \cdot f\big(\sqrt{\lambda_{\min}\lambda_{\max}}\big). \tag{2.3}$$

We remark that the bound (2.2) is, up to the factor C, the same as the standard textbook CG convergence bound for the linear system $(A+t_0 I)\boldsymbol{x} = \boldsymbol{b}$. Also, note that if additional computational work is invested into the computation of error bounds or error estimates during the first pass of the Lanczos method, then this work can of course be avoided in the second pass, as the number of iterations required to reach the desired accuracy is already known.

Remark 2.2. In a massively parallel setting, an alternative to using a two-pass Lanczos approach is to compute $f(A)\boldsymbol{b}$ elementwise, preferably by using the relation

$$[f(A)\boldsymbol{b}]_i = \boldsymbol{e}_i^H f(A)\boldsymbol{b} = \frac{1}{4}(\boldsymbol{b}+\boldsymbol{e}_i)^H f(A)(\boldsymbol{b}+\boldsymbol{e}_i) - \frac{1}{4}(\boldsymbol{b}-\boldsymbol{e}_i)^H f(A)(\boldsymbol{b}-\boldsymbol{e}_i). \tag{2.4}$$

Thus, the full vector $f(A)\boldsymbol{b}$ can be formed by evaluating $2N$ bilinear forms $\boldsymbol{v}^H f(A)\boldsymbol{v}$. If at least $2N$ processors are available—and if energy consumption is not an issue—one can perfectly parallelize the computation of $f(A)\boldsymbol{b}$ in this way. This approach then has several upsides: Most importantly, when approximating bilinear forms $\boldsymbol{v}^H f(A)\boldsymbol{v}$ by the Lanczos process, it is not necessary to store the Lanczos basis. Additionally, due to the relation to Gaussian quadrature it can be expected that the necessary number of iterations is about half the number of iterations that would be required for approximating $f(A)\boldsymbol{v}$, and moreover, this approach is less prone to instability due to rounding errors; see, e.g., [25]. Thus, in the mentioned setting, an approach based on (2.4) can be expected to outperform the other methods that we discuss in this paper. ◇

2.2. Multi-shift CG. The multi-shift CG method [19, 23, 24] for $f(A)\boldsymbol{b}$ is based on first approximating f by a suitable rational function r, $r(A)\boldsymbol{b} \approx f(A)\boldsymbol{b}$, most commonly in partial fraction form

$$r(z) = \sum_{i=1}^{p} \omega_i \frac{1}{t-\zeta_i} \quad \text{or} \quad r(z) = \sum_{i=1}^{p} \omega_i \frac{t}{t^2-\zeta_i}, \tag{2.5}$$

where we assume that all poles ζ_i lie on the negative real axis. Then

$$f(A)\boldsymbol{b} \approx \sum_{i=1}^{p} \omega_i (A-\zeta_i I)^{-1}\boldsymbol{b} \quad \text{or} \quad f(A)\boldsymbol{b} \approx A\sum_{i=1}^{p} \omega_i (A^2-\zeta_i I)^{-1}\boldsymbol{b}, \tag{2.6}$$

respectively, which approximate $f(A)\boldsymbol{b}$ as the weighted sum of solutions of shifted linear systems with A or A^2. Rational functions of the former form in (2.5) arise as $[p-1,p]$-type Padé approximants of Stieltjes functions or the matrix exponential (see, e.g., [29]), while functions of the latter form naturally originate in the Zolotarev approximation of the sign function (see, e.g., [19]).

The main trick for efficiently evaluating (2.6) is the *shift-invariance* of Krylov spaces, that is, $\mathcal{K}_m(A,\boldsymbol{b}) = \mathcal{K}_m(A+\zeta I,\boldsymbol{b})$ for all $\zeta \in \mathbb{C}$. Therefore, when started with an initial guess $\boldsymbol{x}_0(\zeta) = \boldsymbol{0}$ for all shifted systems, the Krylov spaces from which the conjugate gradient method extracts its approximants for the different systems all coincide. This can be exploited by simultaneously solving all shifted systems in the same iterative process, requiring the same number of matrix-vector products and inner products as the solution of a single system. The multi-shift CG implementation proposed in [23] results in the following computational and memory overhead compared to the standard Lanczos method (ignoring negligible scalar operations):

(i) Two vectors of length N need to be stored for each pole of the rational approximation, i.e., $2p$ additional vectors overall.
(ii) In each iteration, two vector additions and three vector scalings are performed (equaling to about two and a half inner products in cost) for each pole of the rational approximation, totaling to $2.5mp$ additional inner products.
(iii) After performing the multi-shift CG method, the iterates of the individual systems need to be combined according to (2.6). This results in an additional number of p vector scalings and $(p-1)$ vector additions (and, in case of a

rational approximation of the later form in (2.5), one additional matrix-vector product), which equals to about p additional inner products.

Remark 2.3. In [19] it is proposed to not store the iterates of the individual systems and then combine them at the end, but instead to directly combine them in each iteration to update the approximation to $f(A)\boldsymbol{b}$. This way, only one additional vector needs to be stored per system, but the computational effort of roughly $2p$ vector additions and scalings is required in each iteration of the multi-shift CG method. ⋄

The (worst-case) speed of convergence of the multi-shift CG method is also described by the classical textbook CG convergence bound, with the worst-conditioned system (i.e., the one corresponding to the pole ζ_i closest to zero) determining the overall necessary number of iterations. When one of the shifts is very close to zero, as it is typically the case for rational approximations of Stieltjes functions, the resulting convergence factor is thus approximately equal to that of the standard Lanczos method given in Theorem 2.1.

The overall computational cost and storage requirements of the multi-shift CG method are determined by the number p of poles of the rational approximation. These of course depend on the overall accuracy to which one wants to approximate $f(A)\boldsymbol{b}$. The overall error of the multi-shift CG approximation $\boldsymbol{f}_m^{MS}$, $\|f(A)\boldsymbol{b} - \boldsymbol{f}_m^{MS}\|$, depends both on the accuracy of the rational approximation and the accuracy to which the shifted linear systems are solved. When aiming for an overall accuracy of ε, a (straightforward) approach is to construct a rational function r such that $|f(z) - r(z)| \leq \varepsilon/(2\|\boldsymbol{b}\|)$ for $z \in \operatorname{spec}(A)$, the spectral interval of A, and then solve the shifted linear systems accurately enough for fulfilling

$$\|r(A)\boldsymbol{b} - \boldsymbol{f}_m^{MS}\| \leq \frac{\varepsilon}{2} \tag{2.7}$$

such that overall

$$\begin{aligned}\|f(A)\boldsymbol{b} - \boldsymbol{f}_m^{MS}\| &\leq \|f(A)\boldsymbol{b} - r(A)\boldsymbol{b}\| + \|r(A)\boldsymbol{b} - \boldsymbol{f}_m^{MS}\| \\ &\leq \max_{z\in\operatorname{spec}(A)} |f(z) - r(z)|\,\|\boldsymbol{b}\| + \frac{\varepsilon}{2} \leq \varepsilon.\end{aligned}$$

Remark 2.4. Typically, systems associated with poles of large magnitude converge significantly faster than those corresponding to poles close to zero. Therefore one can employ a strategy for "removing" already converged systems from the iteration in order to not perform superfluous computations. Strategies for doing this without violating the condition (2.7) are discussed in [19, section 5.3] in the context of approximating the action of the matrix sign function. ⋄

2.3. Restarted Lanczos. The idea of the restarted Lanczos method is to first compute an approximation (1.2) obtained from $m < m_{max}$ iterations of the standard Lanczos process, which we now denote by $\boldsymbol{f}_m^{(1)}$, where the superscript is used to distinguish quantities belonging to different *restart cycles*. The second cycle of the method then consists of an additive update $\boldsymbol{f}_m^{(2)} = \boldsymbol{f}_m^{(1)} + \boldsymbol{e}_m^{(1)}$, where $\boldsymbol{e}_m^{(1)}$ is an approximation of the error $f(A)\boldsymbol{b} - \boldsymbol{f}_m^{(1)}$ obtained by m new Lanczos iterations. Repeatedly applying this approach yields a sequence of approximations

$$\boldsymbol{f}_m^{(k)} = \boldsymbol{f}_m^{(k-1)} + \boldsymbol{e}_m^{(k-1)}, \quad k = 2, 3, \ldots$$

for $f(A)\boldsymbol{b}$. In order to use the Lanczos method for approximating the error $f(A)\boldsymbol{b} - \boldsymbol{f}_m^{(1)}$, it is necessary to be able to represent it in the form

$$e_m^{(1)}(A)\boldsymbol{v}^{(1)},$$

with a new function $e_m^{(1)}(z)$ and a new vector $\boldsymbol{v}^{(1)}$. First results in this direction were given in [18, 33], characterizing the restart function $e_m^{(1)}(z)$ as the mth order divided difference [13] of $f(z)$ with respect to the Ritz values, i.e., the eigenvalues of T_m. However, this error function representation turned out to be numerically unstable. We therefore cite a result from [21, 22] which gives an integral representation for the error which is numerically stable and in addition useful for deriving theoretical results on the convergence of the restarted Lanczos method.

THEOREM 2.5 (see [21, Theorem 2.1]). *Let f be a Stieltjes function as in* (1.3). *Assume* $\operatorname{spec}(A) \cap (-\infty, 0] = \emptyset$, *and denote by $\boldsymbol{f}_m$ the approximation* (1.2) *to $f(A)\boldsymbol{b}$. Assume that* $\operatorname{spec}(T_m) = \{\theta_1, \ldots, \theta_m\}$ *satisfies* $\operatorname{spec}(T_m) \cap (-\infty, 0] = \emptyset$, *and define*

$$e_m(z) := (-1)^{m+1}\|\boldsymbol{b}\|\gamma_m \int_0^\infty \frac{1}{w_m(t)} \cdot \frac{1}{z+t}\, \mathrm{d}\mu(t), \quad z \notin (-\infty, 0],$$

where $w_m(t) = (t+\theta_1)\cdots(t+\theta_m)$ *and* $\gamma_m = \prod_{i=1}^m \beta_{i+1}$. *Then*

$$f(A)\boldsymbol{b} - \boldsymbol{f}_m = e_m(A)\boldsymbol{v}_{m+1},$$

where $\boldsymbol{v}_{m+1}$ is the $(m+1)$st Lanczos vector.

Theorem 2.5 recursively also holds for the Lanczos approximation resulting after a restart cycle because $e_m(z)$ is itself (a scalar multiple of) a Stieltjes function; see [21, Proposition 2.2]. As a consequence, the error representation can be used to obtain a restarted Lanczos method with an arbitrary number of restart cycles. In [22, 45] a version of this method was introduced which evaluates the error function $e_m(z)$ using adaptive numerical quadrature. The following theorem (a more general version of Theorem 2.1) gives an upper bound on the error of the restarted Lanczos method.

THEOREM 2.6 (see [21, Theorem 4.3]). *Let $A \in \mathbb{C}^{N\times N}$ be Hermitian positive definite, $\boldsymbol{b} \in \mathbb{C}^N$, f a Stieltjes function* (1.3), *and $\boldsymbol{f}_m^{(k)}$ the approximation from k cycles of the restarted Lanczos method with restart length m. Further, let $\alpha_m(t)$ be defined as in* (1.1) *and let $t_0 \geq 0$ be the left endpoint of the support of μ. Then*

$$\|f(A)\boldsymbol{b} - \boldsymbol{f}_m^{(k)}\|_A \leq C\alpha_m(t_0)^k,$$

where C is as in (2.3) *and $0 \leq \alpha_m(t_0) < 1$. In particular, the restarted Arnoldi method converges for all restart lengths $m \geq 1$.*

3. Rational Krylov methods for Stieltjes function. In this section, we discuss two rational Krylov methods, namely, the shift-and-invert Lanczos method and the extended Krylov method, for the approximation of Stietljes functions. The convergence of the shift-and-invert method is analyzed in detail. We also briefly comment on iteratively solving the linear systems arising in each iteration.

3.1. Shift-and-invert Lanczos. The shift-and-invert Lanczos method was introduced in [49] for "preconditioning" Lanczos iterations for the matrix exponential times a vector; see also [38] for related work. The main idea is to replace A by a

matrix B with more favorable spectral properties, leading to faster convergence to $f(A)\boldsymbol{b}$. In the following we give a short description of this method.

First, define $B = (A - \xi I)^{-1}$, where $\xi \in \mathbb{R}^-$ is an arbitrary, negative shift. How to best choose this parameter is discussed later in this subsection. Applying m steps of the Lanczos method to B with starting vector $\boldsymbol{b}$, we obtain the relation

$$BV_m^{SI} = V_m^{SI} T_m^{SI} + \beta_m^{SI} \boldsymbol{v}_{m+1}^{SI} \boldsymbol{e}_m^T, \tag{3.1}$$

where V_m^{SI} now contains an orthonormal basis of $\mathcal{K}_m(B, \boldsymbol{b})$. To extract an approximation for $f(A)\boldsymbol{b}$ from $\mathcal{K}_m(B, \boldsymbol{b})$, we consider the transformed function

$$g(y) := f(y^{-1} + \xi) \text{ where } y = (z - \xi)^{-1}. \tag{3.2}$$

We have $f(A)\boldsymbol{b} = g(B)\boldsymbol{b}$ and define the *standard* shift-and-invert approximation as

$$\boldsymbol{g}_m^{SI} := \|\boldsymbol{b}\| V_m^{SI} g(T_m^{SI}) \boldsymbol{e}_1 \approx f(A)\boldsymbol{b}.$$

For reasons that will become apparent when deriving the convergence bounds later in this section, we propose to instead use the so-called *corrected* Lanczos approximation (first introduced in [42] for the approximation of φ-functions)

$$\hat{\boldsymbol{g}}_m^{SI} := \|\boldsymbol{b}\| V_m^{SI} g(T_m^{SI}) \hat{\boldsymbol{e}}_1 + \beta_{m+1}^{SI} \left(\boldsymbol{e}_m^T g\left(T_m^{SI}\right) \boldsymbol{e}_1 \right) \boldsymbol{v}_{m+1}^{SI} \approx f(A)\boldsymbol{b}. \tag{3.3}$$

When f is a Stieltjes function (1.3), then the function g defined in (3.2) clearly admits an integral representation

$$g(y) = y \int_0^\infty \frac{1}{1 + (\xi + t)y} \, \mathrm{d}\mu(t). \tag{3.4}$$

Using (3.4), we find the representation

$$f(A)\boldsymbol{b} - \hat{\boldsymbol{g}}_m^{SI} = e_m(B) \boldsymbol{v}_{m+1}^{SI}, \tag{3.5}$$

$$e_m(y) := (-1)^{m+1} \|\boldsymbol{b}\| \gamma_m y \int_0^\infty \frac{1}{w_m(t)} \cdot \frac{1}{1 + (\xi + t)y} \, \mathrm{d}\mu(t), \quad y \in (0, -1/\xi), \tag{3.6}$$

for the error of the corrected shift-and-invert approximation (3.3), where γ_m and w_m are as in Theorem 2.5. In other words,

$$\begin{aligned} e_m(B) \boldsymbol{v}_{m+1}^{SI} &= (-1)^{m+1} \|\boldsymbol{b}\| \gamma_m B \int_0^\infty \frac{1}{w_m(t)} \cdot (I + (\xi + t)B)^{-1} \boldsymbol{v}_{m+1}^{SI} \, \mathrm{d}\mu(t) \\ &= B \int_0^\infty \boldsymbol{e}_m(t) \, \mathrm{d}\mu(t), \end{aligned} \tag{3.7}$$

where $\boldsymbol{e}_m(t)$ is the error of the mth CG approximation to the linear system $(I + (\xi + t)B)\boldsymbol{x}(t) = \boldsymbol{b}$. This follows from the fact that

$$\frac{(-1)^{m+1} \gamma_m \|\boldsymbol{b}\|}{w_m(t)} \boldsymbol{v}_{m+1}^{SI} = \boldsymbol{r}_m(t)$$

is the residual of the mth CG approximation $\boldsymbol{x}_m(t)$ for that linear system.

In the following, we use the error function representation (3.6) to derive results on the speed of convergence of the shift-and-invert method and on how to choose the shift ξ. Similar results have been obtained in [38], but we provide a different proof here which yields explicit constants in the bounds that were previously unavailable.

LEMMA 3.1. *Let $A \in \mathbb{C}^{N\times N}$ be Hermitian positive definite, $B = (A - \xi I)^{-1}$, $\boldsymbol{b} \in \mathbb{C}^N$, g a function of the form (1.3), and $\hat{\boldsymbol{g}}_m^{SI}$ as defined in (3.3). Let $\lambda_{\min}$ and $\lambda_{\max}$ denote the smallest and largest eigenvalue of A, respectively, and define*

(3.8)
$$\kappa_\xi(t) = \begin{cases} \frac{\lambda_{\max}+t}{\lambda_{\min}+t} \cdot \frac{\lambda_{\min}-\xi}{\lambda_{\max}-\xi} & \text{if } t \le -\xi, \\ \frac{\lambda_{\min}+t}{\lambda_{\max}+t} \cdot \frac{\lambda_{\max}-\xi}{\lambda_{\min}-\xi} & \text{if } t > -\xi, \end{cases} \quad c_\xi(t) = \frac{\sqrt{\kappa_\xi(t)}-1}{\sqrt{\kappa_\xi(t)}+1}, \quad \alpha_m^\xi(t) = \frac{1}{\cosh(m \ln c_\xi(t))}.$$

Then the norm of the error of $\hat{\boldsymbol{g}}_m^{SI}$ is bounded by

$$\begin{aligned} \|f(A)\boldsymbol{b} - \hat{\boldsymbol{g}}_m^{SI}\| \le \|\boldsymbol{b}\| \sqrt{\frac{\lambda_{\max}-\xi}{\lambda_{\min}-\xi}} \int_0^{-\xi} \frac{\alpha_m^\xi(t)}{\lambda_{\min}+t}\, \mathrm{d}\mu(t) \\ + \frac{\lambda_{\max}-\xi}{\lambda_{\min}-\xi} \int_{-\xi}^{\infty} \frac{\alpha_m^\xi(t)}{\sqrt{\lambda_{\min}+t}\sqrt{\lambda_{\max}+t}}\, \mathrm{d}\mu(t). \end{aligned} \tag{3.9}$$

Proof. By using (3.5) and (3.7), we can write $f(A)\boldsymbol{b} - \hat{\boldsymbol{g}}_m^{SI} = B\int_0^\infty \boldsymbol{e}_m(t)\,\mathrm{d}\mu(t)$, where $\boldsymbol{e}_m(t)$ denotes the error of the approximation $\boldsymbol{x}_m(t)$ from m steps of CG for the shifted linear system $(I + (\xi+t)B)\boldsymbol{x} = \boldsymbol{b}$. This yields

$$\begin{aligned} \|f(A)\boldsymbol{b} - \hat{\boldsymbol{g}}_m^{SI}\|_B &\le \|B\|_B \int_0^\infty \|\boldsymbol{e}_m(t)\|_B\, \mathrm{d}\mu(t) \\ &\le \frac{1}{\lambda_{\min}-\xi} \int_0^\infty \frac{1}{\sqrt{\lambda_{\min}+t}} \|\boldsymbol{e}_m(t)\|_{I+(\xi+t)B}\, \mathrm{d}\mu(t), \end{aligned}$$

where we used that $\|\boldsymbol{v}\|_B \le 1/\sqrt{\lambda_{\min}+t} \cdot \|\boldsymbol{v}\|_{I+(\xi+t)B}$ holds for all $t \in [0,\infty)$ and that $\|B\|_B = 1/(\lambda_{\min}-\xi)$. We now apply Theorem 1.1 for the shifted matrices $I + (\xi+t)B$, which are positive definite for $t \in [0,\infty)$. Note that $\kappa_\xi(t)$ is exactly the condition number of the shifted matrix $I + (\xi+t)B$. Applying the CG estimate for all t and using the fact that the initial guess is $\boldsymbol{x}_0(t) = \boldsymbol{0}$ for all t, we conclude that

$$\|f(A)\boldsymbol{b} - \hat{\boldsymbol{g}}_m^{SI}\|_B \le \frac{1}{\lambda_{\min}-\xi} \int_0^\infty \frac{\alpha_m^\xi(t)}{\sqrt{\lambda_{\min}+t}} \|\boldsymbol{x}^*(t)\|_{I+(\xi+t)B}\, \mathrm{d}\mu(t). \tag{3.10}$$

As $\boldsymbol{x}^*(t) = (I + (\xi+t)B)^{-1}\boldsymbol{b}$, a straightforward calculation shows that

$$\|\boldsymbol{x}^*(t)\|_{I+(\xi+t)B} \le \|\boldsymbol{b}\| \begin{cases} \frac{\sqrt{\lambda_{\min}-\xi}}{\sqrt{\lambda_{\min}+t}} & \text{if } t \le -\xi, \\ \frac{\sqrt{\lambda_{\max}-\xi}}{\sqrt{\lambda_{\max}+t}} & \text{if } t > -\xi. \end{cases} \tag{3.11}$$

∎

Inserting (3.11) into (3.10), using the fact that $\|\boldsymbol{v}\| \le \sqrt{\lambda_{\max}-\xi}\,\|\boldsymbol{v}\|_B$ for all $\boldsymbol{v} \in \mathbb{C}^N$, and splitting the integral at $-\xi$ completes the proof.

The error estimate (3.9) shows that the asymptotic convergence factor of the corrected Lanczos approximation for $g(B)\boldsymbol{b}$ will be determined by the largest asymptotic CG convergence factor $\alpha_m^\xi(t)$ across all shifts $t \in [0,\infty)$. According to (3.8), the values $\alpha_m^\xi(t)$ also depend on the shift ξ, and we will therefore now determine the value of ξ for which the maximum of $\alpha_m^\xi(t)$ becomes smallest possible.

For this, first note that c_ξ increases monotonically as a function of $\kappa_\xi(t)$ and α_m^ξ increases monotonically as a function of c. Therefore, $\alpha_m^\xi(t)$ attains its largest value

where $\kappa_\xi(t)$ attains its largest value. The function $\kappa_\xi(t)$ is monotonically decreasing on $[0,-\xi]$ and monotonically increasing on $[-\xi,\infty)$. Therefore,

$$\alpha_m^\xi(t) \le \max\{\kappa_\xi(0), \kappa_{\xi,\infty}\}, \text{ where } \kappa_{\xi,\infty} = \lim_{t\to\infty} \kappa_\xi(t). \tag{3.12}$$

It depends on the choice of the shift ξ which of the two values $\kappa_\xi(0)$ and $\kappa_{\xi,\infty}$ is larger. We have

$$\kappa_\xi(0) = \frac{\lambda_{\max}}{\lambda_{\min}} \cdot \frac{\lambda_{\min}-\xi}{\lambda_{\max}-\xi} \text{ and } \kappa_{\xi,\infty} = \frac{\lambda_{\max}-\xi}{\lambda_{\min}-\xi},$$

i.e., $\kappa_{\xi,\infty}$ is monotonically increasing in ξ and $\kappa_\xi(0)$ is monotonically decreasing in ξ. Hence the bound for $\alpha_m^\xi(t)$ in (3.12) is minimal if ξ is chosen so that $\kappa_\xi(0) = \kappa_{\xi,\infty}$, i.e.,

$$\frac{\lambda_{\max}}{\lambda_{\min}} \cdot \frac{\lambda_{\min}-\xi}{\lambda_{\max}-\xi} = \frac{\lambda_{\max}-\xi}{\lambda_{\min}-\xi}. \tag{3.13}$$

Equation (3.13) is solved by the shift

$$\xi = -\sqrt{\lambda_{\min}\cdot\lambda_{\max}}. \tag{3.14}$$

Using the shift (3.14), we find the following error bound.

THEOREM 3.2. *Let the assumptions of Lemma* 3.1 *hold, and let* $\xi = -\sqrt{\lambda_{\min}\lambda_{\max}}$. *Define the functions*

$$f_1(z) = \int_0^{-\xi} \frac{1}{z+t}\,\mathrm{d}\mu(t) \text{ and } f_2(z) = \int_{-\xi}^{0} \frac{1}{z+t}\,\mathrm{d}\mu(t). \tag{3.15}$$

Then

$$\|g(B)\boldsymbol{b} - \hat{\boldsymbol{g}}_m^{SI}\| \le \|\boldsymbol{b}\|\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}\left(\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} f_1(\lambda_{\min}) + f_2(\sqrt{\lambda_{\min}\lambda_{\max}})\right)\alpha_m(0).$$

Proof. By the considerations above, we have that for $\xi = -\sqrt{\lambda_{\min}\lambda_{\max}}$ the CG convergence factors fulfill $\alpha_m^\xi(0) \ge \alpha_m^\xi(t)$ for all $t \in [0,\infty)$. In addition, we have

$$\frac{\lambda_{\max}-\xi}{\lambda_{\min}-\xi} = \frac{\sqrt{\lambda_{\max}}\left(\sqrt{\lambda_{\max}}+\sqrt{\lambda_{\min}}\right)}{\sqrt{\lambda_{\min}}\left(\sqrt{\lambda_{\max}}+\sqrt{\lambda_{\min}}\right)} = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}. \tag{3.16}$$

Using the fact that $\sqrt{\lambda_{\min}+t}\sqrt{\lambda_{\max}+t} \le \sqrt{\lambda_{\min}\lambda_{\max}} + t$ after inserting (3.16) into (3.9) concludes the proof. ∎

3.2. Extended Krylov. The extended Krylov subspace method [16, 35, 36], like the shift-and-invert method, is a special case of a rational Krylov method. Here the approximations to $f(A)\boldsymbol{b}$ are extracted from an extended Krylov space

$$\mathcal{EK}_m(A,\boldsymbol{b}) = \operatorname{span}\{\boldsymbol{b}, A\boldsymbol{b}, \ldots, A^{m-1}\boldsymbol{b}, A^{-1}\boldsymbol{b}, \ldots, A^{-m}\boldsymbol{b}\}.$$

One iteration of the method involves adding one basis vector from the "positive" and one vector from the "negative" sequence. A five-term recursion for the basis vectors was first derived in [46]. Recursion relations for more general extended Krylov

sequences are treated in [34, 35]. The method yields an *extended Lanczos decomposition*

$$AV_m^{EK} = V_m^{EK} T_m^{EK} + [\boldsymbol{v}_{2m+1}, \boldsymbol{v}_{2m+2}]\tau_m [\boldsymbol{e}_{2m-1}, \boldsymbol{e}_{2m}]^T,$$

where $V_m^{EK} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{2m}] \in \mathbb{C}^{N\times 2m}$ contains an orthonormal basis of $\mathcal{EK}_m(A, \boldsymbol{b})$, $T_m^{EK} = (V_m^{EK})^H A V_m^{EK} \in \mathbb{C}^{2m\times 2m}$, and $\tau_m = [\boldsymbol{v}_{2m+1}, \boldsymbol{v}_{2m+2}]^H A [\boldsymbol{v}_{2m+1}, \boldsymbol{v}_{2m+2}] \in \mathbb{C}^{2\times 2}$. An approximation for $f(A)\boldsymbol{b}$ can then be obtained by projection onto the extended Krylov space in the usual way, i.e.,

$$f(A)\boldsymbol{b} \approx \boldsymbol{f}_m^{EK} = \|\boldsymbol{b}\| V_m^{EK} f(T_m^{EK}) \boldsymbol{e}_1. \tag{3.17}$$

Using the algorithmic approach from [46], one iteration of the extended Krylov subspace method requires one matrix-vector product with A, the solution of one linear system with A, and the computation of six inner products/vector norms in the orthonormalization process. The matrix T_m^{EK} can be cheaply computed from the orthonormalization coefficients without any further inner products [34].

The convergence of the extended Krylov methods for the approximation of Stieltjes matrix functions via (3.17) has been analyzed in [5, 16, 36]. Here we will use the following result.

THEOREM 3.3 (see [5, section 6.1]). *Let $A \in \mathbb{C}^{N\times N}$ be Hermitian positive definite with $\operatorname{spec}(A) = [\lambda_{\min}, \lambda_{\max}] \subset \mathbb{R}^+$, let f be a Stieltjes function* (1.3)*, and let $\boldsymbol{f}_m^{EK}$ be the mth extended Krylov approximation* (3.17) *for $f(A)\boldsymbol{b}$. Then*

$$\|f(A)\boldsymbol{b} - \boldsymbol{f}_m^{EK}\| \le C \left(\frac{\sqrt[4]{\kappa}-1}{\sqrt[4]{\kappa}+1}\right)^{2m} \simeq \mathcal{O}\left(\exp\left(-4m\sqrt[4]{\frac{\lambda_{\min}}{\lambda_{\max}}}\right)\right)$$

for a constant $C > 0$ and with the term on the right asymptotically sharp for large $\kappa = \lambda_{\max}/\lambda_{\min}$.

Remark 3.4. Theorem 3.3 states that the convergence factor of the extended Krylov method when expressed in terms of the subspace dimension $\dim \mathcal{EK}_m = 2m$ (instead of the order m) is the same as

$$\alpha(0) = \frac{\sqrt[4]{\kappa}-1}{\sqrt[4]{\kappa}+1},$$

just like the shift-and-invert method discussed in section 3.1. ◇

3.3. Polynomial solves in the inner iteration. Both the shift-and-invert method and the extended Krylov subspace method require the solution of a linear system in each iteration. They are therefore particularly attractive for matrices for which direct solution methods can be efficiently applied (e.g., not too large matrices with rather small bandwidth, for which it is feasible to compute a Cholesky decomposition). In particular, as the poles of the rational Krylov subspace stay the same across all iterations, it suffices to compute a Cholesky decomposition (of A or $A - \xi I$, depending on the method) once and then use it in all subsequent iterations.

We are interested in the very large scale case, in which only a few vectors of length N can be stored at the same time, meaning particularly that computing a Cholesky decomposition of A is not an option. Therefore, the linear systems occurring in the shift-and-invert or extended Krylov method can only be solved approximately by an iterative method (the *inner iteration*). We deal here with the case in which the inner iteration is again a Krylov subspace method. As A is Hermitian positive definite, a natural choice is the conjugate gradient method.

In cases where a direct solver can be used, one iteration of the shift-and-invert Krylov method and one iteration of the extended Krylov method have approximately the same cost (as there is no difference in the cost of computing a Cholesky factorization of A or $A-\xi I$). When using an iterative method, this dramatically changes, however. In the shift-and-invert method, with the optimal shift $\xi = -\sqrt{\lambda_{\min}\lambda_{\max}}$ it follows from (3.16) that $\kappa(A-\xi I) = \sqrt{\kappa(A)}$. Thus, without preconditioning of the inner iteration, we can expect the linear systems within the extended Krylov method to be much more difficult to solve and require a significantly higher number of iterations.

An important question that arises in the context of using an inner iteration for the solution of the linear systems in a rational Krylov method is to which accuracy these systems need to be solved in order to not negatively influence the convergence of the outer iteration. We discuss this topic in detail in the next section.

4. Relaxing the tolerance in outer-inner rational Krylov methods. In this section, we only discuss the shift-and-invert method. For the extended Krylov method, very similar results can be formulated in a straightforward manner. We have so far only considered the error of the Lanczos approximation for $g(B)\boldsymbol{b}$. If we want to use an inexact version of $B = (A-\xi I)^{-1}$ by solving the involved linear systems iteratively, we need to modify (3.1) to

$$B(V_m - R_m) = V_m H_m + \beta_m \boldsymbol{v}_{m+1}\boldsymbol{e}_m^T, \tag{4.1}$$

where each column $\boldsymbol{r}_j$ of R_m is the residual incurred when solving for $\boldsymbol{w} = (A-\xi I)^{-1}\boldsymbol{v}_j$. Note that we have replaced the tridiagonal matrix T_m^{SI} by a generally dense upper-Hessenberg matrix H_m. Defining $E_m := -BR_mV_m^*$, a matrix of rank at most m, we can further rewrite (4.1) as

$$(B+E_m)V_m = V_m H_m + \beta_m \boldsymbol{v}_{m+1}\boldsymbol{e}_m^T.$$

Therefore, when inexact inner iterations are used, we are effectively computing an Arnoldi approximation $\boldsymbol{g}_m := \|\boldsymbol{b}\|V_m g(H_m)\boldsymbol{e}_1$ to $g(B+E_m)\boldsymbol{b}$, not $g(B)\boldsymbol{b} = f(A)\boldsymbol{b}$ as intended. It is clear that $\|f(A)\boldsymbol{b} - \boldsymbol{g}_m\| = \|g(B)\boldsymbol{b} - \boldsymbol{g}_m\|$ and

$$\|g(B)\boldsymbol{b} - \boldsymbol{g}_m\| \le \|g(B+E_m)\boldsymbol{b} - \boldsymbol{g}_m\| + \|g(B)\boldsymbol{b} - g(B+E_m)\boldsymbol{b}\|, \tag{4.2}$$

and so we need to control both terms on the right-hand side.

4.1. The first term $\|g(B+E_m)\boldsymbol{b} - \boldsymbol{g}_m\|$. The first term corresponds to the error of the (exact) Arnoldi approximation for $g(B+E_m)\boldsymbol{b}$, and it is analyzed in Appendix A. We show that for E_m small enough, this Arnoldi approximation still converges at a rate very close to that of the unperturbed Krylov approximation.

By Theorem 3.2 we know that the exact shift-and-invert method with optimal shift $\xi = -\sqrt{\lambda_{\min}\lambda_{\max}}$, as well as the standard Lanczos and extended Krylov method, converge geometrically as

$$\|f(A)\boldsymbol{b} - \boldsymbol{f}_m\| \le C\alpha^m, \tag{4.3}$$

where C is some constant independent of m, and

$$\alpha = \begin{cases} \frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1} & \text{for standard Lanczos,} \\ \frac{\sqrt[4]{\kappa(A)}-1}{\sqrt[4]{\kappa(A)}+1} & \text{for shift-and-invert Lanczos and extended Krylov.} \end{cases} \tag{4.4}$$

While Theorem 3.2 gives an expression for the constant C needed to strictly satisfy the bound (4.3), a potentially sharper estimate can be obtained by expanding $f(A)\boldsymbol{b} = \sum_{j=1}^{N} \gamma_j^{(N)} \boldsymbol{v}_j$ into a complete orthonormal Krylov basis $\{\boldsymbol{v}_j\}$ of $\mathbb{C}^N$ (assuming that the invariance index of the Krylov space is N; if this is not the case, a reduced expansion of $f(A)\boldsymbol{b}$ can be used for the same argument). Likewise, let us write $\boldsymbol{f}_m = \sum_{j=1}^{m} \gamma_j^{(m)} \boldsymbol{v}_j$. It can be shown that each of the coefficients $\gamma_j^{(m)}$ approaches $\gamma_j^{(N)}$ monotonically as m increases, in the sense that $|\gamma_j^{(1)}| \le |\gamma_j^{(2)}| \le \cdots \le |\gamma_j^{(N)}|$. This is true for the standard Lanczos method, which has been known since [14, 20], but also for the extended Krylov [44] and even the shift-and-invert method. It is now easy to derive an upper bound on $|\gamma_j^{(N)}|$, and thereby all $|\gamma_j^{(m)}|$. Using (4.3) we have

$$|\gamma_m^{(N)}|^2 \le \sum_{j=1}^{m-1} \left|\gamma_j^{(N)} - \gamma_{m-1}^{(N)}\right|^2 + \sum_{j=m}^{N} \left|\gamma_j^{(N)}\right|^2 = \|f(A)\boldsymbol{b} - \boldsymbol{f}_{m-1}\|^2 \le \left(\frac{C}{\alpha}\right)^2 \alpha^{2m},$$

i.e., the Krylov coefficients $\gamma_j^{(m)}$ used to form $\boldsymbol{f}_m$ also decay geometrically at a rate α; see also [41] for related results. Let us use the model that the unperturbed approximation $\boldsymbol{f}_m$ has coefficients that satisfy a geometric series and therefore

$$\|f(A)\boldsymbol{b}\|^2 = \|\boldsymbol{f}_N\|^2 \approx \left(\frac{C}{\alpha}\right)^2 \sum_{j=1}^{N} \alpha^{2j} \approx \left(\frac{C}{\alpha}\right)^2 \frac{\alpha^2}{1-\alpha^2},$$

suggesting the estimate

$$C \approx \sqrt{1-\alpha^2}\, \|f(A)\boldsymbol{b}\|. \tag{4.5}$$

4.2. The second term $\|g(B)\boldsymbol{b} - g(B+E_m)\boldsymbol{b}\|$. In order to analyze the second term on the right-hand side of (4.2), we first consider the simplified function $g_t(y) = (y^{-1} + \xi + t)^{-1}$. Thanks to the Sherman–Morrison–Woodbury formula we have

$$\begin{aligned}(B+E_m)^{-1} &= B^{-1} + B^{-1}BR_m(I - V_m^H B^{-1} B R_m)^{-1} V_m^H B^{-1} \\ &= B^{-1} + R_m(I - V_m^H R_m)^{-1} V_m^H B^{-1},\end{aligned}$$

and therefore

$$\begin{aligned}g_t(B+E_m) &= [(B+E_m)^{-1} + (\xi+t)I]^{-1} \\ &= [B^{-1} + R_m(I - V_m^H R_m)^{-1} V_m^H B^{-1} + (\xi+t)I]^{-1} \\ &= [B^{-1} + \underbrace{R_m(I - V_m^H R_m)^{-1}}_{=:U} \underbrace{V_m^H B^{-1}}_{=:V} + (\xi+t)I]^{-1} \\ &= g_t(B) - g_t(B)U(I + V g_t(B) U)^{-1} V g_t(B).\end{aligned}$$

As a consequence,

$$\|g_t(B+E_m) - g_t(B)\| \le \|g_t(B)\| \|U(I + V g_t(B) U)^{-1}\| \|V g_t(B)\|.$$

Since $B = (A - \xi I)^{-1}$ is Hermitian, it is easy to see that $\|g_t(B)\| = 1/(\lambda_{\min} + t)$ and

$$\|V g_t(B)\| = \|V_m^H (I + (\xi+t)B)^{-1}\| \le \|(I + (\xi+t)B)^{-1}\| = C_t,$$

where

$$C_t := \begin{cases} \frac{\lambda_{\min}-\xi}{\lambda_{\min}+t}, & t < -\xi, \\ \frac{\lambda_{\max}-\xi}{\lambda_{\max}+t}, & t \geq -\xi. \end{cases}$$

To estimate $\|U(I + Vg_t(B)U)^{-1}\|$, we assume that $\|R_m\| \ll 1$ so that upon using a truncated Neumann series $U = R_m(I - V_m^H R_m)^{-1} \approx R_m(I + V_m R_m) \approx R_m$ and $\|U(I + Vg_t(B)U)^{-1}\| \approx R_m$. This gives the approximate error bound

$$\|g_t(B + E_m) - g_t(B)\| \lesssim \|R_m\| \frac{C_t}{\lambda_{\min} + t}. \tag{4.6}$$

As we have $g(z) = \int_0^\infty g_t(z)\,\mathrm{d}\mu(t)$, we obtain an approximate error bound for $g(B + E_m)$ by integrating (4.6):

$$\begin{aligned} \|g(B + E_m) - g(B)\| &= \left\| \int_0^\infty g_t(B + E_m) - g_t(B)\,\mathrm{d}\mu(t) \right\| \\ &\leq \int_0^\infty \|g_t(B + E_m) - g_t(B)\|\,\mathrm{d}\mu(t) \\ &\lesssim \|R_m\| \int_0^\infty \frac{C_t}{\lambda_{\min} + t}\,\mathrm{d}\mu(t). \end{aligned} \tag{4.7}$$

To rewrite the right-hand side of (4.7), we use the same techniques as in the proof of Lemma 3.1 and Theorem 3.2. Let the functions f_1, f_2 be defined as in (3.15); then

$$\begin{aligned} \int_0^\infty \frac{C_t}{\lambda_{\min} + t}\,\mathrm{d}\mu(t) &= (\lambda_{\min} - \xi) \int_0^{-\xi} \frac{1}{(\lambda_{\min} + t)^2}\,\mathrm{d}\mu(t) \\ &\quad + (\lambda_{\max} - \xi) \int_0^{-\xi} \frac{1}{(\lambda_{\min} + t)(\lambda_{\max} + t)} \\ &\leq (\lambda_{\min} - \xi) \int_0^{-\xi} \frac{1}{(\lambda_{\min} + t)^2}\,\mathrm{d}\mu(t) \\ &\quad + (\lambda_{\max} - \xi) \int_0^{-\xi} \frac{1}{(\sqrt{\lambda_{\min}\lambda_{\max}} + t)^2} \\ &= (\lambda_{\min} - \xi)\,|f_1'(\lambda_{\min})| + (\lambda_{\max} - \xi) \left| f_2'\left(\sqrt{\lambda_{\min}\lambda_{\max}}\right) \right|. \end{aligned}$$

Unfortunately, no closed form of the functions f_1, f_2 or their derivatives is available. One can thus either evaluate the integrals numerically or use the trivial upper bound $|f_1'(z)|, |f_2'(z)| \leq f'(z)$. Using the latter approach, we obtain the final estimate:

$$\|g(B + E_m) - g(B)\| \lesssim \|R_m\| \left((\lambda_{\min} - \xi)|f'(\lambda_{\min})| + (\lambda_{\max} - \xi)|f'(\sqrt{\lambda_{\min}\lambda_{\max}})| \right). \tag{4.8}$$

5. Theoretical and practical comparison of the different methods. We now devise recommendations of which methods discussed in section 2 and 3 are best suited for approximating $f(A)\boldsymbol{b}$ in a given situation (available memory, conditioning of A, size of A, availability of spectral information) based purely on the theoretical results available for these methods. Additionally, we summarize several other features and (dis)advantages of the different methods in a concise manner and perform numerical experiments in order to gauge whether the predictions obtained from the theoretical results are trustworthy.

TABLE 5.1

Overview of the general advantages, disadvantages, and prerequisites of two-pass Lanczos (2PL), multi-shift CG (MSCG), restarted Lanczos (R. Lan.), extended Krylov (EKSM), and shift-and-invert Lanczos (SI), together with the quantities governing the (worst-case) speed of convergence of the method (and in case of the outer-inner methods also of the inner iteration).

	2PL	MSCG	R. Lan.	EKSM	SI
Accuracy limited by memory	✗	✓	✗	✓	✓
Requires spectral information	✗	✓	✗	✗	✓
Preconditioning possible	✗	✗	✗	✓	✓
Additional overhead	✓	✓	✗	✗	✗
Inner conv. factor det. by	—	—	—	$\sqrt{\kappa(A)}$	$\sqrt[4]{\kappa(A)}$
Conv. factor det. by	$\sqrt{\kappa(A)}$	$\sqrt{\kappa(A)}$	$\sqrt{\kappa(A)}$	$\sqrt[4]{\kappa(A)}$	$\sqrt[4]{\kappa(A)}$

5.1. Advantages, disadvantages, and prerequisites. We briefly discuss general properties of the different methods, which go beyond the comparison of error bounds in section 5.2. This comparison is compactly summarized in Table 5.1, together with the quantities determining the asymptotic speed of convergence.

Limited accuracy due to available memory. Without further countermeasures, the accuracy of some of the presented methods is still limited by the available memory. For the multi-shift CG method, the available memory dictates the maximum number of poles which can be used for the rational approximation, as one or two additional vectors of length N (depending on the specific implementation; cf. Remark 2.3) need to be stored. If the number of poles necessary for reaching the target accuracy exceeds the available memory, an alternative is to run the multi-shift CG method several times for subsets of the poles, which then increases the number of matrix-vector and inner products.

For the extended Krylov and shift-and-invert Lanczos method, the attainable accuracy is limited by the available memory because the outer iteration still requires the storage of the full orthonormal basis in order to construct the final approximation to $f(A)\boldsymbol{b}$. If this becomes a limiting factor, restarting techniques could be employed for the outer iteration. However, the restarting of rational Krylov methods and the interaction between restarts and inexact inner solves are largely unexplored topics so far.

The two-pass Lanczos and restarted Lanczos method can reach any desired accuracy independent of the available memory (ignoring numerical effects like round-off error and assuming that the tridiagonal $m \times m$ matrix in the two-pass Lanczos method does not grow beyond memory).

Reliance on a priori spectral information. In two of the discussed methods, spectral information on the matrix A is required. In the multi-shift CG method, when constructing a suitable rational approximation $r \approx f$, one requires (bounds on) the largest and smallest eigenvalue of $\lambda_{\max}$ and $\lambda_{\min}$ of A. In the same way, computing the "optimal" shift $\xi = -\sqrt{\lambda_{\min}\lambda_{\max}}$ in the shift-and-invert method requires knowledge of these extremal eigenvalues. It is, however, possible to choose an arbitrary pole ξ independent of spectral information of A. For certain functions such alternative strategies have been shown to be successful, see, e.g. [39].

In contrast, the two-pass Lanczos, restarted Lanczos, and extended Krylov method do not require any a priori spectral information.

Possibility for preconditioning. By applying a suitable preconditioner, the number of iterations necessary in the outer-inner methods, i.e., the extended Krylov and

shift-and-invert Lanczos method, can potentially be greatly reduced. This can make these methods much more competitive than what one would expect from the inner convergence factors shown in Table 5.1. However, with preconditioning, the extracted approximations are no longer elements of a polynomial Krylov space $\mathcal{K}_m(A, \boldsymbol{b})$, and hence a fair comparison is no longer possible.

Additional overhead. Besides matrix-vector products, there are also other arithmetic operations that add to the computational complexity of the considered methods. For the two-pass Lanczos method, $f(T_m)$ needs to be evaluated, and, when m gets large, this can be a challenge in its own right. While this is in principle also true for the extended Krylov and shift-and-invert method, the number of (outer) iterations in these methods will typically be significantly smaller. In the multi-shift CG method, additional vector operations for each pole beyond the first have to be performed. When the target accuracy is increased, this will also lead to an increase in the degree of the rational approximation and thus the number of poles. Depending on how the cost of a matrix-vector product compares to a vector operation, this additional work can become nonnegligible; see also Experiment 5.3 below.

5.2. Predicting the number of matrix-vector products. We now use the convergence results from above to estimate the number of matrix-vector products that are needed to achieve a certain relative error when approximating $f(A)\boldsymbol{b}$ and thus obtain recommendations for which methods are most suitable under which circumstances. Let us stress that all the bounds presented so far are worst-case predictions that only take the extremal eigenvalues into account and therefore cannot, e.g., predict superlinear convergence effects due to spectral adaption [4]. Therefore, our predictions cannot be expected to be accurate for all matrices with $\operatorname{spec}(A) \subseteq [\lambda_{\min}, \lambda_{\max}]$, but rather only for matrices whose spectra are close to a worst-case distribution. Put another way, our predictions can be expected to be good in regimes where m/N is small so that superlinear convergence has not set in. As most of the discussed methods are influenced by this in a similar manner, we still hope that the recommendations derived from this worst-case analysis are also valid in other cases. This is indeed confirmed by the numerical experiments reported in section 5.3.

For obtaining the predictions in the nonrestarted methods, we first estimate the number of (outer) iterations via the relation

$$\|\boldsymbol{f}_m - f(A)\boldsymbol{b}\| \leq C\alpha^m, \tag{5.1}$$

where α is given by (4.4) and C by (4.5). Here $\boldsymbol{f}_m$ can be the mth iterate of the two-pass Lanczos, shift-and-invert Lanczos, or extended Krylov method or the mth conjugate gradient iterate for the system $(A - \zeta_1 I)\boldsymbol{x}_1 = \boldsymbol{b}$, where ζ_1 is the pole with smallest absolute value. From (5.1) we then obtain the prediction

$$m^* = \left\lceil \log_\alpha \left(\frac{\varepsilon}{C} \right) \right\rceil.$$

For the two-pass Lanczos method, the estimated number of matrix-vector products is $2m^*$, while for the multi-shift CG method it is m^* (the value of m^* differs here, as the conditioning of the matrix $A - \zeta_i I$ is slightly better than that of A). For the outer-inner rational-polynomial methods, after computing the necessary number of outer iterations, we use the same approach for estimating the inner iterations. By summing over all inner iterations, we then obtain an estimate of the total number of matrix-vector products.

For the restarted Lanczos method, after k cycles with restart length m_{re} we have

$$\|\boldsymbol{f}_{m_{re}}^{(k)} - f(A)\boldsymbol{b}\| \leq C(\alpha_{m_{re}}(t_0))^k$$

with the convergence factor $\alpha_{m_{re}}(t_0)$ given in (2.1). We obtain the estimate

$$k^* = \left\lceil \log_{\cosh(m_{re}\log(\alpha))}(\cosh(m^*\log_\alpha)) \right\rceil, \tag{5.2}$$

where α and m^* correspond to the standard Lanczos method and where we have used the representation of the Lanczos and restarted Lanczos convergence factor in terms of the hyperbolic cosine. This is due to the fact that the estimate

$$\frac{1}{\cosh m_{re} \ln c} \approx 2c^{m_{re}}$$

is rather rough for m_{re} small. As the restart length is typically a small value, we thus obtain much better estimates using (5.2). The final estimate for the number of matrix-vector products is then given by $m_{re} \cdot k^*$.

5.3. Experimental confirmation of the predictions. We now perform numerical experiments to illustrate how reliable the predictions from the previous section are in practice. It turns out that choosing the target residual norms as suggested by (4.8) in the inner iterations of the rational methods is much stricter than necessary to reach the desired accuracy. Experimentally, we found the following strategy to yield sufficient accuracy in our experiments: when the overall target accuracy is ε, we solve the first linear system to a residual norm below,

$$\varepsilon_1 = \frac{\varepsilon}{2((\lambda_{\min} - \xi)|f'(\lambda_{\min})| + (\lambda_{\max} - \xi)|f'(\sqrt{\lambda_{\min}\lambda_{\max}})|)},$$

and let the residual norm grow geometrically, $\varepsilon_j = \frac{\varepsilon_1}{\alpha(0)^{j-1}}, j = 2, 3, \ldots$. Apart from this deviation from our theoretical basis, all methods are executed as described before.

Experiment 5.1. In our first experiment, the matrix $A \in \mathbb{C}^{1,000\times 1,000}$ is diagonal with Chebyshev eigenvalues in the interval $[0.1, 200.1]$ and $\boldsymbol{b}$ is a normalized vector of all ones, and we aim to approximate $A^{-1/2}\boldsymbol{b}$ with a relative accuracy of 10^{-6}. In the multi-shift CG method, we use the optimal Zolotarev rational approximation [51] for the inverse square root, which requires 15 poles for the target accuracy. As one needs to store two additional vectors per pole in the multi-shift CG method we choose a restart length of $m_{re} = 30$ in the restarted Lanczos method in order to compare methods with roughly the same memory consumption.

For a matrix with Chebyshev eigenvalues, we expect our predictions to be rather accurate as no superlinear convergence takes place. This is confirmed by Table 5.2 which shows the predicted number of matrix-vector products according to the approach outlined in section 5.2 as well as the actual number of matrix-vector products required by our implementations. We find that the two outer-inner rational-polynomial methods are vastly outperformed by the polynomial Krylov methods.

TABLE 5.2
Predicted number of matrix-vector products for two-pass Lanczos (2PL), multi-shift CG (MSCG), restarted Lanczos (R. Lan.), extended Krylov (EKSM), and shift-and-invert Lanczos (SI), together with the number actually required. The matrix A has Chebyshev eigenvalues in $[0.1, 200.1]$.

	2PL	MSCG	R. Lan.	EKSM	SI
Predicted matrix-vector products	564	215	501	1800	5729
Required matrix-vector products	552	240	480	1883	6903

This is in particular true for the inexact extended Krylov method which needs by far the most matrix-vector products. This is not too surprising as already the solution of one linear system with A requires about the same number of matrix-vector products as the multi-shift CG method. Among the polynomial methods, the multi-shift CG method needs the fewest matrix-vector products (as expected), while the restarted Lanczos method needs about 13% fewer matrix-vector products than the two-pass Lanczos approach for this example.

Experiment 5.2. The matrix in Experiment 5.1 is deliberately chosen so that the actual convergence is very close to what is predicted by the worst-case bounds, informed only by the extremal eigenvalues of A. In reality, all eigenvalues of A have an influence on the convergence of (rational) Krylov methods. Thus, one might get the impression that our approach for theoretically comparing the different methods holds little value in practice. While it is true that the predicted number of matrix-vector products cannot be expected to be accurate, it actually turns out that the prediction of the *ratio between the numbers of matrix-vector products of the different methods* is quite accurate for very different eigenvalue distributions, in particular between the three polynomial methods. To illustrate this, we now consider a diagonal matrix with eigenvalues in $[\lambda_{\min}, \lambda_{\max}]$ given by

$$\lambda_j = \lambda_{\min} + \frac{j-1}{N-1}(\lambda_{\max} - \lambda_{\min})\gamma^{N-j}, \quad j = 2, \ldots, N-1 \tag{5.3}$$

with $\gamma \in (0,1)$ and where we choose $\lambda_{\min} = 0.1, \lambda_{\max} = 200.1$, and $N = 1000$ as before. The lower the parameter γ, the more the eigenvalues in (5.3) are clustered at one end of the spectrum, making the distribution more "favorable" for Krylov methods, as fast spectral adaptation and hence superlinear convergence can be expected. However, as the spectral interval for (5.3) stays the same, irrespective of the value of γ, the worst-case convergence bounds we used in our predictions will be less and less sharp when γ is decreased.

Figure 5.1 depicts the results of applying the discussed methods to matrices with eigenvalue distributions (5.3) with $\gamma \in [0.65, 0.99]$. On the left-hand side, the absolute number of matrix-vector products is shown. As expected, this number decreases for

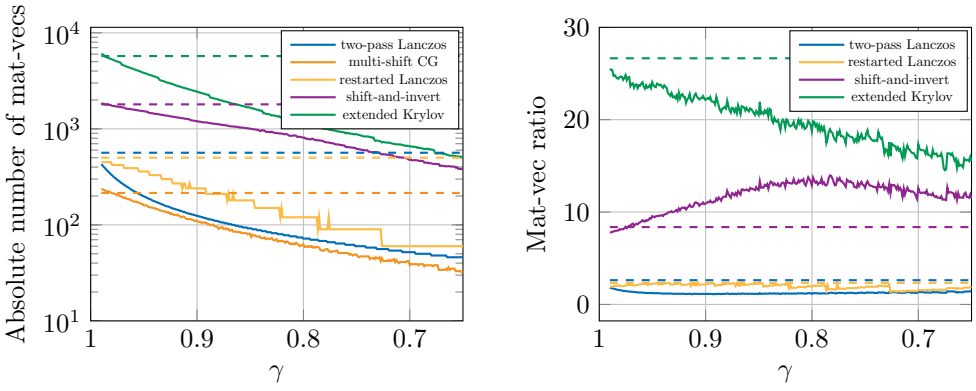

FIG. 5.1. *Left: Number of matrix-vector products required for approximating $A^{-1/2}\boldsymbol{b}$ for a diagonal matrix A with eigenvalues* (5.3) *for different values of γ. Our predictions according to section* 5.2 *are shown as dashed lines. Right: The same data as on the left-hand side, but shown as a relative number compared to the number of iterations needed by multi-shift CG.*

decreasing γ, and thus the distance to our prediction (shown by the dashed lines) grows larger and larger. On the right-hand side, we show the *relative* number of matrix-vector products of the methods, with the number needed by multi-shift CG as a baseline. For the polynomial methods, this results approximately in a horizontal line, revealing that the number of matrix-vector products these methods need in comparison to multi-shift CG stays almost constant for all the different eigenvalue distributions. Thus, although the prediction cannot be used to get a realistic estimate of the amount of work that is needed to solve a given problem, this experiment indicates that it gives a good idea of how different methods compare.

Experiment 5.3. In terms of matrix-vector products, the multi-shift CG method always outperforms the other methods, which is to be expected. In this experiment, we use a very simple model of overall computational complexity to get a rough estimate of how multi-shift CG and restarted Lanczos compare in overall computation time for a given problem. To do so, we count vector operations in addition to matrix-vector products. We count the cost of one simple vector operation (addition or scaling) as one unit of work, written $1\mathcal{V}$. Thus, an inner product has a cost of $\approx 2\mathcal{V}$. The cost $\mathcal{M}$ of a matrix-vector product in these units of work depends on A. For example, for the discretization of a differential operator on a regular two-dimensional lattice, a matrix-vector product has a cost of $\mathcal{M} \approx 10\mathcal{V}$, while for a discretization on a three-dimensional lattice we have $\mathcal{M} \approx 14\mathcal{V}$. We ignore the cost of all operations that are independent of N.

In the multi-shift CG method, the computational cost of one iteration for the "seed system" (i.e., the system with smallest shift) is $1\mathcal{M} + 12\mathcal{V}$, and each additional system requires an effort of $5\mathcal{V}$.

For restarted Lanczos, one iteration has a cost of $1\mathcal{M} + 9\mathcal{V}$, and forming the iterate at the end of each restart cycle has a cost of $\approx 2m_{re}\mathcal{V}$, where m_{re} is the restart length (we ignore the fact that the matrix-vector product $V_{m_{re}}\boldsymbol{y}_{m_{re}}$ can typically be executed faster than $2m_{re}-1$ individual vector operations). Using these formulas, we can estimate the overall number of work units required to compute $f(A)\boldsymbol{b}$ to a certain target accuracy by combining them with our estimates from section 5.2. We again use the optimal Zolotarev rational approximation for the multi-shift CG method and choose the cycle length in restarted Lanczos as $m_{re} = 2p$, where p is the required number of Zolotarev poles.

To obtain a realistic comparison, we do not assume that the additional work of $5\mathcal{V}$ per pole is performed in all iterations for all poles, but instead use the strategy from [19] for removing already converged systems from the iteration in such a way that the overall error in the approximation for $f(A)\boldsymbol{b}$ is still guaranteed to be below the target accuracy.

Figure 5.2 shows the resulting estimates for the same setting as in Experiment 5.1 for varying target accuracies. For high target accuracies, the estimate for the restarted Lanczos method is lower than that of the multi-shift CG method, as the number of poles necessary to construct an accurate enough rational approximation increases. For the three-dimensional discretization, the break-even point comes earlier, as the cost of a matrix-vector product is higher compared to a vector operation.

6. Conclusions. Our theoretical results along with the practical comparisons in section 5 indicate that, among the considered outer-inner polynomial Krylov methods for approximating Stieltjes matrix functions $f(A)\boldsymbol{b}$, both the multi-shift CG and restarted Lanczos methods are the most favorable. We find that the inexact (with

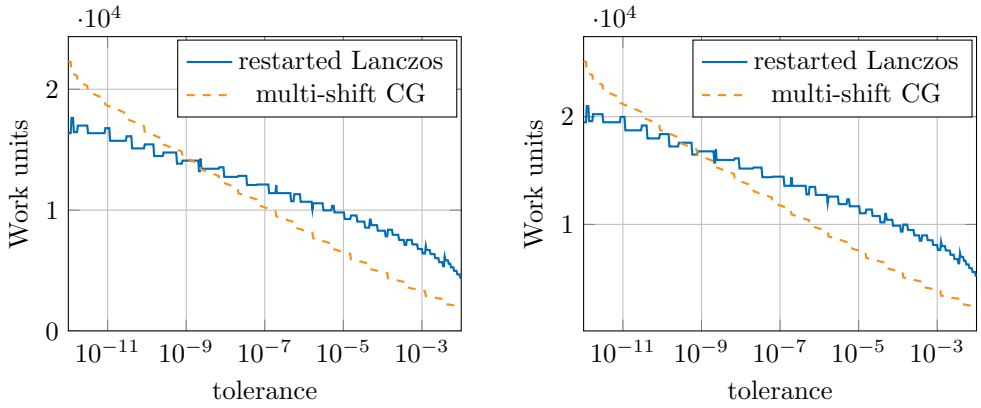

FIG. 5.2. *Estimated amount of work needed to approximate* $A^{-1/2}\boldsymbol{b}$ *to the target accuracy. Left: Two-dimensional discretization. Right: Three-dimensional discretization.*

polynomial inner solves) versions of the shift-and-invert method as well as the extended Krylov method are generally not competitive.

The choice between multi-shift CG and restarted Lanczos should be informed by further considerations. In particular, the multi-shift CG method crucially requires inclusion intervals for the spectrum of A, and if these are not available or difficult to estimate (e.g., using a restarted Krylov method), then restarted Lanczos should be the method of choice. Note that restarted Lanczos can be implemented with deflation strategies that can further speed up the convergence. The multi-shift CG method, on the other hand, generally requires the smallest number of total matrix-vector products and might be more amendable to parallel implementation.

Appendix A. Convergence of the shift-and-invert method with inexact solves. When using inexact solves in the shift-and-invert method, we obtain the "inexact Arnoldi decomposition"

$$(B + E_m)V_m = V_m H_m + \beta_m \boldsymbol{v}_{m+1}\boldsymbol{e}_m^T$$

with the perturbation matrix $E_m := -BR_mV_m^*$. As E_m and thus also $B + E_m$ are non-Hermitian, the results on the speed of convergence derived for Hermitian matrices are no longer applicable. We now explain why, as long as $\varepsilon := \|E_m\|$ is sufficiently small, we can still expect these results to hold in practice. Many of the techniques used in the following closely resemble the approach used in [41] to derive residual estimates for the inexact Arnoldi method for the matrix exponential.

We denote by $\mathcal{W}(M)$ the field of values (or numerical range) of a matrix M, by $\Delta(c, r)$ the closed disk with center c and radius r, by $\mathcal{E}(c, a, b)$ the horizontal, axis-aligned ellipse with semiaxes $a > b$ and center c, and by $\mathcal{S}(c_1, c_2, r)$ the *Bunimovich stadium* with semicircle radius r and semicircle centers $c_1, c_2 \in \mathbb{R}$; see Figure A.1. As $B = (A - \xi I)^{-1}$ is Hermitian, we have $\mathcal{W}(B) = \left[(\lambda_{\max} - \xi)^{-1}, (\lambda_{\min} - \xi)^{-1}\right]$ and clearly $\mathcal{W}(E_m) \subseteq \Delta(0, \varepsilon)$. Therefore, we find

$$\begin{aligned}\mathcal{W}(B + E_m) &\subseteq \mathcal{W}(B) + \mathcal{W}(E_m)\\ &= \mathcal{S}\left(\frac{1}{\lambda_{\max} - \xi}, \frac{1}{\lambda_{\min} - \xi}, \varepsilon\right).\end{aligned}$$

In order to derive a convergence rate for a matrix with field of values inside the Bunimovich stadium, a conformal mapping from $\mathbb{C} \setminus \mathcal{S}(\frac{1}{\lambda_{\max}-\xi}, \frac{1}{\lambda_{\min}-\xi}, \varepsilon)$ onto

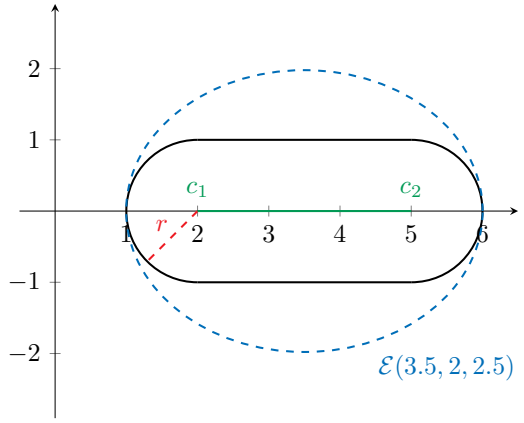

FIG. A.1. *A Bunimovich stadium* $\mathcal{S}(2, 5, 1)$ *and its enclosing ellipse* $\mathcal{E}(3.5, 2, 2.5)$.

$\mathbb{C} \setminus \Delta(0, 1)$ is required. Unfortunately, no closed form for this mapping is known; see [50] for a treatment of this topic, where several numerical approximations for the conformal mapping of the Bunimovich stadium are proposed. We therefore embed $\mathcal{S}(\frac{1}{\lambda_{\max}-\xi}, \frac{1}{\lambda_{\min}-\xi}, \varepsilon)$ into an ellipse as illustrated in Figure A.1. Specifically, we use

$$\mathcal{S}\left(\frac{1}{\lambda_{\max}-\xi}, \frac{1}{\lambda_{\min}-\xi}, \varepsilon\right) \subseteq \mathcal{E}(a+\varepsilon, \sqrt{\varepsilon}\cdot\sqrt{2a+\varepsilon}, c)$$

with $c = \frac{1}{-2\xi}$ and $a = c \cdot \beta$, where $\beta = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$.

The conformal mapping from $\mathbb{C} \setminus \mathcal{E}(a+\varepsilon, \sqrt{\varepsilon}\sqrt{2a+\varepsilon}, c)$ onto $\mathbb{C} \setminus \Delta(0, 1)$ is given by the scaled and shifted Zhukovsky map

$$\psi(z) = \frac{a}{2}\left(Rz + \frac{1}{Rz}\right) + \frac{1}{\lambda_{\max}-\xi} + \frac{1}{2\xi}, \quad \text{where} \quad R = 1 + \frac{\varepsilon}{a} + \sqrt{\frac{2\varepsilon+\varepsilon^2}{\sqrt{a}}}.$$

Further, let $\tau > 1$ be such that $\psi(\tau) < -\frac{1}{\xi}$. Note that such τ is guaranteed to exist as long as $\varepsilon < \frac{1}{\lambda_{\max}-\xi}$, which we assume from here on.

We can then write $g(B + E_m)$ in terms of Faber polynomials Φ_j as

$$g(B+E_m) = \sum_{j=0}^{\infty} \eta_j \Phi_j(B+E_m), \quad \text{with} \quad \eta_j = \frac{1}{2\pi i}\oint_{|z|=\tau} \frac{g(\psi(z))}{z^{j+1}}\,\mathrm{d}z.$$

Defining a polynomial approximation $p_{m-1}(z) = \sum_{j=0}^{m-1} \eta_j \Phi_j(z)$, we have

$$\|g(B+E_m) - p(B+E_m)\| \le \sum_{j=m}^{\infty} \|\eta_j \Phi_j(B+E_m)\| \le 2\sum_{j=m}^{\infty} |\eta_j|$$

because $\|\Phi_j(B+E_m)\| \le 2$ due to $\mathcal{W}(B+E_m) \subseteq \mathcal{E}(a+\varepsilon, \sqrt{\varepsilon}\sqrt{2a+\varepsilon}, c)$; see [3, Theorem 1.1].

Using the quasi-optimality of the Arnoldi approximation, we can conclude that

$$\|g(B+E_m)\boldsymbol{b} - \boldsymbol{g}_m\| \le 2(1+\sqrt{2})\|\boldsymbol{b}\| \sum_{j=m}^{\infty} |\eta_j|$$

with the Crouzeix–Palencia constant $1+\sqrt{2}$. Bounding the Faber coefficients as $|\eta_j| \le \frac{1}{\tau^j} \max_{|z|=\tau} |g(\psi(z))|$ and noting that $\max_{|z|=\tau} |g(\psi(z))| = |g(\psi(\tau))|$, we obtain the bound

$$\|g(B+E_m)\boldsymbol{b} - \boldsymbol{g}_m\| \le 4(1+\sqrt{2})\|\boldsymbol{b}\| \frac{\tau}{\tau-1} |g(\psi(\tau))| \cdot \left|\frac{1}{\tau}\right|^m.$$

Thus, if $\|E_m\| < (\lambda_{\max} - \xi)^{-1}$, so that we can find a suitable ellipse on which g is analytic, we can expect convergence with a rate τ^{-1} for the perturbed problem. Similar arguments can be made concerning the decay of the Arnoldi coefficients of the perturbed iteration. It remains to investigate how τ^{-1} compares to the rate

$$\alpha(0) = \frac{\sqrt[4]{\kappa}-1}{\sqrt[4]{\kappa}+1}$$

of the unperturbed problem in dependence of ε. Solving $\psi(\tau) = 0$ yields

$$\tau^* = -\frac{c}{aR} - \sqrt{\frac{c^2}{(aR)^2} - \frac{1}{R^2}}$$

which, after straightforward algebraic manipulations, gives

$$\left|\frac{1}{\tau^*}\right| = \frac{\beta - 2\xi\varepsilon + 2\sqrt{\varepsilon}\sqrt{-\beta\xi+\varepsilon\xi^2}}{1+\sqrt{1-\beta^2}}.$$

By noting that

$$\frac{\beta}{1+\sqrt{1-\beta^2}} = \frac{\sqrt[4]{\kappa}-1}{\sqrt[4]{\kappa}+1} = \alpha(0),$$

we can further rewrite this as

$$\left|\frac{1}{\tau^*}\right| = \alpha(0) + \frac{-2\xi\varepsilon + 2\sqrt{\varepsilon}\sqrt{-\beta\xi+\varepsilon\xi^2}}{1+\sqrt{1-\beta^2}}, \tag{A.1}$$

which more clearly reveals how the convergence rate deteriorates with growing ε. For $\varepsilon \ll 1$, the term involving $\sqrt{\varepsilon}$ dominates the perturbation given in (A.1). Thus, we can expect the convergence rate to deteriorate approximately like $\sqrt{\varepsilon}$.

In Figure A.2, we illustrate how $|\tau^*|^{-1}$ evolves with growing ε for a diagonal matrix A with $N = 1000$ Chebyshev eigenvalues in the interval $[0.1, 200.1]$. For this example, the smallest eigenvalue of B is $(\lambda_{\max} - \xi)^{-1} \approx 0.0049$. As predicted by our arguments above, for small ε, the convergence rate is essentially the same as that of the unperturbed problem. With growing ε, the convergence rate deteriorates until it reaches the value 1 for $\varepsilon = (\lambda_{\max} - \xi)^{-1}$. In that case, we cannot guarantee convergence as any longer $-\frac{1}{\xi} \in \mathcal{E}(a+\varepsilon, \sqrt{\varepsilon}\cdot\sqrt{2a+\varepsilon}, c)$, which is a singularity of g.

Finally, we address how the tolerance of the inner iteration affects $\varepsilon = \|E_m\|$. Fortunately, this is rather easy. From the definition of E_m, we have

$$\|E_m\| = \| - BR_m V_m^T\| = \|BR_m\| \le \|B\| \sqrt{\sum_{j=0}^{m} \|\boldsymbol{r}_j\|^2},$$

so that we can control it via the residual norms of the inner iterations.

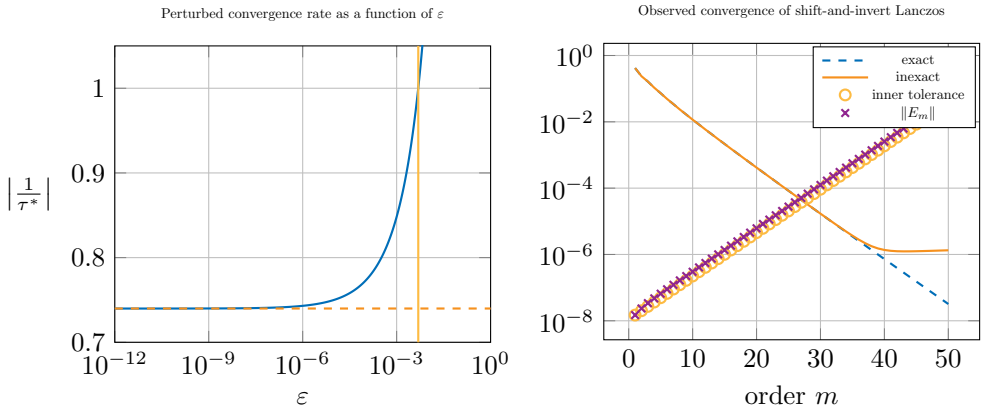

FIG. A.2. *Left: Convergence rate for the perturbed problem as a function of* $\varepsilon = \|E_m\|$. *The horizontal, dashed line shows the convergence rate from Theorem* 3.2 *for the unperturbed problem, and the vertical dash-dotted line shows* $(\lambda_{\max} - \xi)^{-1}$. *Right: Convergence of the exact and inexact shift-and-invert method for the same matrix and vector as in Experiment* 5.1 *when aiming for an overall tolerance of* $\texttt{tol} = 10^{-6}$ *and using the relaxation strategy outlined in section* 4.

In Figure A.2 (right), we apply the shift-and-invert method to the diagonal example matrix already used in Experiment 5.1 and also show the norm of the error matrix $\|E_m\|$ in each step. Comparing the norm ε of the error matrix in Figure A.2 (right) with the convergence rate given for these values of in ε in Figure A.2 (left), we would expect the convergence rate of the inexact method to deteriorate much earlier than it does in reality.

There are different factors playing a role in the explanation of this effect: Our results are valid for matrices with field of values in an ellipse which encloses the Bunimovich stadium, i.e., we have chosen a set containing the field of values of $B+E_m$ which is larger than necessary, as otherwise we have not been able to construct a conformal mapping. In addition, the convergence rate is only a worst-case estimate. While in the unperturbed case we know that convergence for a matrix with Chebyshev eigenvalues will closely follow this worst-case bound, it is not clear how closely the perturbed matrix follows the worst-case bound for the perturbed case. Finally, we used the trivial inclusion $\mathcal{W}(E_m) \subseteq \Delta(0, \|E_m\|)$, which often overestimates the actual diameter of $\mathcal{W}(E_m)$.

Acknowledgments. We thank the two anonymous referees who have provided valuable comments. In particular, the approach in Remark 2.2 was suggested by one of the referees. We would like to thank Kathryn Lund for fruitful discussions and interesting ideas which led to an improvement of the manuscript, in particular the content of section 4. We also thank Leonid Knizhnerman and Valeria Simoncini for some clarifying discussions.

REFERENCES

[1] M. AFANASJEW, M. EIERMANN, O. G. ERNST, AND S. GÜTTEL, *Implementation of a restarted Krylov subspace method for the evaluation of matrix functions*, Linear Algebra Appl., 429 (2008), pp. 2293–2314.

[2] W. E. ARNOLDI, *The principle of minimized iteration in the solution of the matrix eigenvalue problem*, Quart. Appl. Math., 9 (1951), pp. 17–29.

[3] B. BECKERMANN, *Image numérique, GMRES et polynômes de Faber*, C. R. Math. Acad. Sci. Paris., 340 (2005), pp. 855–860.

[4] B. Beckermann and S. Güttel, *Superlinear convergence of the rational Arnoldi method for the approximation of matrix functions*, Numer. Math., 121 (2012), pp. 205–236.
[5] B. Beckermann and L. Reichel, *Error estimates and evaluation of matrix functions via the Faber transform*, SIAM J. Numer. Anal., 47 (2009), pp. 3849–3883.
[6] C. Berg, *Stieltjes-Pick-Bernstein-Schoenberg and their connection to complete monotonicity*, in Positive Definite Functions: From Schoenberg to Space-Time Challenges, J. Mateu and E. Porcu, eds., Department of Mathematics, University Jaume I, Castellón de la Plana, Spain, 2008.
[7] C. Berg and G. Forst, *Potential Theory on Locally Compact Abelian Groups*, Springer, Berlin, 1975.
[8] M. Berljafa and S. Güttel, *Generalized rational Krylov decompositions with an application to rational approximation*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 894–916.
[9] J. Bloch, A. Frommer, B. Lang, and T. Wettig, *An iterative method to compute the sign function of a non-Hermitian matrix and its application to the overlap Dirac operator at nonzero chemical potential*, Comput. Phys. Commun., 177 (2007), pp. 933–943.
[10] A. Boriçi, *Fast methods for computing the Neuberger operator*, in Numerical Challenges in Lattice Quantum Chromodynamics, A. Frommer, T. Lippert, B. Medeke, and K. Schilling, eds., Springer, Berlin, 2000, pp. 40–47.
[11] J. Brannick, A. Frommer, K. Kahl, B. Leder, M. Rottmann, and A. Strebel, *Multigrid preconditioning for the overlap operator in lattice QCD*, Numer. Math., 132 (2016), pp. 463–490.
[12] E. Carson and Z. Strakoš, *On the cost of iterative computations*, Philos. Trans. Roy. Soc. A, 378 (2020), 20190050.
[13] C. de Boor, *Divided differences*, Surv. Approx. Theory, 1 (2005), pp. 46–69.
[14] V. Druskin, *On monotonicity of the Lanczos approximation to the matrix exponential*, Linear Algebra Appl., 429 (2008), pp. 1679–1683.
[15] V. Druskin and L. Knizhnerman, *Two polynomial methods of calculating functions of symmetric matrices*, U.S.S.R. Comput. Math. Math. Phys., 29 (1989), pp. 112–121.
[16] V. Druskin and L. Knizhnerman, *Extended Krylov subspaces: Approximation of the matrix square root and related functions*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 755–771.
[17] V. Druskin, C. Lieberman, and M. Zaslavsky, *On adaptive choice of shifts in rational Krylov subspace reduction of evolutionary problems*, SIAM J. Sci. Comput., 32 (2010), pp. 2485–2496.
[18] M. Eiermann and O. G. Ernst, *A restarted Krylov subspace method for the evaluation of matrix functions*, SIAM J. Numer. Anal., 44 (2006), pp. 2481–2504.
[19] 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 (2002), pp. 203–224.
[20] A. Frommer, *Monotone convergence of the Lanczos approximations to matrix functions of Hermitian matrices*, Electron. Trans. Numer. Anal., 35 (2009), pp. 118–128.
[21] A. Frommer, S. Güttel, and M. Schweitzer, *Convergence of restarted Krylov subspace methods for Stieltjes functions of matrices*, SIAM J. Matrix Anal. Appl., 35 (2014), pp. 1602–1624.
[22] A. Frommer, S. Güttel, and M. Schweitzer, *Efficient and stable Arnoldi restarts for matrix functions based on quadrature*, SIAM J. Matrix Anal. Appl., 35 (2014), pp. 661–683.
[23] A. Frommer and P. Maass, *Fast CG-based methods for Tikhonov–Phillips regularization*, SIAM J. Sci. Comput., 20 (1999), pp. 1831–1850.
[24] A. Frommer and V. Simoncini, *Matrix functions*, in Model Order Reduction: Theory, Research Aspects and Applications, W. H. A. Schilders, H. A. van der Vorst, and J. Rommes, eds., Springer, Berlin, 2008, pp. 275–303.
[25] G. H. Golub and G. Meurant, *Matrices, Moments and Quadrature with Applications*, Princeton University Press, Princeton, NJ, 2010.
[26] S. Güttel, *Rational Krylov approximation of matrix functions: Numerical methods and optimal pole selection*, GAMM-Mitt., 36 (2013), pp. 8–31.
[27] S. Güttel and L. Knizhnerman, *A black-box rational Arnoldi variant for Cauchy–Stieltjes matrix functions*, BIT, 53 (2013), pp. 595–616.
[28] S. Güttel, D. Kressner, and K. Lund, *Limited-memory polynomial methods for large-scale matrix functions*, GAMM-Mitt., 43 (2020), e202000019.
[29] S. Güttel and Y. Nakatsukasa, *Scaled and squared subdiagonal Padé approximation for the matrix exponential*, SIAM J. Matrix Anal. Appl., 37 (2016), pp. 145–170.
[30] P. Henrici, *Applied and Computational Complex Analysis, Vol.* 2, John Wiley & Sons, New York, 1977.

[31] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Res. Natl. Bur. Stand., 49 (1952), pp. 409–436.
[32] M. Ilić, I. W. Turner, and A. N. Pettitt, *Bayesian computations and efficient algorithms for computing functions of large, sparse matrices*, ANZIAM J., 45 (2004), pp. C504–C518.
[33] M. Ilić, I. W. Turner, and D. P. Simpson, *A restarted Lanczos approximation to functions of a symmetric matrix*, IMA J. Numer. Anal., 30 (2010), pp. 1044–1061.
[34] C. Jagels and L. Reichel, *The extended Krylov subspace method and orthogonal Laurent polynomials*, Linear Algebra Appl., 431 (2009), pp. 441–458.
[35] C. Jagels and L. Reichel, *Recursion relations for the extended Krylov subspace method*, Linear Algebra Appl., 434 (2011), pp. 1716–1732.
[36] L. Knizhnerman and V. Simoncini, *A new investigation of the extended Krylov subspace method for matrix function evaluations*, Numer. Linear Algebra Appl., 17 (2010), pp. 615–638.
[37] C. Lanczos, *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*, J. Res. Natl. Bur. Stand., 45 (1950), pp. 255–282.
[38] I. Moret, *Rational Lanczos approximations to the matrix square root and related functions*, Numer. Linear Algebra Appl., 16 (2009), pp. 431–445.
[39] I. Moret and P. Novati, *Krylov subspace methods for functions of fractional differential operators*, Math. Comp., 88 (2019), pp. 293–312.
[40] I. Moret and M. Popolizio, *The restarted shift-and-invert Krylov method for matrix functions*, Numer. Linear Algebra Appl., 21 (2014), pp. 68–80.
[41] S. Pozza and V. Simoncini, *Inexact Arnoldi residual estimates and decay properties for functions of non-Hermitian matrices*, BIT, 59 (2019), pp. 969–986.
[42] Y. Saad, *Analysis of some Krylov subspace approximations to the matrix exponential operator*, SIAM J. Numer. Anal., 29 (1992), pp. 209–228.
[43] Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd edition, SIAM, Philadelphia, 2000.
[44] M. Schweitzer, *Monotone convergence of the extended Krylov subspace method for Laplace-Stieltjes functions of Hermitian positive definite matrices*, Linear Algebra Appl., 507 (2016), pp. 486–498.
[45] M. Schweitzer, *Restarting and Error Estimation in Polynomial and Extended Krylov Subspace Methods for the Approximation of Matrix Functions*, Ph.D. thesis, Bergische Universität Wuppertal, Wuppertal, Germany, 2016.
[46] V. Simoncini, *A new iterative method for solving large-scale Lyapunov matrix equations*, SIAM J. Sci. Comput., 29 (2007), pp. 1268–1288.
[47] D. P. Simpson, I. W. Turner, and A. N. Pettitt, *Fast Sampling from a Gaussian Markov Random Field Using Krylov Subspace Approaches*, tech. report, Queensland University of Technology, Brisbane, Australia, 2008.
[48] H. Tal-Ezer, *On restart and error estimation for Krylov approximation of $w = f(A)v$*, SIAM J. Sci. Comput., 29 (2007), pp. 2426–2441.
[49] J. van den Eshof and M. Hochbruck, *Preconditioning Lanczos approximations to the matrix exponential*, SIAM J. Sci. Comput., 27 (2006), pp. 1438–1457.
[50] V. K. Varma, *Conformal Map and Harmonic Measure of the Bunimovich Stadium*, preprint, 2014, https://arxiv.org/abs/1410.4932.
[51] G. Zolotarev, *Application of elliptic functions to the problem of functions which vary the least or the most from zero*, Abh. St. Petersb., 30 (1877), pp. 1–59.