mathematics

MDPI

Article

A Flexible Extended Krylov Subspace Method for Approximating Markov Functions of Matrices

Shengjie Xu † and **Fei Xue** *,†

School of Mathematical and Statistical Sciences, Clemson University, O-110 Martin Hall, Box 340975, Clemson, SC 29634, USA; shengjx@g.clemson.edu

* Correspondence: fxue@clemson.edu

† These authors contributed equally to this work.

Abstract: A flexible extended Krylov subspace method ($\mathcal{F}$-EKSM) is considered for numerical approximation of the action of a matrix function $f(A)$ to a vector b, where the function f is of Markov type. $\mathcal{F}$-EKSM has the same framework as the extended Krylov subspace method (EKSM), replacing the zero pole in EKSM with a properly chosen fixed nonzero pole. For symmetric positive definite matrices, the optimal fixed pole is derived for $\mathcal{F}$-EKSM to achieve the lowest possible upper bound on the asymptotic convergence factor, which is lower than that of EKSM. The analysis is based on properties of Faber polynomials of A and $(I - A/s)^{-1}$. For large and sparse matrices that can be handled efficiently by LU factorizations, numerical experiments show that $\mathcal{F}$-EKSM and a variant of RKSM based on a small number of fixed poles outperform EKSM in both storage and runtime, and usually have advantages over adaptive RKSM in runtime.

Keywords: Markov-type functions; rational Krylov subspace; extended Krylov subspace

MSC: 65F50; 65F60; 65E10

Citation: Xu, S.; Xue, F. A Flexible Extended Krylov Subspace Method for Approximating Markov Functions of Matrices. *Mathematics* **2023**, *11*, 4341. https://doi.org/10.3390/math11204341

Academic Editor: Michael Voskoglou

Received: 21 September 2023
Revised: 15 October 2023
Accepted: 17 October 2023
Published: 19 October 2023

1. Introduction

Consider a large square matrix $A \in \mathbb{R}^{n\times n}$ and a function f such that the matrix function $f(A) \in \mathbb{R}^{n\times n}$ is well-defined [1,2]. The numerical approximation of

$$q = f(A)b, \tag{1}$$

where $b \in \mathbb{R}^n$ is a vector, is a common problem in scientific computing. It arises in numerical solutions of differential equations [3–6], matrix functional integrators [7,8], model order reduction [9,10], and optimization problems [11,12]. Note that approximating the action of $f(A)$ to a vector b and approximating the matrix $f(A)$ are different. For a large sparse matrix A, $f(A)$ is usually fully dense and infeasible to form explicitly.

Numerical methods for approximating the action of $f(A)$ to a vector have been extensively studied in recent years, especially for large-scale problems; see, e.g., [13–15] and references therein. Existing algorithms often construct certain polynomial or rational approximations to f over the spectrum of A and apply such approximations directly to the vector b without forming any dense matrices of the same size as A. A class of mostly common projection methods are based on Krylov subspaces $\mathcal{K}_m(A, b)$; however, for many large matrices this may require a very large dimension of approximation spaces. Rational Krylov subspace methods have been investigated to decrease the size of subspaces for approximations; see, e.g., [16–20]. Two well-known examples are the extended rational Krylov subspace method (EKSM) [14,21,22] and the adaptive rational Krylov subspace method (adaptive RKSM) [23,24].

In this paper, we explore a generalization of EKSM that uses one fixed nonzero pole alternately with the infinite pole for approximating the action of Markov-type (Cauchy–Stieltjes) functions [25]. Markov-type functions can be written as

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

where μ is a measure that ensures that the integral converges absolutely. Note that this definition can be generalized to integrals defined on $[\alpha, \beta]$, where $-\infty \leq \alpha < \beta < \infty$. We consider the interval $(-\infty, 0]$ here, as this is sufficient for the most widely-studied Markov-type functions $f(z) = z^{\gamma}$ with $-1 < \gamma < 0$, $f(z) = \frac{e^{\theta\sqrt{z}}-1}{z}$ and for their simple modifications, such as $z^{\ell} f(z)$ with $\ell \in \mathbb{Z}^{+}$. Our analysis can be extended to integrals on $(-\infty, \beta]$ as long as the measure $\mu(\zeta)$ satisfies $\int_{-\infty}^{\beta} |d\mu(\zeta)| < \infty$, as assumed in [21].

This study is motivated by [21], which provided an upper bound on the convergence factor of EKSM for approximating $f(A)b$. Our work concerns a generalization of EKSM, replacing the zero pole used in EKSM with a fixed *nonzero* pole s; hence, we call it the *flexible extended Krylov subspace method* ($\mathcal{F}$-EKSM). This algorithm can apply the three-term recurrence to enlarge the subspace for symmetric matrices such as EKSM, while full orthogonalization process may be necessary for adaptive RKSM regardless of the symmetry of matrices. Beckermann and Reichel [26] studied the asymptotic convergence rate of RKSM with different pole selections for approximating $f(A)b$ of Markov functions via Faber transform; however, they did not provide explicit expressions of the optimal cyclic 2 poles or the corresponding rate of convergence, which could be done by solving a quartic equation analytically. In this paper, we derive explicit expressions of the optimal pole s and the corresponding convergence factor using the different analytic tool from [21]. While our bounds on the convergence factor seem to not be as tight as the bounds in [26] numerically, our poles and bounds are provided in explicit expressions; in addition, our pole usually leads to faster convergence for discrete Laplacian matrices based on finite difference discretizations and many practical nonsymmetric matrices.

We explore the optimal pole s to achieve the lowest upper bound on the asymptotic convergence factor, which is guaranteed to outperform EKSM on symmetric positive definite (SPD) matrices. For nonsymmetric matrices with an elliptic numerical range, we provide numerical evidence to demonstrate the advantage of $\mathcal{F}$-EKSM over EKSM in convergence rate. Numerical experiments show that $\mathcal{F}$-EKSM converges at least as rapidly as the upper bound suggests. In practice, if the linear systems needed for constructing rational Krylov subspaces can be solved efficiently by LU factorizations, then $\mathcal{F}$-EKSM outperforms EKSM in both time and storage cost over a wide range of matrices, and it could run considerably faster than adaptive RKSM for many problems. In this paper, we only consider factorization-based direct methods for solving the inner linear systems; for the use and implications of iterative linear solvers see, e.g., [27].

Rational Krylov subspace methods may exhibit superlinear convergence for approximating $f(A)b$. As the algorithms proceed and more rational Ritz values converge to the exterior eigenvalues of A, the 'effective spectrum' of A not covered by the converged Ritz values shrinks, leading to gradual acceleration of the convergence. This analysis has been performed for EKSM applied to the 1D discrete Laplacian ([28], Section 5.2) based on the result from [21], leading to a sharper explicit bound on the convergence. The same idea could be explored with $\mathcal{F}$-EKSM; however, it is not considered here, as we did not observe superlinear convergence in our experiments. This was probably because the effective spectrum of our large test matrices did not shrink quickly enough to exhibit convergence speedup before the stopping criterion was satisfied.

Though $\mathcal{F}$-EKSM is closely connected to EKSM, we emphasize that the convergence of $\mathcal{F}$-EKSM *cannot* be derived directly from that of EKSM applied to a shifted matrix. Admittedly, with the same vector b it is the case that $\mathcal{F}$-EKSM with a pole s applied to A and EKSM applied to $I - A/s$ both generate the same subspaces $\mathcal{Q}_{2m+1}^{(s)}(A, b) =$

$(I - A/s)^{-m}\text{span}\{b, Ab, ..., A^{2m}b\}$, however, the existing theory of EKSM [21] can only provide a bound on the convergence factor for approximating $f(I - A/s)b$, which is not what is needed and has no obvious relationship with the convergence for approximating $f(A)b$ from the same subspace. Our analysis is based on a special min–min optimization instead of the results of EKSM applied to a shifted matrix.

The remainder of this paper is organized as follows. In Section 3, we discuss the implementation of $\mathcal{F}$-EKSM. In Section 4, we analyze the linear convergence factor of $\mathcal{F}$-EKSM and provide the optimal pole with which the lowest upper bound on the convergence factor can be achieved for SPD matrices. In addition, we numerically explore the optimal pole and the convergence factor of $\mathcal{F}$-EKSM for nonsymmetric matrices with an elliptic numerical range. In Section 5, we consider a variant of RKSM that applies a few fixed cyclic poles to provide faster approximations than $\mathcal{F}$-EKSM for certain challenging nonsymmetric matrices. In Section 6, we show the results of numerical experiments for different methods on a variety of matrices. Our conclusions are provided in Section 7, followed by several proofs of Lemmas in the Appendices.

2. Rational Krylov Subspace Methods and $\mathcal{F}$-EKSM

For a wide range of matrix function approximation problems, polynomial Krylov subspace methods converge very slowly [29,30]. To speed up convergence, a more efficient approach is to apply rational Krylov subspace methods; see, e.g., [31,32] and references therein.

The procedure of RKSM is outlined as follows. Starting with an initial nonzero vector b that generates $\mathcal{Q}_1(A, b) = \text{span}\{v_1\}$, where $v_1 = b/||b||$, RKSM keeps expanding the subspaces to search for increasingly more accurate approximate solutions to our problem of interest. In order for RKSM to expand the current subspace $\mathcal{Q}_m(A, b)$ to $\mathcal{Q}_{m+1}(A, b)$, we apply the linear operator $(\gamma_m I - \eta_m A)^{-1}(\alpha_m I - \beta_m A)$ to a vector $u \in \mathcal{Q}_m(A, b) \setminus \mathcal{Q}_{m-1}(A, b)$. To build an orthonormal basis $\{v_1, v_2, ..., v_{m+1}\}$ of the enlarged subspace $\mathcal{Q}_{m+1}(A, b)$, we may choose $u = v_m$ and adopt the modified Gram–Schmidt orthogonalization, obtaining

$$(\gamma_m I - \eta_m A)^{-1}(\alpha_m I - \beta_m A)v_m = \sum_{i=1}^{m+1} h_{im} v_i, \tag{2}$$

where $h_{im} = v_i^*(\gamma_m I - \eta_m A)^{-1}(\alpha_m I - \beta_m A)v_m$. To ensure that the linear operator is well-defined and nonsingular, we require that $|\gamma_m|^2 + |\eta_m|^2 \neq 0$, $|\alpha_m|^2 + |\beta_m|^2 \neq 0$, $(\gamma_m, \eta_m) \neq (\alpha_m, \beta_m)$ up to a scaling factor, and that $\frac{\alpha_m}{\beta_m}$ and $\frac{\gamma_m}{\eta_m}$ (if β_m and η_m are nonzeros) not be an eigenvalue of A. The use of four parameters $(\alpha_m, \beta_m, \gamma_m, \eta_m)$ provides the flexibility to accommodate both the zero ($\gamma_m = 0$, $\eta_m = -1$) and the infinity ($\gamma_m = 1$, $\eta_m = 0$) poles in a unified framework. The expansion of rational Krylov subspaces does not have to be based on the last orthonormal basis vector $u = v_m$, as in (2). There are alternative ways to choose the *continuation vector* to expand the subspaces; see, e.g., [33].

The shift-inverse matrix vector product $(\gamma_m I - \eta_m A)^{-1}(\alpha_m I - \beta_m A)v_m$ is equivalent to solving $(\gamma_m I - \eta_m A)w = (\alpha_m I - \beta_m A)v_m$ (the inner linear system) for w. Multiplying both sides of (2) by $\gamma_m I - \eta_m A$, moving all terms containing A to the left-hand side and all other terms to the right-hand side, we have

$$A\left(\sum_{i=1}^{m+1} \eta_m h_{im} v_i - \beta_m v_m\right) = -\alpha_m v_m + \gamma_m \sum_{i=1}^{m+1} h_{im} v_i. \tag{3}$$

Note that the above relation should hold for each index value $m = 1, 2, \ldots$, thus, it is not hard to see that (3) can be written in the following matrix form:

$$\begin{aligned} AV_{m+1}\underline{F}_m &\equiv AV_{m+1}\left(\underline{H}_m \text{diag}(\eta_1, \ldots, \eta_m) - \begin{bmatrix} \text{diag}(\beta_1, \ldots, \beta_m) \\ 0_{1\times m} \end{bmatrix}\right) \\ &= V_{m+1}\left(\underline{H}_m \text{diag}(\gamma_1, \ldots, \gamma_m) - \begin{bmatrix} \text{diag}(\alpha_1, \ldots, \alpha_m) \\ 0_{1\times m} \end{bmatrix}\right) \equiv V_{m+1}\underline{G}_m, \end{aligned}$$

where $V_{m+1} = [v_1, v_2, ..., v_{m+1}]$ contains the orthonormal basis vectors of the rational Krylov subspace:

$$\begin{aligned} &\mathcal{Q}_{m+1}(A, v_1) = q_m(A)^{-1}\mathcal{K}_{m+1}(A, v_1) \\ = &\left(\prod_{k=1}^{m}(\gamma_k I - \eta_k A)^{-1}\right)\text{span}\Big\{v_1, Av_1, A^2v_1, ..., A^m v_1\Big\}, \end{aligned}$$

and $\underline{H}_m$, $\underline{F}_m$, and $\underline{G}_m \in \mathbb{R}^{(m+1)\times m}$ are all upper Hessenberg matrices. Specifically,

$$\begin{aligned} \underline{H}_m = \begin{bmatrix} H_m \\ h_{(m+1)m}e_m^* \end{bmatrix}, \underline{F}_m = \begin{bmatrix} F_m \\ f_{(m+1)m}e_m^* \end{bmatrix} = \begin{bmatrix} H_m\text{diag}(\eta_1, ..., \eta_m) - \text{diag}(\beta_1, ..., \beta_m) \\ h_{(m+1)m}\eta_m e_m^* \end{bmatrix}, \\ \underline{G}_m = \begin{bmatrix} G_m \\ g_{(m+1)m}e_m^* \end{bmatrix} = \begin{bmatrix} H_m\text{diag}(\gamma_1, ..., \gamma_m) - \text{diag}(\alpha_1, ..., \alpha_m) \\ h_{(m+1)m}\gamma_m e_m^* \end{bmatrix}, \end{aligned} \quad (4)$$

where $e_m = [0, \ldots, 0, 1]^* \in \mathbb{R}^m$.

The idea of RKSM as a projection method is the same as standard Krylov: first, solve the Galerkin projected problem defined for $V_m^* A V_m = F_m G_m^{-1}$, then project the solution back to the m-dimensional subspace as an approximate solution for the original problem defined for A.

The parameters $\alpha_m, \beta_m, \gamma_m, \eta_m$ can be changed in each iteration to determine the poles. For example, if we set $\alpha_k = \eta_k = 0, \gamma_k = 1$ and $\beta_k = -1$ for all $k \in \mathbb{Z}^+$, it is the standard Krylov subspace method; if we set $\beta_{2k} = \eta_{2k-1} = -1$, $\alpha_{2k-1} = \gamma_{2k} = 1$, and $\alpha_{2k} = \beta_{2k-1} = \gamma_{2k-1} = \eta_{2k} = 0$ for $k \in \mathbb{Z}^+$, it becomes EKSM. A special variant of EKSM [34,35] constructs the following extended subspaces:

$$\mathcal{K}^{l,m}(A, b) = \text{span}\Big\{A^{-l+1}b, ..., A^{-1}b, b, Ab, ..., A^{m-1}b\Big\}. \quad (5)$$

A practical choice for the two indices l and m leads to subspaces of the form $\mathcal{K}^{m,im+1}(A, b)$ for some $i \in \mathbb{N}^+$, which requires an orthonormal basis for the Krylov subspaces with vectors

$$b, Ab, A^2b, ..., A^ib, A^{-1}b, A^{i+1}b, ..., A^{2i}b, A^{-2}b, A^{2i+1}b, ...$$

However, there is no convergence theory for this special variant.

The general rational Krylov space of order m is provided by [36,37]:

$$\begin{aligned} &\mathcal{Q}_m(A, b) = q_{m-1}(A)^{-1}\text{span}\{b, Ab, ..., A^{m-1}b\}, \\ \text{where} \quad &q_{m-1}(z) = (\gamma_1 - \eta_1 z)(\gamma_2 - \eta_2 z)...(\gamma_{m-1} - \eta_{m-1}z), \end{aligned}$$

with γ_i, η_i prescribed in (2). For EKSM, the rational Krylov subspace of dimension is

$$\mathcal{Q}_{2m+1}^{(E)}(A, b) = \mathcal{K}_{m+1}(A, b) \cup \mathcal{K}_{m+1}(A^{-1}, b) = A^{-m}\text{span}\{b, Ab, ..., A^{2m}b\}.$$

EKSM applies the operators A^{-1} and A in an alternating manner in each iteration.

For adaptive RKSM, the operation at step m can be written as follows:

$$(I - A/s_m)^{-1}(A - \sigma_m I)v_m = \textstyle\sum_{i=1}^{m+1} h_{im}v_i,$$

where s_m is a nonzero pole and σ_m is a zero of the underlying rational function. To find the optimal poles and zeros at each step, we first restrict the poles and zeros to disjoint sets Ξ and Σ, respectively, where $\Sigma \supseteq W(A)$ and $\Xi \subseteq \mathbb{C}\backslash W(A)$ [38] and where $W(A) = \{x^*Ax \mid x \in \mathbb{C}^n, \|x\|_2 = 1\}$ is the numerical range of A. The pair (Σ, Ξ) is called a condenser [39,40]. An analysis of RKSM considers a sequence of rational nodal functions

$$r_m(z) = \textstyle\prod_{j=1}^{m} \frac{z-\sigma_j}{1-z/s_j},$$

where the zeros $\sigma_j \in \Sigma$ and the poles $s_j \in \Xi$. Adaptive RKSM tries to obtain asymptotically optimal rational functions by defining σ_{j+1} and s_{j+1} recursively with the following conditions: after choosing σ_1 and s_1 of minimal distance, define [38]:

$$\sigma_{j+1} = \arg\max_{z\in\Sigma} |r_j(z)|, \qquad s_{j+1} = \arg\min_{z\in\Xi} |r_j(z)|. \tag{6}$$

The points $\{(\sigma_j, s_j)\}$ are called generalized Leja points [41,42]. In practice, we compute approximations with respect to the poles and zeros defined in (6) during the progress of iteration. Adaptive RKSM usually converges with fewer iterations than EKSM while using a smaller approximation subspace [24,38,43]. While usually converging in fewer iterations than the variants with a few cyclic poles [32], each step of adaptive RKSM requires a solution to a shifted linear system with a new shift, which is more expensive than using existing LU factorizations to solve the linear system with the same coefficient matrix that has been factorized. If the linear system at each RKSM step is solved by a direct method, adaptive RKSM tends to require longer runtimes than variants with a few cyclic poles based on reusing LU factorizations for each distinct pole. Adaptive RKSM is most competitive if the linear systems arising from each step need to be solved approximately by an iterative method and if effective preconditioning can be structured for each linear system with different shift.

In this paper, we consider generating rational Krylov subspaces with cyclic poles s, $+\infty$, s, $+\infty$, ... ($s \neq 0$), which we call the *flexible extended Krylov subspace method* ($\mathcal{F}$-EKSM). The corresponding linear operators are provided by $(I - A/s)^{-1}$ and A, which are applied in an alternating manner. To this end, we set $\beta_{2k} = -1$, $\eta_{2k-1} = 1/s$, $\alpha_{2k-1} = \gamma_k = 1$, and $\alpha_{2k} = \beta_{2k-1} = \eta_{2k} = 0$ for $k \in \mathbb{Z}^+$. The approximation space of $\mathcal{F}$-EKSM is

$$\mathcal{Q}^{(s)}_{2m+1}(A, b) = (I - A/s)^{-m}\text{span}\{b, Ab, ..., A^{2m}b\}.$$

The choice of the repeated pole s influences the convergence rate of $\mathcal{F}$-EKSM. Our goal is to find the optimal pole s^* for Markov-type functions of matrices such that $\mathcal{F}$-EKSM achieves the lowest upper bound on the linear convergence factor. This subspace is identical to the one generated by EKSM applied to the shifted matrix $I - A/s$; the convergence theory of EKSM [21] would provide a convergence factor bound for approximating $f(I - A/s)b$ instead of $f(A)b$, however, this is not our concern here, as our results are derived with a special min–min optimization analysis, not from the results of EKSM applied to a shifted matrix.

3. Implementation of $\mathcal{F}$-EKSM for Approximating $f(A)b$

Without loss of generality, suppose that $||b||_2 = 1$ in (1) and let the initial subspace be $\text{span}\{b, (I - A/s)^{-1}b\}$. The approximation to $f(A)b$ after $(m-1)$ steps is

$$q_m = V_{2m} f(A_m) V_{2m}^* b = V_{2m} f(A_m) e_1,$$

where $V_{2m} \in \mathbb{C}^{n\times(2m)}$ is an orthonormal set of basis vectors of the subspace

$$\mathcal{Q}^{(s)}_{2m+1}(A, b) = \mathcal{K}_m(A, b) \cup \mathcal{K}_m((I - A/s)^{-1}, (I - A/s)^{-1}b)$$

and $A_m = V_{2m}^* A V_{2m}$ denotes the restriction of matrix A in $\mathcal{Q}^{(s)}_{2m+1}(A, b)$.

Because $\mathcal{K}_m(A, b) = \mathcal{K}_m((I - A/s), b)$, we have $\mathcal{Q}^{(s)}_{2m+1}(A, b) = \mathcal{Q}^{(E)}_{2m+1}(I - A/s, b)$. To construct this subspace, we can apply EKSM to the matrix $I - A/s$, obtain $V_{2m} = [v_1, ..., v_{2m}]$ as the orthonormal set of basis vectors of $\mathcal{Q}^{(E)}_{2m+1}(I - A/s, b)$, and obtain

$$\mathcal{T}_{2m} = V_{2m}^*(I - A/s)V_{2m} \in \mathbb{R}^{2m\times 2m}, \tag{7}$$

which is a block upper Hessenberg matrix (see Section 3 in [22]). From (7), we obtain $A_m = V_{2m}^* A V_{2m} = s(I_{2m} - \mathcal{T}_{2m})$. Following the derivation of EKSM [22], we can derive a similar Arnoldi relation $AV_{2m} = V_{2m+2}\underline{T}_{2m}$ for our proposed $\mathcal{F}$-EKSM, where $\underline{T}_{2m} \in \mathbb{R}^{(2m+2)\times 2m}$ is a block upper Hessenberg matrix with 2×2 blocks, and we obtain $A_m = V_{2m}^* A V_{2m} = T_{2m}$. More implementation details of RKSM can be found in [33,44]. Notice that similar to EKSM, for symmetric problems, the orthogonalization cost of $\mathcal{F}$-EKSM can be saved with the block three-term recurrence to enlarge the subspace.

The residual norm of $\mathcal{F}$-EKSM is $\|f(A)b - V_{2m}f(A_m)e_1\|$; however, it is not directly computable because $f(A)b$ is unknown. One stopping criterion for the Arnoldi approximation is to compute $|h_{m+1,m}e_m^* f(A_m)e_1|$ [30,45,46]; however, this may not be valid for RKSM. Another possibility is to compute $\|q_m - q_{m-1}\|$, which is the norm of the difference between two computed approximations; see, e.g., [21,47]. Alternatively, it is possible to monitor the angle $\angle(q_m, q_{m-1})$ between the approximations [48] in two consecutive iterations. This convergence criterion is sometimes used in the literature on eigenvalue computations; see, e.g., [49,50]. The two criteria usually exhibit very similar behavior. In this section, we choose the latter.

For all RKSM, linear system solvers are required in common, as the action of $(\gamma_m I - \eta_m A)^{-1}$ to vectors is needed in (2). For EKSM and $\mathcal{F}$-EKSM, respectively, we need the action of A^{-1} and $(I - A/s)^{-1}$ to the vectors. If these linear systems can be solved efficiently by direct methods, both of them need only one LU factorization performed one time and applicable to all linear solves. However, because of the adaptive poles, for adaptive RKSM it is necessary to solve a linear system with a different coefficient matrix for every iteration. Although adaptive RKSM achieves an asymptotically optimal convergence rate, it can be more time-consuming than EKSM and $\mathcal{F}$-EKSM, as a new LU factorization in each step is usually much more expensive than a linear solution using existing LU factors.

4. Convergence Analysis of $\mathcal{F}$-EKSM

Next, we study the optimal pole s for $\mathcal{F}$-EKSM to achieve the lowest upper bound on the convergence factor through min–min optimization.

4.1. General Convergence Analysis

In this section, we explore the asymptotic convergence of $\mathcal{F}$-EKSM. Consider the class of Markov-type functions $f(z)$ in (8). For any $a \geq 0$, a Markov-type function can be split into the sum of two integrals [14]:

$$f(z) = \int_{-\infty}^{0} \frac{d\mu(\zeta)}{z-\zeta},\ z \in \mathbb{C}\backslash(-\infty, 0], \tag{8}$$

$$f(z) = f_1(z) + f_2(z), \quad \text{where} \quad f_1(z) = \int_{-\infty}^{-a} \frac{d\mu(\zeta)}{z-\zeta},\ f_2(z) = \int_{-a}^{0} \frac{d\mu(\zeta)}{z-\zeta}. \tag{9}$$

Here, we let $W_1 := W(A) = \{w^* A w : \|w\|_2 = 1\}$ be the numerical range of matrix A and define $W_2 := \{\frac{s}{z-s} | z \in W_1\}$, where s is the repeated pole for $\mathcal{F}$-EKSM. We assume that W_1 is symmetric with respect to the real axis $\mathbb{R}$ and lies strictly in the right half of the complex plane. Then, for $s < 0$, W_2 is symmetric with respect to the real axis $\mathbb{R}$ and lies in the left half of the complex plane. We define $\phi_i : \bar{\mathbb{C}}\backslash W_i \to \bar{\mathbb{C}}\backslash D$ as the direct Riemann mapping [51] for W_i $(i = 1, 2)$, where D is the unit disk, and define $\psi_i = \phi_i^{-1}$ as the inverse Riemann mapping.

Our convergence analysis initially follows the approach in [21], then analyzes a special min–min optimization. It first uses the Faber polynomials [52], providing a rational expansion of functions for investigating the approximation behavior of $\mathcal{F}$-EKSM. The main challenge is to find how the fixed real pole s impacts the convergence and to determine the optimal s to achieve the lowest upper bound on the convergence factor.

Lemma 1. *For the Markov-type function defined by (8) (where μ is a measure such that the integral converges absolutely) and some given $a > 0$, the following inequalities hold for any $m \in \mathbb{N}, m > 1$:*

$$\left| f_1(z) - \sum_{k=0}^{m-1} \gamma_{1,k} F_{1,k}(z) \right| \leq c_1 |\phi_1(-a)|^{-m}, \qquad z \in W_1, \tag{10}$$

$$\left| f_2(z) - \sum_{k=0}^{m-1} \gamma_{2,k} F_{2,k}\left(\frac{1}{z/s-1}\right) \right| \leq c_2 |\phi_2(z_0)|^{-m}, \qquad z \in W_1, \tag{11}$$

where $|\phi_2(z_0)| = \min\{|\phi_2(-1)|, |\phi_2(\frac{s}{-a-s})|\}$ and where $|\phi_1(-a)|$, $|\phi_2(-1)|$, and $|\phi_2(\frac{s}{-a-s})|$ are all greater than 1. Here, for $i = 1, 2$, $\gamma_{i,k}$ are some real numbers and $F_{i,k}$ denotes the Faber polynomials of degree k associated with the Riemann mapping ϕ_i, while c_1 and c_2 are constant positive real numbers independent of m.

Proof. The proof is provided in Appendix A. □

Lemma 2. *Assume that $\|b\|_2 = 1$. For any given $a \geq 0$ used in (9), the error of approximating $f(A)b$ by $\mathcal{F}$-EKSM with cyclic poles $s, +\infty, s, +\infty, \ldots$ satisfies*

$$\|f(A)b - V_{2m} f(A_m) e_1\| \leq c_8 \min\left\{ |\phi_1(-a)|, |\phi_2(-1)|, \left|\phi_2\left(\frac{s}{-a-s}\right)\right| \right\}^{-m}.$$

Proof. Let us define

$$g(z) = f_1(z) - \sum_{k=0}^{m-1} \gamma_{1,k} F_{1,k}(z), \quad h(z) = f_2(z) - \sum_{k=0}^{m} \gamma_{2,k} F_{2,k}\left(\frac{1}{z/s-1}\right).$$

Because both g and h are analytic in W_1, and as $W(T_{2m}) \subset W_1$, from Theorem 2 in [53] we have

$$\|g(A)\|, \|g(A_m)\| \leq 11.08 \max_{z \in W_1} |g(z)|, \ \|h(A)\|, \|h(A_m)\| \leq 11.08 \max_{z \in W_1} |h(z)|.$$

Next, we follow the proof in Section 3, Theorem 3.4 in [21], and use the above inequality:

$$\begin{aligned}
&\|f(A)b - V_{2m} f(A_m) e_1\| = \Big\| f_1(A)b - \textstyle\sum_{k=0}^{m-1} \gamma_{1,k} F_{1,k}(A)b - V_{2m} f_1(A_m) e_1 \\
&\quad + V_{2m} \textstyle\sum_{k=0}^{m-1} \gamma_{1,k} F_{1,k}(A_m) e_1 + f_2(A)b - \sum_{k=0}^{m} \gamma_{2,k} F_{2,k}\left(s(A - sI)^{-1}\right) b \\
&\quad - V_{2m} f_2(A_m) e_1 + V_{2m} \textstyle\sum_{k=0}^{m} \gamma_{2,k} F_{2,k}\left(s(A_m - sI)^{-1}\right) e_1 \Big\| \\
&= \|g(A)b - V_{2m} g(A_m) e_1 + h(A)b - V_{2m} h(A_m) e_1\| \\
&\leq \|g(A)\| + \|g(A_m)\| + \|h(A)\| + \|h(A_m)\| \\
&\leq 2 \times 11.08 \left(\max_{z \in W_1} |g(z)| + \max_{z \in W_1} |h(z)| \right) \\
&\leq 22.16 \left(c_1 |\phi_1(-a)|^{-m} + c_2 \min\{|\phi_2(-1)|, |\phi_2(\tfrac{s}{-a-s})|\}^{-m} \right) \\
&\leq c_8 \min\{|\phi_1(-a)|, |\phi_2(-1)|, |\phi_2(\tfrac{s}{-a-s})|\}^{-m}.
\end{aligned}$$

□

Remark 1. *The results of our analysis can additionally be applied to a linear combination of several Markov-type functions with monomials z^l, $l \in \mathbb{Z}^+$. One example of these functions is $f(z) = z^\nu$, $\nu \in (0, 1)$, as $z^\nu = z z^{\nu-1}$ and $z^{\nu-1}$ is a Markov function. In addition, if the support of the underlying measure of the Markov function is a proper subset of $(-\infty, 0]$, the error bound may not*

be sharp. The asymptotic convergence might be superlinear as well; see, e.g., [28]. While this idea could be explored with $\mathcal{F}$-EKSM, it is not considered here because we did not observe superlinear convergence in our experiments. This was probably because the effective spectrum of our large test matrices did not shrink quickly enough to exhibit convergence speedup before the stopping criterion was satisfied.

To find the optimal pole to achieve the lowest upper bound on the asymptotic convergence factor of $\mathcal{F}$-EKSM, we need to determine $s \leq 0$ such that

$$\min\left\{ |\phi_1(-a)|, |\phi_2(-1)|, \left|\phi_2\left(\frac{s}{-a-s}\right)\right| \right\}$$

is maximized. Let us define

$$\rho(s,a) = 1/\min\left\{ |\phi_1(-a)|, |\phi_2(-1)|, \left|\phi_2\left(\tfrac{s}{-a-s}\right)\right| \right\}. \tag{12}$$

Therefore, $\mathcal{F}$-EKSM converges at least linearly with a convergence factor $\rho(s,a) < 1$, where ρ depends on s and a. For any given pole $s \leq 0$, we can find the artificial parameter $a > 0$ used in (9) such that it minimizes $\rho(s,a)$. Let $\tilde{\rho}(s)$ be the minimized $\rho(s,a)$; then, we need to find $s \leq 0$ that minimizes $\tilde{\rho}(s)$. We denote the minimized $\tilde{\rho}(s)$ by ρ^*, which is the lowest upper bound on the asymptotic convergence factor.

In summary, to find the optimal pole s needed to obtain the lowest upper bound on the asymptotic convergence factor of $\mathcal{F}$-EKSM, we can solve the following optimization problem:

$$\rho^* = \min_{s\leq 0} \tilde{\rho}(s) = 1/\max_{s\leq 0} \max_{a\geq 0} \min\left\{ |\phi_1(-a)|, |\phi_2(-1)|, \left|\phi_2\left(\tfrac{s}{-a-s}\right)\right| \right\}. \tag{13}$$

The asymptotic convergence factor of $\mathcal{F}$-EKSM in (13) is dependent on the Riemann mapping ϕ. The formula of ϕ is different for matrices with different numerical ranges, which leads to different values of ρ^*. In the following section, we show that this problem has an analytical solution if A is symmetric positive definite.

4.2. The Symmetric Positive Definite Case

To explore the optimal pole s and the corresponding bound on the convergence factor of $\mathcal{F}$-EKSM, we can consider a symmetric positive definite matrix A. Assume that $\alpha, \beta > 0$ are the smallest and the largest eigenvalues of A, respectively. The Riemann mappings ϕ_1, ϕ_2 are

$$\phi_1(z) = \begin{cases} \frac{z-c}{d} + \sqrt{\left(\frac{z-c}{d}\right)^2 - 1}, & \Re(z-c) > 0 \\ \frac{z-c}{d} - \sqrt{\left(\frac{z-c}{d}\right)^2 - 1}, & \Re(z-c) < 0, \end{cases} \qquad z \in \bar{\mathbb{C}}\backslash W_1 \tag{14}$$

$$\phi_2(z) = \begin{cases} \frac{z-\widehat{c}}{\widehat{d}} + \sqrt{\left(\frac{z-\widehat{c}}{\widehat{d}}\right)^2 - 1}, & \Re\left(\frac{z-\widehat{c}}{\widehat{d}}\right) > 0 \\ \frac{z-\widehat{c}}{\widehat{d}} - \sqrt{\left(\frac{z-\widehat{c}}{\widehat{d}}\right)^2 - 1}, & \Re\left(\frac{z-\widehat{c}}{\widehat{d}}\right) < 0, \end{cases} \qquad z \in \bar{\mathbb{C}}\backslash W_2, \tag{15}$$

where $c = \frac{\alpha+\beta}{2}$, $d = \frac{\beta-\alpha}{2}$, $\widehat{c} = \frac{1}{2}(\frac{s}{\alpha-s} + \frac{s}{\beta-s})$ and $\widehat{d} = \frac{1}{2}(\frac{s}{\beta-s} - \frac{s}{\alpha-s})$. It follows that

$$\begin{aligned} |\phi_1(-a)| &= M + \sqrt{M^2-1}, && \text{where } M = \frac{c+a}{d} > 0, \\ |\phi_2(-1)| &= N_1 + \sqrt{N_1^2-1}, && \text{where } N_1 = \frac{1+\widehat{c}}{\widehat{d}} > 0, \text{ and} \\ \left|\phi_2\left(\frac{s}{-a-s}\right)\right| &= |N_2| + \sqrt{N_2^2-1}, && \text{where } N_2 = \frac{\frac{s}{-a-s} - \widehat{c}}{\widehat{d}}. \end{aligned} \tag{16}$$

Note that all the three expressions are the values of the function $q(t) = t + \sqrt{t^2 - 1}$ at different values of t. Therefore, to compare $|\phi_1(-a)|$, $|\phi_2(-1)|$, and $\left|\phi_2\left(\frac{s}{-a-s}\right)\right|$, it is sufficient to compare M, N_1, and $|N_2|$, which is much easier.

Lemma 3. *Using the notation provided in* (14)–(16),

$$\max_{a\geq 0} \min\{M, N_1, |N_2|\} = \begin{cases} \dfrac{\alpha+\beta-2\alpha\beta/s}{\beta-\alpha}, & \text{if } s \in (-\infty, s_0] \\ \dfrac{\alpha+\beta-2s+2\sqrt{(\alpha-s)(\beta-s)}}{\beta-\alpha}, & \text{if } s \in (s_0, 0], \end{cases}$$

where $s_0 = -\frac{\sqrt{\alpha\beta}}{\kappa^{1/6}+\kappa^{-1/6}}$ *and* $\kappa = \beta/\alpha$ *is the condition number of matrix* A.

Proof. The proof is provided in Appendix B. □

We are now ready to show the major result regarding the optimal pole and the corresponding lowest upper bound on the asymptotic convergence factor of $\mathcal{F}$-EKSM for approximating $f(A)b$ of Markov-type functions.

Theorem 1. *Let* $\tilde{\rho}(s) = \min_{a\geq 0} \rho(s, a)$ *be the convergence factor of* $\mathcal{F}$*-EKSM for approximating* $f(A)b$ *as defined in* (12)*, where the matrix* A *is symmetric positive definite. Then, for the optimization problem*

$$\rho^* = \min_{s\leq 0} \min_{a\geq 0} \rho(s, a) = \min_{s\leq 0} \tilde{\rho}(s),$$

the optimal solution is

$$s^* = s_0 = -\frac{\sqrt{\alpha\beta}}{\kappa^{1/6}+\kappa^{-1/6}} \tag{17}$$

and the optimal objective function value is

$$\rho^* = \frac{1}{Z^* + \sqrt{Z^{*2}-1}}, \text{ where } Z^* = \frac{\kappa+1+2\sqrt{\kappa}\left(\kappa^{1/6}+\kappa^{-1/6}\right)}{\kappa-1}.$$

Proof. It is equivalent to find the optimal s that solves the following problem:

$$\max_{s\leq 0} T, \quad \text{where} \quad T = \max_{a\geq 0} \min\{M, N_1, |N_2|\}.$$

From Lemma 3, T is a piecewise function with variable s, and we only need to find its maximum value for $s \in (-\infty, 0]$.

For $s \in (-\infty, s_0]$, $T(s) = \frac{\alpha+\beta-2\alpha\beta/s}{\beta-\alpha}$ is a monotonically increasing function; therefore, when $s = s_0$, $T(s)$ has its maximum value on this interval.

For $s \in (s_0, 0]$, $T(s) = \frac{\alpha+\beta-2s+2\sqrt{(\alpha-s)(\beta-s)}}{\beta-\alpha}$. We find the first derivative of $T(s)$ to be

$$T'(s) = -\frac{1}{\beta-\alpha} \frac{\left(\sqrt{\alpha-s}+\sqrt{\beta-s}\right)^2}{\sqrt{(\alpha-s)(\beta-s)}} < 0.$$

Because $T(s)$ decreases monotonically on $(s_0, 0]$, it has its maximum value when $s = s_0$. Therefore,

$$\max_{s\leq 0} T(s) = T(s_0) = \frac{\alpha+\beta-2\alpha\beta/s_0}{\beta-\alpha} = \frac{\kappa+1+2\sqrt{\kappa}\left(\kappa^{1/6}+\kappa^{-1/6}\right)}{\kappa-1}. \tag{18}$$

To sum up, for both $s \in (-\infty, s_0]$ and $s \in (s_0, 0]$, the maximizer of $T(s)$ is $s = s_0$. Consequently, it is the global optimal solution for $s \in (-\infty, 0]$.

With $s = s_0$, we can now return to (13) and (16) and obtain

$$\rho^* = \frac{1}{\max_{s\le 0}\max_{a\ge 0}\min\{|\phi_1(-a)|, |\phi_2(-1)|, |\phi_2(\frac{s}{-a-s})|\}} = \frac{1}{T(s_0)+\sqrt{T(s_0)^2-1}},$$

where $T(s_0)$ is provided in (18). The proof is established. □

4.3. Nonsymmetric Case

Similar to the SPD case, to explore the lowest upper bound on the convergence factor of $\mathcal{F}$-EKSM we can consider a nonsymmetric matrix A with eigenvalues located in the right half of the complex plane. Let α, β, $\gamma > 0$ and assume that the numerical range of matrix A can be covered by an ellipse centered at point $c = \frac{\alpha+\beta}{2}$ with a semi-major axis $d = \frac{\beta-\alpha}{2}$ and semi-minor axis γ.

The Riemann mapping ϕ_1 is provided by

$$\phi_1(z) = \frac{c-z}{d}\frac{1+\eta}{2} + \sqrt{\left(\frac{c-z}{d}\frac{1+\eta}{2}\right)^2 - \eta}, \quad z \in \bar{\mathbb{C}}\backslash W_1,$$

where $\eta = \frac{d-\gamma}{d+\gamma}$. Although the Riemann mapping ϕ_2 is not easy to derive explicitly, for a given s we can first approximate W_2 by a polygon, then use the Schwarz–Christoffel mapping toolbox [54] to approximate ϕ_2 numerically. Then, we can compare $|\phi_1(-a)|$, $|\phi_2(-1)|$, and $|\phi_2(\frac{s}{-a-s})|$ for different values of a. Based on (13), we tested different values of s to find the optimal pole such that $\max_{a\ge 0}\min\{|\phi_1(-a)|, |\phi_2(-1)|, |\phi_2(\frac{s}{-a-s})|\}$ is maximized. Table 1 shows the optimal pole s and the upper bound on the asymptotic convergence factor ρ for matrices with different elliptic numerical ranges.

Table 1. Convergence factor of EKSM and $\mathcal{F}$-EKSM for matrices with different elliptical numerical ranges ($\alpha = 1$).

η	**1**	**1**	**0.75**	**0.75**	**0.5**	**0.5**	**0.25**	**0.25**	**0**	**0**
β	$\mathbf{10^2}$	$\mathbf{10^4}$	$\mathbf{10^2}$	$\mathbf{10^4}$	$\mathbf{10^2}$	$\mathbf{10^4}$	$\mathbf{10^2}$	$\mathbf{10^4}$	$\mathbf{10^2}$	$\mathbf{10^4}$
$\rho^*_{\mathcal{F}-EK}$	0.37	0.65	0.46	0.84	0.53	0.87	0.59	0.89	0.64	0.91
ρ_{EK}	0.52	0.82	0.63	0.95	0.71	0.96	0.77	0.97	0.82	0.98
$\frac{\log\rho^*}{\log\rho}$	1.53	2.18	1.69	3.45	1.86	3.77	2.02	4.11	2.17	4.46
s^*	−3.82	−20.6	−2.96	−11.0	−2.92	−14.5	−3.27	−17.6	−3.83	−20.9

In Table 1, $\eta = 1$ indicates the SPD case; with $\eta = 0$, the numerical range becomes a disk. It can be seen that when η decreases, the convergence factor ρ for both EKSM and $\mathcal{F}$-EKSM increases, which implies that in the case of an elliptic numerical range both methods converge significantly more slowly than in the SPD case. In particular, when $\beta = 10^4$ and $\eta = 0$, it takes about 4.5 times as many steps as are needed for the corresponding SPD case ($\beta = 10^4, \eta = 1$). It is worthwhile to compare these two methods with adaptive RKSM to determine whether the slowdown is severe.

Another observation from Table 1 is that the optimal pole s^* in the nonsymmetric case is not far away from that in the SPD case. Hence, it is reasonable to approximate the optimal shift s^* for the nonsymmetric case using the one for the SPD case (see (17)), as the actual optimal s^* based on an accurate estimate of the numerical range is generally difficult to evaluate, if it is even possible at all. Actually, for a nonsymmetric matrix $A \in \mathbb{R}^{n\times n}$, the approximation of s^* using (16) is exactly the optimal pole for its symmetric part $(A + A^*)/2$. Because

$$\begin{aligned} W(A) &= \left\{ x^H A x \mid x \in \mathbb{C}^n, x^* x = 1 \right\} \\ &= \left\{ (p+qi)^H A (p+qi) \mid p, q \in \mathbb{R}^n, p^* p + q^* q = 1 \right\} \\ &= \{ (p^* A p + q^* A q) + (p^* A q - q^* A p) i \mid p, q \in \mathbb{R}^n, p^* p + q^* q = 1 \}, \\ W\left(\frac{A + A^*}{2} \right) &= \left\{ (p+qi)^H \frac{A + A^*}{2} (p+qi) \mid p, q \in \mathbb{R}^n, p^* p + q^* q = 1 \right\} \\ &= \{ p^* A p + q^* A q \mid p, q \in \mathbb{R}^n, p^* p + q^* q = 1 \}, \end{aligned}$$

it is clear that $W\left(\frac{A+A^*}{2}\right) = \Re(W(A))$. If $W(A)$ has an ellipse boundary centered at point $c = \frac{\alpha+\beta}{2}$ with a semi-major axis $d = \frac{\beta-\alpha}{2}$ and semi-minor axis γ, it follows that $W(\frac{A+A^*}{2}) = \{x \mid \alpha \leq x \leq \beta\}$. To obtain such an approximation with respect to s^*, we only need to run a modest number of Arnoldi steps within an acceptable amount of time in order to obtain the approximations with respect to α and β that are needed in (17).

4.4. Convergence Analysis with Blaschke Product

Another convergence analysis for approximating functions of matrices can be seen in [26]. Using the same notation as above and combining Theorem 5.2 with Equation (6.4) in [26], we obtain a bound with the following form:

$$\|f(A)b - V_{2m} f(A_m) e_1\| \leq c \max_{y \in \phi([-\infty, 0])} \frac{1}{|B(y)|} = c \max_{y \in \phi_1([-\infty, 0])} \left| \prod_{j=1}^{2m} \frac{y - w_j}{1 - w_j y} \right|,$$

where $B(y)$ is called the Blaschke product and $w_j = \phi_1(s_j)$. Using the cyclic poles s, ∞ as $\mathcal{F}$-EKSM, we find $w \in [\phi_1(0), \infty]$, $(\phi_1(0) \geq 1)$ to minimize

$$\max_{y \in \phi_1([-\infty, 0])} \left| \frac{y - w}{y(1 - wy)} \right|. \tag{19}$$

Note that for $y \in [\phi_1(0), \infty]$, $\left| \frac{y-w}{y(1-wy)} \right|$ achieves its maximum either when $y = \phi_1(0)$ or when $y = w + \sqrt{w^2 - 1}$. The problem then becomes the following optimization problem for w:

$$\widetilde{\rho^*} = \min_{w \in [\phi_1(0), \infty]} \max \left\{ \frac{\phi_1(0) - w}{\phi_1(0)(1 - w\phi_1(0))}, \left(w - \sqrt{w^2 - 1} \right)^2 \right\}.$$

It can be shown that the minimum is achieved when $\frac{\phi_1(0)-w}{\phi_1(0)(1-w\phi_1(0))} = \left(w - \sqrt{w^2 - 1} \right)^2$. The optimal w is then one root of a fourth-order equation, which is greater than $\phi_1(0)$:

$$-4w_1^2 w^4 + 4w_1(w_1^2 + 1)w^3 + (w_1^2 - 1)^2 w^2 - 4w_1(w_1^2 + 1)w + 4w_1^2 = 0, \tag{20}$$

where $w_1 = \phi_1(0)$.

For the symmetric positive definite case, where $\phi_1(z)$ is defined as in (14), $w_1 = \frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}$; thus, the optimal w in (20) only depends on the condition number of the matrix A.

The convergence analysis for the optimal pole based on [26] involves a quartic function in (20), and it is difficult to to find an explicit formula for the optimal pole. On the other hand, our analysis based on [21] provides an explicit formula for the optimal pole in Theorem 1. Next, in Section 6, we compare the theoretical convergence rates and actual performance for these two optimal poles to different benchmarks.

5. RKSM with Several Cyclic Poles for Approximating $f(A)b$

In our problem setting, the shift-inverse matrix/vector operations for RKSM are performed by factorization-based direct linear solvers; $\mathcal{F}$-EKSM usually outperforms EKSM in both space size and runtime. Compared with adaptive RKSM, $\mathcal{F}$-EKSM often takes more steps but less time to converge for large sparse SPD matrices, although its performance in both space and time can become inferior to adaptive RKSM for certain challenging nonsymmetric problems. To improve the performance of $\mathcal{F}$-EKSM, we consider using a few more fixed repeated poles. The rationale for this strategy is to take a balanced tradeoff between $\mathcal{F}$-EKSM and the adaptive variants of RKSM, ensuring that this variant of RKSM has modest storage and runtime costs.

For example, we can consider such a method based on four repeated poles. Starting with the optimal pole $s_1 < 0$ of $\mathcal{F}$-EKSM (17) and $s_2 = -\beta$ (the negative of the largest real part of all eigenvalues), we apply several steps of adaptive RKSM to find and use new poles until we find at least one pole smaller than s_1 and one pole greater than s_1 (both in terms of modulus). For all poles obtained adaptively during this procedure, we let s_3 be the smallest (in modulus) and s_4 be the largest one. It is not hard to see that the adaptive RKSM steps terminate with the last pole s_f being either s_3 or s_4. Our numerical experience suggests that additional simple adjustment to s_3 or s_4 can help to improve convergence. Specifically, if $s_f = s_3$, then s_3 is divided by a factor of μ; otherwise, s_4 is multiplied by the same factor. Experimentally, we found that $\mu = \sqrt{10}$ provides the best overall performance. Thus, we keep the LU factorizations associated with the four poles, and in each step we choose the pole cyclically from the set $\{s_1, s_2, s_3, s_4\}$.

In fact, we can use convergence analysis with Blaschke product in Section 4.4. If we want to use four cyclic poles, we can solve the following optimization problem:

$$\min_{w_1,w_2,w_3,w_4 \in [\phi_1(0),\infty]} \max_{y \in [\phi_1(0),\infty]} \left| \prod_{j=1}^{4} \frac{y - w_j}{1 - w_j y} \right|.$$

It takes time to compute the optimal w_1, w_2, w_3, w_4 numerically for the specific problem setting, and our numerical experience shows that it takes a similar number of iterations to converge compared to the above $\mathcal{F}$-EKSM variant with four poles.

6. Numerical Experiments

We tested different variants of RKSM for approximating $f(A)b$, where the functions were $f_1(z) = z^{-1/2}$, $f_2(z) = e^{-\sqrt{z}}$, $f_3(z) = \frac{\tanh\sqrt{z}}{\sqrt{z}}$, $f_4(z) = z^{1/4}$, and $f_5(z) = \log(z)$. The first four consist of Markov-type functions and a Markov-type function multiplied with monomials z^l, $l \in \mathbb{Z}$; while the last function f_5 is non-Markov type, our algorithms exhibit similar behavior when approximating $f_5(A)b$ as on the other functions. All experiments were carried out in MATLAB R2019b on a laptop with 16 GB DDR4 2400 MHz memory, Windows 10 operating system, and 2.81 GHz Intel dual-core CPU.

6.1. Asymptotic Convergence of EKSM and $\mathcal{F}$-EKSM

For a real symmetric positive definite matrix A, EKSM with cyclic poles $0, \infty, 0, \infty, \ldots$ converges at least linearly as follows:

$$\|f(A)b - V_{2m} f(A_m) e_1\| \le C\rho^m,$$

where $\rho = \frac{1}{Z+\sqrt{Z^2-1}}$, $Z = \frac{\kappa+1+2\sqrt{\kappa}}{\kappa-1}$ (and see Proposition 3.6 in [21]). Similarly, $\mathcal{F}$-EKSM with cyclic poles $s^*, \infty, s^*, \infty, \ldots$ converges at least linearly with factor

$$\rho^* = \frac{1}{Z^* + \sqrt{(Z^*)^2 - 1}}, \text{ where } Z^* = \frac{\kappa + 1 + 2\sqrt{\kappa}\left(\kappa^{1/6} + \kappa^{-1/6}\right)}{\kappa - 1}$$

because $Z^* > Z$, $\mathcal{F}$-EKSM has a smaller upper bound on the convergence factor than EKSM.

For the optimal pole from (20), using the Blaschke product technique we can denote the method as $\mathcal{F}$-EKSM* and its optimal pole as $\widetilde{s^*} = \phi_1^{-1}(w_1)$, with the convergence factor $\widetilde{\rho^*}$ in (19). Because $\widetilde{s^*}$ is a root of a fourth-order equation, it is difficult to explicitly find its value; thus, we list several examples to compute the poles and convergence factors for both single pole methods. In addition, we list the convergence factors for the shift-inverse Arnoldi method (SI) based on one fixed nonzero pole:

$$\rho_{SI} = \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1 + 2\kappa^{1/4}}$$

(see [26], Corollary 6.4a).

For a matrix A with the smallest eigenvalue $\alpha = 1$ and largest eigenvalue $\beta = \kappa$, Table 2 shows the difference in the upper bounds of their convergence factors; note that the asymptotic convergence factors are independent of the function f.

Table 2. Bounds on the asymptotic convergence factor for EKSM, $\mathcal{F}$-EKSM, and $\mathcal{F}$-EKSM* with optimal pole s^*.

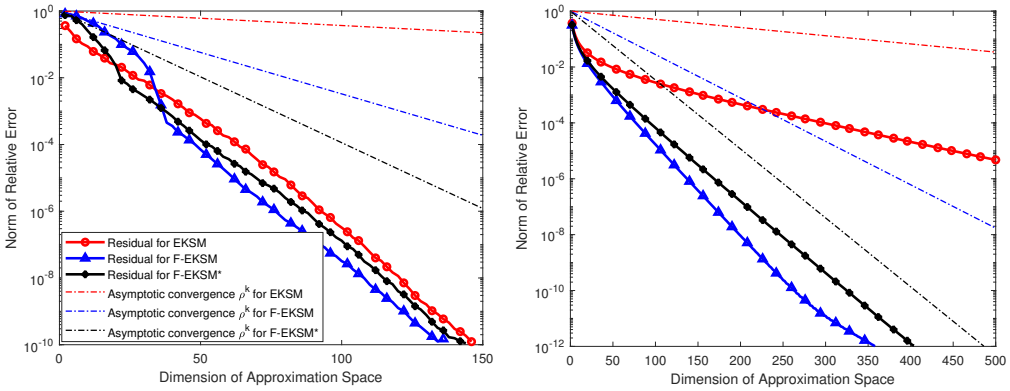
κ	**10**	$\mathbf{10^2}$	$\mathbf{10^3}$	$\mathbf{10^4}$	$\mathbf{10^5}$	$\mathbf{10^6}$	$\mathbf{10^8}$	$\mathbf{10^{10}}$
$\rho^*_{\mathcal{F}EK}$	0.1896	0.3660	0.5195	0.6455	0.7440	0.8182	0.9113	0.9578
$\widetilde{\rho^*}_{\mathcal{F}EK^*}$	0.0537	0.1853	0.3435	0.4945	0.6235	0.7265	0.8628	0.9339
ρ_{EK}	0.2801	0.5195	0.6980	0.8182	0.8935	0.9387	0.9802	0.9937
ρ_{SI}	0.4365	0.6076	0.7370	0.8333	0.8989	0.9405	0.9804	0.9937
$\frac{\log \rho^*}{\log \rho_{EK}}$	1.3069	1.5349	1.8218	2.1814	2.6264	3.1718	4.6448	6.8140
$\frac{\log \rho^*}{\log \widetilde{\rho^*}}$	0.5686	0.5962	0.6128	0.6216	0.6260	0.6281	0.6296	0.6299
s^*	−1.4714	−3.8188	−9.0909	−20.589	−45.4370	−99.010	−463.16	−2153.4
$\widetilde{s^*}$	−0.6058	−1.5527	−3.6568	−8.2269	−18.0917	−39.3540	−183.87	−854.7

It can be seen from Table 2 that $\mathcal{F}$-EKSM has a lower upper bound on the asymptotic convergence factor than EKSM, with $\mathcal{F}$-EKSM* having an even lower upper bound. The optimal pole s^* for $\mathcal{F}$-EKSM is roughly two to three times that of $\widetilde{s^*}$ for $\mathcal{F}$-EKSM*. The shift-inverse Arnoldi has the largest converegence factor; thus, we did not compare it with the other methods in our later tests.

To check the asymptotic convergence factor for each method, it is necessary to know the exact vector $f(A)b$ for a given matrix A and vector b in order to calculate the norm of the residual at each step. For relatively large matrices it is only possible to directly evaluate the exact $f(A)b$ for diagonal matrices within a reasonable time. Because each SPD matrix is orthogonally similar to the diagonal matrix of its eigenvalues, our experiment results can be expected to reflect the behavior of EKSM, $\mathcal{F}$-EKSM, and $\mathcal{F}$-EKSM* applied to more general SPD matrices.

We constructed two diagonal matrices A_1, A_2. The diagonal entry d_j for A_1 is $d_j = (\alpha_1 + \beta_1)/2 + \cos(\theta_j)(\beta_1 - \alpha_1)/2$, $1 \le j \le 10{,}000$, where the θ_js are uniformly distributed on the interval $[0, 2\pi]$ and $\alpha_1 = 10^{-7}$, $\beta_1 = 1$. The diagonal entry d_j for A_2 is $d_j = 10^{-8}\rho_2^{(j-1)}$, $1 \le j \le 20{,}000$, $\rho_2 = 1.001$. We approximated $f_i(A_i)b$ using EKSM, $\mathcal{F}$-EKSM, and $\mathcal{F}$-EKSM*, where b is a fixed vector with entries of standard normally distributed random numbers. The experimental results are shown in Figure 1.

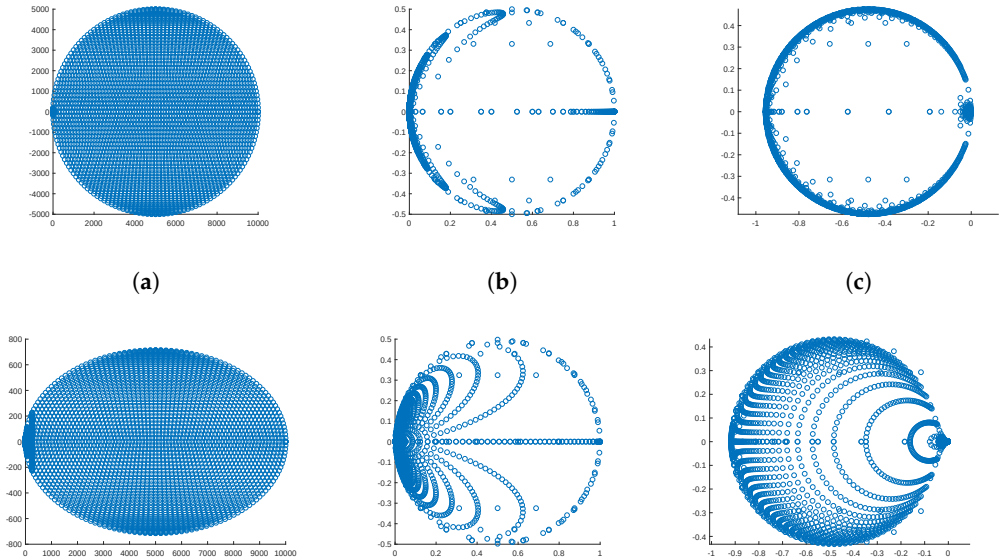
From the figures for different Markov-type functions, it can be seen that all methods converge with factors no worse than their theoretical bounds, which verifies the validity of the results shown in Theorem 1. Moreover, $\mathcal{F}$-EKSM and $\mathcal{F}$-EKSM* always converge faster than EKSM for all the test functions and matrices. In particular, for approximating $f_2(A_2)b$, our $\mathcal{F}$-EKSM only takes about one quarter as many iterations to achieve a relative error less than 10^{-5} as compared to EKSM. Furthermore, the theoretical bounds are not always sharp for those methods.



(**a**) (**b**)

Figure 1. Actual and asymptotic convergence of EKSM, $\mathcal{F}$-EKSM, and $\mathcal{F}$-EKSM* for SPD matrices: (**a**) $f_1(A_1)b$, $\kappa = 10^8$ and (**b**) $f_2(A_2)b$, $\kappa = 4.799 \times 10^8$.

For nonsymmetric matrices with elliptic numerical ranges, the theoretical convergence rate cannot be derived by an explicit formula; Table 1 shows the numerical results. To verify these results, we constructed a block diagonal matrix $A_3 \in \mathbb{R}^{4901\times 4901}$ with 2×2 diagonal blocks with eigenvalues that lie on the circle centered at $z = 5000.5$ with radius 4999.5. We then constructed another block diagonal matrix $A_4 \in \mathbb{R}^{4901\times 4901}$ with eigenvalues that lie in the ellipse centered at $z = 5000.5$, with a semi-major axis 4999.5 and semi-minor axis 714.2. The optimal pole for $\mathcal{F}$-EKSM can be computed using the strategy described in Section 4.2 ($s_3^* = -20.87$ and $s_4^* = -11.02$). Figure 2 shows the spectra of the matrices A_i, A_i^{-1} and $(I - A_i/s_i^*)^{-1}$ for $i = 3, 4$. Figure 3 shows the results for the nonsymmetric matrices. Table 1 shows the asymptotic convergence factors of EKSM and $\mathcal{F}$-EKSM for matrices A_3 and A_4, where $\beta = 10^4$, $\eta = 1$, $\rho^*_{\mathcal{F}EK} = 0.65$, and $\rho_{EK} = 0.82$ for A_3, while $\beta = 10^4$, $\eta = 0.75$, $\rho^*_{\mathcal{F}EK} = 0.84$, and $\rho_{EK} = 0.95$ for A_4.



(**a**) (**b**) (**c**)

(**d**) (**e**) (**f**)

Figure 2. Spectra of several matrices: (**a**) spectrum of A_3; (**b**) spectrum of A_3^{-1}; (**c**) spectrum of $(I - A_3/s_3^*)^{-1}$; (**d**) spectrum of A_4; (**e**) spectrum of A_4^{-1}; (**f**) spectrum of $(I - A_4/s_4^*)^{-1}$.

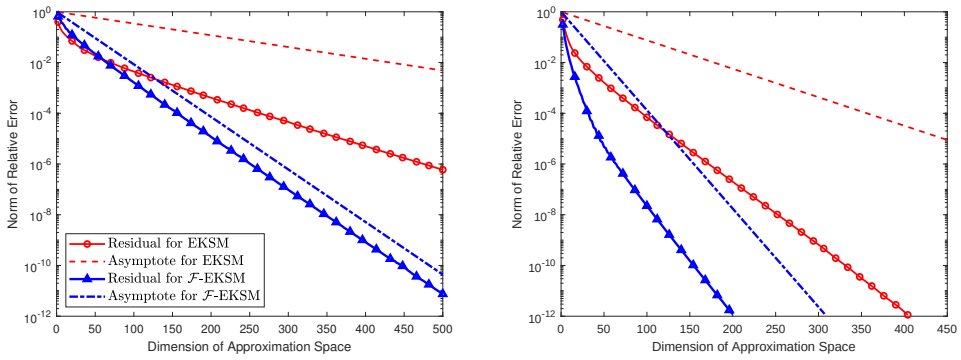


Figure 3. Performance and theoretical asymptotic convergence of EKSM and $\mathcal{F}$-EKSM for nonsymmetric matrices with elliptic spectra: (**a**) approximating $f_1(A_3)b$; (**b**) approximating $f_4(A_4)b$.

Similar to the results for SPD matrices, $\mathcal{F}$-EKSM converges faster than EKSM for the two artificial nonsymmetric problems. The upper bounds on the convergence factor in Table 1 match the actual convergence factor quite well.

6.2. Test for Practical SPD Matrices

Next, we tested EKSM, $\mathcal{F}$-EKSM, $\mathcal{F}$-EKSM*, and adaptive RKSM on several SPD matrices and compared their runtimes and the dimension of their approximation subspaces. Note that for general SPD matrices the largest and smallest eigenvalues are usually not known in advance. The Lanczos method or its restarted variants can be applied to estimate them, and this computation time should be taken into consideration for $\mathcal{F}$-EKSM and $\mathcal{F}$-EKSM*. The variant of EKSM in (5) is not considered in this section, as there is no convergence theory to compare with the actual performance and the convergence rate largely depends on the choice of l and m in (5).

For the SPD matrices, we used Cholesky decomposition with approximate minimum degree ordering to solve the linear systems involving A or $I - A/s$ for all four methods. The stopping criterion of all methods was to check whether the angle between the approximate solutions obtained at two successive steps fell below a certain tolerance. EKSM and $\mathcal{F}$-EKSM, and $\mathcal{F}$-EKSM* all apply the Lanczos three-term recurrence and perform local re-orthogonalization to enlarge the subspace, whereas adaptive RKSM applies a full orthogonalization process with global re-orthogonalization.

We tested four 2D discrete Laplacian matrices of orders 128^2, 256^2, 512^2, and 1024^2 based on standard five-point stencils on a square. For all problems, the vector b was a vector with entries of standard normally distributed random numbers, allowing the behavior of all four methods to be compared for matrices with different condition numbers.

Table 3 reports the runtimes and the dimensions of the rational Krylov subspaces that the four methods entail when applied to all test problems; in the table, EK, FEK, and ARK are abbreviation of EKSM, $\mathcal{F}$-EKSM, and adaptive RKSM, respectively. The stopping criterion was that the angle between the approximate solutions from two successive steps was less than $\tau = 10^{-9}$. The single pole s^* results for $\mathcal{F}$-EKSM are -352.26, -569.84, -915.56, and -1464.6, respectively, while for $\mathcal{F}$-EKSM* the $\widetilde{s^*}$ results are -140.89, -227.29, -364.53, and -582.43, respectively. The shortest CPU time appearing in each line listed in the table is marked in bold.

With only one exception, that of $f_2(z) = e^{-\sqrt{z}}$, it is apparent that $\mathcal{F}$-EKSM converges the fastest of the four methods in terms of wall clock time for all the test functions and matrices with different condition numbers. While $\mathcal{F}$-EKSM takes more steps than adaptive RKSM to converge, it requires fewer steps than EKSM. Furthermore, the advantage of $\mathcal{F}$-EKSM becomes more pronounced for matrices with a larger condition number. Notably, the advantage of $\mathcal{F}$-EKSM in terms of computation time is stronger than in terms of the

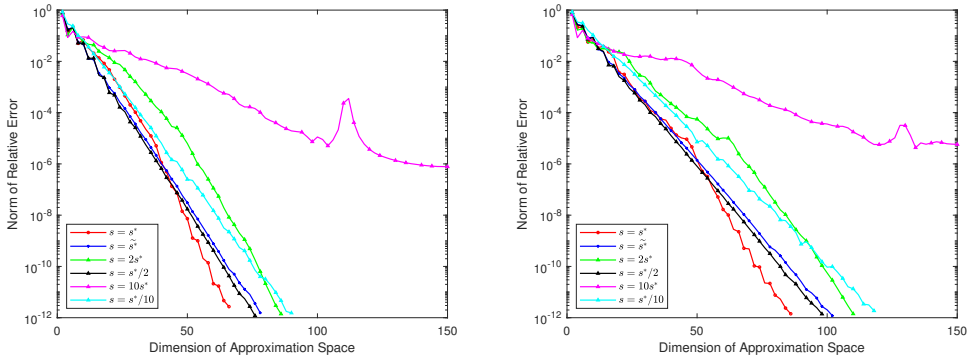
spatial dimension, which is due to the orthogonalization cost being proportional to the square of the spatial dimension. $\mathcal{F}$-EKSM* takes slightly more steps than $\mathcal{F}$-EKSM to converge in these examples, and both methods have similar computation times.

Table 3. Performance of EKSM, $\mathcal{F}$-EKSM, $\mathcal{F}$-EKSM*, and adaptive RKSM on SPD problems.

		Time (s)				Space Dimension			
Function	**Problem**	**EK**	**FEK**	**FEK***	**ARK**	**EK**	**FEK**	**FEK***	**ARK**
f_1	Lap. A	0.28	**0.19**	0.23	0.70	52	42	46	22
	Lap. B	1.10	**0.70**	0.88	3.70	76	52	60	25
	Lap. C	7.35	**4.49**	5.21	19.30	102	66	76	28
	Lap. D	43.36	**24.52**	27.82	100.96	138	84	96	30
f_2	Lap. A	**0.18**	0.20	**0.18**	0.94	32	48	34	28
	Lap. B	**0.48**	0.87	0.63	4.32	30	60	40	32
	Lap. C	**2.66**	5.55	3.62	27.72	38	80	54	38
	Lap. D	**10.44**	27.60	18.10	135.63	36	96	64	39
f_3	Lap. A	0.29	**0.22**	0.26	0.69	52	42	46	22
	Lap. B	1.17	**0.71**	0.88	3.26	76	52	60	25
	Lap. C	7.55	**4.49**	5.28	19.36	102	66	76	28
	Lap. D	43.15	**23.81**	27.61	101.28	138	84	96	30
f_4	Lap. A	0.21	**0.18**	0.20	0.71	50	38	42	23
	Lap. B	0.88	**0.60**	0.74	3.55	66	46	54	27
	Lap. C	6.29	**3.90**	4.63	20.71	90	58	68	30
	Lap. D	36.09	**19.76**	23.76	114.46	120	72	84	33
f_5	Lap. A	0.32	**0.17**	0.24	0.67	48	36	42	21
	Lap. B	1.04	**0.65**	0.77	3.20	66	46	52	24
	Lap. C	6.28	**3.81**	4.53	18.60	88	56	66	27
	Lap. D	35.06	**19.34**	23.90	97.79	116	70	84	29

The unusual behavior of all methods for $f_2(z) = e^{-\sqrt{z}}$ can be explained as follows. For these Laplacian matrices, the largest eigenvalues λ_{max} range from 1.3×10^5 to 8.4×10^6. Because $f_2(\lambda_{\min}) \approx 0.0118$ and $f_2(\lambda_{\max}) \leq f_2(10^5) \approx 4.6 \times 10^{-138}$ (because f_2 decreases monotonically on $[0, \infty)$), the eigenvector components in vector b associated with relatively large eigenvalues would be eliminated in vector $f_2(A)b$ in double precision. In fact, because $\frac{f_2(10^3)}{f_2(\lambda_{min})} \approx 1.6 \times 10^{-12} \approx \tau$, all eigenvalues of A greater than 10^3 are essentially 'invisible' for f_2 under tolerance $\tau = 10^{-12}$, and the *effective* condition number of all four Laplacian matrices is about $\frac{10^3}{\lambda_{min}} \approx 51$. As a result, it takes EKSM the same number of steps to converge for all these matrices; the shift for $\mathcal{F}$-EKSM and $\mathcal{F}$-EKSM* computed using λ_{min} and λ_{max} of these matrices is in fact not optimal for matrices with such a small effective condition number.

The pole s^* in (17) for the SPD matrix is *optimal*, as we have proved that it has the smallest asymptotic convergence factor among all choices of the single pole. In order to numerically compare the behaviors for different setting of the single pole, we tested matrices Lap. B and Lap. C in Table 3 with f_1 and f_3, respectively. For each problem, we tested $\mathcal{F}$-EKSM by setting different single poles s to s^*, $2s^*$, $s^*/2$, $10s^*$, $s^*/10$ and setting $\widetilde{s^*}$ for $\mathcal{F}$-EKSM*. Figure 4 shows the experimental results. It can be seen that $\mathcal{F}$-EKSM has the fastest asymptotic convergence rate among all the 6 different values of single poles when setting the optimal single pole s^*, which confirms that s^* is indeed optimal in our experiments.



(**a**) (**b**)

Figure 4. Convergence of $\mathcal{F}$-EKSM for different setting of single poles: (**a**) f_1 on Lap. B and (**b**) f_3 on Lap. C.

6.3. Test for Practical Nonsymmetric Matrices

We consider 18 nonsymmetric real matrices, all of which have all eigenvalues on the right half of the complex plane. While these are all real sparse matrices, they all have complex eigenvalues with positive real parts. Half are in the form of $M^{-1}K$, where both M and K are sparse and $A = M^{-1}K$ is not formed explicitly. Table 4 reports several features for each matrix A: the matrix size is n, the smallest and largest eigenvalues in terms of absolute value are $|\lambda_{\rm sm}|$ and $|\lambda_{\rm lm}|$, respectively, the smallest and largest real parts of the eigenvalues are $\mathrm{Re}(\lambda_{\rm sr})$ and $\mathrm{Re}(\lambda_{\rm lr})$, respectively, and the largest imaginary part of the eigenvalues is $\mathrm{Im}(\lambda_{\rm li})$. Note that all these original matrices have spectra strictly in the left half of the complex plane; we simply switched their signs to make $f(A)$ well-defined for Markov-type functions.

Table 4. Selected features of the test problems.

Problem	**Size** n	$\lvert\lambda_{\rm sm}\rvert$	$\lvert\lambda_{\rm lm}\rvert$	$\mathrm{Re}(\lambda_{\rm sr})$	$\mathrm{Re}(\lambda_{\rm lr})$	$\mathrm{Im}(\lambda_{\rm li})$	$-s^*$	**Tol**
aerofoilA	16,388	8.41×10^{-2}	1.02×10^{3}	1.04×10^{-2}	1.01×10^{3}	8.12×10^{2}	1.185	10^{-11}
aerofoilB	23,560	2.73×10^{-1}	2.63×10^{2}	2.73×10^{-1}	1.43×10^{2}	2.60×10^{2}	2.447	10^{-13}
matRe500A	3595	4.00×10^{-1}	1.21×10^{2}	-2.01×10^{-1}	1.21×10^{2}	1.12×10^{2}	1.805	10^{-12}
matRe500B	9391	3.00×10^{-1}	5.22×10^{2}	-4.04×10^{-1}	5.22×10^{2}	1.98×10^{2}	3.331	10^{-12}
matRe500C	22,385	2.78×10^{-1}	1.56×10^{3}	-4.49×10^{-1}	1.56×10^{3}	2.70×10^{2}	4.666	10^{-12}
matRe500D	50,527	2.84×10^{-1}	5.11×10^{3}	-3.28×10^{-1}	5.11×10^{3}	3.76×10^{2}	7.163	10^{-12}
matRe500E	110,620	2.62×10^{-1}	1.38×10^{4}	-3.07×10^{-1}	1.38×10^{4}	5.37×10^{2}	9.548	10^{-12}
obstacle	37,168	2.84×10^{-1}	2.95×10^{5}	2.91×10^{-2}	2.95×10^{5}	1.50×10^{2}	6.273	10^{-10}
plate	37,507	1.00×10^{-2}	8.71×10^{4}	7.35×10^{-4}	8.71×10^{4}	1.32×10^{2}	2.021	10^{-9}
tolosa	4000	$1.18 \times 10^{+1}$	4.84×10^{3}	1.56×10^{-1}	1.45×10^{3}	4.62×10^{3}	6.772	10^{-7}
raefsky3	21,200	6.63×10^{-6}	7.99×10^{5}	6.63×10^{-6}	7.99×10^{5}	1.42×10^{0}	0.033	10^{-6}
step	96,307	5.87×10^{-3}	2.18×10^{4}	5.87×10^{-3}	2.18×10^{4}	6.61×10^{1}	0.903	10^{-9}
cavity	37,507	7.53×10^{-3}	2.51×10^{7}	4.99×10^{-3}	2.51×10^{7}	7.16×10^{2}	11.25	10^{-8}
convdiffA	146,689	$2.37 \times 10^{+1}$	5.69×10^{2}	unknown	5.62×10^{2}	3.73×10^{2}	41.32	10^{-13}
convdiffB	146,689	$7.88 \times 10^{+1}$	1.35×10^{3}	unknown	3.98×10^{2}	1.33×10^{3}	317.2	10^{-13}
convdiffC	146,689	$2.14 \times 10^{+2}$	3.84×10^{3}	unknown	3.63×10^{2}	3.83×10^{3}	117.4	10^{-13}
gt01r	7980	5.87×10^{-1}	1.96×10^{4}	1.33×10^{-1}	2.84×10^{3}	1.96×10^{4}	18.09	10^{-11}
venkat	62,424	5.50×10^{-5}	1.08×10^{1}	5.50×10^{-5}	1.08×10^{1}	2.21×10^{0}	0.003	10^{-11}

For the single pole of $\mathcal{F}$-EKSM and the initial pole of the four-pole variant (4Ps) in Section 5, we used the optimal pole for the SPD case matrices (17) by setting $\alpha = \min\{|\lambda_{\rm lm}|, \mathrm{Re}(\lambda_{\rm sr})\}$ and $\beta = \max\{|\lambda_{\rm lm}|, \mathrm{Re}(\lambda_{\rm lr})\}$. The same setting of α and β for building the Riemann mapping $\phi_1(z)$ in (14) was applied to compute the optimal pole for $\mathcal{F}$-EKSM*

in (20). In particular, because precise evaluation of α and β is time-consuming, we approximated them using the 'eigs' function in MATLAB, which is based on the Krylov–Schur Algorithm; see, e.g., [55]. We set the residual tolerance to equal 10^{-3} for 'eigs', ensuring that all the test matrices could find the largest and smallest eigenvalues within a reasonable time. Higher accuracy in computing the eigenvalues is not required when determining the optimal pole, as the convergence performance for $\mathcal{F}$-EKSM does not change noticeably with tiny changes of the value of the pole. Note that the single pole we used for each problem is independent of the Markov-type function; see Table 4. For nonsymmetric matrices, we used LU factorizations to solve the linear systems involving coefficient matrices of A or $I - A/s$ for all methods. The stopping criterion was either when the angle between the approximate solutions was less than a tolerance for two successive steps, or when the dimension of the Krylov subspaces reached 1000. There have been a few discussions about restarting for approximating $f(A)b$, though only for polynomial approximation based on Arnoldi-like methods; see, e.g., [17,56]. In this paper, we only focus on the comparison of convergence rates for several different Krylov methods without restarting. Here, we need to choose a proper tolerance for each problem such that it is small enough to fully exhibit the rate of convergence for all methods while not being too small to satisfy. The last column of Table 4 reports the tolerances, which are fixed regardless of the different Markov-type functions.

In Tables 5–9, we report the runtime and dimension of the approximation spaces that the four methods entail for approximating $f_i(A)b$ to the specified tolerances. The runtime *includes* the time spent on the evaluation of optimal poles for $\mathcal{F}$-EKSM, $\mathcal{F}$-EKSM*, and the four-pole variant. The "$-$" symbol indicates failure to find a sufficiently accurate solution when the maximum dimension of approximation space (1000) was reached. The shortest CPU time appeared in each line of the listed tables is marked in bold. Figure 5 shows an example plot for each method, with the relative error $\sin\angle(q_{k+1}, q_k)$ (where q_k is the approximation to $f(A)b$ at step k) plotted against the dimension of the approximation space for each function.

In the ninety total cases for eighteen problems and five functions shown in Tables 5–9, the four-pole variant is the fastest in runtime in sixty cases and $\mathcal{F}$-EKSM is the fastest in fourteen cases. Among all cases when the four-pole variant is not the fastest, it is no more than 10% slower than the fastest in twelve cases and 10–20% slower in eight cases.

Overall, the four-pole variant is the best in terms of runtime, though there are several exceptions. The first is *tolosa*, which is the only problem on which adaptive RKSM ran the fastest for all functions. For this problem, the dimension of the matrix is relatively small; this makes it more efficient to perform a new linear solver at each step, as the LU cost is cheap. Moreover, for *tolosa*, $\mathrm{Im}(\lambda_{\mathrm{li}})$ is close to $|\lambda_{\mathrm{lm}}|$, meaning that the algorithms based on repeated poles converge slower; see Table 1. The second exception is *venkat*, where the four-pole variant is not the fastest for f_3. In fact, for f_3, EKSM, $\mathcal{F}$-EKSM, and $\mathcal{F}$-EKSM* converge within a much smaller dimension of the approximation space than for the other functions. A possible explanation is that in computer arithmetic it is difficult to accurately capture the relative change in function values for f_3 at small variables, and *venkat* has majority of eigenvalues that are small in terms of absolute value. For example, $\frac{f_3(6\times10^{-5})-f_3(6\times10^{-5}(1+10^{-7}))}{f_3(6\times10^{-5})} = 2.0\times10^{-11}$, which means that a relative change of 10^{-7} in the independent variable of f_3 near 6×10^{-5} can lead to a relative change of 2.0×10^{-11} in function value; thus, f_3 fails to observe such a difference in input above the given tolerance 10^{-10}. The third exception is *convdiffA*, for which EKSM is fastest for four out of all five functions. In fact, EKSM takes fewer steps to converge than $\mathcal{F}$-EKSM, which can be seen at the bottom right of Figure 5. A possible reason for this is that $|\lambda_{\mathrm{lm}}|$ and $\mathrm{Re}(\lambda_{\mathrm{sr}})$ can only be evaluated approximately by several iterations of the Arnoldi method, and sometimes their values cannot be found accurately. The 'optimal' pole based on inaccurate α can sometimes be far away from the real optimal pole. Notable, for this exception $\mathcal{F}$-EKSM* takes fewer steps to converge than EKSM, though it requires more computation time. This is because EKSM uses infinite poles; for the matrix in the form of $M^{-1}K$, where M is an identity matrix, it is not necessary to apply a linear solver for infinite poles, only a simple matrix

vector multiplication. For the other cases in which $\mathcal{F}$-EKSM or $\mathcal{F}$-EKSM* runs fastest, the four-pole variant usually runs only slightly slower, as in those cases it takes less time to enlarge the approximation space than to compute more LU factorizations for using more repeated poles.

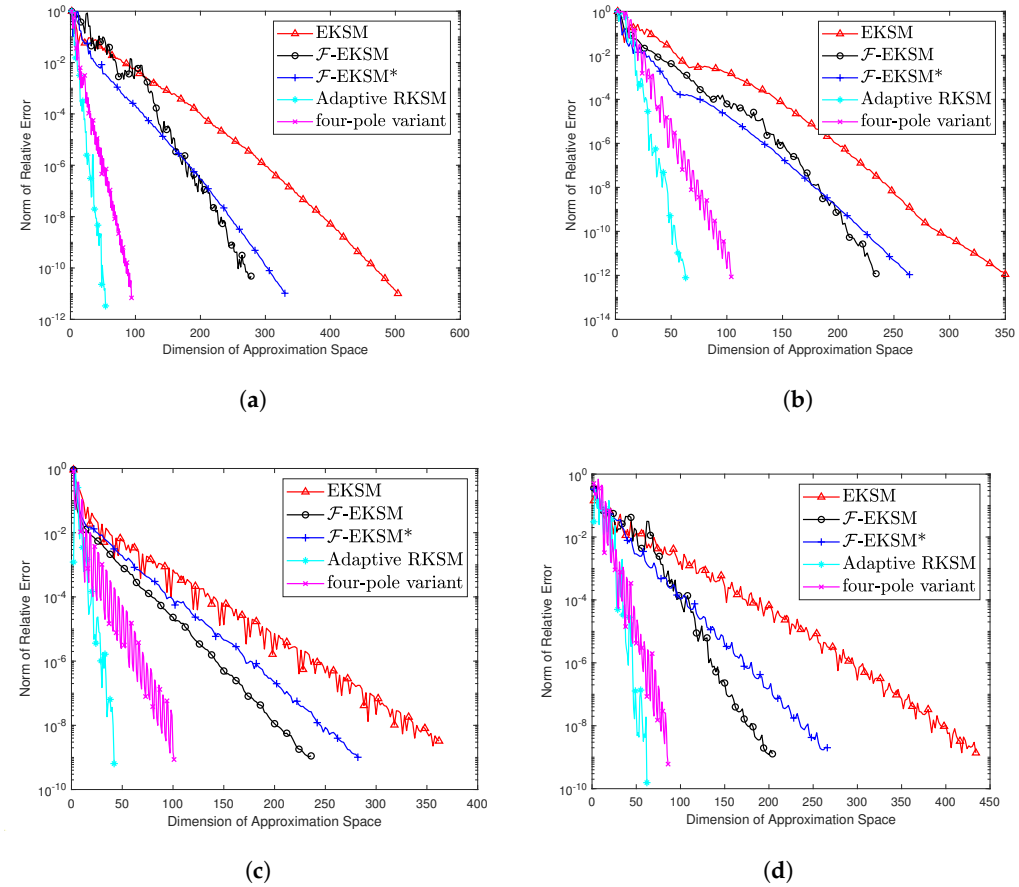


Figure 5. Decay of $\sin\angle(q_k, q_{k+1})$ as the approximation space dimension increases: (**a**) f_1 on aerofoilA; (**b**) f_2 on matRe500E; (**c**) f_3 on plate; (**d**) f_4 on step.

It is important to underscore that the runtime needed for $\mathcal{F}$-EKSM with our optimal pole is less than that for $\mathcal{F}$-EKSM* with an optimal pole derived based on [26] for a majority of the nonsymmetric test matrices, which is similar to the minor advantage in runtime of our $\mathcal{F}$-EKSM shown in Table 3 for Laplacian matrices. In addition, the runtime of the four-pole variant suggests that if sparse LU factorization is efficient for the shifted matrices needed for RKSM, then using a small number of near optimal poles seems to be an effective way to achieve the lowest overall runtime.

In terms of the dimension of the approximation space, adaptive RKSM always need the smallest subspace to converge, with the four-pole variant in second place except for f_2 for *tolosa*. In most, cases EKSM needs the largest subspace to converge, while in others $\mathcal{F}$-EKSM needs the largest subspace. In the cases where $\mathcal{F}$-EKSM and the four-pole variant converge, the four-pole variant takes 7.8% to 87.6% fewer steps.

In summary, our experiments suggest that $\mathcal{F}$-EKSM, $\mathcal{F}$-EKSM*, and the four-pole variant are competitive in reducing the runtime of rational Krylov methods based on direct linear solvers for approximating $f(A)b$; on the other hand, if the goal is to save storage cost, adaptive RKSM is preferable.

Table 5. Performance of five rational Krylov subspace methods for the function $f_1(z) = z^{-1/2}$.

	Time (s)					Space Dimension				
Problem	**EK**	**FEK**	**FEK***	**ARK**	**4Ps**	**EK**	**FEK**	**FEK***	**ARK**	**4Ps**
aerofoilA	29.59	10.43	13.22	23.36	**5.63**	504	278	330	54	94
aerofoilB	6.34	**4.70**	5.34	26.03	5.59	206	138	156	44	75
matRe500A	0.98	0.80	0.95	1.60	**0.77**	140	100	120	41	67
matRe500B	2.81	**2.22**	2.73	6.21	2.37	170	116	142	45	80
matRe500C	7.17	5.39	6.56	17.48	**5.21**	214	146	172	46	88
matRe500D	19.06	16.37	16.18	51.95	**12.80**	264	200	200	50	92
matRe500E	54.03	45.71	43.07	139.99	**30.97**	324	248	236	53	92
obstacle	28.04	14.58	19.74	36.10	**10.65**	376	226	290	50	104
plate	25.16	19.91	23.43	51.20	**15.61**	316	198	240	35	85
tolosa	1.44	1.14	1.47	**0.18**	0.59	174	154	164	31	76
raefsky3	–	–	–	15.79	**6.36**	–	–	–	28	48
step	67.41	45.12	52.46	110.92	**36.01**	364	192	224	40	69
cavity	–	173.99	59.17	39.21	**19.12**	–	938	540	42	134
convdiffA	**16.75**	32.12	19.06	36.87	20.05	108	128	78	39	61
convdiffB	19.47	23.25	14.80	33.62	**14.22**	118	120	86	39	67
convdiffC	17.70	21.86	14.18	34.58	**13.59**	112	116	82	39	63
GT01R	32.07	18.49	8.74	12.58	**4.64**	600	472	328	63	151
venkat	60.62	37.67	13.51	30.22	**8.72**	528	400	210	40	69

Table 6. Performance of five rational Krylov subspace methods for the function $f_2(z) = e^{-\sqrt{z}}$.

	Time (s)					Space Dimension				
Problem	**EK**	**FEK**	**FEK***	**ARK**	**4Ps**	**EK**	**FEK**	**FEK***	**ARK**	**4Ps**
aerofoilA	36.69	8.55	15.60	22.82	**5.77**	528	250	354	55	98
aerofoilB	7.26	**4.41**	5.94	28.71	5.34	218	126	168	50	71
matRe500A	1.14	0.88	1.08	1.82	**0.75**	150	110	130	45	68
matRe500B	3.43	**2.37**	2.95	6.75	2.65	184	126	150	49	88
matRe500C	8.50	**5.57**	7.56	21.21	5.72	228	150	192	57	100
matRe500D	22.45	15.13	17.15	57.99	**13.44**	290	186	210	56	100
matRe500E	61.06	42.70	49.31	168.76	**33.15**	350	234	264	63	104
obstacle	31.66	15.23	22.72	40.76	**9.16**	408	234	312	56	83
plate	35.45	24.05	26.59	64.00	**15.42**	388	244	268	44	81
tolosa	2.41	2.08	2.22	**0.32**	6.92	210	194	198	46	233
raefsky3	–	42.25	110.27	17.21	**8.10**	–	548	804	31	96
step	89.53	46.87	61.80	119.95	**36.88**	434	204	278	44	82
cavity	–	164.15	56.39	46.52	**16.65**	–	824	504	50	102
convdiffA	21.92	31.98	20.91	42.67	**20.55**	126	128	86	44	64
convdiffB	23.26	25.79	15.45	43.14	**14.01**	130	128	92	48	64
convdiffC	22.93	23.13	15.55	41.18	**13.64**	128	120	92	46	64
GT01R	50.13	17.77	10.52	12.87	**3.24**	650	440	352	68	115
venkat	60.80	15.39	14.05	33.22	**8.58**	516	228	214	43	67

Table 7. Performance of five rational Krylov subspace methods for the function $f_3(z) = \frac{\tanh\sqrt{z}}{\sqrt{z}}$.

	Time (s)					Space Dimension				
Problem	**EK**	**FEK**	**FEK***	**ARK**	**4Ps**	**EK**	**FEK**	**FEK***	**ARK**	**4Ps**
aerofoilA	55.22	10.67	21.92	22.07	**5.14**	496	236	340	54	76
aerofoilB	8.53	**4.41**	6.37	27.94	5.70	210	118	158	49	75
matRe500A	1.54	1.09	1.49	1.70	**0.79**	140	104	124	42	59
matRe500B	3.91	**2.52**	3.28	6.53	2.68	172	112	138	46	80
matRe500C	8.77	**5.59**	7.52	17.73	5.75	208	136	174	47	92
matRe500D	24.04	13.09	17.81	55.73	**12.67**	266	150	200	54	88
matRe500E	60.38	33.71	45.53	143.02	**30.38**	322	174	238	54	92
obstacle	35.95	16.59	23.63	35.95	**11.04**	376	226	290	50	104
plate	39.00	25.73	32.25	61.05	**17.28**	362	236	282	42	101
tolosa	2.57	1.99	2.30	**0.19**	0.80	174	154	164	31	76
raefsky3	–	75.56	173.31	17.84	**9.05**	–	556	732	31	100
step	86.80	46.22	62.19	142.57	**36.27**	382	192	268	51	77
cavity	–	45.50	96.82	46.38	**20.37**	–	376	548	50	138
convdiffA	**17.06**	29.98	19.30	35.20	18.25	108	120	78	37	51
convdiffB	20.37	23.29	**14.13**	35.40	14.61	118	120	86	39	67
convdiffC	18.79	21.79	**13.39**	33.17	13.73	112	114	82	38	63
GT01R	83.07	7.72	15.42	12.26	**5.42**	606	242	328	66	151
venkat	**2.48**	5.13	5.08	39.07	8.27	66	64	64	50	59

Table 8. Performance of five rational Krylov subspace methods for the function $f_4(z) = z^{1/4}$.

	Time (s)					Space Dimension				
Problem	**EK**	**FEK**	**FEK***	**ARK**	**4Ps**	**EK**	**FEK**	**FEK***	**ARK**	**4Ps**
aerofoilA	42.22	10.18	17.33	23.64	**5.65**	514	254	344	58	94
aerofoilB	7.24	**4.51**	5.65	28.03	6.04	204	122	156	49	87
matRe500A	1.12	0.88	1.08	1.81	**0.82**	138	100	120	44	71
matRe500B	3.14	**2.35**	2.88	6.98	2.42	168	114	140	50	79
matRe500C	7.63	**5.17**	6.72	20.17	5.27	210	134	170	54	87
matRe500D	19.84	14.03	16.77	58.09	**12.62**	258	166	198	56	88
matRe500E	53.91	38.53	43.11	167.40	**29.82**	316	206	232	63	88
obstacle	30.11	16.19	22.36	47.98	**11.67**	376	234	294	65	116
plate	39.41	26.07	32.72	82.89	**18.73**	388	254	308	56	121
tolosa	1.87	1.68	1.86	**0.15**	0.68	164	150	158	30	68
raefsky3	–	48.82	103.27	23.58	**7.60**	–	526	726	42	86
step	91.72	46.64	59.19	168.31	**36.99**	434	204	266	62	86
cavity	–	128.72	81.85	69.43	**25.41**	–	712	576	75	190
convdiffA	**15.64**	29.43	18.71	36.12	19.66	104	118	76	38	59
convdiffB	19.18	19.98	**14.31**	37.46	16.27	116	110	86	42	75
convdiffC	17.22	19.49	**13.40**	36.65	15.15	110	106	82	41	71
GT01R	51.81	16.50	11.90	13.62	**5.27**	608	388	336	70	163
venkat	82.40	25.55	17.10	37.61	**8.95**	552	294	226	48	67

Table 9. Performance of five rational Krylov subspace methods for the function $f_5(z) = \log(z)$.

	Time (s)					Space Dimension				
Problem	**EK**	**FEK**	**FEK***	**ARK**	**4Ps**	**EK**	**FEK**	**FEK***	**ARK**	**4Ps**
aerofoilA	40.81	10.75	18.15	23.09	**5.89**	516	256	344	55	94
aerofoilB	7.39	**4.53**	5.87	27.81	5.98	206	124	156	49	83
matRe500A	1.38	1.01	1.25	1.79	**0.86**	140	102	122	44	64
matRe500B	3.45	**2.45**	3.07	6.76	2.60	168	114	140	48	80
matRe500C	7.87	**5.26**	7.00	19.28	5.41	210	134	170	52	88
matRe500D	20.17	14.70	16.95	50.55	**12.73**	260	174	198	49	88
matRe500E	54.72	41.43	43.04	145.27	**30.49**	318	222	232	54	92
obstacle	30.77	16.28	22.89	40.04	**11.73**	376	230	294	55	112
plate	34.92	25.61	32.11	72.58	**18.98**	362	250	302	49	117
tolosa	2.63	1.94	2.15	**0.23**	0.92	170	154	160	30	76
raefsky3	–	–	–	17.76	**8.07**	–	–	–	31	88
step	90.39	46.88	59.46	137.08	**37.45**	434	204	266	51	85
cavity	–	157.56	80.11	56.39	**24.96**	–	796	582	61	178
convdiffA	**15.85**	29.75	18.80	35.57	19.58	104	120	76	38	59
convdiffB	18.64	20.67	**13.72**	35.90	15.65	114	112	84	40	71
convdiffC	17.59	19.75	**13.43**	33.00	14.42	110	108	82	38	67
GT01R	47.93	20.42	12.02	12.34	**5.83**	612	426	336	63	163
venkat	72.81	28.59	15.32	36.06	**8.78**	542	324	218	47	68

7. Conclusions

In this paper, we have studied an algorithm called the flexible extended Krylov subspace method ($\mathcal{F}$-EKSM) for approximating $f(A)b$ for Markov-type functions. The central idea is to find an optimal pole to replace the zero pole in EKSM such that $\mathcal{F}$-EKSM needs only the same single LU factorization as EKSM while converging more rapidly.

In the main theoretical contribution of this work, Theorem 1, we prove that there exists a unique optimal pole for a symmetric positive-definite matrix that guarantees the fastest convergence of $\mathcal{F}$-EKSM, which always outperforms EKSM. The theorem provides a formula for both the optimal pole and an upper bound on the convergence factor. Numerical experiments show that $\mathcal{F}$-EKSM is more efficient than EKSM and that it is competitive in runtime compared with adaptive RKSM if the shifted linear systems needed for rational Krylov methods are solved using a direct linear solver.

$\mathcal{F}$-EKSM may lose its advantages for challenging nonsymmetric matrices because of possible failure to compute the optimal poles numerically and due to its relatively slow convergence rate for these problems. This performance can be improved by using four fixed poles chosen flexibly in the early stage of computation. Our numerical results show that the four-pole variant is the most efficient in terms of runtime for many problems.

Author Contributions: Conceptualization, S.X. and F.X.; methodology, S.X. and F.X.; software, S.X.; validation, S.X.; formal analysis, S.X. and F.X.; investigation, S.X.; resources, F.X.; data curation, F.X.; writing—original draft preparation, S.X.; writing—review and editing, F.X.; visualization, S.X.; supervision, F.X.; project administration, F.X.; funding acquisition, F.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the U.S. National Science Foundation under grants DMS-1719461, DMS-1819097, and DMS-2111496.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank Jeong-Rock Yoon for useful conversations and suggestions on the proof of the bounded rotation of $f(z) = \frac{1}{z}$. In addition, we thank the two anonymous reviewers for their helpful comments and suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proof of Lemma 1

Because $\phi_i : \bar{\mathbb{C}}\backslash W_i \to \bar{\mathbb{C}}\backslash D$ $(i = 1, 2)$ and $-a \in \bar{\mathbb{C}}\backslash W_1$, $-1 \in \bar{\mathbb{C}}\backslash W_2$, $\frac{s}{-a-s} \in \bar{\mathbb{C}}\backslash W_2$, we can conclude that $|\phi_1(-a)|$, $|\phi_2(-1)|$ and $\left|\phi_2\left(\frac{s}{-a-s}\right)\right|$ are all greater than 1.

Equation (10) has been proved in Section 3, Lemma 3.1 in [21]. For (11) involving f_2, our proof is analogous to what was done for f_1.

Because ϕ denotes the Riemann mapping, for a fixed $z \in W_1$ we can write the Faber polynomials by means of their generating function [57]:

$$\frac{1}{z-\zeta} = -\frac{1}{\psi'[\phi(\zeta)]} \sum_{k=0}^{\infty} \frac{F_k(z)}{\phi(\zeta)^{k+1}}, \quad z \in W_1, \quad \zeta \notin W_1.$$

We define a new variable $y = \frac{1}{z/s-1}$. Then, from (9), we have

$$\begin{aligned} f_2(z) = f_2\left(s\left(1+\frac{1}{y}\right)\right) &= \int_{-a}^{0} \frac{d\mu(\zeta)}{s(1+\frac{1}{y})-\zeta} = y\int_{-a}^{0} \frac{1}{s-\zeta}\frac{d\mu(\zeta)}{y-\frac{s}{\zeta-s}} \\ &= -y\int_{-a}^{0} \frac{1}{s-\zeta}\frac{1}{\psi_2'\left[\phi_2\left(\frac{s}{\zeta-s}\right)\right]} \sum_{k=0}^{\infty} \frac{F_{2,k}(y)}{\phi_2\left(\frac{s}{\zeta-s}\right)^{k+1}} d\mu(\zeta) \\ &= -y\sum_{k=0}^{\infty} F_{2,k}(y) \int_{-a}^{0} \frac{d\mu(\zeta)}{(s-\zeta)\psi_2'\left[\phi_2\left(\frac{s}{\zeta-s}\right)\right]\phi_2\left(\frac{s}{\zeta-s}\right)^{k+1}}. \end{aligned} \tag{A1}$$

Because W_2 is symmetric with respect to the real axis $\mathbb{R}$ and lies in the left half of the complex plane, ψ_2 monotonically maps $(-\infty, -1)$ onto $(-\infty, \min\{\mathbb{R} \cap W_2\})$ and $(1, +\infty)$ onto $(\max\{\mathbb{R} \cap W_2\}, +\infty)$. We then have the following properties for ψ_2 and ϕ_2:

$$\begin{aligned} &|\phi_2(\zeta)| \geq c_3|\zeta|, \text{ and } \left|\psi_2'[\phi_2(\zeta)]\right| \geq c_4, \quad \zeta \in \mathbb{R}\backslash W_2, \\ &|\phi_2(\zeta)| \geq |\phi_2(z_0)|, \quad \zeta \in (-\infty, z_0) \quad \text{for} \quad z_0 \in (-\infty, \min\{\mathbb{R} \cap W_2\}), \\ &|\phi_2(\zeta)| \geq |\phi_2(z_0)|, \quad \zeta \in (z_0, \infty) \quad \text{for} \quad z_0 \in (\max\{\mathbb{R} \cap W_2\}, +\infty), \end{aligned}$$

where $c_3, c_4 > 0$ are constants independent of ζ.

It follows that the integral in the last expression of (A1) satisfies

$$\begin{aligned} &\left|\int_{-a}^{0} \frac{d\mu(\zeta)}{(s-\zeta)\psi_2'\left[\phi_2\left(\frac{s}{\zeta-s}\right)\right]\phi_2\left(\frac{s}{\zeta-s}\right)^{k+1}}\right| \leq \int_{-a}^{0} \frac{|d\mu(\zeta)|}{|(s-\zeta)|\left|\psi_2'\left[\phi_2\left(\frac{s}{\zeta-s}\right)\right]\right|\left|\phi_2\left(\frac{s}{\zeta-s}\right)^{k+1}\right|} \\ \leq \quad &\int_{-a}^{0} \frac{|d\mu(\zeta)|}{|(s-\zeta)|c_4c_3\left|\frac{s}{\zeta-s}\right|\left|\phi_2\left(\frac{s}{\zeta-s}\right)\right|^{k}} \leq c_5 \min_{\zeta\in(-a,0)}\left|\phi_2\left(\frac{s}{\zeta-s}\right)\right|^{-k}. \end{aligned}$$

Note that for any $\zeta \in (-\infty, 0)$ it is the case that $\frac{s}{\zeta-s} \in (-\infty, -1) \cup (0, \infty) \subset \bar{\mathbb{C}}\backslash W_2$. It follows that $\left|\phi_2\left(\frac{s}{\zeta-s}\right)\right| > 1$.

We note that the map $f(z) = \frac{s}{z-s}$ $(s < 0)$ from W_1 to W_2 can be written as a composition of three maps $f = f_3 \circ f_2 \circ f_1$, where $f_1(z) = z - s$, $f_2(z) = 1/z$, and $f_3(z) = sz$, all of which are bijective. We denote ∂W_1 and ∂W_2 as the boundary of W_1 and W_2, respectively. Let Γ_1 be the image of ∂W_1 under f_1 and let Γ_2 be the image of Γ_1 under f_2 (which is the preimage of ∂W_2 under f_3). Because W_1 is the numerical range of A, it is convex and compact per the Hausdorff–Toeplitz theorem; therefore, ∂W_1 has a boundary rotation of 2π. As f_1 translates ∂W_1 horizontally to the direction of the positive real axis, it preserves the shape of the preimage such that Γ_1 is of bounded rotation as well. In addition, it can be shown that $f_2 = \frac{1}{z}$ maps Γ_1 to Γ_2 with bounded rotation (see details in Lemma A1 shown below). Finally, because $f_3(z) = sz$ preserves the shape of the preimage, ∂W_2 has bounded

rotation. From Chapter IX, Section 3, Theorem 11 in [52], we have $\max_{y\in W_2}|F_{2,k}(y)| \le c_6$, and the following inequality holds for $y, z \in W_2$:

$$\sum_{k=m}^{\infty} F_{2,k}(y)|\phi_2(z)|^{-k} \le c_6 \sum_{k=m}^{\infty} |\phi_2(z)|^{-k} = c_7|\phi_2(z)|^{-m},$$

where c_7 is some real positive constant independent of m.

In light of the above observations, if we denote

$$\gamma_{2,k} = -y\int_{-a}^{0} \frac{d\mu(\zeta)}{(s-\zeta)\psi_2'\left[\phi_2\left(\frac{s}{\zeta-s}\right)\right]\phi_2\left(\frac{s}{\zeta-s}\right)^{k+1}},$$

then

$$\left|f_2(z) - \sum_{k=0}^{m-1}\gamma_{2,k}F_{2,k}\left(\frac{1}{z/s-1}\right)\right| \le \left|y\sum_{k=m}^{\infty}F_{2,k}(y)\int_{-a}^{0}\frac{d\mu(\zeta)}{(s-\zeta)\psi_2'\left[\phi_2\left(\frac{s}{\zeta-s}\right)\right]\phi_2\left(\frac{s}{\zeta-s}\right)^{k+1}}\right|$$
$$\le |y|c_5\sum_{k=m}^{\infty}F_{2,k}(y)\min_{\zeta\in(-a,0)}\left|\phi_2\left(\frac{s}{\zeta-s}\right)\right|^{-k} \le |y|c_7c_5\min_{\zeta\in(-a,0)}\left|\phi_2\left(\frac{s}{\zeta-s}\right)\right|^{-m}$$
$$\le c_2\min_{\zeta\in(-a,0)}\left|\phi_2\left(\frac{s}{\zeta-s}\right)\right|^{-m},$$

where $c_2 = c_7c_5$ is an upper bound of $|y|c_7c_5$ due to $y = \frac{1}{z/s-1} \in [-1, 0)$.

For $\zeta \in (-a, 0)$, there are two cases to derive the minimum of $\left|\phi_2\left(\frac{s}{\zeta-s}\right)\right|$.

Case 1: if $-a \ge s$, then $\frac{s}{\zeta-s} \in [\frac{s}{-a-s}, -1]$. Because W_1 lies strictly in the right half of the plane, by definition W_2 lies between the vertical lines $real(z) = 0$ and $real(z) = -1$; thus, $\min\{\mathbb{R}\cap W_2\} \ge -1$ and we have

$$\min_{\zeta\in(-a,0)}\left|\phi_2\left(\frac{s}{\zeta-s}\right)\right| = |\phi_2(-1)|.$$

Case 2: if $-a < s$, then $\frac{s}{\zeta-s} \in (-\infty, -1]\cup[\frac{s}{-a-s}, +\infty)$; clearly, $\frac{s}{-a-s} \ge 0$, and because $\max\{\mathbb{R}\cap W_2\} \le 0$, we have

$$\min_{\zeta\in(-a,0)}\left|\phi_2\left(\frac{s}{\zeta-s}\right)\right| = \min\left\{|\phi_2(-1)|, \left|\phi_2\left(\frac{s}{-a-s}\right)\right|\right\} = |\phi_2(z_0)|.$$

Note that the conclusion of Case 2 is valid for Case 1; therefore,

$$\left|f_2(z) - \sum_{k=0}^{m-1}\gamma_{2,k}F_{2,k}\left(\frac{1}{z/s-1}\right)\right| \le c_2|\phi_2(z_0)|^{-m}.$$

Lemma A1. *Define the mapping $f : \Gamma_1 \mapsto \Gamma_2$, $f(z) = \frac{1}{z}$, where $\Gamma_1 \subset \mathbb{C}$ is the boundary of a compact convex domain lying strictly in the right half of the complex plane and is symmetric with respect to the real axis. Let Γ_1 be the image of the interval $[0, 2\pi)$ under the injection $\gamma(t)$, which is assumed to be absolutely continuous. Then, Γ_2 has bounded rotation.*

Proof. We define $I_1 \subset [0, 2\pi)$ as the subset where $\gamma'(t)$ is continuous. Note that the directional angle of the tangent line to $\gamma(t)$ at t is $\theta(t) = \arg[\gamma'(t)]$. The boundary rotation of Γ_1 is defined in Page 270 of reference [58] as follows:

$$\mathrm{BD}(\Gamma_1) = \int_0^{2\pi}|d\theta(t)| = \int_{I_1}|d\theta(t)| + \int_{\gamma^{-1}(\Gamma_1)\setminus I_1}|d\theta(t)| = 2\pi,$$

which is due to the convexity of the domain enclosed by Γ_1.

Note that Γ_2 is the image of the interval $[0, 2\pi)$ under the injection $(f \circ \gamma)(t)$. Similarly, the directional angle of the tangent line to $(f \circ \gamma)(t)$ at t is

$$\phi(t) = \arg[(f \circ \gamma)'(t)] = \arg[f'(\gamma(t))\gamma'(t)] = \arg[f'(\gamma(t))] + \arg[\gamma'(t)] = \arg[f'(\gamma(t))] + \theta(t).$$

Because f is a conformal mapping that preserves the angles, the variation of the directional angle at the discontinuities of $\gamma'(t)$ (if any) are preserved. This can be written as $\int_{\gamma^{-1}(\Gamma_1)\setminus I_1} |d\phi(t)| = \int_{\gamma^{-1}(\Gamma_1)\setminus I_1} |d\theta(t)|$. The boundary rotation of Γ_2 can be written as

$$\begin{aligned}
&\mathrm{BD}(\Gamma_2) = \int_{I_1} |d\phi(t)| + \int_{\gamma^{-1}(\Gamma_1)\setminus I_1} |d\phi(t)| = \int_{I_1} |d\phi(t)| + \int_{\gamma^{-1}(\Gamma_1)\setminus I_1} |d\theta(t)| \\
&\leq \int_{I_1} |d\arg[f'(\gamma(t))]| + \int_{I_1} |d\theta(t)| + \int_{\gamma^{-1}(\Gamma_1)\setminus I_1} |d\theta(t)| = \int_{I_1} |d\arg[f'(\gamma(t))]| + \mathrm{BD}(\Gamma_1),
\end{aligned}$$

where

$$|d\arg[f'(\gamma(t))]| = |d\arg[-1/\gamma^2(t)]| = 2|d\arg[\gamma(t)]|.$$

Because Γ_1 is the boundary of a compact convex domain, lies strictly in the right half of the complex plane, and is symmetric with respect to the real axis, there exists $\theta_0 \in (0, \frac{\pi}{2})$ such that

$$\max_{t\in[0,2\pi)} \arg[\gamma(t)] = \theta_0, \quad \min_{t\in[0,2\pi)} \arg[\gamma(t)] = -\theta_0,$$

and we can select $t^+, t^- \in [0, 2\pi)$ such that

$$\arg[\gamma(t^+)] = \theta_0, \quad \arg[\gamma(t^-)] = -\theta_0.$$

Note that while these may not be unique, they split Γ_1 into two disjoint continuous branches, denoted as Γ_1^+ and Γ_1^-, on both of which $\arg[\gamma(t)]$ is monotonic (though not necessarily strictly) with respect to t. It follows that

$$\begin{aligned}
&\int_{I_1} |d\arg[f'(\gamma(t))]| = 2\int_{I_1} |d\arg[\gamma(t)]| \\
=&2\left(\int_{\gamma(t)\in\Gamma_1^+} |d\arg[\gamma(t)]| + \int_{\gamma(t)\in\Gamma_1^-} |d\arg[\gamma(t)]|\right) = 2(2\theta_0 + 2\theta_0) = 8\theta_0,
\end{aligned}$$

because the two end points of Γ_1^+ and Γ_1^- are $\gamma(t^+)$ and $\gamma(t^-)$, respectively. Here, $\theta_0 < \frac{\pi}{2}$; thus, we have

$$\int_{I_1} |d\arg[f'(\gamma(t))]| < 4\pi.$$

Moreover, because $\mathrm{BD}(\Gamma_1) = 2\pi$, we have $\mathrm{BD}(\Gamma_2) < 4\pi + 2\pi = 6\pi$. The claim is established. □

Appendix B. Proof of Lemma 3

We first define $t = \min\{M, N_1, |N_2|\}$ and $T = \max_{a\geq 0} t$. For a fixed pole $s \leq 0$, M, N_1, N_2 are functions of the variable a; specifically, M is a linear function of a, N_1 is a constant independent of a, and N_2 is linear to the reciprocal of a shifted value of a (see (16)). Here, we are interested in their absolute values.

We can set up a Cartesian coordinate system to illustrate M, N_1, and $|N_2|$ as functions of $-a$ and compare their values. The horizontal asymptote of the function N_2 is $f = -\frac{\widehat{c}}{\widehat{d}}$. First, we need to compare this with N_1.

Case 1: $-\frac{\widehat{c}}{\widehat{d}} \geq \frac{1+\widehat{c}}{\widehat{d}} \Rightarrow s \in \left(-\infty, -\sqrt{\alpha\beta}\right]$.

The illustration is shown in Figure A1. Because N_1 is constant, $T \leq N_1$, and for $-a \to -\infty$, $N_1 < N_2 < M$; thus, $t = N_1$ such that $T \geq N_1$. To sum up, in this case we have $T = N_1 = \frac{1+\widehat{c}}{\widehat{d}} = \frac{\alpha+\beta-2\alpha\beta/s}{\beta-\alpha}$.

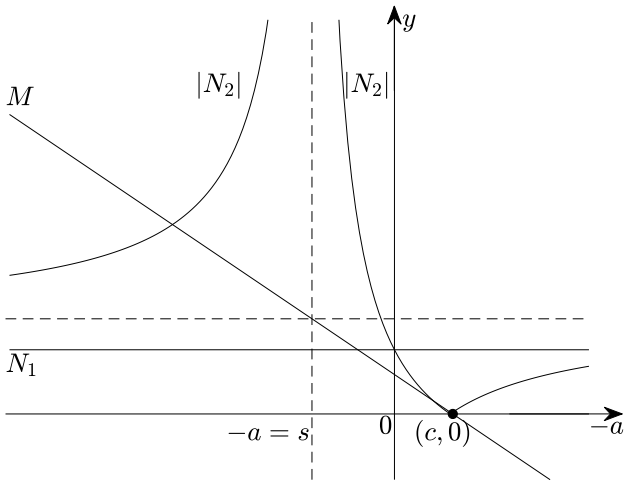


Figure A1. Sketch map of Case 1 in proof of Lemma 3.

Case 2: $-\frac{\widehat{c}}{\widehat{d}} < \frac{1+\widehat{c}}{\widehat{d}} \Rightarrow s \in \left(-\sqrt{\alpha\beta}, 0\right]$.

There is an intersection between N_1 and N_2 for $-a < s$, which we denote as p_1. Solving the equation $N_1 = N_2$ for a, we obtain $-a = s\left(\frac{1}{2\widehat{c}+1} + 1\right)$. Note that p_1 can be either above or below the line of M.

First, if p_1 is below or on the line of M, then

$$\frac{c - s\left(\frac{1}{2\widehat{c}+1} + 1\right)}{d} \geq \frac{1+\widehat{c}}{\widehat{d}} \implies (\alpha+\beta)s^3 - 3\alpha\beta s^2 + \alpha^2\beta^2 \leq 0.$$

Letting $\theta(s) = (\alpha+\beta)s^3 - 3\alpha\beta s^2 + \alpha^2\beta^2$, we need to find the interval of s such that $\theta(s) \leq 0$. Here, $\theta(s)$ is a cubic function and its two stationary points have positive function values; thus, it only has one real root. We can use the Cardano formula [59] to obtain its real root:

$$s_0 = -\frac{\sqrt{\alpha\beta}}{\kappa^{1/6} + \kappa^{-1/6}} \geq -\sqrt{\alpha\beta}.$$

Therefore, for $s \in \left(-\sqrt{\alpha\beta}, s_0\right]$ we have $\theta(s) \leq 0$.

We denote the intersection between M and N_2 as p_2. As shown in Figure A2, when $-a$ is between p_1 and p_2, $T = N_1 = \frac{1+\widehat{c}}{\widehat{d}} = \frac{\alpha+\beta-2\alpha\beta/s}{\beta-\alpha}$.

Second, if p_1 is above the line of M, then $s \in (s_0, 0]$; see Figure A2.

When $M = N_2$, we have

$$\frac{a+c}{d} = \frac{\frac{s}{-a-s} - \widehat{c}}{\widehat{d}} \implies \quad a^2 + 2sa + s(\alpha+\beta) - \alpha\beta = 0.$$

We only need the root for $-a \leq 0$, which is $-a = s - \sqrt{(\alpha-s)(\beta-s)}$. Therefore,

$$\min\{M, N_1, |N_2|\} = \begin{cases} N_2, & -a \in \left(-\infty, s - \sqrt{(\alpha-s)(\beta-s)}\right] \\ M, & -a \in \left(s - \sqrt{(\alpha-s)(\beta-s)}, 0\right] \end{cases}.$$

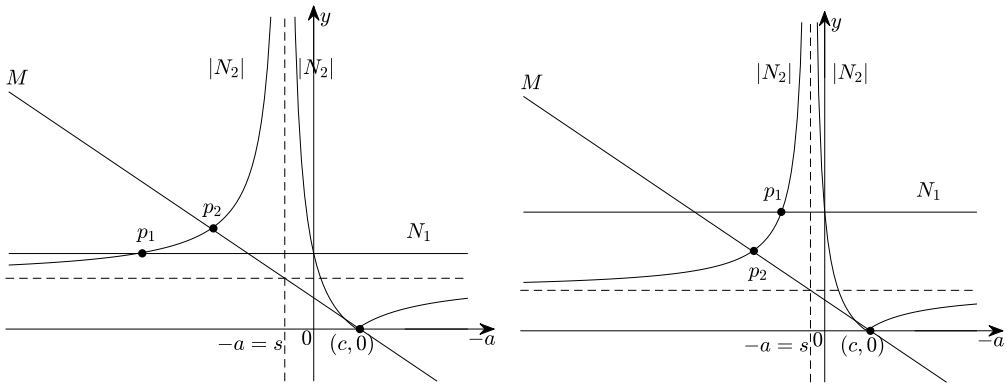


Figure A2. Sketch map of Case 2 in proof of Lemma 3: (**a**) when p_1 is below or on the line of M and (**b**) when p_1 is above the line of M.

It is clear from Figure A2 that the maximum of t is achieved at point p_2, that is, when $-a = s - \sqrt{(\alpha - s)(\beta - s)}$,

$$T = \frac{c - s + \sqrt{(\alpha - s)(\beta - s)}}{d} = \frac{\alpha + \beta - 2s + 2\sqrt{(\alpha - s)(\beta - s)}}{\beta - \alpha}.$$

References

1. Higham, N.J. *Functions of Matrices: Theory and Computation*; Society for Industrial and Applied Mathematics (SIAM): Philadelphia, PA, USA, 2008; pp. xx+425.
2. Schilders, W.H.A.; van der Vorst, H.A.; Rommes, J. (Eds.) *Model Order Reduction: Theory, Research Aspects and Applications*; Mathematics in Industry; Springer: Berlin/Heidelberg, Germany, 2008; Volume 13, pp. xii+471.
3. Grimm, V.; Hochbruck, M. Rational approximation to trigonometric operators. *BIT* **2008**, *48*, 215–229. [CrossRef]
4. Burrage, K.; Hale, N.; Kay, D. An efficient implicit FEM scheme for fractional-in-space reaction-diffusion equations. *SIAM J. Sci. Comput.* **2012**, *34*, A2145–A2172. [CrossRef]
5. Bloch, J.C.; Heybrock, S. A nested Krylov subspace method to compute the sign function of large complex matrices. *Comput. Phys. Commun.* **2011**, *182*, 878–889. [CrossRef]
6. Kressner, D.; Tobler, C. Krylov subspace methods for linear systems with tensor product structure. *SIAM J. Matrix Anal. Appl.* **2010**, *31*, 1688–1714. [CrossRef]
7. Hochbruck, M.; Ostermann, A. Exponential Runge-Kutta methods for parabolic problems. *Appl. Numer. Math.* **2005**, *53*, 323–339. [CrossRef]
8. Hochbruck, M.; Lubich, C. Exponential integrators for quantum-classical molecular dynamics. *BIT* **1999**, *39*, 620–645. [CrossRef]
9. Bai, Z. Krylov subspace techniques for reduced-order modeling of large-scale dynamical systems. *Appl. Numer. Math.* **2002**, *43*, 9–44. [CrossRef]
10. Freund, R.W. Krylov-subspace methods for reduced-order modeling in circuit simulation. *J. Comput. Appl. Math.* **2000**, *123*, 395–421. [CrossRef]
11. Wang, S.; de Sturler, E.; Paulino, G.H. Large-scale topology optimization using preconditioned Krylov subspace methods with recycling. *Internat. J. Numer. Methods Engrg.* **2007**, *69*, 2441–2468. [CrossRef]
12. Biros, G.; Ghattas, O. Parallel Lagrange-Newton-Krylov-Schur methods for PDE-constrained optimization. I. The Krylov-Schur solver. *SIAM J. Sci. Comput.* **2005**, *27*, 687–713. [CrossRef]
13. Simoncini, V. Computational methods for linear matrix equations. *SIAM Rev.* **2016**, *58*, 377–441. [CrossRef]
14. Druskin, V.; Knizhnerman, L. Extended Krylov subspaces: Approximation of the matrix square root and related functions. *SIAM J. Matrix Anal. Appl.* **1998**, *19*, 755–771. [CrossRef]
15. Bergamaschi, L.; Caliari, M.; Martínez, A.; Vianello, M. Comparing Leja and Krylov approximations of large scale matrix exponentials. In Proceedings of the Computational Science–ICCS 2006, Reading, UK, 28–31 May 2006; Springer: Berlin/Heidelberg, Germany, 2006; pp. 685–692.
16. Knizhnerman, L.; Simoncini, V. Convergence analysis of the extended Krylov subspace method for the Lyapunov equation. *Numer. Math.* **2011**, *118*, 567–586. [CrossRef]
17. Eiermann, M.; Ernst, O.G. A restarted Krylov subspace method for the evaluation of matrix functions. *SIAM J. Numer. Anal.* **2006**, *44*, 2481–2504. [CrossRef]

18. Frommer, A.; Simoncini, V. Stopping criteria for rational matrix functions of Hermitian and symmetric matrices. *SIAM J. Sci. Comput.* **2008**, *30*, 1387–1412. [CrossRef]
19. Popolizio, M.; Simoncini, V. Acceleration techniques for approximating the matrix exponential operator. *SIAM J. Matrix Anal. Appl.* **2008**, *30*, 657–683. [CrossRef]
20. Bai, Z.Z.; Miao, C.Q. Computing eigenpairs of Hermitian matrices in perfect Krylov subspaces. *Numer. Algorithms* **2019**, *82*, 1251–1277. [CrossRef]
21. Knizhnerman, L.; Simoncini, V. A new investigation of the extended Krylov subspace method for matrix function evaluations. *Numer. Linear Algebra Appl.* **2010**, *17*, 615–638. [CrossRef]
22. Simoncini, V. A new iterative method for solving large-scale Lyapunov matrix equations. *SIAM J. Sci. Comput.* **2007**, *29*, 1268–1288. [CrossRef]
23. Güttel, S.; Knizhnerman, L. A black-box rational Arnoldi variant for Cauchy-Stieltjes matrix functions. *BIT* **2013**, *53*, 595–616. [CrossRef]
24. Druskin, V.; Simoncini, V. Adaptive rational Krylov subspaces for large-scale dynamical systems. *Syst. Control Lett.* **2011**, *60*, 546–560. [CrossRef]
25. Benzi, M.; Simoncini, V. Decay bounds for functions of Hermitian matrices with banded or Kronecker structure. *SIAM J. Matrix Anal. Appl.* **2015**, *36*, 1263–1282. [CrossRef]
26. Beckermann, B.; Reichel, L. Error estimates and evaluation of matrix functions via the Faber transform. *SIAM J. Numer. Anal.* **2009**, *47*, 3849–3883. [CrossRef]
27. Xu, S.; Xue, F. Inexact rational Krylov subspace methods for approximating the action of functions of matrices. *Electron. Trans. Numer. Anal.* **2023**, *58*, 538–567. [CrossRef]
28. Beckermann, B.; Güttel, S. Superlinear convergence of the rational Arnoldi method for the approximation of matrix functions. *Numer. Math.* **2012**, *121*, 205–236. [CrossRef]
29. Hochbruck, M.; Lubich, C. On Krylov subspace approximations to the matrix exponential operator. *SIAM J. Numer. Anal.* **1997**, *34*, 1911–1925. [CrossRef]
30. Druskin, V.; Knizhnerman, L. Krylov subspace approximation of eigenpairs and matrix functions in exact and computer arithmetic. *Numer. Linear Algebra Appl.* **1995**, *2*, 205–217. [CrossRef]
31. Druskin, V.; Knizhnerman, L.; Zaslavsky, M. Solution of large scale evolutionary problems using rational Krylov subspaces with optimized shifts. *SIAM J. Sci. Comput.* **2009**, *31*, 3760–3780. [CrossRef]
32. Druskin, V.; Knizhnerman, L.; Simoncini, V. Analysis of the rational Krylov subspace and ADI methods for solving the Lyapunov equation. *SIAM J. Numer. Anal.* **2011**, *49*, 1875–1898. [CrossRef]
33. Berljafa, M.; Güttel, S. Generalized rational Krylov decompositions with an application to rational approximation. *SIAM J. Matrix Anal. Appl.* **2015**, *36*, 894–916. [CrossRef]
34. Jagels, C.; Reichel, L. Recursion relations for the extended Krylov subspace method. *Linear Algebra Appl.* **2011**, *434*, 1716–1732. [CrossRef]
35. Jagels, C.; Reichel, L. The extended Krylov subspace method and orthogonal Laurent polynomials. *Linear Algebra Appl.* **2009**, *431*, 441–458. [CrossRef]
36. Ruhe, A. Rational Krylov sequence methods for eigenvalue computation. *Linear Algebra Appl.* **1984**, *58*, 391–405. [CrossRef]
37. Ruhe, A. Rational Krylov algorithms for nonsymmetric eigenvalue problems. In *Recent Advances in Iterative Methods*; The IMA Volumes in Mathematics and Its Applications; Springer: New York, NY, USA, 1994; Volume 60, pp. 149–164.
38. Güttel, S. Rational Krylov approximation of matrix functions: Numerical methods and optimal pole selection. *GAMM-Mitt.* **2013**, *36*, 8–31. [CrossRef]
39. Bagby, T. The modulus of a plane condenser. *J. Math. Mech.* **1967**, *17*, 315–329. [CrossRef]
40. Gončar, A.A. The problems of E. I. Zolotarev which are connected with rational functions. *Mat. Sb.* **1969**, *78*, 640–654.
41. Caliari, M.; Vianello, M.; Bergamaschi, L. Interpolating discrete advection-diffusion propagators at Leja sequences. *J. Comput. Appl. Math.* **2004**, *172*, 79–99. [CrossRef]
42. Caliari, M.; Vianello, M.; Bergamaschi, L. The LEM exponential integrator for advection-diffusion-reaction equations. *J. Comput. Appl. Math.* **2007**, *210*, 56–63. [CrossRef]
43. Druskin, V.; Lieberman, C.; Zaslavsky, M. On adaptive choice of shifts in rational Krylov subspace reduction of evolutionary problems. *SIAM J. Sci. Comput.* **2010**, *32*, 2485–2496. [CrossRef]
44. Ruhe, A. Rational Krylov: A practical algorithm for large sparse nonsymmetric matrix pencils. *SIAM J. Sci. Comput.* **1998**, *19*, 1535–1551. [CrossRef]
45. Botchev, M.A.; Grimm, V.; Hochbruck, M. Residual, restarting, and Richardson iteration for the matrix exponential. *SIAM J. Sci. Comput.* **2013**, *35*, A1376–A1397. [CrossRef]
46. Saad, Y. Analysis of some Krylov subspace approximations to the matrix exponential operator. *SIAM J. Numer. Anal.* **1992**, *29*, 209–228. [CrossRef]
47. van den Eshof, J.; Hochbruck, M. Preconditioning Lanczos approximations to the matrix exponential. *SIAM J. Sci. Comput.* **2006**, *27*, 1438–1457. [CrossRef]
48. Björck, A.; Golub, G.H. Numerical methods for computing angles between linear subspaces. *Math. Comp.* **1973**, *27*, 579–594. [CrossRef]

49. De Sturler, E. Truncation strategies for optimal Krylov subspace methods. *SIAM J. Numer. Anal.* **1999**, *36*, 864–889. [CrossRef]
50. Elman, H.C.; Su, T. Low-rank solution methods for stochastic eigenvalue problems. *SIAM J. Sci. Comput.* **2019**, *41*, A2657–A2680. [CrossRef]
51. Bak, J.; Newman, D.J. *Complex Analysis*, 3rd ed.; Undergraduate Texts in Mathematics; Springer: New York, NY, USA, 2010; pp. xii+328.
52. Suetin, P.K. *Series of Faber Polynomials*; Analytical Methods and Special Functions; Gordon and Breach Science Publishers: Amsterdam, The Netherlands, 1998; Volume 1, pp. xx+301.
53. Crouzeix, M. Numerical range and functional calculus in Hilbert space. *J. Funct. Anal.* **2007**, *244*, 668–690. [CrossRef]
54. Driscoll, T.A. Algorithm 843: Improvements to the Schwarz-Christoffel toolbox for MATLAB. *ACM Trans. Math. Softw.* **2005**, *31*, 239–251. [CrossRef]
55. Stewart, G.W. A Krylov-Schur algorithm for large eigenproblems. *SIAM J. Matrix Anal. Appl.* **2001**, *23*, 601–614. [CrossRef]
56. Frommer, A.; Güttel, S.; Schweitzer, M. Convergence of restarted Krylov subspace methods for Stieltjes functions of matrices. *SIAM J. Matrix Anal. Appl.* **2014**, *35*, 1602–1624. [CrossRef]
57. Curtiss, J.H. Faber polynomials and the Faber series. *Am. Math. Mon.* **1971**, *78*, 577–596. [CrossRef]
58. Duren, P.L. *Univalent Functions*; Fundamental Principles of Mathematical Sciences; Springer: New York, NY, USA, 1983; Volume 259, pp. xiv+382.
59. Bronshtein, I.N.; Semendyayev, K.A.; Musiol, G.; Mühlig, H. *Handbook of Mathematics*, 6th ed.; Springer: Berlin/Heidelberg, Germany, 2015; pp. xliv+1207.